Chapter 1

Cell and Tissue Organisation

Emma Marshman, Catherine Booth and Christopher S. Potten
Paterson Institute for Cancer Research, Manchester, UK

CONTENTS

To understand the changes that take place during cancer development, it is important to understand the basic principles of cell and tissue organisation and the mechanisms that control growth and structure.

TISSUE CLASSIFICATION

Groups of cells that are similar in structure, function and embryonic origin are referred to as tissues. The tissues of the body can be divided into four main groups as follows:

Epithelial Tissue

Epithelial tissue covers most of the free surfaces of the body, both internal and external, and often invaginates to form specialized structures such as glands. For example, it forms the outer layer of skin and the lining of the gastrointestinal tract and breast ducts. In addition to providing physical protection, epithelial cells control permeability, provide sensation and produce specialized secretions from glands, e.g. mucus, hormones and enzymes. Taking all the surface linings and their associated glands and structures together, epithelial tissues make up the major part of total body mass.

Connective Tissue

Connective tissue, or mesenchyme, protects and supports the body and its organs. Types of mesenchymal tissue include cartilage, bone and adipose tissue. The reticuloendothelial system is often considered a type of connective tissue. Reticuloendothelial cells are the defensive and oxygen-supplying cells of the body and are mostly derived from bone marrow precursor cells. The reticuloendothelial cells or haematopoietic cells are distributed throughout the body as free cells in blood and lymph or make up organs such as the spleen and lymph nodes.

Muscle Tissue

Muscle tissue is responsible for movement, such as skeletal movement, but also movement of food, blood and secretions. To carry out this function, muscle cells possess organelles and properties distinct from those of other cells which makes them capable of powerful contractions that shorten the cell along the longitudinal axis. There are three types of muscle tissue: skeletal, cardiac and smooth muscle. The contraction mechanism is similar in all three, but they differ in their internal organisation.

Nervous Tissue

Nervous tissue is specialized for the conduction of electrical impulses from one region of the body to another. Neural tissue consists of two basic cell types, neurons and supporting cells called glial cells. About 98% of the neural tissue in the body is concentrated in the brain and spinal chord with the rest making up the peripheral nervous system.

Since each tissue is made up of a number of specialized cell types that maintain tissue structure and function, there must be exquisite control over cell numbers to maintain the integrity of the tissue. The ability to respond to cell loss (via damage or senescence) varies in the different tissues, since not all cells have the same capacity for regeneration. Tissues can therefore be classified into

three groups depending on this cell replacement capability, as follows:

Rapidly Self-renewing Tissues

In tissues such as the skin, the intestine and the haematopoietic system, there is continuous cell loss either by surface abrasion, by damage or because the cell has aged. This cell loss has to be compensated for by cell production (proliferation), otherwise the tissue would begin to shrink (or expand if proliferation exceeds cell loss). Thus, the number of cells produced by cell division precisely balances cell loss in order for the tissue to maintain its size and mass.

Conditionally Renewing Tissues

In tissues such as the liver, breast, prostate and connective tissue, there is little or no replacement under normal circumstances. However, there is potential for regenerative proliferation under conditions in which the tissue's integrity is significantly compromised, e.g. damage or disease, or in response to hormonal influences.

Essentially Non-renewing Tissues

In some tissues, e.g. the female germ line and the central nervous system, there is little or no cell replacement or capacity for regeneration in the adult.

CELL PROLIFERATION AND ITS CONTROL

Control of cell division within a tissue is particularly important in rapidly self-renewing tissues when proliferation must balance cell loss. Although the exact mechanisms used by tissues to sense the need to increase or decrease cell division are unclear, it is obvious that proliferation must be regulated by a complex network of signals and messages including growth factors, cytokines and hormones. These messages can be produced by the cells themselves (autocrine regulation), may be produced by neighbouring cells of either similar or unrelated cell types (paracrine regulation, e.g. epithelial–mesenchymal interactions), and by circulating hormones (systemic endocrine regulation) as illustrated in **Figure 1**.

Some of the network of signals that control tissue homeostasis may prevent overproduction where necessary or arrest the cell cycle if a cell is damaged. In the latter scenario DNA damage is detected and the cell cycle arrested as it reaches specific checkpoints. These checkpoints will be discussed in more detail in the chapter Regulation of the Cell Cycle. At the checkpoint, the defective DNA can either be repaired or, if too severe, the cell may commit suicide in a process referred to as apoptosis. Apoptosis will be discussed in detail in the chapter *Apoptosis*.

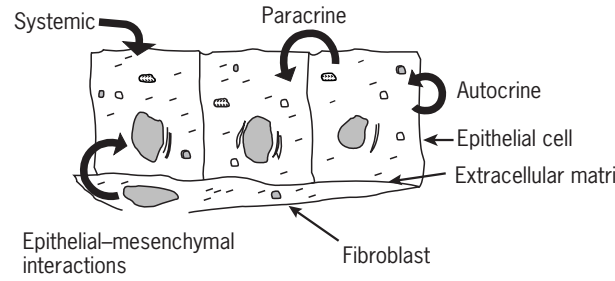


Figure 1 Signal mechanisms involved in cellular communication.

Malfunctions of Control

The first stages of cancer formation are associated with malfunctions in the control mechanisms described above, in such a way that the critical balance between cell proliferation and cell loss by differentiation or apoptosis is disturbed or deregulated. If this balance is shifted in favour of proliferation, the tissue will expand in a progressive and eventually uncontrolled manner, distorting the tissue structure and function. The balance may only need to be shifted slightly in favour of proliferation for a cancer to develop. Cancer development will be further described towards the end of this chapter.

CELLULAR HIERARCHIES

At the bottom of all the hierarchies in the body are the embryonic stem cells. Embryonic stem cells are referred to as totipotent, i.e. they are capable of differentiating into all types of tissue. Embryonic cells may separate and each form complete embryos, e.g. twins. Human embryonic stem cells have recently been isolated from embryonic tissue and can be maintained as undifferentiated cells in laboratory cultures under certain carefully controlled conditions. Even after 4–5 months in culture, these cells are still able to form types of cells from all three embryonic germ layers; including gut epithelium (endoderm), cartilage, bone, smooth muscle and striated muscle (mesoderm) and neural epithelium and embryonic ganglia (ectoderm) (Thomson *et al*., 1998). Studies with embryonic stem cells will give valuable information about the mechanisms controlling differentiation and organisation and may ultimately allow us to grow replacements for tissues or even organs that have been damaged by disease.

In development, these embryonic stem cells are abundant; however, as the animal ages the cellular potency becomes more and more restricted (the capability for wide gene expression becomes more restricted) until ultimately stem cells only remain in tissues capable of regeneration.

Such stem cells are found at the point of origin of cell production within an adult tissue and can produce a steady stream of cells (Potten, 1992). These daughter cells, termed dividing transit cells, can expand their numbers via further cell divisions and mature into functional differentiated cells, called simple transit cells. Simple transit cells are eventually lost from the tissue at the end of their functional lifespan. The linear evolution in the adult animal tissue is therefore organized into a hierarchy or 'family tree' with the cells responsible for cell production at the bottom and the functional cells at the top. The specialization process involved in the progression from the bottom to the top of the hierarchy, termed differentiation, represents a change in the pattern of gene expression which may be the consequence of changes either in the internal programming of the cell or of the external stimuli that affect the cell.

At the bottom of the adult hierarchy, and ultimately responsible for cell replacement in renewing tissues, are the pluripotent stem cells (capable of producing many but not all differentiated cell lineages, i.e. they are not totipotent). In many cases, these cells cannot be identified by a common marker or a single property. Instead, cells are classed as stem cells if they exhibit or have the potential to exhibit the following properties:

1. stem cells are undifferentiated (relative to the cells in the tissue);
2. stem cells are capable of proliferation;
3. stem cells are capable of self-maintenance;
4. stem cells can produce differentiated progeny;
5. stem cells can regenerate the tissue after damage.

When a stem cell divides, under normal circumstances, it is thought to generate a daughter that is another stem cell (thereby maintaining itself) and one daughter that will move up the hierarchy towards differentiation. Although this situation remains to be conclusively proven, it is certainly the average situation that must occur in an adult tissue. Whether the determinants of such division are intrinsic to the stem cell itself or are influenced by the surrounding environment also remains to be determined.

If stem cell numbers need to increase or decrease in response to external stimuli, this asymmetric form of cell division will switch to symmetrical division in which either two stem cell daughter or two nonstem cell daughters are produced. Stem cell expansion will inevitably increase cellular production (i.e. speed up regeneration, generate hyperplasia), whereas stem cell removal will reduce or remove cellular production (depending on how many stem cells remain in the tissue), e.g. generate aplasia or hypoplasia.

The next steps in the life of a nonstem cell daughter, particularly in a rapidly renewing tissue, are the amplification of cell numbers. The daughter cells divide a number of times and are known as transit amplifying cells. During this time the cells gradually appear to lose their stem cell properties and acquire a more mature phenotype until, after a given number of divisions, they are fully differentiated cells. These transit amplifying cells are therefore generally a short-lived phenotype, although during the early cell generations they may be called upon to behave as stem cells in a trauma situation in some tissues. With successive divisions they eventually lose this ability. The later-generation differentiated cells then perform the function for which they were generated, gradually senesce and die. This is also therefore a form of programmed cell death.

The advantage of such an organisation is that only a few stem cells are needed to maintain a whole tissue. Generally, these stem cells have a slow cell cycle time which allows for genetic housekeeping, i.e. time to repair any genetic damage. Small numbers of stem cells followed by around five generations of transit amplifying cells create an environment in which the greatest risk of introducing a mutation (during division) is in the transit cells (which are ultimately lost from the tissue) rather than in the long-lived stem cells. In conditionally renewing tissues the organisation is less clear. Although stem cells must exist, it is possible that they are normally quiescent or are cycling very slowly, and are only activated by trauma or hormonal stimuli.

The progression from stem cell to differentiated cell could be preprogrammed but is more likely to be controlled by extrinsic factors. An organized hierarchy obviously experiences (and/or is able to respond to) different control signals at different stages. This can be aided by a physical organisation, such that there is a spatial distribution within the hierarchy controlled by a series of microenvironments or niches. A gradient of controlling factors probably exists along the maturation axis.

Particularly important in the microenvironment is the basement membrane upon which epithelial cells sit. This basement membrane is a highly organized extracellular matrix (ECM) made up of proteins such as collagen and laminins. The effects of the matrix are primarily mediated by cell adhesion molecules such as integrins and cadherins which are families of cell surface receptors. Cell adhesion molecules help to connect the exterior of the cell with the interior of the cell in two ways: by transducing signals initiating from the extracellular interactions and by mediating structural linkages between the cytoskeleton and the ECM of other cells (Horwitz and Werb, 1998). These processes will be further described in the chapters *Wnt Signal Transduction* and *Extracellular Matrix: The Networking Solution.*

CELL ORGANISATION IN SPECIFIC TISSUES

To illustrate the points made in the previous section, the stem cells and hierarchies of a number of tissues will be described in more detail.

Haematopoietic System

The hierarchical organisation of the continually renewing cells in the bone marrow has been extensively studied. All mature blood cells in the body are derived from a small number of stem cells that reside in the bone marrow in a process called haematopoiesis. Over 10^{11} new cells are produced daily to maintain homeostasis since the majority of mature blood cells are short-lived. In addition, normal daily cell replacement must also be sporadically increased to fight infection or to compensate for blood loss.

The haematopoietic lineage is shown in **Figure 2**. The most primitive stem cell of the bone marrow is the pluripotent stem cell which has the capability to produce all the different cell types of the blood. To add another level of complexity, this pluripotent stem cell may itself be part of a stem cell hierarchy. Myeloid and lymphoid stem cells are produced from the pluripotent stem cell population. The myeloid stem cell then goes on to produce a number of progenitor cells which are the precursors of the six types of mature functional myeloid cells: erythrocytes, thrombocytes, eosinophils, macrophages, mast cells and neutrophils. These cells have different functions within the immune system and in the blood. There may be further as yet unknown subdivisions in the stem cell hierarchy. The lymphoid stem cell produces a number of lymphoid progenitors which mature into B and T lymphocytes to provide defence against pathogens or toxins.

Although mature blood cells can be distinguished from each other, stem cells and progenitor cells have no specific distinguishing features under the microscope. Identification of early progenitor cells and stem cells is also made difficult by the low incidence of these cells in blood. For example, pluripotent stem cells are thought to make up only 0.01–0.1% of total bone marrow cells (Heyworth *et al*., 1997). Functional assays have been devised, the first of which was described by Till and McCullogh (1961). This method involves transplantation of some healthy bone marrow cells into mice whose own bone marrow has been destroyed by irradiation. The transplanted cells produce colonies of differentiated haematopoietic cells in the spleen which can be counted. In addition to functional assays, external markers have been used to identify progenitor cells. Myeloid and lymphoid stem cells and early progenitor cells can be separated from blood by antibodies that react to specific antigens only present on these cell types, e.g. CD34 antigen which is expressed on 0.5–5% of human bone marrow cells. Methods for separation of pluripotent stem cells using specific markers are under development.

In the bone marrow, stem cells and their progeny are exposed to a number of different stimuli including physical interactions with other cells mediated by cell adhesion molecules, interactions with extracellular matrix molecules such as collagen and fibronectin and exposure to growth-stimulatory and growth-inhibitory chemicals called cytokines. There are over 15 cytokines involved in haematopoiesis and these are produced by a number of cell types including the mature cells themselves, e.g. neutrophils, B and T cells, as well as by fibroblasts and bone marrow stromal cells providing autocrine and paracrine regulation (Heyworth *et al*., 1997). All these signals coordinate the self-renewal and differentiation of the stem cells and the formation of the mature cell types.

The role of cytokines in determining which type of cell (e.g. mast cell or neutrophil) an early progenitor cell differentiates into is highly complex. Some cytokines have many target cells, whereas others are much more restricted. Interleukin (IL-3), for example, can stimulate stem cells

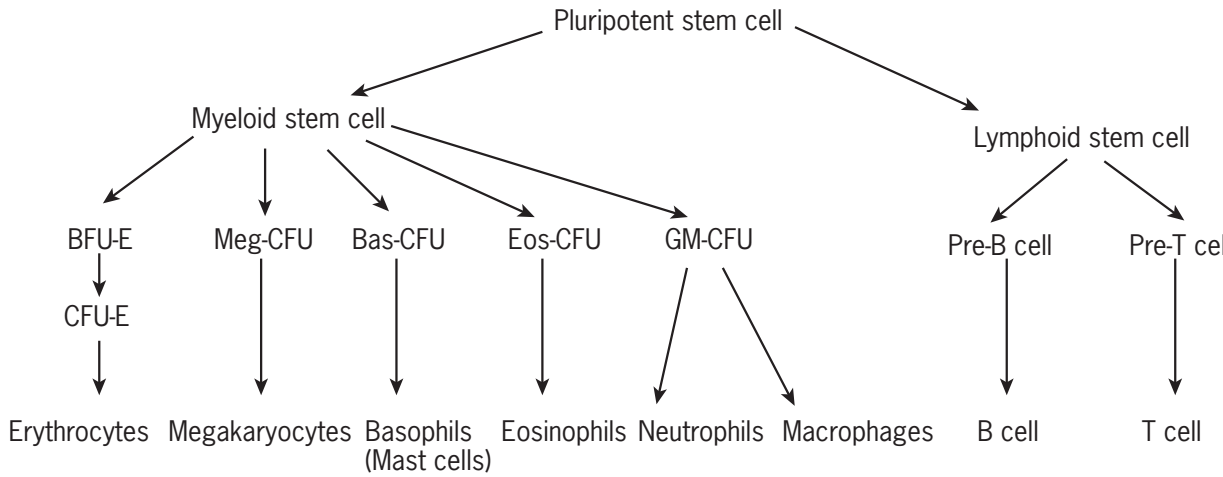


Figure 2 Haematopoietic cell lineage.
Abbreviations: BFU-E, erythroid burst-forming unit; CFU-E, erythroid colony-forming unit; Meg-CFU, megakaryocytic colony-forming unit; Bas-CFU, basophilic colony-forming unit; Eos-CFU, eosinophilic colony-forming unit; GM-CFU, granulocyte macrophage colony-forming unit.

to produce myeloid progenitor cells and can also stimulate myeloid progenitor cells to produce a number of mature cell types (Dexter, 1993). Another example is granulocyte–macrophage colony-stimulating factor (GM-CSF) which acts on the granulocyte–macrophage progenitor cell and the eosinophil progenitor cell to produce neutrophils, macrophages and eosinophils. In contrast, some growth factors have direct effects on only one cell population, e.g. erythropoietin, which acts only on the erythroid progenitor cell to produce erythrocytes. Other cytokines mainly influence the maturation of cells rather than the proliferation of progenitor cells, e.g. IL-5 and eosinophil development.

Originally it was thought that haematopoiesis was regulated solely by modulation of the production of these directly acting cytokines, e.g. stem cells would be acted upon by IL-3 to produce progenitor cells and then specific cytokines would be made to induce maturation of the progenitor cells into whichever specific cells were required by the bone marrow. It is now known that control is exerted at a more complex level such that a certain growth factor alone will not have effects on a particular cell type; however, when it is combined with another factor proliferation or maturation can be induced. For example, lymphoid stem cells will not respond to macrophage colony-stimulating factor or IL-1 alone, but are stimulated in the presence of a combination of these two growth factors.

The haematopoietic cell lineage has illustrated the complex communication network required for the differentiation of relatively unknown stem cells into the specific cells of the blood.

Small Intestine

The epithelium of the small intestine provides another example of a self-renewing tissue which has been studied for many years. The tissue organisation of the small intestine is different from that described for the haematopoietic system, being highly polarized and structured. However, the regulation mechanisms are equally complicated and largely unknown at present.

In the small intestine, epithelium covers finger-like projections called villi and flask-shaped crypts located at the base of the villi which are embedded in the connective tissue (see **Figure 3**). Epithelial cells are produced in the lower part of the crypt and migrate up the crypt on to the villi and are continuously shed from the villus tip. In common with the haematopoietic system, migration from the early precursor cells is accompanied by differentiation and specialization. Cells differentiate into three functional cell types as they move up the crypt: the predominant enterocyte, the mucus-secreting goblet cell and the peptide hormone-secreting enteroendocrine cells. In addition, a number of cells migrate down to the base of the crypt to become the fourth cell type, the Paneth cells. Paneth cells secrete a number of proteins including lysozyme, which is thought to play a role in fighting bacterial infection.

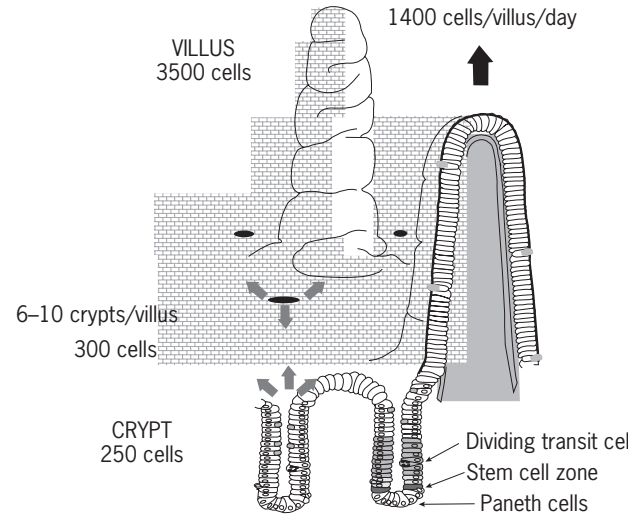


Figure 3 Organisation of small intestinal epithelium.

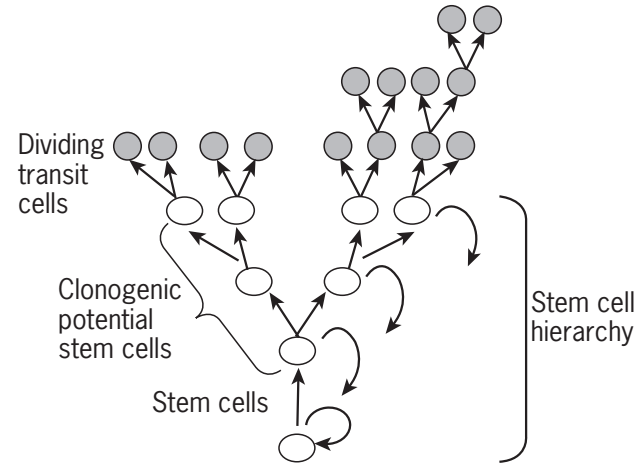


Figure 4 Proposed stem cell model for the small intestine.

Replacement of cells shed at the villus tip must be balanced by cell production in the crypt, at a rate of about 10^{10} cells per day in humans (**Figure 3**) (Potten, 1992). Cell replacement is achieved by stem cells located amongst or just above the Paneth cells at the base of the crypt. Unfortunately, there are no markers for intestinal stem cells and at present, characterization studies can only be carried out by disturbing the system and observing the outcome. A stem cell model has been proposed based on clonal regeneration studies following radiation or drug exposure (**Figure 4**). The proposed model suggests that there are 4–6 ancestor or functioning stem cells per crypt (Potten, 1998). These stem cells are very sensitive to toxic insults (e.g. radiation and some chemotherapeutic agents) and are unable to repair damaged DNA. If damaged they readily initiate apoptosis and die. This sensitivity may reflect the need to avoid repopulation of the crypt with cells containing damaged DNA, and thereby preserves the integrity of the tissue. Stem cells that die, however, are

easily replaced by the other surviving stem cell members or by their immediate daughter cells, which make up the second tier of the hierarchy. The second tier stem cells have a better repair capacity and, if not required to regenerate the first tier (such as in a normal situation), they are displaced into the transit compartment. If this second tier is destroyed, a third tier may also exist that contains about 20 even more resistant stem cells with the best repair capacity. These three tiers therefore make up a population of around 30–40 potential stem cells–cells that are acting as stem cells or retain the ability to act as a stem cell if required. Since each of these cells can regenerate a clonal population (a crypt), they are also termed clonogenic cells. Above the level of clonogenic stem cells there are about 124 dividing transit cells which have no stem-cell attributes. These proliferative cells move or are displaced at a rate of 1–2 cell positions per hour from the crypt on to the villus (Potten, 1992, 1998).

Regulation of cell proliferation in the gut is not fully understood. However, a large number of factors are known to be involved, including growth factors, cytokines and ECM molecules. The epidermal growth factor (EGF) family is one group of substances known to stimulate proliferation and includes epidermal growth factor itself and TGF-α (Potten *et al*., 1997). In contrast, the TGF-β family of growth factors have been associated with negative regulation or inhibition of crypt cell proliferation (see also the chapter *Signalling by TGF-β*). In common with growth factors, *in vitro* studies suggest that some interleukins have stimulatory effects (e.g. IL-4) and some have inhibitory effects (e.g. IL-11 and IL-6).

The ECM underlying the epithelium plays a role in a number of key processes, one of which is cell migration. The process of migration is not fully understood and it was initially thought that cells moved in tandem with underlying connective tissue. More recent studies suggest that cells 'walk' over stationary ECM which contains a number of adhesion molecules such as E-cadherin, laminin, fibronectin, tenascin and collagen. Migration is thought to involve decreased cell attachment to one or more of these adhesion molecules, since adhesion molecule expression patterns vary along the crypt/villus axis. The stationary nature of stem cells may be due to their strong anchorage to the stroma. For example, fibronectin, which is a particularly 'sticky' adhesion molecule, is abundant in the crypt whereas tenascin which is less adhesive is predominant on the villus. In addition, movement is controlled by the expression/availability of integrins, epithelial cell receptors for these adhesion molecules – a cell can only be influenced by adhesion molecule levels within the basement membrane if it expresses the appropriate receptors.

Epidermis

The outermost layer of the skin, the epidermis, is another example of a self-renewing tissue. The epidermis is predominantly made up of keratinocytes (about 80% of the total). Other epidermal cell types exist with specific functions: melanocytes give the skin its pigmentation and afford some protection against ultraviolet light, Merkel cells sense fine mechanical events and Langerhans cells form part of the body's immune system.

The outer surface of the epidermis is called the stratum corneum and is composed of a layer of thin, dead keratinocytes. These cells bear little resemblance to normal keratinocytes, since by the time they reach the surface their nucleus and internal organelles have disappeared and they are reduced to thin plates of keratin. Keratins are a family of insoluble proteins that form intermediate filaments within cells and confer mechanical strength. Surface keratinocytes represent the final mature functional differentiated cells of the skin. These cells are continually being shed or lost and therefore perpetual cell replacement is required to maintain epidermal function.

Below the stratum corneum are three other epidermal cell layers: the granular layer, the spinous layer and the basal layer. These epidermal regions are depicted in **Figure 5**, although there are many more layers of cells than are shown in the diagram. In common with the small intestine and the haematopoietic system, stem cells are responsible for the regenerative potential of skin. These stem cells are located within the basal layer (Lavkar and Sun, 1983). Studies carried out on mouse epidermis suggest that 5–12% of cells in the basal layer are stem cells (Potten, 1992). Transitory dividing cells produced from these stem cells make up about 50% of the basal layer with the remaining basal layer cells being postmitotic and having no proliferative characteristics. These cells are committed to terminal differentiation and achieve this as they slip out of the basal layer and migrate into the spinous layer, where they flatten. From the spinous layer, cells progress up into the granular layer until they reach the stratum corneum where they are eventually shed. The stem-cell progeny generate a discrete column of cells, from basal cell to keratinized cell, arranged in a hexagonal pattern and called an epidermal proliferative unit (Potten, 1981). It has been estimated that it takes the human keratinocyte between 26 and 42 days to travel from the basal layer to the outermost cornified layer and therefore it takes 1–2 months for the epidermis to replace itself completely.

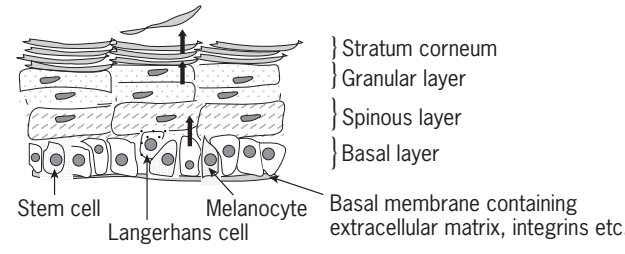


Figure 5 The murine epidermal proliferative unit.

In common with the small intestine, the underlying ECM plays a key role in basal layer processes. It has been suggested that the ECM mediates adhesion, regulates terminal differentiation and aids cell movement upward from the basal layer. When basal keratinocytes become committed to undergo terminal differentiation, their ability to adhere to components of the ECM decreases and upward cell migration occurs (Jones and Watt, 1993). Populations of putative stem cells that are greater than 90% pure have been isolated on the basis of their adhesive properties.

The epithelial cells of the skin, and indeed other sites of the body, are able to form a barrier due to a number of functionally and structurally distinct epithelial cell junctions, including tight junctions, gap junctions, desmosomes and hemidesmosomes. Tight junctions seal neighbouring cells together to stop water-soluble molecules leaking between the cells and confine transport proteins either to the outward-facing membrane (apical) or to the inner membranes (basolateral) to control the passage of certain chemicals (e.g. glucose transport in the small intestine). In contrast, gap junctions are involved in cell–cell signalling. Gap junctions are intercellular channels made up of connexin proteins that allow inorganic ions and other small water-soluble molecules to pass directly from the cytoplasm of one cell to the cytoplasm of another, thereby coupling the cells both metabolically and electrically. To maintain mechanical strength, cells are linked together with desmosomes. Desmosomes consist of a dense plaque of intracellular attachment proteins (including plakoglobin and desmoplakins) which are associated with rope-like intermediate keratin filaments that form a continuous network throughout the tissue. Hemidesmosomes, or half-desmosomes, connect the basal surface of epithelial cells to the underlying basement membrane such as that which separates the epidermis and the dermis. In addition to aiding attachment, hemidesmosomes have also been found to be important in modulating the organisation of the cytoskeleton, proliferation and differentiation. These effects are mediated by integrins which transduce signals from the ECM to the interior of the cell as described earlier. Absence or defects of hemidesmosomal proteins can result in devastating blistering skin diseases.

Breast

The breast or mammary gland is an example of a conditionally renewing tissue in that cell replacement is generally limited except under certain conditions, e.g. pregnancy.

The organisation of the breast changes during three developmental phases. The first stage occurs in the foetus where mammary glands arise as buds from the epidermis which elongate to form simple, branched ducts. At puberty, there is rapid extension and branching of the ducts which terminate in globular structures called terminal end buds. These terminal end buds and terminal ducts then go on to form lobules of alveolar buds.

The third phase of mammary development occurs during pregnancy and lactation and at this stage the breast can be considered to be morphologically mature and functionally active. The alveolar buds and lobes subdivide further, giving rise to large clusters of alveolar lobes. During lactation, the clusters of alveolar lobes become distended and form secretory alveoli lined with alveolar cells which produce milk. After cessation of lactation, involution of the breast occurs where the secretory cells of the alveoli degenerate and disappear. Similarly, after the menopause, there is progressive involution of the ductal and glandular components of the breast. The connective tissue supporting the breast also degenerates with loss of stromal cells and collagen fibres.

In the normal breast, the ducts and lobes of the mammary gland are separated from the stroma by a basement membrane. This basement membrane is lined with two cell types, an outer lining of myoepithelial cells containing myofilaments and an inner lining of epithelial cells. As described in the section regarding the epidermis, the epithelial cells of the breast are connected together with desmosomes whereas myoepithelial cells connect to the basement membrane with hemidesmosomes.

Studies in rodent mammary glands indicate that epithelial cell types and alveolar cells arise from stem cell populations capable of generating the fully differentiated lactating mammary gland. These stem cells are thought to be present in the basal cell layer of ducts and end buds, although little more is known about their identity (Rudland *et al*., 1997). It has been suggested that stem cells can give rise to either ductal epithelial cells in a reversible manner or myoepithelial cells in an irreversible manner. Alveolar cells are thought to be derived from ductal epithelial cells.

As with the other tissues described, the differentiation of cells produced by breast stem cells is strictly controlled. Unlike the haematopoietic system where differentiation is controlled mainly by paracrine and autocrine secretions, the breast is also subject to control by circulating hormones secreted by the pituitary, ovary and adrenal glands. For example, during each menstrual cycle at about the time of ovulation, there is an increase in lobular size and epithelial cell vacuolization under the influence of oestrogens and rising progesterone. When menstruation occurs, the fall in hormone levels causes lobular regression. Similarly in pregnancy, oestrogens and progesterone stimulate proliferation and development, and prolactin released by the pituitary gland activates the production of alveolar cells. Additionally, lactation is triggered by the release of oxytocin, which causes contraction of the smooth muscle components of the myoepithelial cells surrounding the alveoli leading to milk expulsion.

Local growth hormones are also important since the growth promoting effects of oestrogen are believed to be mediated by TGF-α and insulin-like growth factor-1 (IGF-1) which increase epithelial cell growth and inhibit

myoepithelial cell differentiation (Rudland *et al*., 1997). Additionally, production of basic fibroblast growth factor (bFGF) by breast stem cells may regulate their own growth and that of myoepithelial cells in an autocrine/paracrine fashion.

Liver

The liver is another example of a conditionally renewing tissue. Epithelial cells make up the majority of the liver, particularly hepatocytes and also biliary duct epithelium. The hierarchial organisation of the liver, however, is not fully understood. Unlike the tissues mentioned above, where one population of stem cells is responsible for cell replacement, it seems likely that a two-compartment system is operative in this tissue (Alison, 1998). First, in the event of damage, hepatocytes are able to regenerate themselves very efficiently. This is unusual because the ability to regenerate is normally characteristic of undifferentiated stem cells, and hepatocytes are thought to be fully differentiated. A second compartment is activated if the function of surviving hepatocytes is drastically impaired and involves generation of new hepatocytes from stem cells. Liver stem cells, believed to be located in biliary ductules, produce oval cells. Oval cells are then thought to differentiate into the functional cell types, e.g. hepatocytes. The proposed model is illustrated in **Figure 6**.

Others

As indicated above, knowledge about the organisation of tissues ranges from the haematopoietic system and small intestine, which have been extensively studied, to the liver and breast, where information is more limited. Recently, a number of developments have been made in understanding the hierarchy of other tissue types. For example, mesenchymal stem cells have been isolated from human bone marrow. These cells replicate as undifferentiated cells and have the potential to differentiate into lineages of mesenchymal tissues including bone, cartilage, fat, tendon, muscle and marrow stroma (Pittenger *et al*., 1999). Neuronal stem cells have also been discovered in foetal brain which can develop into neurons and glial cells if certain growth factors are present *in vitro* (Vogel, 1999).

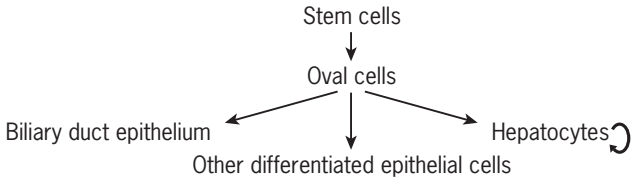


Figure 6 A proposed model for liver cell generation and regeneration.

CONTROL OF TISSUE ORGANISATION IN DEVELOPMENT

Regulation of tissue organisation in development is likely to involve a number of highly complicated mechanisms; however, one group of genes called homeobox genes has already been found to play a significant role in this process. Homeobox genes are a family of regulatory genes encoding transcription factors (homeoproteins) that can activate or repress the expression of a large number of target genes and so determine cell fate and general pattern formation. One major family of homeobox genes termed Hox genes control the identity of various regions along the body axis. These Hox genes are activated in sequence such that early Hox genes that control hindbrain development, followed by activation of Hox genes that control the thoracic region and late genes that control the lumbo-sacral region. Perturbing the expression of these factors can induce gross changes in tissue, organ and even limb development. For example, synpolydactyly, an inherited disease characterized by hand and foot malformation, is caused by expansions of the *HOXD13* gene.

In addition to developmental regulation, certain homeobox genes are also involved in inducing differentiation in renewing tissues. For example, the *cdx* genes are involved in controlling intestinal epithelial cell differentiation, possibly by transducing signals from laminin-1 in the underlying mesenchyme, and *HOXA9* and *PBX1* are some of the many genes involved in the control of haematopoietic differentiation. Altered expression of any of these can suppress differentiation and ultimately lead to tumour formation.

CANCER DEVELOPMENT AND TISSUE ORGANISATION

The chapter *Overview of Oncogenesis* will discuss the mechanisms involved in carcinogenesis in detail, but briefly the process is thought to involve a number of steps. First, a cell experiences a mutation that may or may not influence its immediate behaviour. This cell may then be more susceptible to subsequent mutations and, over time, gradually accumulate enough damage such that the normal control or 'braking' mechanisms is perturbed. This gradual accumulation of mutations is therefore known as the multistage model of carcinogenesis and explains why cancer is generally a disease of old age unless, for example, the primary mutation is an inherited disorder.

Within a tissue experiencing this process, the first observable histological stage is hyperplasia or cellular overgrowth, although this term must be used carefully since tissue regeneration in response to wounding is itself a form of hyperplasia. Hyperplasia can therefore be benign in addition to cancerous. Since in the adult hyperplasia can only occur in proliferating tissues it is not surprising that

almost all cancers arise in rapidly renewing or conditionally renewing tissues. In each case there is a malfunction in cellular homeostasis and cell production exceeds cell loss.

The origin of cell production, and the only permanent resident of a renewing tissue, is the stem cell. Cancers can therefore be thought of as stem cell diseases (transformation of a maturing cell would have no long-term effect since even if it divides a few times, each cell is ultimately lost from the tissue in a relatively short time frame). An expansion of stem cell numbers can therefore lead to hyperplasia. Normally such an expansion would be detected by the tissue and the excess stem cell removed, via apoptosis. However, if this does not occur, cellular output will be dramatically increased. For example, in the colon expression of the anti-apoptotic gene *bcl-2* may allow the survival of a single extra stem cell in an intestinal crypt (Potten *et al*., 1997). This alone can lead to 128 extra cells being produced by that one crypt (owing to the expansion by the transit amplifying cells). As the animal ages these excess stem cells persist and may experience further mutations (e.g. in apoptosis regulation such as by *p53*, growth factor signal transduction such as in *SMAD* and *ras*, DNA repair by mismatch repair enzymes such as MSH2, or in cellular adhesion such as changed integrin or E-cadherin expression), thereby increasing cancer risk. These mutations generally occur in three vital areas – regulation of cell division in the renewing population (restraint), DNA repair (such that the normal DNA is not maintained) and interactions with the extracellular environment (cells or matrix). Together these will subvert the normal differentiation process and allow unrestrained tissue growth without the accompanying levels of cell death, followed by invasion and metastasis into other tissue sites.

CONCLUSIONS

The organisation of cells and tissues has been discussed in development and in the normal adult and we have attempted to highlight the complex nature of the regulation processes that control cell proliferation, differentiation and regeneration. Cancer development provides us with an excellent example of the devastating effects observed when these processes are subverted and emphasizes the need for such exquisitely controlled mechanisms.

REFERENCES

Alison, M. (1998). Liver stem cells: a two compartment system. *Current Opinion in Cell Biology*, **10**, 710–715.

Dexter, T. M. (1993). Synergistic interactions in haemopoiesis: biological implications and clinical use. *European Journal of Cancer*, **29A**, S6–S9.

Graham, G. J. and Pagnell, I. B. (1992). The haematopoietic stem cell: properties and control mechanisms. *Seminars in Cellular Biology*, **3**, 423–434.

Heyworth, C. M., *et al*. (1997). Growth factors and the regulation of haematopoietic stem cells. In: Potten, C. S. (ed.), *Stem Cells*. 423–446 (Academic Press, London).

Horwitz, A. R. and Werb, Z. (1998). Cell adhesion and the extracellular matrix: recent progress and emerging themes. *Current Opinion in Cell Biology*, **10**, 563–565.

Jones, P. H. and Watt, F. M. (1993). Separation of human epidermal stem cells from transit amplifying cells on the basis of differences in integrin function and expression. *Cell*, **73**, 713–724.

Lavker, R. M. and Sun, T.-T. (1983). Epidermal stem cells. *Journal of Investigative Dermatology*, **81**(1S), 121–127.

Pittenger, M. F., *et al*. (1999). Multilineage potential of adult human mesenchymal stem cells. *Science*, **284**, 143–147.

Potten, C. S. (1981). Cell replacement in epidermis (keratopoiesis) via discrete units of proliferation. *International Review of Cytology*, **69**, 271–317.

Potten, C. S. (1992). Cell lineages. In: McGee, J. O'D., *et al*. (eds), *Oxford Textbook of Pathology*, Vol. 1, 43–52 (Oxford University Press, Oxford).

Potten, C. S. (1998). Stem cells in gastrointestinal epithelium: numbers, characteristics and death. *Philosophical Transactions of the Royal Society of London*, **353**, 821–830.

Potten, C. S., *et al*. (1997). The intestinal epithelial stem cell: the mucosal governor. *International Journal of Experimental Pathology*, **78**, 219–243.

Rudland, P. S., *et al*. (1997). Mammary stem cells in normal development and cancer. In: Potten, C. S. (ed.), *Stem Cells*. 147–232 (Academic Press, London).

Thomson, J. A., *et al*. (1998). Embryonic stem cell lines derived from human blastocysts. *Science*, **282**, 1145–1147.

Till, J. E. and McCulloch, E. A (1961). A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. *Radiation Research*, **14**, 213–222.

Vogel, G. (1999). Harnessing the power of stem cells. *Science*, **283**, 1432–1134.

FURTHER READING

Cillo, C., Faiella, A., Cantile, M., and Boncinelli, E. (1999). Homeobox genes and cancer. *Experimental Cell Research*, **248(1)**, 1–9.

D'Andrea, A. D. (1994). Haematopoietic growth factors and the regulation of differentiative decisions. *Current Opinion in Cell Biology*, **6**, 804–808.

Loeffler, M. and Potten, C. S. (1997). Stem cells and cellular pedigrees – a conceptual introduction. In: Potten, C. S. (ed.), *Stem Cells*. 423–446 (Academic Press Ltd., London).

Schwarzbauer, J. (1999). Basement membrane: putting up the barrier. *Current Biology*, **9**, R242–244.

Stappenbeck, T. S., Wong, M. H., Saam, J. R., Mysorekar, I. U. and Gordon, J. I. (1998). Notes from some crypt watchers: regulation of renewal in the mouse intestinal epithelium. *Current Opinion in Cell Biology*, **10**, 702–709.

Chapter 2

Regulation of the Cell Cycle

Arthur B. Pardee
Dana-Farber Cancer Institute, Boston, MA, USA

CONTENTS

OVERVIEW

In each of us are about 50 trillion living cells, all of which originated from only one cell, a fertilized egg. As we developed into adults this cell divided into two cells, these into four, and so forth, at least 45 times. The orderly process by which one cell becomes two is named the cell cycle. This cycle is fundamental not only for understanding cell growth, but also for replacement of cells lost by damage, as in wound healing and from the normal wear and tear of our bodies. The cell cycle is evidently tightly regulated, because we usually make new cells only when they are needed. Indeed, cancers arise when cell growth control is defective. 'Cancer is a wound that does not heal.'

One should remember that cells in most tissues are not usually progressing through the cycle, but are at rest, happily performing their specialized functions in support of the whole organism. But as exceptions, bone marrow, intestinal epithelial and some other cells are constantly dividing. A cell has a life cycle. It is formed, eventually becomes worn and dies by a programmed cell-death mechanism called apoptosis. Thereafter, nearby cells grow and divide to replace it. Cell numbers are balanced by proliferation versus apoptosis. After a cell becomes cancerous the balance is perturbed in favour of proliferation. These facts can be overlooked because much research is performed with cells put into culture and under conditions that permit proliferation.

THE NORMAL CELL CYCLE

History of Cell Cycle Biology

Before the cell cycle, microscopy revealed an interval of about 1 day between successive divisions of one cell into two cells. Until about 50 years ago, no changes could be observed during most of this interval, until about 1 hour before division when chromosomes which contain the hereditary material become visible and are equally partitioned between the two daughter cells, a complex process termed mitosis that takes place through M phase.

Production of two cells from one requires duplication of all of the myriad molecules that compose each cell. The most evidently duplicated molecules are deoxyribonucleic acid (DNA), the heredity-containing material in chromosomes. DNA does not duplicate throughout the cycle, but only during several hours in mid-cycle. This period is named the S phase, for DNA synthesis. Other molecules are duplicated at different times throughout the cycle. These findings about DNA synthesis (Howard and Pelc, 1951) showed that the cycle is organized as a series of events, and created the present framework of its four phases: a 'gap' (G_1 phase) during which a cell prepares for DNA synthesis, DNA synthesis (S phase), preparation for mitosis (G_2 phase) and the mitotic M phase, after which the cell divides and two new cycles commence (**Figure 1**). For a historical summary of biology of the cycle, see Baserga (1985).

Quiescence

Commencing by considering normal animal cells, most of the cells within us are in a quiescent state (G_0 phase). They have left their cycling during the G_1 state, so in quiescent cells DNA has not yet duplicated. But quiescent cells differ from G_1 cells in many other properties, in particular lacking molecules required for growth. This fact told us that the molecular switch that controls growth versus quiescence, and that is defective in cancers, is to be found in G_1 phase (Pardee, 1989).

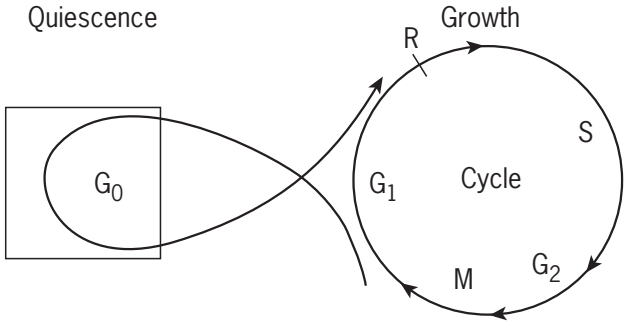


Figure 1 The basic cell cycle. The consecutive phases of the cycle, entry from G_0 and exit from G_1 to quiescence and differentiation are indicated.

G_1 Phase

When cells are activated to proliferate they advance from G_0 to G_1 phase, during several hours after several growth factor proteins are provided in their environment. These include epidermal growth factor (EGF) and insulin-like growth factor (IGF-1), which must overcome inhibitions by crowding of cells and the negative factor TGF-β. Growth factors and nutrients must be supplied from the blood in an organism. To grow cells outside the body, in tissue culture, a nutrient medium is required, in which growth factors are usually supplied by adding blood serum. Cells complete their cycle and then become quiescent after growth factors have been removed.

The length of time that cells in a culture spend in G_1 phase is highly variable, e.g. from 6 to 24 h, unlike the fairly uniform time they spend in the other phases. Many other synthetic biochemical processes take place in G_1 phase (see below).

S Phase

The requirement for growth factors to pass through G_1 phase is lost at the restriction point (R), located shortly before cells start to synthesize DNA. At the beginning of S phase, enzymes involved in DNA duplication increase, and they move into the nucleus where DNA is duplicated, from the surrounding cytoplasm where proteins are synthesized. Then at specific times during the next 6–8 h the DNAs of the perhaps 40 000 genes located on 23 pairs of chromosomes are replicated, each according to a timed program. For example the dihyrofolate reductase gene replicates quickly in very early S phase, but other genes are duplicated at other specific times throughout S phase.

G_2 Phase

After DNA synthesis is completed, several hours are required before initiation of mitosis, presumably to produce needed enzymatic machinery. Many G_2 products are unknown; a terminal one is the maturation promoting factor (MPF).

M Phase and Cell Division

Mitosis requires less than 1 h, and is subdivided into four main stages, in which the duplicate chromosomes pair and condense, and a mitotic 'machinery' consisting mainly of microtubule proteins segregates them equally between the two daughter cells. At completion of M phase, proteins of the mitotic apparatus are destroyed. The daughter cells then become separated, and each can repeat the cycle processes.

MOLECULAR BIOLOGY OF THE CYCLE

Signalling Molecules

Comparisons of growth of mammalian cancer and normal cells in culture revealed in 1974 that the basis of cancer's deranged growth control is located in G_1 phase, shortly before initiation of DNA synthesis (Pardee, 1989). In the same year, genetic studies of the cycle were intiated; research with cycle-controlling yeast mutants led to the discovery of numerous cycle-regulatory genes (Hartwell and Kastan, 1994). Biochemistry and molecular biology soon followed, with the identification of new genes and key enzymes; in particular proteins named cyclins that activate these kinases were discovered by Hunt and Ruderman (see review by Murray and Hunt). These rise and fall during the cycle because of periodic changes in their synthesis and destruction (Minshull *et al*., 1989). Cyclin-dependent kinases (cdks) that phosphorylate proteins required for cell cycle progression were identified (Nurse *et al*., 1998). Several proteins that inhibit these kinases and that vary during the cycle were discovered later. This involvement of both positively and negatively acting molecules illustrates the Ying–Yang principle of dynamic opposing actions, frequently seen in biology.

G_1 Phase Kinases, Cyclins and Inhibitors

We will outline the main steps of growth activation and control in G_1 phase, but this process is too complex to describe here fully (**Figure 2**) (see Murray and Hunt; Andreef). In summary, a biochemical network regulates the critical process of controlling cell growth during G_1 phase. Numerous nutrients including sugars, salts, vitamins and essential amino acids are required for cell growth (Baserga, 1985). Externally supplied growth factors start the cell cycle, from G_0 into G_1 phase. They initiate a multi-step cascade of signals that ultimately

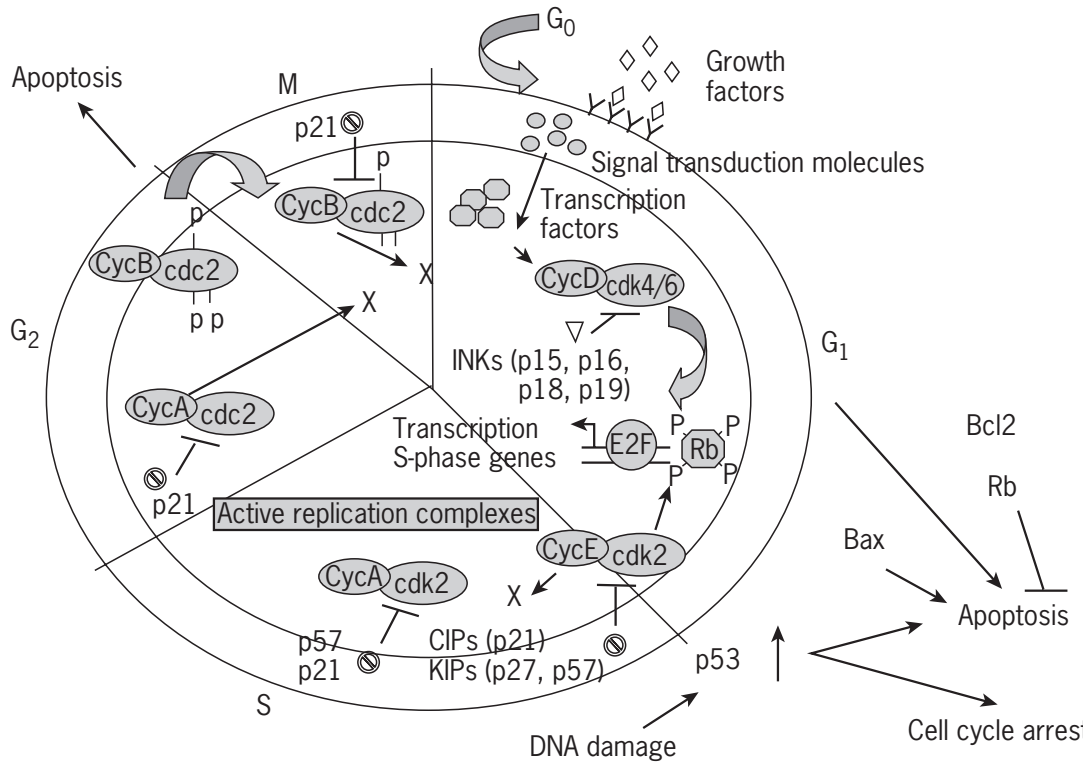


Figure 2 Cell-cycle control molecules. Some of the many molecules that provide growth-regulating signals throughout the cycle are shown, and are discussed throughout this chapter. (Adapted from Ford and Pardee, 1999.)

activates genes to produce their messenger RNAs and proteins, and which culminates in the starting up of DNA synthesis.

The growth factors bind extracellulary to their specific receptor proteins that traverse the membrane surrounding each cell (**Figure 3**). These receptors conduct the external signal to the interior of the cell, and there they activate the receptor's special kinase. These then turn on a cascade of signals involving other proteins including Ras, Fos, Myc and MAP kinases. The Ying–Yang principle is again involved, as illustrated by phosphorylations catalysed by PI-3 kinase that are balanced by dephosphorylations catalysed by the PTEN phosphatase enzyme. The activation of G_1 phase results in expression of at least 100 genes.

The discovery of cyclins, which are the key proteins regulating transition through the cycle (Roberts, 1999), was soon followed by discoveries of multiple cdks. Their complexes with cyclins catalyse stages of cell cycle progression (**Figure 4**). As cells proceed through the cycle, four major cyclins are produced sequentially (D, E, A and B), and they activate several cyclin-dependent kinases. Central is cyclin D which increases in early to mid G_1 phase and regulates cyclin-dependent kinases cdk4 and cdk6 (Sherr, 1996). Cyclin D/cdk triggers the synthesis of cyclin E in late G_1 phase, which in turn activates cdk2, cyclin A production and DNA synthesis.

Phosphorylations are also regulatory, in addition to the synthesis of cyclins. Yet another kinase, CAK, activates the cyclin-dependent kinases. Furthermore, a major role is played by relocalization of cyclin–cdk to the active nuclear compartment within a cell during the cell cycle.

Further investigations revealed yet other proteins whose role is to block activities by binding to cyclin–cdk complexes. These are a family named inhibitors of kinases (INKs). They counterbalance the cyclin's activation of cdks, to affect cycling, development and tumorigenesis (Sherr, 1996). The inhibitory proteins block cyclin D–cdk activities. p27 blocks cell progression, is high in quiescent cells and decreases during late G_1 to release cdk–cyclin activities. Inhibition of cyclins by the cdk inhibitor p21 has often been demonstrated to be induced under various growth-arresting conditions.

In the next step, activated cdks phosphorylate proteins that are essential for progression of the cell cycle. The retinoblastoma tumour suppressor (pRb), absent in retinoblastomas, releases a gene-activating protein named E2F when it is phosphorylated. If this is prevented, E2F is not active, cyclin E is not synthesized, and cells cannot pass through the R point. Additionally, proteasomes' activity of destruction of key inhibitory proteins is vital for passing each checkpoint in the cycle (Koepp *et al*., 1999). The proteasome is a biochemical machine, composed of protein subunits, that chews up proteins including cyclins

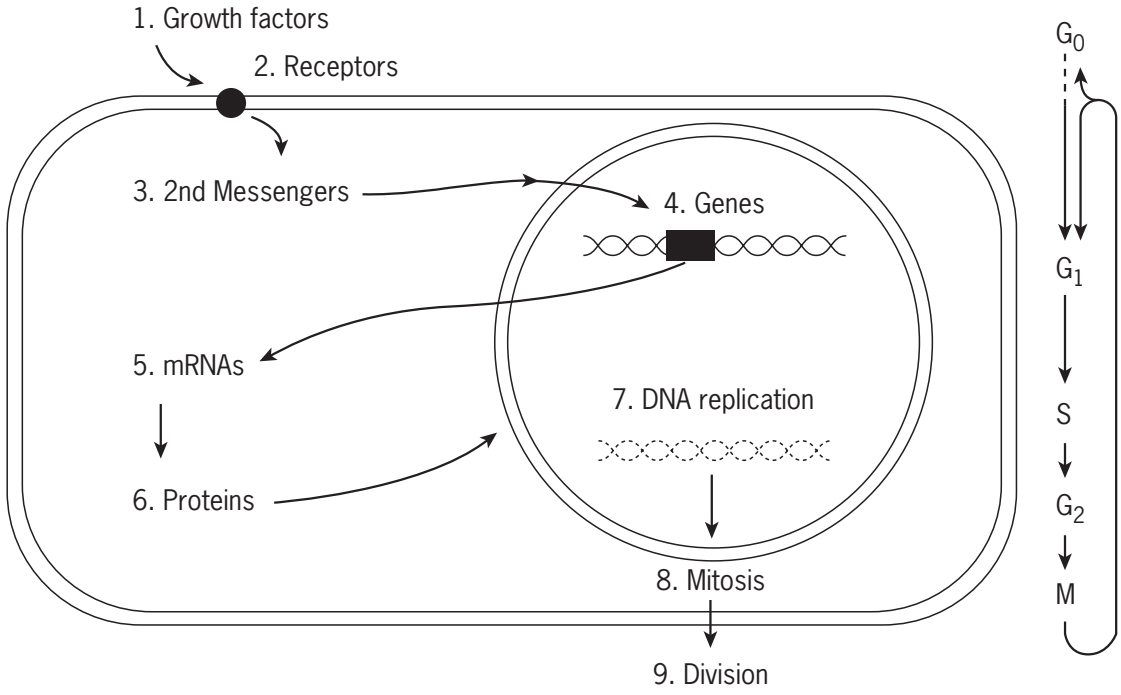


Figure 3 The path to cell proliferation. Growth factors initiate a signalling cascade that takes a cell through the cycle, indicated at the right.

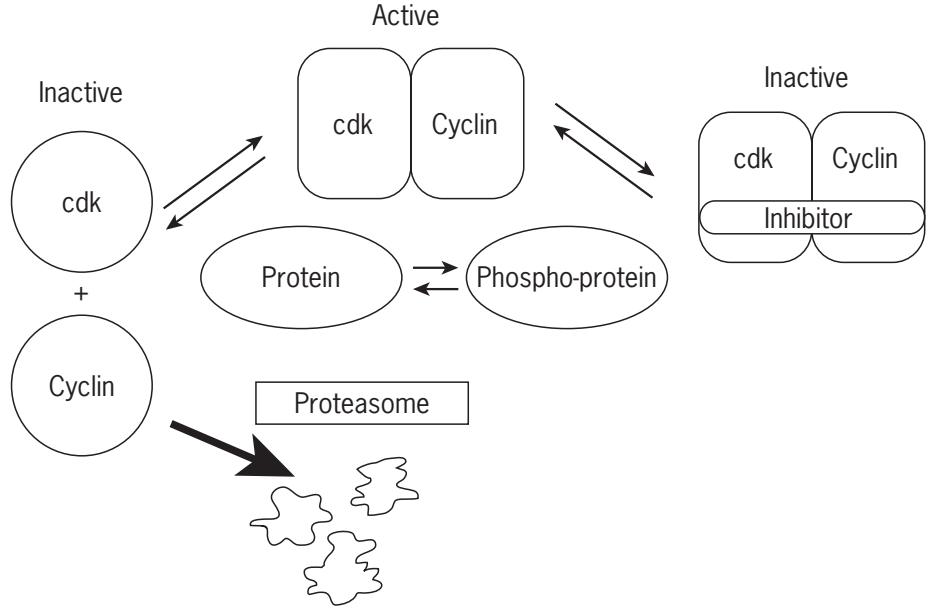


Figure 4 Interactions of cdk with cyclin and inhibitory protein. A cyclin-dependent kinase (cdk) acquires the enzymatic activity to phosphorylate substrate proteins when it binds a cyclin, and this protein complex is inactivated by binding of an inhibitory protein. These reactions are readily reversible and they depend upon phosphorylation of cdk. Cyclins are irreversibly degraded by proteasomes after they are no longer needed.

after they become chemically labelled and targeted for removal.

Unlike peptide growth factors, steroid hormones do not initiate cytoplasmic signalling pathways, but move directly to the nucleus where they activate genes. The sex hormone oestrogen binds to its receptor protein in the nucleus of breast cells, and this in turn binds to and activates growth-stimulating target genes.

Entry into S Phase

Increased cyclin D and E overcome inhibition of cdk activity, and pRb is phosphorylated. This releases E2F and activates genes involved in initiating S phase, including enzymes of DNA synthesis. An example is DNA polymerase, whose transcription is regulated at G_1/S phase by a complex of proteins that contains pRb-like p107, cyclin A and kinase. It is worth noting that most major cell cycle processes are catalysed by large complexes composed of many proteins.

Progression through S phase depends upon cyclin A kinase. Early in S phase, cyclins D and E are degraded by proteasomes. Degradation also removes E2F, which is necessary to prevent programmed cell death (apoptosis) of S phase cells (Lees and Weinberg, 1999).

G_2 Phase and Entry into Mitosis

Mitosis depends upon completion of S phase, and events in G_2 phase are preparatory for it. The complex molecular basis for onset of mitosis was explosively discovered in the early 1980s. Ruderman found that fertilization of oocytes triggers activation of cyclin mRNAs. Hunt discovered that the amounts of cyclin proteins oscillate during the cell cycle, rising during DNA replication and early mitosis and falling at the end of mitosis. Injection of isolated cyclin A into quiescent oocytes drove the cells into M phase. At this time also, Nurse identified the cdk kinase cdc2 as essential for entry into M phase. Unbound cdc2 by itself was inactive. Newly formed cyclin B was shown to bind to and activate cdc2, establishing the first molecular mechanism to explain cell-cycle progression. Then destruction of cyclin B, involving a specialized multi-subunit anaphase-promoting complex, is essential for completion of the cycle.

Research with cell free systems has permitted detailed biochemical investigations of mitosis, showing for example that cyclin B binds to cdc2 (**Figure 2**). This activation of cdc2 kinase is necessary for progression into and through mitosis. The kinase is regulated by a variety of proteins that include cyclin B, phosphatases and kinases and by its subcellular localization. Cyclin B1 begins to accumulate in S phase and increases through G_2. It forms a complex with cdc2, which primes cdc2 phosphorylation. The complex is, however, still inactive, owing to other phosphorylations on cdc2. During G_2 phase, a kinase's (wee1) activity is greater than that of the phosphatase cdc25, and this imbalance keeps cyclin B–cdc2 inactive. At the G_2–M boundary, wee1 is degraded, allowing cdc25 to activate the complex. Furthermore, during G_2 the cyclin B–cdc2 complex resides in the cytoplasm, and at the G_2–M boundary it is rapidly relocated to the nucleus, where it phosphorylates the nuclear membrane protein laminin, which causes the nuclear membrane to break down. Thereafter, chromosomes condense and mitosis proceeds. These many phosphorylations are important for the massive morphological changes that are necessary for a cell to divide.

The Next Cycle – Licensing for DNA Synthesis

A process named licensing permits only one DNA replication per cycle. DNA synthesis cannot be reinitiated until after mitosis is completed. pRb is a critical determinant in preventing DNA reduplication. Perhaps related is the breakdown and reformation during mitosis of the membrane around the nucleus. This permits interaction of molecules from the nucleus and cytoplasm. Degradation of cyclin B by proteasomes is necessary for the start of S phase in the following cycle. Licensing can be disrupted: cells that have lost the cdk inhibitor p21 undergo multiple rounds of DNA synthesis without mitosis, and this process is also activated by anticancer agents. Staurosporin can eliminate the dependence of DNA synthesis on the prior M phase.

Cell Ageing

The normal cell cycle outlined above is modified by various conditions. One of these is cell age. The cycle in early embryo cells is very rapid. It lacks G_1 phase and the corresponding growth-controlling G_1 checkpoint. Mature human cells slow their cycle as they become older, and they cease growing, in G0 or G_1 phase, after about 50 cycles, as initially shown by Hayflick (Baserga). A cdk inhibitor was first discovered in ageing cells by its increase before final arrest of cycling. A progressive shortening of the telomeric DNA, located at the ends of chromosomes, after each cycle is proposed to provide a biological 'clock' for cell ageing. (See chapter on *Telomerase*.)

REGULATION OF CYCLE PHASES

Checkpoints

Entry into and exit from S and M phases are very carefully regulated events. Checkpoint is a name given (Hartwell and Kastan, 1994) to the set of identified cycle-regulatory steps: G_1 restriction point (and the similar START in yeast) and the G_1/S and G_2/M blocks resulting from DNA

damage. Cell-cycle checkpoints are based upon pathways and feedback mechanisms ensuring that a phase of cell cycle does not begin until the preceding phase has been completed with high fidelity. If a checkpoint fails, programmed cell death (apoptosis) or genomic instability ensues. Such failures are important steps in the progression from normal to cancerous cells.

A surveillance system is engaged to make the choice between cell growth and quiescence (Pardee, 1989). When extracellular stimulation by growth factors or nutrients is inadequate, cells cannot pass beyond a specific point in late G_1 phase, in mammalian cells named the restriction point (R). Instead they revert to quiescence (G_0). The final steps that are needed to pass R require synthesis of an unstable protein, proposed to be cyclin E. Under inadequate conditions, synthesis does not keep up with loss, and so this protein cannot be accumulated to an amount sufficient to move the cell into S phase. Once beyond the R point, cells are committed to divide and they no longer require the extracellular growth factors during the remainder of the cell cycle. Restriction point control is defective in cancer cells, and this independence releases cancer cells to continue growing under conditions that keep normal cells in the quiescent state (Pardee, 1989).

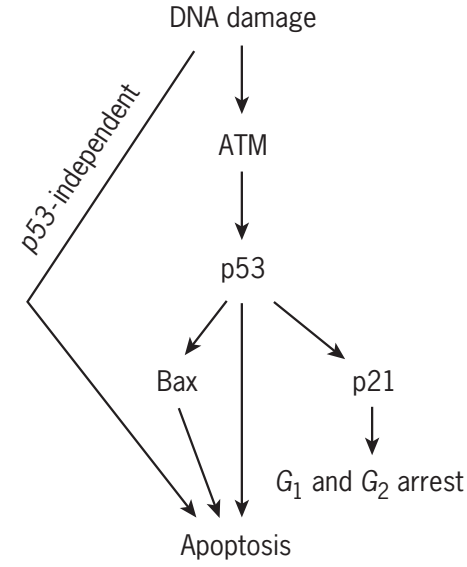


Figure 5 A molecular sequence from DNA damage to apoptosis. A variety of conditions that make a cell unnecessary, such as irreversible damage to it, initiate p53-dependent and independent signalling pathways that lead to apoptosis.

The DNA Damage-induced G_1 Checkpoint

After DNA is damaged, other checkpoint controls delay entry into the next phase of the cell cycle. One such major checkpoint is at the G_1 to S transition, which prevents cells from beginning DNA synthesis until the damaged DNA is repaired. Several proteins, in particular p53, have been implicated in this checkpoint mechanism (**Figure 5**). Individuals who are mutated in the ataxia telangiectasia gene, *ATM*, are very sensitive to X-rays and have a high incidence of tumorigenesis. In response to DNA damage, ATM phosphorylates and increases the level of the p53 protein, a tumour suppressor that is mutated in more than 50% of cancers (Levine, 1997). p53 causes cells to arrest at the G_1–S boundary, which is at least partly due to its production of p21, one of the proteins that inhibits cyclin–cdk complexes.

The DNA Damage-induced S and G_2–M Checkpoints

Within several minutes of exposure to DNA-damaging agents, such as X-rays, mammalian cells in S phase exhibit a dose-dependent reduction in DNA synthesis. Less is known about the mechanism of this S phase checkpoint than about those in G_1 and G_2.

DNA damage also induces a G_2–M checkpoint, as described by Tolmach. This checkpoint delay gives time for DNA repair before the cell goes through mitosis. If repair is not completed in this interval, the cells progress into mitosis without repairing all the DNA damage (Fingert *et al*., 1988), and this results in death or mutations of surviving daughter cells which can thereby become cancerous. This molecular G_2–M checkpoint mechanism is a complex network of phosphorylations and dephosphorylations catalysed by several enzymes and proteins that are moved between cytoplasm and nucleus. Basically, a block in activation of cyclin B–cdc2 prevents the movement of cells into mitosis.

M phase Checkpoints

Mitosis properly segregates chromosomes into the daughter cells. Accurate segregation depends on proper chromosome alignments on and attachment to the mitotic spindle, which is composed of microtubule proteins. A checkpoint ensures that this segregation process occurs correctly. As little as one double strand break in DNA, or depletion of deoxynucleotide building blocks, activates the checkpoint control and stops cells at the G_2–M boundary. This control mechanism delays completion of mitosis until all the chromosomes are attached to the mitotic spindle. The mechanism blocks progression through mitosis if chromosomes are misaligned, and assembly of the microtubules that guide the chromosomes can be inhibited by anticancer drugs such as taxol. Mutations of mitotic checkpoint genes are found in human cancers.

Checkpoints and Programmed Cell Death (Apoptosis)

Apoptosis is a highly regulated process that eliminates physiologically unneeded cells and those that are damaged beyond repair (see the chapter *Apoptosis*). Activated checkpoints give time for a cell to repair its damaged DNA, but if the damage is not soon corrected the cells will initiate apoptosis. This mechanism therefore may prevent the mutations that cause cancer (Sellers and Fisher, 1999). Checkpoint genes, including *p53*, called 'the guardian of the genome,' are involved in causing apoptosis, as is *bax* and other members of the *bcl-2* family (**Figure 5**). Different cells show various responses to damage and drugs, partly because they express different members of the Bcl-2 family (see the chapter *Apoptosis*).

The cyclin A–kinase complex necessary for S phase progression is inhibited when cells are treated with X-rays, and this can result in apoptosis because of the inability of this complex to remove the apoptotic G_1–S factor E2F (Lees and Weinberg, 1999).

Cell Ageing and the Cycle

The elimination of cell ageing is named immortalization. It is an important step in cancer progression, although it does not cause cancer-associated changes (see Hanahan and Weinberg). One way in which human cells can be immortalized is by inserting the gene for the enzyme telomerase, which restores the ageing cell's telomere lengths. Telomerase is also involved in the G_2–M checkpoint.

Cancer is a major cause of death in the elderly. Its incidence increases rapidly, killing about 10% of people between ages 75 and 85 versus 1% between ages 45 and 55. Yet in spite of these epidemiological facts, there is 60% under-representation of cancer patients 65 years or older in treatment trials, few studies have specifically focused on persons over age 65 and many pathological and molecular investigations do not include age as a determinant variable.

CHECKPOINTS, MUTATIONS AND CANCER

The general sequential organisation and duration of the cycle are preserved in cancer, but checkpoint controls are defective (Pardee, 1989; Hartwell and Kastan, 1994). Modifications in cancers are found at many levels of growth regulation, some of which have already been mentioned. The main defect is misregulation of growth initiation at the R point. Furthermore, since checkpoints ensure that mutations are kept low in normal cells, defective checkpoints increase the mutation rate in cancer cells and result in progressive loss of control and emergence of neoplastic disease.

Mutations are causal for cancer; the disease is based upon them. Mutations are found in many genes in advanced cancers. Some of these change cell-cycle controls, including creating a supply of nutrients through angiogenesis, modulating DNA repair, apoptosis, immortalization and metastatic capability (see Hanahan and Weinberg).

The minority of cancer-prone mutations are hereditary. In these cases, a mutated gene on one of a pair of chromosomes is inherited. If, later in life, a mutation occurs of this gene in its partner chromosome, a cancer cell can be produced. Several inherited diseases that are associated with cancer susceptibility have defective checkpoint control. Li–Fraumeni syndrome is a hereditary disease characterized by cancers arising in close relatives. It is a result of a germline mutation in the *p53* gene that abrogates the G_1 checkpoint. Ataxia telangiectasia is characterized by acute cancer predisposition and also other major dysfunctions. Cells from AT patients in culture exhibit severely impaired G_1, S and G_2 checkpoint functions. As mentioned, the *ATM* gene is activated in response to DNA damage and is necessary for activating p53. Another cancer, retinoblastoma, involves mutations in the *Rb* gene, and produces childhood retinal tumours. Survivors have a high risk of developing secondary cancers, particularly osteosarcoma. The BRCA-1 and -2 mutations are associated with hereditary breast cancer; they modify cell cycling and DNA repair. Several other genetic diseases, including Bloom's syndrome, Fanconi anaemia and Nijmegen breakage syndrome, are associated with defects in cell cycle checkpoints and cancer susceptibility.

The majority of cancer-related mutations arise throughout life. For example, cyclin A levels often become abnormally high in cancer cells, and contribute to tumorigenesis. The cyclin D1 and E genes are amplified and over-expressed in many human cancers.

Carcinogenesis can also be caused by viruses such as SV40 and papillomavirus. They introduce their genes that produce proteins that bind to and eliminate the functions of p53 and Rb, thereby bypassing G_1–S, and to a lesser extent G_2, checkpoint controls.

Cancers are often associated with environmental mutagens, such as are produced by smoking. Repeated exposures can produce the several different mutations that are required to cause a cancer. Master mutations can activate growth-promoting oncogenes or loss or inactivation of the tumour-suppressor genes that limit growth. As an example, many cancers have lost or mutated the *p53* gene. One consequence of this mutation is survival of the cancer cell, because p53-dependent checkpoints are eliminated and the programmed cell death mechanism is diminished. Another consequence is that the mutation rate is increased, termed genomic instability (see the chapter *Genomic Instability and DNA Repair*).

The mutations of a half-dozen or more cellular genes is required for tumour formation (Kinzler and Vogelstein, 1996). This number of events is very unlikely in normal cells, whose rate of mutation is approximately 10^{-7} per gene

per cell duplication. Therefore, mutations of genes that increase the overall mutation rate are frequent early events in tumorigenesis. Defects of checkpoint controls in cancers, including hereditary ones, create mutations which are likely to be misrepaired and are progressive because of error-prone repair mechanisms in cancer cells. For example, mutations of BRCA1 or BRCA2 cause *p53* abnormality, which leads to breast cancer (Tseng *et al*., 1997). Other good examples are mutations in colon cancer that cause defective repair of damaged DNA and thereby create genomic instability.

Substances that modify checkpoint controls can change the rate of appearance of mutations, and therefore the progression of cancer. For cells in tissue culture, very high doses of caffeine or related compounds bypass the G_2–M checkpoint, and as a consequence most damaged cells die. These results demonstrate the protective role of the G_2–M checkpoint against damage-induced chromosomal aberrations (Fingert *et al*., 1988). However, chromosomal abnormalities may appear in the few surviving cells.

CANCER THERAPY AND THE CYCLE

Classical Chemotherapies

Currently applied therapies are aimed at killing cancer cells with cytotoxic agents that are applied in combinations. They can prolong the lives of patient with some kinds of cancer but have little effect against others, and all too often the cancer reappears within a few years. One drug provided alone is generally ineffective, because some cancer cells survive this treatment and so the cancer reappears. Multiple drugs are necessary for effectiveness, but this multi-targeting is limited by toxicity to normal cells.

Molecular differences between cancer and normal cells are subtle (see Hanahan and Weinberg). They are mainly related to defective controls of cell growth and survival. Many clinically applied drugs preferentially kill the cycling cancer cells relative to the generally quiescent normal cells, which are essential for survival of the individual. But some kinds of normal cells are cycling, and so the drugs are toxic to the patient. Several cell cycle events provide targets for therapy (**Figure 6**). (See the section on *The Treatment of Human Cancer*.)

Antagonists of Growth Factors

Sex hormones stimulate the growth of some breast, ovarian and prostate cancer cells. Blocking these hormones' action can kill these cells. Tamoxifen is chemically related to the sex hormone oestrogen, with which it competes for binding to oestrogen receptors in a cell. Since tamoxifen blocks the stimulation by oestrogen and does not activate growth, it is in fact inhibitory.

Blocking S Phase

Cornerstones of standard chemotherapy are inhibitors of DNA synthesis. These are small molecule antagonists

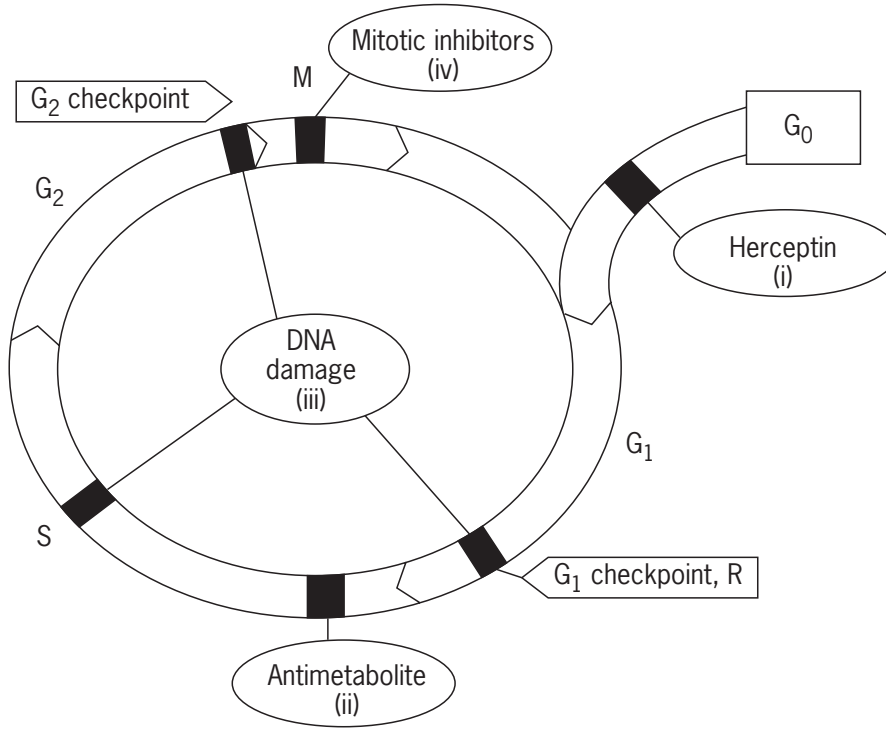


Figure 6 Current therapeutic approaches. Various therapeutic methods that are discussed in the text are summarized here.

structurally similar to metabolic compounds required in the synthesis of DNA and cell survival. As examples, fluorouracil is structurally very similar to uracil, which is needed for DNA synthesis, and methotrexate is an analogue of the vitamin folic acid, also essential for DNA synthesis.

DNA-damaging Agents

Agents that damage DNA are lethal. Examples are X-rays and clinically applied alkylating compounds such as cytoxan and cisplatin. They are more effective against cancer versus noncancer cells because the latter generally are not growing. Also, the normal cells more effectively repair damage during checkpoint delays, before the lethal event of passage of the damaged cell through mitosis and consequent partitioning of damaged DNA between the daughter cells.

Mitotic Inhibitors

Several clinically applied drugs upset the mitotic mechanism in cycling cells, and thereby are lethal. These include taxol from the yew tree and alkaloid toxins from the vinca plant. These currently used compounds, and also experimental epothilones derived from microorganisms, cause lethal mitotic arrest of cycling cells. Their targets are the microtubule proteins, which guide chromosomes through their mitotic separation. Purified plant and microbial products such as these very frequently are starting points for finding anticancer drugs.

A novel drug such as epothilone enters the clinic every few years. Another current example is the antimetabolite gemcytabine, which during S phase is incorporated into newly forming DNA where it arrests continuation of lengthening of the molecule. Difficulties in introducing novel drugs have roots not only in drug discovery, but to a great extent in complex legal requirements for meeting safety standards. These require extremely extensive clinical trials, which with the many costs of doing business, require hundreds of millions of dollars to develop one drug.

Cycle Activators as New Targets

Discovery in cancer cells of over-activated growth-signalling pathways provides possibilities for chemotherapy at every step. Drugs targeted against these reactions are being applied clinically and are in clinical trials.

Tuning Down External Stimuli

Some tumours secrete self-stimulating growth factors into their environment, which also can affect nearby cells. A fascinating example is stimulation by a tumour of the production of new blood vessels, angiogenesis. This process creates the blood supply essential for nourishment of the tumour. Secretion by the tumour of a growth factor VEGF stimulates this production of blood vessel cells, and of blood vessels from their assembly. Both antiangiogenic drugs and also antibodies that neutralize VEGF are being developed as anticancer agents (Boehm-Viswanathan, 2000).

About one third of breast tumour cells lack oestrogen receptors. Unlike those discussed above these are not stimulated by oestrogen or inhibited by tamoxifen, and so they are generally treated with classical anticancer agents following surgery. They are stimulated to grow by EGF, because too many EGF family receptors are on their surface. The monoclonal antibody (herceptin) made against these receptors is effective against some of these cancers, especially when applied in combination with the drugs taxol and doxorubicin.

Targets in the G_1 Phase Signal Transduction Pathway

The molecules that transmit growth signals from a cell's membrane receptors to its nucleus during G_1 phase (**Figure 2**) provide numerous targets for cancer treatment, now under investigation (Adams and Kaelin, 1998; Kaelin, 1999). One major participant is Ras, a small protein that must be positioned against the inner surface of the cell membrane to interact with growth factor receptors. Enzymes must chemically modify Ras for it to occupy this position, and so drugs are being developed that prevent this modification and thereby block the signalling pathway.

Signalling events require numerous kinases that modify the activities of other proteins by addition of phosphates to them. Kinase inhibitors can arrest cell growth and cause death of tumour cells (Shapiro and Harper, 1999). Specific inhibitors of critical cyclin-dependent kinases are being developed. An inhibitor has already demonstrated high efficacy in the treatment of chronic myelogenous leukaemia, a malignancy characterized by the activation of Abl kinase (Drucker and Lydon, 2000).

S Phase Lethality

A cell initiates a sequence of molecular events culminating in apoptosis during S phase unless certain molecules that initiated DNA synthesis are first inactivated. Applying a molecular analogue of part of the G_1–S factor E2F blocks the degradation of E2F and causes apoptosis (Lees and Weinberg, 1999).

Modulating Checkpoints

Several novel potential therapies are being developed (**Figure 7**).

Mitotic catastrophe

The loss of G_2 cell cycle checkpoints can increase tumour-cell sensitivity to chemotherapy. Furthermore, these cells often cannot take refuge at the G_1 checkpoint owing to the loss of p53 or other G_1 checkpoint molecules, whereas

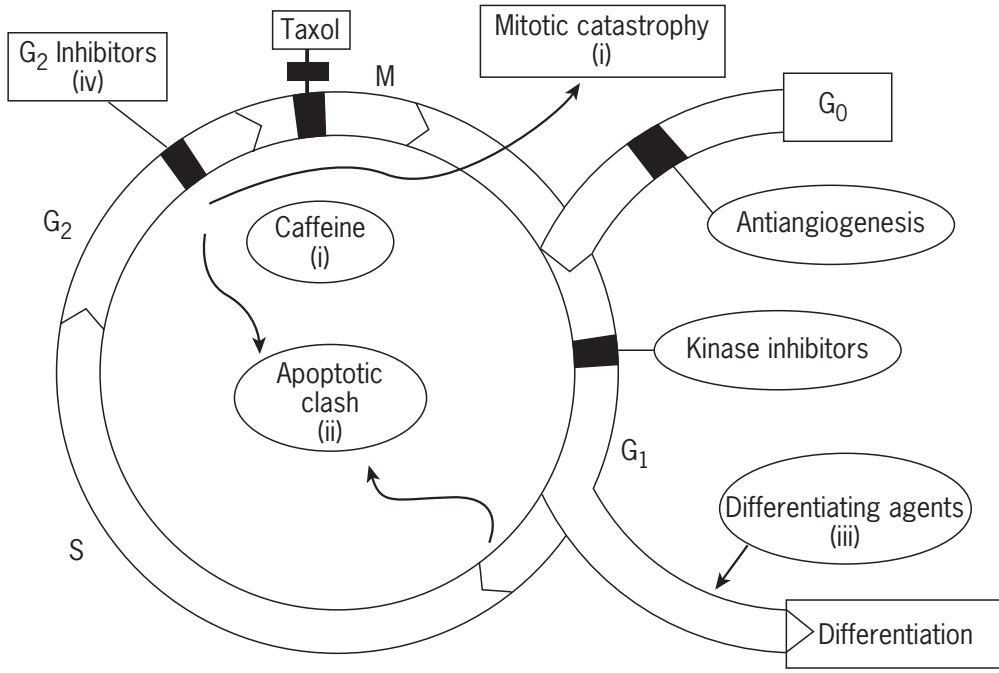


Figure 7 Potential therapies. Several potential therapies described in the text are illustrated.

cells with the normal checkpoint may still recover. Following DNA damage and the arrest at the G_2 checkpoint, some compounds can cause cycling cancer cells to move on through a lethal mitotic cell division. This process was recently aptly renamed 'mitotic catastrophe'. Post-treatment with a caffeine analogue enhanced cytotoxicity of drugs to cancer cells implanted in mice, but not to the mice (Fingert *et al*., 1988). These agents, however, proved to be toxic to humans at doses that abrogate the G_2 checkpoint, which limits their therapeutic use. Other inhibitors that can eliminate the G_2 checkpoint such as the inhibitors of cyclin-dependent kinases, flavopiridol and UCN-0101, are currently undergoing clinical trials (Shapiro and Harper, 1999).

Clash Hypothesis

Remarkably effective synergistic killing of a variety of human cancer cells was found with the combined application together of two molecules derived from plant sources, β-lapachone and taxol (Li *et al*., 1999). These combined drugs killed nearly all of several kinds of cancer cells in culture, at concentrations that did not show major lethality when the drugs were applied singly. Several kinds of human cancer cells growing in mice were destroyed when the drugs were applied together but not if they were applied separately. The tumours did not reappear for at least 2 months. Very importantly, the mice showed no signs of toxicity under these conditions, so there was a very high therapeutic index. Hence there are strong indications of clinical utility of this drug combination.

The mechanism of this powerful synergistic and tumour-specific lethality is being investigated. One hypothesis is based upon the proposal of 'clashing' checkpoint signals; apoptosis is caused by the production of two simultaneous molecular checkpoint signals created by growth conditions (Evan and Littlewood, 1998; Blagosklonny, 1999). Since β-lapachone causes G_1 arrest whereas taxol gives a G_2–M arrest, such clashes of conflicting molecular signals might also be created by this combination of drugs. They could selectively cause apoptotic death of cancer cells which already have defective checkpoint and apoptotic mechanisms.

Differentiation Therapy

An alternative to toxic cancer therapies is to restore the normal cell's properties. Cells in the blood are created by growth and then differentiation of precursor stem cells, followed by R point arrest of the matured cells and their eventual death. Leukaemic cells are mutated blood cells that do not undergo this terminal differentiation and death, and instead they continue to proliferate. Drugs have been discovered that recreate this differentiation-growth arrest process. Retinoids (vitamin A derivatives) are used in this way to treat promyelocytic leukaemia.

Newer drugs such as Saha show potential to inhibit leukaemias through differentiation therapy (Richon *et al*., 1999). This approach is being generalized to make solid tumours differentiate as well. The drugs function by turning on differentiation-related genes, a process that is activated by addition of acetyl groups to the histone

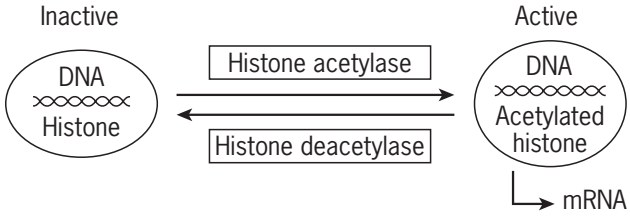


Figure 8 Control by histone acetylation. Histones that surround the DNA in chromatin block activities of genes. Acetylation of the histones permits them to produce their messenger RNAs.

proteins associated with DNA in chromatin (**Figure 8**). This process is catalysed by the enzyme histone acetylase, and it is reversed by deacetylase enzymes, which produce an inactive structure. These changes from the acetylated to the deacetylated state function as an on–off switch for regulation of gene expression. The differentiating drugs shift this balance by blocking deacetylase, thereby the activating acetylation dominates. Thus, drugs that permit histone acetylation, or that decrease the closely connected DNA methylation, are approaches for re-expression of tumour-suppressor genes such as *BRCA1*, *p16* and *p21* that are silenced in cancers by these processes.

Selective Protection of Normal Cells

Chemotherapy of cancer is limited by toxicity to normal cells. With traditional chemotherapy, dose-limiting side effects emerge, including toxicity to bone marrow and gastrointestinal tract, dermatological toxicity and cardiotoxicity. Therefore, selective protection of normal cells against chemotherapeutic drugs could improve the therapeutic index (the ratio of doses that affect cancer versus host), permitting the application of higher drug concentrations (see Blagosklonny and Pardee). Defective checkpoint mechanisms in cancer cells can be the basis of such a selective survival of normal cells (**Figure 6**).

Until recently, the mainstream approach for cancer treatment was directed to finding synergistic combinations in which all the drugs are toxic against cancer cells, and so combinations with an independently inactive drug were considered inappropriate. However, a high therapeutic index, with less toxicity to normal cells, was found in clinical trials when drugs that blocked entry of normal cells into M phase were administered before subsequently adding taxol. This prevented taxol lethality in M phase. This antagonism was translated in the clinic as a decrease in side effects to normal cells.

G_1 checkpoint arrest in normal but not in cancer cells is produced by low, nonlethal concentrations of compounds such as cycloheximide that slow protein synthesis. Later addition of a toxic S-phase-specific agent cannot kill them. In contrast, independence of cancer cells from this cycle arrest causes them to enter S phase, where they are killed while they are synthesizing DNA. This idea, proposed in 1975, has been revived in a new form. Low doses of doxorubicin or etoposide induce p53- and p21-dependent growth arrest of normal cells without cell death, but these drugs do not arrest cancer cells. This pretreatment thereby abolished the cytotoxicity otherwise caused by later addition of microtubule-active drugs (paclitaxel, vincristine, epothilones). Protection of cells with normal checkpoint was achieved, whereas no protection was observed in cancer cells lacking p53 or p21 (see Blagosklonny and Pardee).

Novel inhibitors of the cell cycle are being developed as lethal drugs against solid cancers and leukaemia cells. However, these are active also against normal cells; proliferating bone marrow and epithelial cells are particularly vulnerable. Thus, searches for compounds that reversibly inhibit proliferation of these normal cells will be especially valuable to protect the individual. For example, although two compounds similarly inhibited protein kinase C, UCN-01 was selected as the drug to develop because it had higher cytotoxicity to cancer. For a selective growth arrest of normal cells, dependent on protein kinase C, one would choose the less toxic inhibitor GF109203X. Other strategies utilizing the retention of checkpoints in normal cells to protect them versus tumour cells are discussed (see Blagosklonny and Pardee). Since proliferation of normal cells is highly regulated, the search for such inhibitors should produce surprises.

In summary, defects in cancers of various molecular mechanisms that control cell growth, differentiation and apoptosis have recently been discovered. These differences from normal cells provide novel targets for therapy, some of which are being developed and tested.

ACKNOWLEDGEMENT

This work was supported by NIH Grant CA RO1 61253.

REFERENCES

Adams, P. D. and Kaelin, W. G., Jr (1998). Negative control elements of the cell cycle in human tumors. *Current Opinion in Cell Biology*, **10**, 791–797.

Blagosklonny, M. V. (1999). A node between proliferation, apoptosis, and growth arrest. *Bioessays*, **21**, 704–709.

Boehm-Viswanathan, T. (2000). Is angiogenesis inhibition the Holy Grail of cancer therapy? *Current Opinion in Oncology*, **12**, 89–94.

Druker, B. J. and Lydon, N. B. (2000). Lessons learned from the development of an Abl tyrosine inhibitor for chronic myelogenous leukemia. *Journal of Clinical Investigation*, **105**, 3–7.

Evan, G. and Littlewood, T. (1998). A matter of life and cell death. *Science*, **281**, 1317–1322.

Fingert, H. J., *et al*. (1988). *In vivo* and *in vitro* enhanced antitumor effects by pentoxifylline in human cancer cells treated with thiotepa. *Cancer Research*, **48**, 4375–4381.

Hartwell, L. H. and Kastan, M. B. (1994). Cell cycle control and cancer. *Science*, **266**, 1821–1828.

Howard, A. and Pelc, S. R. (1951). Nuclear Incorporation of P32 as demonstrated by autoradiography. *Experimental Cell Research*, **2**, 178–187.

Kaelin, W. G., Jr (1999). Choosing anticancer drug targets in the postgenomic era. *Journal of Clinical Investigation*, **104**, 1503–1506.

Kinzler, K. W. and Vogelstein, B. (1996). Lessons from hereditary colorectal cancer. *Cell*, **87**, 159–170.

Koepp, D. M., *et al*. (1999). How the cyclin became a cyclin: regulated proteolysis in the cell cycle. *Cell*, **97**, 431–434.

Lees, J. A. and Weinberg, R. A. (1999). Tossing monkey wrenches into the clock: new ways of treating cancer. *Proceedings of the National Academy of Sciences of the USA*, **96**, 4221–4223.

Levine, A. J. (1997). p53, the cellular gatekeeper for growth and division. *Cell*, **88**, 323–331.

Li, C. J., *et al*. (1999). Potent inhibition of tumor survival *in vivo* by β-lapachone plus taxol: combining drugs imposes different artifical checkpoints. *Proceedings of the National Academy of Sciences of the USA*, **96**, 13369–13374.

Minshull, J., *et al*. (1989). The role of cyclin synthesis, modification and destruction in the control of cell division. *Journal of Cell Science, Supplement*, **12**, 77–97.

Nurse, P., *et al*. (1998). Understanding the cell cycle. *Nature Medicine*, **4**, 11030–1106.

Pardee, A. B. (1989). G_1 events and regulation of cell proliferation. *Science*, **246**, 603–608.

Richon, V. M., *et al*. (1999). A class of hybrid polar inducers of transformed cell differentiation inhibits histone deactylase. *Proceedings of the National Academy of Sciences of the USA*, **95**, 3003–3007.

Roberts, J. M. (1999). Evolving ideas about cyclins. *Cell*, **97**, 129–132.

Shapiro, G. I. and Harper, J. W. (1999). Anticancer drug targets: cell cycle and checkpoint control. *Journal of Clinical Investigation*, **104**, 1645–1653.

Sherr, C. J. (1996). Cancer cell cycles. *Science*, **274**, 1672–1677.

Tseng, S. L., *et al*. (1997). Allelic loss at BRCA1, BRCA2, and adjacent loci in relation to TP53 abnormality in breast cancer. *Genes Chromosomes Cancer*, **20**, 377–382.

FURTHER READING

Andreeff, M., *et al*. (2000). Cell proliferation, differentiation and apoptosis. In: Holland, J., *et al*. (eds), *Cancer Medicine*, 5th edn. 17–32. (B. C. Decker, Hamilton, Ontario).

Baserga, R. (1985). *The Biology of Cell Reproduction.* (Harvard University Press, Cambridge, MA).

Blagosklonny, M. V. and Pardee, A. B. (2001). Exploiting cancer cell cycling for selective protection of normal cells. *Cancer Research*, **11**, 4301–4305.

Ford, H. L. and Pardee, A. B. (1999). Cancer and the cell cycle. *Journal of Cellular Biochemistry*, **75**, 166–172.

Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, **100**, 57–70.

Murray, A. W. and Hunt, T. (1993). *The Cell Cycle, an Introduction.* (Freeman, San Francisco).

Sellers, W. R. and Fisher, D. E. (1999). Apoptosis and cancer drug targeting. *Journal of Clinical Investigation*, **104**, 1655–1661.

Chapter 3

Overview of Oncogenesis

Julie L. Boerner, Jacqueline S. Biscardi and Sarah J. Parsons
University of Virginia, Charlottesville, VA, USA

CONTENTS

PROPERTIES OF NEOPLASTIC CELLS

Normal cells are exquisitely attuned to their environment and respond to external cues via tightly regulated signalling pathways that either trigger or repress growth. In order for a cell to undergo mitogenesis, a growth-promoting signal from the extracellular environment must first initiate a cascade of events within the cell that results in activation of genes that stimulate cell division. With few exceptions, most of the cell populations within an adult organism are terminally differentiated and no longer proliferate. Cancer arises when a cell, for a variety of reasons, escapes the normal constraints placed on its growth and begins to divide in an unregulated fashion.

Factors that Promote Growth

Extracellular factors that stimulate growth include peptide growth factors such as epidermal growth factor (EGF) and platelet-derived growth factor (PDGF), which bind tyrosine kinase receptors located on the cell surface. Cytokines such as growth hormone, interleukins and prolactin also bind cell surface receptors which are not kinases themselves but are able to transduce their signals via interaction with separate tyrosine kinase molecules. Another class of growth factors bind serpentine receptors that couple to intracellular pathways via heterotrimeric G proteins. Lastly, steroid hormones such as oestrogen, which bind intracellular receptors, also have mitotic activity. All classes of receptors are capable of triggering a cascade of signalling events culminating in mitotic activity, or proliferation, of the target cell.

Mutation of Growth Regulatory Genes in Cancer

In tumour cells, molecules that regulate signalling pathways which stimulate growth are often mutated, resulting in a constant 'on' signal to the cell. These molecules can be positive for growth regulation, as cancer-causing oncogenes, or negative for growth regulation, as protective tumour-suppressor genes. In addition, tumour cells often develop their own autocrine loops for growth, wherein a growth factor required for the activation of a pro-growth signalling pathway is constantly produced and secreted into the extracellular milieu. One example of such an autocrine loop is found in breast tumour cells, which are able to produce the growth factor TGF-α. TGF-α binds and activates the EGF receptor, thereby triggering mitotic pathways within the cell.

Cell Death

Apoptosis, or programmed cell death, is yet another process that is subverted by a tumour cell (Jacotot *et al*., 2000). In a normal cell, a series of 'checkpoints' must be met before the cell permits itself to divide. If irreparable damage to its DNA is present, the cell undergoes apoptosis, thus ensuring that its mutated DNA is not transmitted to progeny cells. The molecules that regulate this process of apoptosis are often themselves mutated in cancer cells, which are then able to escape the checks and balances that a normal cell must undergo before it can divide. (See chapter on *Apoptosis*.)

Cell–Cell Interaction

Cells can also respond mitogenically to cues from other cells. Normal cells are growth-inhibited by contact with other cells and form a monolayer when grown in culture. Cancer cells, on the other hand, form foci, or piled-up accumulations of constantly dividing cells; foci result as a consequence of loss of contact inhibition. Molecules called cellular adhesion molecules (CAMs) and cadherins are expressed on the surface of cells and negatively regulate growth. Cadherin molecules on adjacent cells bind one another in a calcium-dependent manner; this binding prevents cells from entering the mitotic cycle (Christofiori

and Semb, 1999). Further intracellular signalling occurs via the catenin family of molecules, which link the cadherins to the cytoskeleton and to the transcription machinery. The negative regulation normally provided by the interaction of these molecules is frequently lost in tumour cells. For example, epithelial cell cadherin (E-cadherin) is mutated, absent, or reduced in expression in a variety of human tumours. In cell culture systems, loss of the E-cadherin gene leads to loss of cell–cell contacts and increases in cell motility and invasiveness. Aberrant phosphorylation of the catenins can lead to loss of proper cell–cell contacts, which is thought to be a step in the acquisition of invasive properties of the cancer cell. Moreover, the *APC* tumour-suppressor gene, which is mutated in human cancer, is known to associate with β-catenin. Cancer-causing mutations in *APC* involve the portion of the molecule that binds β-catenin. Thus, loss of these tumour-suppressor genes and their appropriate interactions with cadherins and catenins relieves the constraints of contact inhibition, a hallmark of tumour cells.

Cell–Substratum Interactions

Pro-growth cues can also come from the extracellular matrix or substratum (such as the basement membrane) on which cells grow (Miyamoto *et al*., 1998). Proteins such as fibronectin, a component of the ECM, bind integrin receptors on the cell surface. The integrin receptors then cooperate with growth factor receptors to trigger mitogenic pathways via activation of signalling cascades involving several different kinase molecules. Alternative, tumour-specific isoforms and unique combinations of integrins are often present in tumour cells, thereby providing additional means by which growth signals can be initiated. (See also chapter *Extracellular Matrix: The Networking Solution*.)

Angiogenesis

Tumours also exhibit extensive vascularization which increases as the tumour grows (Folkman, 1992). This outgrowth of new blood vessels is termed angiogenesis and is not seen in normal adult animals except in the cases of wound healing and pregnancy, where new tissues such as placenta are formed. In the absence of new blood vessels, a tumour is able to grow to a maximum size of approximately 1 mm in diameter, the distance that oxygen and nutrients are able to diffuse into the tumour (Kurschat and Mauch, 2000). Vascularization thus allows the tumour to grow larger. (See also chapter on *Angiogenesis*.)

Migration and Metastasis

Tumour cells often have the ability to migrate away from the original tumour site and grow in distant parts of the body (Kurschat and Mauch, 2000). This ability to metastasize requires that the tumour cells produce proteases that degrade the basement membrane of blood vessels through which the tumour cells will travel. Metastatic tumour cells are also able to escape immune surveillance of the host organism and then to grow again in another part of the body. Factors produced at these distal sites are thought to provide a favourable environment for the continued proliferation of the tumour cells. Evidence for this comes from the finding that specific types of tumours have a propensity to metastasize to the same sites, e.g. prostate cancer commonly targets bone. Moreover, tumour cells also have the ability to affect the underlying layer of stroma, or fibroblast cells, stimulating them to produce growth factors and cytokines that enhance tumour growth. **Figure 1** depicts how a cell breaks away from the primary tumour and generates a secondary neoplasm within the organism. (See also chapter on *Invasion and Metastasis*.)

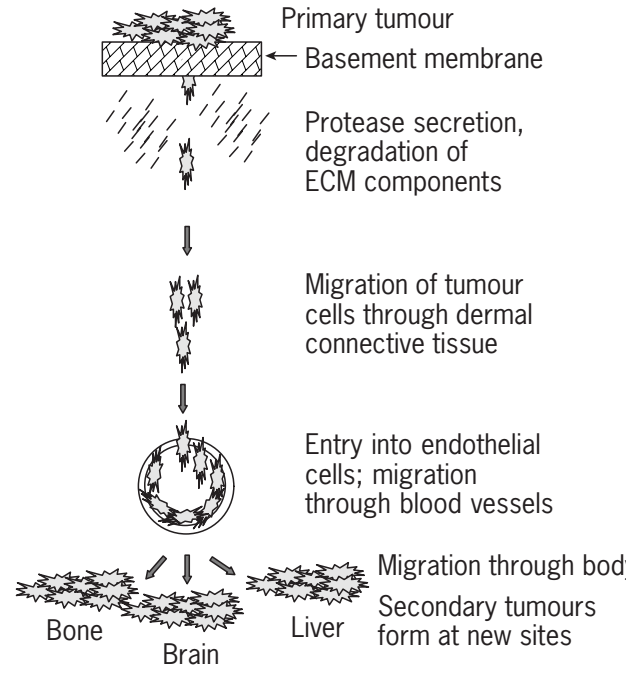


Figure 1 Metastatic progression. Individual cells within the primary tumour upregulate production of specific proteases, which gives the cell the ability to degrade ECM or basement membrane. Cells then break away from the primary tumour and begin to migrate. Migrating cells adhere to connective tissue and invade dermal tissue. Cells thus enter the vascular system by migrating between endothelial cells and moving through blood vessels. Lastly, tumour cells reach secondary sites where conditions are favourable for their continued growth.

Conversion of a Normal Cell to a Tumour Cell

A cell becomes converted from a normal to a neoplastic cancer cell when the regulation of one or more of the above

processes is lost. Loss of regulation occurs when mutations arise in two broad families of genes that regulate growth: oncogenes, which act as positive signals for growth, and tumour-suppressor genes, which act as brakes or checkpoints on a cell's progression through the cell cycle. These mutations may be caused by environmental, chemical or biological agents or events that result in irreversible alterations in the genome of a cell, so that progeny cells also carry the same mutations that allow for uncontrolled growth. This is the first step on a pathway that can eventually lead to an aggressive, metastatic tumour. Fortunately, organisms possess several means of dealing with environmental insults and genetic alterations. More than one genetic 'hit,' or error, is required before an actual tumour is able to arise, as will be discussed later.

ONCOGENES

Historical Perspective

In 1911, Peyton Rous laid the groundwork for the oncogene theory of cancer, a theory that became the basis for all modern cellular signalling and genetic research. He identified a spindle-celled sarcoma in chickens that was transplantable from one bird to another, using a filtrate of the tumour (Rous, 1911; Weiss *et al*., 1985). The infectious agent responsible for the tumours was later found to be the Rous sarcoma virus (RSV). Thus, a cancer causing agent had been discovered, but the means by which the virus induced tumours was still unclear.

Further insight into the process of oncogenesis was provided in 1914, by Theodor Boveri who hypothesized that cells in cancer tissue contain defective chromosomes. Working with double-fertilized sea urchin embryos, Boveri observed that the resulting aberrant chromosomes were passed on to progeny cells (Boveri, 1925). This finding led him to believe that cells of malignant tumours have damaged chromosomes and that a neoplastic cell can arise from a normal cell and pass its altered genome on to progeny cells. Thus, a tumour cell is in some way defective and has lost the properties of a normal cell. Environmental insults were also implicated in causing cancers. In 1918, Yamagiwa and Ichikawa showed that continual irritation of the normal epithelium of rabbit ears caused papilloma-like growths and metastasis.

Building on the early work of Rous, Shope provided further evidence for the viral basis of oncogenesis by his demonstration that a papilloma-like growth was transmissible from animal to animal (Weiss *et al*., 1985). In 1951, Gross showed that mice inoculated with leukaemic extracts developed neoplasms. From these extracts, the Gross murine leukaemia virus was isolated (Weiss *et al*., 1985). Seven years later, Temin and Rubin showed that infection of cultured chicken fibroblasts with the Rous sarcoma virus caused neoplastic transformation of the cells (Weiss *et al*., 1985). Martin and others later identified the oncogenic portion of the RSV genome as v-*src*, the viral *src* oncogene. These early results suggested a transmissible mechanism for tumour initiation. Thus, as early as the beginning of the twentieth century, a cellular/genetic model of oncogenesis had been postulated.

The Oncogene Hypothesis

The best-known theory of oncogenesis, however, is a relatively recent one. In 1982, Bishop and Varmus hypothesized that cancer-causing genes, or oncogenes, that are carried by tumour-inducing viruses have normal counterparts that are present in the genomes of all vertebrate cells (Bishop, 1982). These normal genes are termed proto-oncogenes. Evidence for this hypothesis came from hybridization studies, where radiolabelled v-*src* DNA was found to bind, or hybridize, to its complementary counterpart (c-*src*) in normal avian cellular DNA. The v-*src* and c-*src* genes encode a tyrosine kinase, an enzyme that transfers phosphate from ATP to the amino acid tyrosine found in cellular proteins. These phosphorylations have profound effects on cell growth. Similar studies eventually led to the identification of a family of viral oncogenes, which can be transmitted by either DNA or RNA viruses. DNA viruses either can cause lytic infection leading to the death of the cellular host, or can replicate their DNA along with that of the host genome and promote neoplastic transformation of the cell. DNA viruses encode various proteins which, along with environmental and genetic factors, help to initiate and maintain the neoplastic state. RNA tumour viruses, on the other hand, integrate DNA copies of their RNA genomes into the genome of the host cell. Since the viral genomes contain transforming oncogenes, they induce cancerous transformation of the host cell.

Mechanism of Acquisition of Cellular Sequences by RNA Tumour Viruses

Multiple lines of evidence indicate that viral oncogenes arise when an RNA virus integrates its genome near the coding sequence for a proto-oncogene and incorporates the proto-oncogene's DNA into its own genetic material during the virus replication cycle. Through multiple rounds of infection and genome replication, deletions and other mutations occur in the proto-oncogene, conferring on the gene tumorigenic properties. Ensuing infection of a normal cell by an RNA tumour virus carrying such an oncogene causes malignant transformation of that cell. Although this process rarely occurs in human tumours, many of the same

Table 1 Examples of human oncogenes

Oncogene	Originally identified in	Mechanism of activation in human tumours	Location	Associated human cancers
src	Rous sarcoma virus	Overexpression, C-terminal deletion	Cytoplasmic	Breast, colon, lung carcinomas
myc	Avian myelocytomatosis virus	Translocation	Nuclear	Burkitt lymphoma
abl	Abelson murine leukaemia virus	Translocation	Cytoplasmic	Chronic myeloid leukaemia
Ha-*ras*	Harvey murine sarcoma virus	Point mutation	Cytoplasmic	Bladder cancer
K-*ras*	Kirsten murine sarcoma virus	Point mutation	Cytoplasmic	Colon, lung carcinomas
erbB (EGFR)	Avian erythroblastosis virus	Overexpression, deletion	Cytoplasmic	Breast carcinoma, glioblastoma

genes 'captured' by animal retroviruses are altered in human cancers. These alterations take the form of base pair changes, insertions, amplifications and translocations, which result in a protein product that no longer responds normally to growth-regulatory cues. Only one allele of the gene needs to be mutated for the oncogenic effect. Thus, oncogenes are described as carrying dominant mutations. For example, Ras in human tumours is often found to be mutated at a single amino acid residue (Wittinghofer, 1998), whereas the oncogene *abl* is activated by chromosomal translocation (Heisterkamp *et al*., 1985). In the case of c-Src, a negative regulatory site present in normal c-*src* is mutated in a small subset of colon cancers, thus rendering the protein constitutively active (Irby *et al*., 1999), whereas in human breast cancer, overexpression of the normal c-Src protein appears to play a role in the deregulation of cell growth (Biscardi *et al*., 1999).

Proto-oncogenes can be classified as either cytoplasmic or nuclear, depending on where in the cell they are localized. Many of the cytoplasmic proto-oncogenes code for tyrosine kinase molecules, enzymes that are able to phosphorylate substrate proteins on tyrosine residues and that are known to be essential for controlling the signalling cascades that regulate mitosis. Others, such as Ras, transmit cellular growth signals by binding guanine nucleotides in the form of GTP or GDP. Ras is often found mutated at single sites such that it is constantly bound to GTP, which causes the molecule to be constitutively active. Mutations in Ras are found in approximately 30% of human cancers (Wittinghofer, 1998). Serine–threonine kinases, such as the Raf family of kinases, are the targets of Ras and constitute another family of proto-oncogenes that regulate proliferation. Nuclear oncogenes such as *myc* regulate gene transcription. **Table 1** lists a few examples of the better-known oncogenes, their subcellular localization and mechanism of oncogenic activation. Although these oncogenes are defined as cancer causing genes, it is important to note that the introduction of a single activated oncogene into a cell does not result in neoplastic transformation. At least two active oncogenes, or an activated oncogene and an inactivated tumour suppressor, are required for tumour formation.

TUMOUR SUPPRESSORS

Tumour-suppressor genes are defined as recessive genes, i.e. they must sustain mutations or deletions of both alleles in order to contribute to cancer formation and progression. This definition implies that one functional allele of the tumour-suppressor gene is sufficient for normal cell function. Patients with familial cancers frequently inherit one normal and one abnormal allele of the tumour-suppressor gene from their parents. If the second, normal allele is lost, the protective effect of the gene product no longer exists. Therefore, introduction of a wild-type copy of the gene back into the tumour should inhibit further tumour growth. Unfortunately, putative tumour-suppressor genes shown to be inactivated in cancer are not sufficient by themselves to restore normal cell function. Thus, whether such genes are actually tumour suppressors remains a debated question.

Discovery and Identification

The origin of the concept of tumour-suppressor genes (or anti-oncogenes) came from cell fusion studies dating back to the early 1900s. These studies revealed that when one tumour cell is fused with another and the fused product is introduced into mice, tumour formation results (Sager, 1989). However, when a tumour cell is fused with a normal cell and introduced into mice, the fusion blocks tumour formation. These observations suggested that some activity must be present in the normal cell that inhibits transformation.

Retinoblastoma (Rb) Gene

The first tumour-suppressor gene identified was the *Rb* gene, which is associated with the childhood illness of retinoblastoma (Knudson, 1971). In an epidemiological study, Knudson and colleagues noticed that bilateral retinoblastoma occurred frequently within the same family, whereas unilateral retinoblastoma did not appear to be a

genetically inherited disease. In families with bilateral retinoblastoma, karyotyping techniques were used to detect homozygous loss of chromosome 13q, a defect that was transmitted to offspring. Homozygous loss was found to be necessary but not sufficient for the formation of retinoblastoma, since not every family member with the loss of both alleles developed the disease. Later, the gene responsible for development of the disease was cloned and termed *Rb* for retinoblastoma. Reintroduction of this gene into cultured retinoblastoma tumour cells reversed the malignant phenotype, suggesting that the gene was indeed a tumour suppressor (Bookstein *et al*., 1989).

Tumour Suppressors in Colon Cancer

Since the cloning of *Rb*, many other tumour-suppressor genes have been identified (**Table 2**). Several of the most notable are a group of tumour-suppressor genes that were identified by studying progressive stages of colon cancer. They include the 'adenomatous polyposis coli' (*APC*) gene, the 'deleted in colon cancer' (*DCC*) gene and the 'mutated in colon cancer' (*MCC*) gene (Peddanna *et al*., 1996). *APC* maps to chromosome 5q21 and is mutated in 70% of patients with a hereditary form of colon cancer, termed familial adenomatous polyposis (FAP). Also mapping to chromosome 5q is the *MCC* gene, which was found to be mutated in 55% of all colon cancers studied. *DCC* was mapped to chromosome 18 and is deleted in 73% of colon cancers. *APC* and *DCC* code for proteins that play roles in regulating cell adhesion in normal cells. It is speculated that loss of these genes can lead to increases in cell motility, a key characteristic of metastasis.

p53

The p53 protein is involved in sensing DNA damage and regulating cell death (Marx, 1993). In normal cells, when DNA damage is sensed by p53, the cell cycle is arrested to permit DNA repair. Upon completion of this process, the cell progresses through the mitotic cycle. If repair fails to occur, p53 initiates the process of apoptosis, or programmed cell death. Thus, normal cells with genetic defects die. If p53 is not present in the cell (via gene deletion) or is mutated to be nonfunctional, DNA damage is not repaired, and the cell progresses through the cell cycle, transmitting its damaged DNA to its progeny. p53 is so important to the maintenance of 'healthy' DNA that it is mutated or deleted in over 70% of human cancers, including osteosarcomas, rhabdomyosarcomas and carcinomas of the breast, colon, lung and prostate.

BRCA1 and *BRCA2*

Another more recently identified tumour-suppressor gene, *BRCA1*, was found to be linked to an increased risk of hereditary breast cancer (Zheng *et al*., 2000). Loss of chromosome 17q had long been known to occur in familial breast cancer. The *BRCA1* gene mapped to chromosome 17q, but it was not until 1993 that it was identified and cloned. Many heritable mutations were identified in *BRCA1* from breast cancer patients and include an 11-bp deletion, a 1-bp insertion, a stop codon and a missense substitution. However, this may be an underestimation of its involvement in oncogenesis, as mutations and inactivating events, such as promoter methylation, also may regulate *BRCA1*

Table 2 Tumour-suppressor genes and their function and associated cancers

Name	Function in normal cells	Associated cancers
p53	Cell cycle regulator	Colon and others
BRCA1	Cell cycle regulator, genomic integrity and chromatin structure	Breast, ovarian, prostate and others
BRCA2	Genomic integrity	Breast, ovarian, prostate and others
PTEN	Tyrosine and lipid phosphatase	Prostate, glioblastomas
APC	Cell adhesion	Colon
DCC	Cell adhesion	Colon
MCC	Undetermined	Colon
p16-INK4A	Cell cycle regulator	Colon and others
MLH1	Mismatch repair	Colon and gastric cancers
MSH2	Mismatch repair	Colon and gastric cancers
DPC4	Cell death regulator	Pancreatic
Wt1	Cell death regulator	Wilms' tumour
NF1	Regulator of GTPases	Astrocytomas
NF2	Cell adhesion	Astrocytomas
VHL	Ubiquitination	Renal
PTC	Regulator of hedgehog signalling	Thyroid
TSC2	Cell cycle regulator	Breast and renal
TSG101	Cell cycle regulator	Renal and leukaemia

expression. Such gene regulation events are still being defined and are difficult to identify by screening techniques. Another *BRCA* family member, *BRCA2*, also has been cloned. This gene localizes to 13p12–13, and mutations within it correlate with breast cancer occurrence. *BRCA1* and 2 also are mutated or deleted in about 33 and 34% of sporadic breast tumours, respectively.

PTEN

PTEN, a gene encoding a phosphoprotein and phospholipid phosphatase, was first identified in glioblastoma patients who had sustained deletions of chromosome 10q23 (Li *et al*., 1997). *PTEN* is mutated in 31% of glioblastomas, 100% of prostate cancers and 6% of breast cancers. Interestingly, deletion of *PTEN* in gliomas segregates independently of mutations in p53, i.e. tumours containing *PTEN* mutations do not contain p53 mutations (Liu *et al*., 1997). However, *PTEN* deletions/mutations do correlate with amplification of the EGF receptor, a known oncogene. In normal cells, it is thought that PTEN downregulates phosphorylation events that promote cell growth. Its loss, therefore, allows for unregulated and unhindered proliferation.

Other Tumour-suppressor Genes

Another tumour-suppressor gene is the *p16-INK4A* gene, which negatively regulates cell cycle events. It is lost from chromosome 9 in a wide range of cancers (Kamb *et al*., 1994). Genes involved in the efficacy of DNA replication, *MLH1* and *MSH2*, are found deleted in 50% of hereditary non-polyposis colorectal cancers (Konishi *et al*., 1996). *DPC4* (deleted in pancreatic cancer) is lost from chromosome 18q in pancreatic cancer (Hahn *et al*., 1996). Still other tumour-suppressor genes include the Wilms' tumour-associated tumour suppressor *Wt1*, the human astrocytoma-associated tumour suppressors *NF1* and *NF2*, the von Hippel–Lindau syndrome tumour suppressor *VHL*, the papillary thyroid cancer tumour-suppressor gene *PTC* and tumour-suppressor genes associated with breast and renal cancer, *TSG101* and *TSC2*, respectively. Each of these genes encode protein products that negatively regulate the acquisition of a malignant phenotype by a normal cell.

Tumour-suppressor Genes, Normal Cellular Function and Carcinogenesis

In normal cells, products of tumour-suppressor genes have been shown to regulate negatively cell growth and proliferation. For example, the *Rb* gene product sequesters transcription factors that are required for normal cell cycle progression. The ability of Rb to function as a block to cell cycle progression is regulated by phosphorylation of the Rb protein on multiple serine residues (Harbour and Dean, 2000). In quiescence, Rb is hypo- or under-phosphorylated and binds members of the E2F transcription factor family. Upon growth factor stimulation of the cell, Rb becomes phosphorylated by cyclin-regulated kinases and releases E2F, which then induces gene transcription events necessary for cell division. In cancers, deletion or inactivation of Rb results in constitutively 'free' E2F, which in turn leads to unfettered gene transcription and oncogenic transformation. Similarly, as described above, p53 has been shown to sense DNA damage, cause cell cycle arrest, regulate transcription and stimulate apoptotic cell death pathways in normal cells (Marx, 1993). Loss of this function increases the chance of damaged DNA being transmitted to subsequent generations of cells. The exact role of BRCA1 is still unclear, but studies using mice that lack the *BRCA1* gene show that it is essential for cellular proliferation during early embryonic development (Zheng *et al*., 2000). BRCA1 may also regulate transcriptional events, since it is capable of acting as a coactivator of p53 and a corepressor of c-Myc. Recent studies also implicate a role for BRCA1 in chromatin remodelling, which is required for DNA transcriptional and replication events. PTEN regulates the phosphorylation status of phospholipids that are involved in regulating apoptotic pathways within the cell (Di Cristofano and Pandolfi, 2000). Taken together, these findings indicate that tumour-suppressor gene products act by negatively controlling cell growth in normal cells and that their loss contributes to the unregulated cell growth seen in tumour cells.

Mismatch Repair Genes

Critical regulators of genomic integrity, as exemplified by mismatch repair genes, also have been implicated as tumour-suppressor genes. The microsatellite instability genes described above, *MLH1* and *MSH2*, are important to the maintenance of genomic integrity by repairing mismatched base pairs that arise with a stable frequency during DNA replication (Kolodner and Marsischky, 1999). Mismatched base pairs are recognized and cleared by mismatch repair enzymes, and new bases are added in their place. Without such genes, repairs are not made and mutations are introduced into newly synthesized DNA. Alternatively, the stress of the mismatch structure may fragment the DNA. Both of these possibilities can lead to changes in the sequence of genes critical to cell growth or death. Although the alteration of mismatch repair genes may seem like a key event for all cancers, it has been determined that only 13% of gastric/colorectal cancers and less than 2% of other cancers have mutations in mismatch repair genes. Furthermore, the 2% occurrence is thought to reflect the normal rate of DNA mutation. These considerations therefore suggest that mismatch repair defects may be more specific for gastric cancers and not a general phenomenon associated with cancer development.

New Techniques for Identification of Tumour-suppressor Genes

RFLP

Knudson's original method of analysing karyotypes of tumour cells is still in use today for examining large, consistent chromosomal alterations, but new methods have evolved in the last 20 years. One such technique, restriction fragment length polymorphism (RFLP), utilizes bacterial restriction enzymes that cleave DNA at specific sites. DNA encoding a normal gene has a characteristic DNA fragment pattern, while tumour DNA shows an abnormal pattern. RFLP DNA fragments have been linked together to span an entire normal human genome. At a frequency of approximately every 10 million base pairs, a known gene has been mapped to specific RFLP fragments. This approach has yielded a crude map of the genome, which is more sensitive than karyotyping methods and allows one to map a loss of specific regions of a chromosome.

Comparative Genomic Hybridization (CGH)

Comparative genomic hybridization compares the ability of tumour RNA labelled with one fluorophore (i.e. green) and normal RNA labelled with another fluorophore (i.e. red) to hybridize to a chromosome spread from a given tissue type. The spread is analysed by fluorescence microscopy after hybridization of the RNAs. Losses (or gains) in a chromosome can be observed by the colour of the fluorophore hybridized to the region on the chromosome. For example, if a region of chromosome 13 is deleted in the tumour DNA, chromosome 13 will appear red since there would be no green-labelled (tumour-derived) RNA to hybridize with that region. Gains in the tumour DNA score green whereas equal expression in both normal and tumour cells scores yellow (a merging of red and green). Thus a complete genomic map of genetic changes that occur in a tumour cell can be obtained. Other more sensitive techniques, such as DNA microarrays, are currently being perfected to identify additional tumour-suppressor genes. It should be noted that in comparison with oncogenes, a very small number of tumour-suppressor genes have been discovered. The techniques that detect tumour-suppressor genes are somewhat insensitive, since even the most accurate screening approaches localize the region of loss only to a megabase or more.

Methylation

The techniques described above tend to identify tumour-suppressor genes that are grossly mutated in cancer. However, there are other mechanisms of tumour-suppressor gene inactivation, including point substitutions, small insertions and deletions. Almost half of all tumour-suppressor genes are also methylated in their promoter regions, preventing gene transcription (Baylin, 1997). Abnormally high levels of methylation appear in cancer cells that have a loss in the *p21/WAF1* gene. In normal cells, $p21^{WAF1}$ protein negatively regulates the ability of DNA-methyltransferase to add a methyl group to CpG islands, thereby protecting these sites in the DNA from methylation. Inactivation or loss of $p21^{WAF1}$ allows these sites to be methylated and transcriptionally silenced. Some tumour-suppressor genes shown to be methylated in tumours include *BRCA1*, *VHL*, and *p16INK4A*.

MOLECULAR MECHANISMS OF CANCER

Tumorigenesis *in vivo* is actually a multistep process requiring the alterations of two or more genes (Knudson, 1971). **Figure 2** depicts a single cell bearing a mutation or genetic 'hit' in gene A. This mutation is passed on to progeny cells, which, at a defined probability, sustain a second 'hit' in gene B. The figure depicts the mutation in gene A as a dominant 'oncogene-like' mutation and the mutations in both alleles of gene B as a recessive 'tumour-suppressor-like' mutation. Such alterations provide the initial steps in tumour formation. Every cell in the tumour carries the identical mutations that initiated tumour development.

The Two-hit Hypothesis of Knudson and Hereditary Cancers

One of the first concepts to arise regarding the molecular mechanism of tumours was suggested by Knudson and colleagues, who developed the two-hit hypothesis (Knudson, 1985). The assumptions of this hypothesis are threefold: malignant transformation of a single cell is sufficient to give rise to a tumour; any cell in a specific tissue is as likely to be transformed as any other of the same type; and once a malignant cell is generated, the mean time to tumour detection is generally constant. Once these assumptions are met, the model suggests that at least two events are necessary for carcinogenesis and that the cell with the first event must survive in the tissue long enough to sustain a second event.

Multistep Carcinogenesis Models

Land and Weinberg Model

At about the same time that Knudson proposed the two-hit hypothesis, Weinberg and Barrett independently suggested models of carcinogenesis based on the activation of a series of oncogenes. Weinberg suggested that the activation of two or more oncogenes is required for tumorigenesis and that the right combination must be activated in

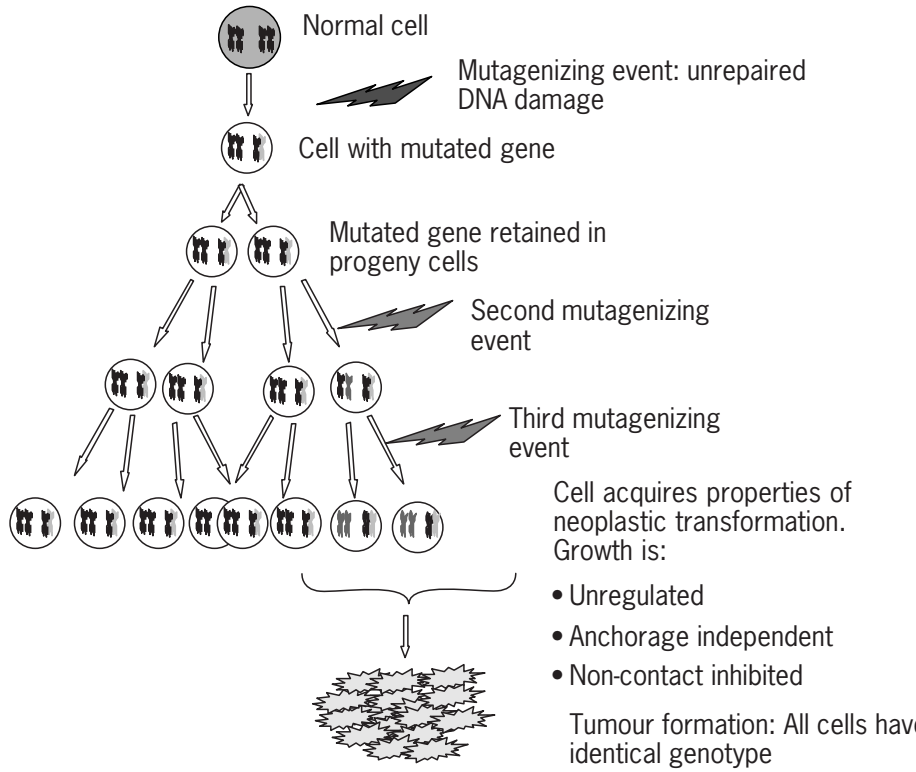


Figure 2 Acquisition of tumorigenic phenotype and clonality. Once a cell has acquired a mutation in an oncogene (depicted as the chromosome in light grey), that mutation is passed on to subsequent generations of progeny cells. These cells are still phenotypically normal, however. Cellular transformation occurs when a second and third mutation arise in a tumour-suppressor gene, e.g. in one of the previously mutated cells (depicted in dark grey). This cell now harbours three mutations in at least two different genes, and displays the hallmarks of neoplastic growth in culture.

the right context (Weinberg, 1983). Which oncogenes are activated is dependent on the signalling events each regulates. For example, Ras (a cytoplasmic oncoprotein) was shown to cooperate with Myc (a nuclear oncoprotein) to form tumours. Other combinations of cytoplasmic and nuclear oncoproteins also cause tumours to form, but one oncoprotein from each group must be activated.

Barrett's Model

Barrett's model further divides the process into tumour initiation vs tumour-promotion events (Boyd and Barrett, 1990). These investigators suggested that initiation is most often a mutational event, including mutations in a proto-oncogene, such as *ras*. Promotion, on the other hand, can be a mutational or an epigenetic change, and is defined as a series of '... qualitative, heritable changes in a sub-population of initiated cells, resulting in malignancy or an increased potential to progress to malignancy.' In this model, morphological transformation (or initiation) occurs upon treatment with a mutagen or carcinogen. This event is followed by a loss or inactivation of a gene controlling cell senescence along with activation of immortalizing genes. With such changes, an immortal cell line is generated. The subsequent loss or inactivation of a tumour-suppressor gene or activation of a transforming oncogene then leads to the formation of a tumorigenic cell.

Vogelstein's Model

Studies by Vogelstein and colleagues led to a progression model in colon cancer which includes both the activation of oncogenes and the loss of tumour suppressors (Vogelstein and Kinzler, 1993). This model, dubbed the Vogelgram, is based on several observations. The first is that cancer cells contain 3–7 somatic mutations per cell. Second, benign tissue surrounding the malignant tissue frequently contains many of the same set of mutations found in the tumour but lacks at least one mutation that is found in tumour tissue. Third, certain genes have a high probability of mutating at each definable stage of colon cancer progression. Based on these and other genetic data, a model for colon cancer progression was formulated. **Figure 3** suggests that the loss of the tumour-suppressor gene *APC* occurs early in the process of transformation, converting colonic epithelial cells to a hyperproliferative state. Hypomethylation of DNA then occurs in the early ademona stage, followed by activation of the oncogene Ki-*ras* in carcinoma *in situ*. The tumour-suppressor genes *DCC* and *p53* are lost later in the disease,

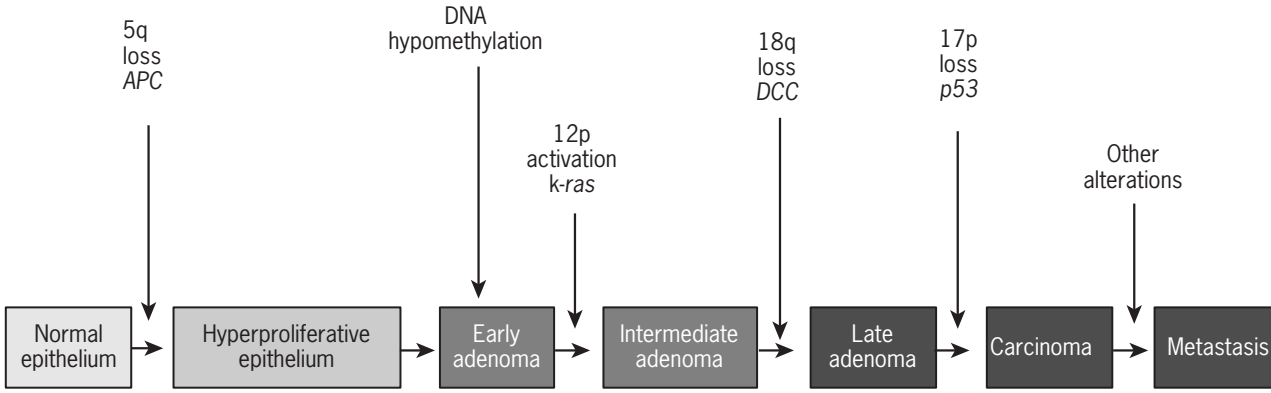


Figure 3 The Vogelstein model of multistep carcinogenesis. The progression of a normal colonic epithelium to metastatic colon cancer can be observed as it passes through several distinct stages. Chromosomal loss can be noted at different steps of progression. For example, chromosomal loss at 5q appears to occur prior to development of hyperproliferative epithelium whereas loss at 17p does not occur until the late adenoma transitions to a full carcinoma.

with the eventual development of a metastatic colon cancer.

CLINICAL CORRELATIONS

There are many ways in which mutations in cancer-promoting genes can occur. The predisposition to cancer can be inherited, as in patients with Li–Fraumeni syndrome, whose cells contain a germ-line mutation of *p53*, one of the cell cycle checkpoint regulators described above. Cells from patients with chronic myelogenous leukaemia often contain an abnormal chromosome resulting from a translocation between chromosomes 9 and 22, the so-called Philadelphia chromosome (Rowley, 1973). This abnormal fusion juxtaposes two genes, which code for the proteins BCR and the Abl tyrosine kinase, and results in aberrant activity and subcellular localization of the Abl protein. In breast cancer, *BRCA1* is mutated at specific sites in the gene. Such mutations are largely inherited.

In contrast to germ-line or inherited mutations, 'sporadic' cancers also can arise when a mutation occurs in a previously normal somatic cell. In this regard, environmental factors are thought to play major roles as mutagens or carcinogenic agents. For example, the relationship between tobacco smoke and lung cancer is well documented (Henderson *et al*., 1991). Anilines used in rubber tyre production are linked to the development of bladder cancer, while exposure to solar ultraviolet rays can cause melanoma (Case *et al*., 1993; Armstrong *et al*., 1997). Hundreds of chemical carcinogens that exist in food and products in daily use can either directly or through the production of secondary metabolites irreversibly alter a normal cell's DNA. So-called 'lifestyle' factors can also play a role. A link has been made between consumption of a diet high in animal fats and prostate cancer (Tzonou *et al*., 1999). In women, reproductive history and the resulting cumulative lifetime exposure to oestrogen correlate with an increased risk of breast cancer (Hankinson *et al*., 1995). How environmental factors trigger the activation and mutation of cancer-causing genes is, in many cases, still unclear. Subsequent chapters will detail what is known about a very complicated and intricate process.

REFERENCES

Armstrong, B. K., *et al*. (1997). Sun exposure and skin cancer. *Australasian Journal of Dermatology*, **38**, S1–S6.

Baylin, S. B. (1997). Tying it all together: epigenetics, genetics, cell cycle, and cancer. *Science*, **277**, 1948–1949.

Biscardi, J. S., *et al*. (1999). c-Src, receptor tyrosine kinases, and human cancer. *Advances in Cancer Research*, **76**, 61–119.

Bishop, J. M. (1982). Oncogenes. *Scientific American*, **246**, 80–92.

Bookstein, R., *et al*. (1989). Human retinoblastoma gene: long-range mapping and analysis of its deletion in a breast cancer cell line. *Molecular and Cellular Biology*, **9**, 1628–1634.

Boveri, T. (1929). The Origin of Malignant Tumors. Williams Wilkins, Baltimore, pp. 1–119.

Boyd, J. A. and Barrett, J. C. (1990). Genetic and cellular basis of multistep carcinogenesis. *Pharmacology and Therapeutics*, **46**, 469–486.

Case, R. A., *et al*. (1993). Tumours of the urinary bladder in workmen engaged in the manufacture and use of certain dyestuff intermediates in the British chemical industry. Part I. The role of aniline, benzidine, alpha-naphthylamine, and beta-naphthylamine. 1954. *British Journal of Industrial Medicine*, **50**, 389–411.

Christofiori, G. and Semb, H. (1999). The role of the cell adhesion molecule E-cadherin as a tumor-supressor gene. *Trends in Biochemical Sciences*, **217**, 801–806.

Di Cristofano, A. and Pandolfi, P. P. (2000). The multiple roles of PTEN in tumor suppression. *Cell*, **100**, 387–390.

Folkman, J. (1992). The role of angiogenesis in tumor growth. *Seminars in Cancer Biology*, **3**, 65–71.

Hahn, S. A., *et al*. (1996). DPC4, a candidate tumor suppressor gene at human chromosome 18q21.1. *Science*, **271**, 350–353.

Hankinson, S. E., *et al*. (1995). Reproductive factors and family history of breast cancer in relation to plasma estrogen and prolactin levels in postmenopausal women in the Nurses' Health Study (United States). *Cancer Causes and Control*, **6**, 217–224.

Harbour, J. W. and Dean, D. C. (2000). Rb function in cell-cycle regulation and apoptosis. *Nature Cell Biology*, **2**, E65–E67.

Heisterkamp, N., *et al*. (1985). Structural organization of the bcr gene and its role in the Ph′ translocation. *Nature*, **315**, 758–761.

Henderson, B. E., *et al*. (1991). Toward the primary prevention of cancer. *Science*, **254**, 1131–1138.

Irby, R. B., *et al*. (1999). Activating SRC mutation in a subset of advanced human colon cancers. *Nature Genetics*, **21**, 187–190.

Jacotot, E., *et al*. (2000). Apoptosis and cell cycle: distinct checkpoints with overlapping upstream control. *Pathologie et Biologie*, **48**, 271–279.

Kamb, A., *et al*. (1994). Analysis of the p16 gene (CDKN2) as a candidate for the chromosome 9p melanoma susceptibility locus. *Nature Genetics*, **8**, 23–26.

Knudson, A. G., Jr (1971). Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences of the USA*, **68**, 820–823.

Knudson, A. G., Jr (1985). Hereditary cancer, oncogenes, and antioncogenes. *Cancer Research*, **45**, 1437–1443.

Kolodner, R. D. and Marsischky, G. T. (1999). Eukaryotic DNA mismatch repair. *Current Opinions in Genetics and Development*, **9**, 89–96.

Konishi, M., *et al*. (1996). Molecular nature of colon tumors in hereditary nonpolyposis colon cancer, familial polyposis, and sporadic colon cancer. *Gastroenterology*, **111**, 307–317.

Kurschat, P. and Mauch, C. (2000). Mechanisms of metastasis. *Clinical and Experimental Dermatology*, **25**, 482–489.

Li, J., *et al*. (1997). PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. *Science*, **275**, 1943–1947.

Liu, W., *et al*. (1997). PTEN/MMAC1 mutations and EGFR amplification in glioblastomas. *Cancer Research*, **57**, 5254–5257.

Marx, J. (1993). How p53 suppresses cell growth. *Science*, **262**, 1644–1645.

Miyamoto, S., *et al*. (1998). Fibronectin and integrins in cell adhesion, signaling, and morphogenesis. *Annals of the New York Academy of Sciences*, **857**, 119–129.

Peddanna, N., *et al*. (1996) Genetics of colorectal cancer. *International Journal of Oncology*, **9**, 327–335.

Rous, P. (1911). A sarcoma of the fowl transmissible by an agent separable from the tumor cells. *Journal of Experimental Medicine*, **13**, 397–411.

Rowley, J. D. (1973). A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature*, **243**, 290–293.

Sager, R. (1989). Tumor suppressor genes: the puzzle and the promise. *Science*, **246**, 1406–1412.

Tzonou, A., *et al*. (1999). Diet and cancer of the prostate: a case-control study in Greece. *International Journal of Cancer*, **80**, 704–708.

Vogelstein, B. and Kinzler, K. W. (1993). The multistep nature of cancer. *Trends in Genetics*, **9**, 138–141.

Weinberg, R. A. (1983). Alteration of the genomes of tumor cells. *Cancer*, **51**, 1971–1975.

Weiss, R., *et al*. (eds) (1985). *RNA Tumor Viruses*, 2nd edn (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY).

Wittinghofer, F. (1998). Ras signalling. Caught in the act of the switch-on. *Nature*, **394**, 317, 319–320.

Zheng, L., *et al*. (2000). Lessons learned from BRCA1 and BRCA2. *Oncogene*, **19**, 6159–6175.

FURTHER READING

Bishop, J. M. (1992). Oncogenes. *Scientific American*, **246**, 80–92.

Knudson, A. G. (1985). Hereditary cancer, oncogenes, and antioncogenesis. *Cancer Research*, **48**, 1437–1442.

Land, H., *et al*. (1983) Cellular oncogenes and multistep carcinogenesis. *Science*, **222**, 771–778.

Sager, R. (1989). Tumor suppressor genes: the puzzle and the promise. *Science*, **246**, 1406–1412.

Vogelstein, B. and Kinzler, K. W. (1993). The multistep nature of cancer. *Trends in Genetics*, **9**, 138–141.

Weiss, R., *et al*. (eds) (1985). *RNA Tumor Viruses*, 2nd edn (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY).

Chapter 4

Inherited Predispositions to Cancer

Gareth Evans
St. Mary's Hospital, Manchester, UK

CONTENTS

INTRODUCTION

There has been increasing evidence of familial predisposition to cancer since the classic model of hereditary retinoblastoma was outlined (Knudson, 1971). The notion that some cancer is hereditary has long been held by more than just a few diehard clinicians. The earliest reports of cancer families date back more than 180 years to a large cluster of breast cancer in the wife and family of a French physician named Broca and the cluster of gastric cancer in Napoleon's family. Despite the pioneering work of clinicians and researchers such as Henry Lynch and Mary-Claire King in the USA in the 1960s to 1980s, demonstrating the hereditary nature of at least a proportion of cancers such as those affecting the breast and colon, the hereditary element was not proven until the advent of molecular biology when abnormalities were demonstrated in cancer-predisposing genes. It is, therefore, only since 1987 that developments in molecular biology have proven the hereditary nature of a small proportion of certain common cancers. That cancer is now indisputably 'genetic' at the cellular level is beyond dispute. All tumours result from mutations or deletions of two types of gene (Eeles *et al*., 1996): the tumour suppressor gene, which needs to be inactivated to allow growth (like the brakes on a car), and the oncogene, which requires activation to promote growth (like the accelerator pedal of a car being stuck down). The great majority of these events are acquired whether through replication error (simple copying of DNA during cell division) or due to external agents (chemical mutagens, radiation, viruses). Occasionally, mutations in tumour suppressor genes can be inherited rather than acquired. Identifying the genes which cause hereditary disease has given an insight into many cancers. The role of cancer-predisposing genes in the causation of sporadic cancer is still the subject of much research, and we can still learn from the more obscure cancer-prone syndromes.

Broadly, the predisposition can be subdivided into rare genetic syndromes which have malignancy as a high-risk side effect and a larger group which cannot be easily identified clinically, and which have a strong family history of one or more common malignancies.

RETINOBLASTOMA

Retinoblastoma has been the model from which much of our current knowledge of tumour suppressor genes was fashioned. This early childhood eye malignancy was recognized as having a familial tendency in the nineteenth century. About 50% of cases are due to the inheritance of a gene defect in one copy of the retinoblastoma gene (*RB* on chromosome 13), and over 90% of individuals who carry a mutation will develop retinoblastoma, usually bilaterally. In 1971, Knudson (1971) proposed that the disorder was caused by mutational events in both copies of the gene. Those cases that inherited a mutated copy, only need one further mutation and are far more likely to develop the malignancy, which occurs at a younger age and is usually bilateral. The sporadic cases require two mutations ('hits') in a retinal cell as opposed to one in the familial case (**Figure 1**) and so bilateral tumours are extremely unlikely to occur and the unilateral tumours present later. This hypothesis, which has since been proven to be true, now bears the conceiver's name. Familial retinoblastoma may even be present in foetal life, as can be seen in **Figure 2 (see colour plate section)**. This case had a 13q deletion as a result of a maternal chromosomal translocation.

The discovery of retinoblastoma cases with cytogenetically visible constitutional deletions in the long arm of chromosome 13 (Francke and King, 1976; Knudson *et al*., 1976) concentrated research on that region. One of the genes deleted in these cases, esterase D, then acted as a marker for further studies. One study showed that although

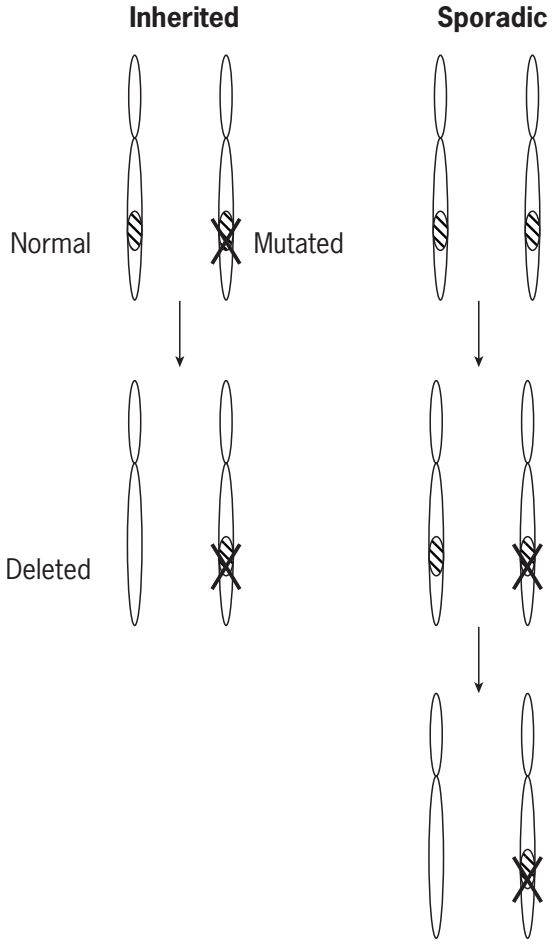


Figure 1 Ideogram of the 'two-hit' hypothesis. The first hit is usually a mutation (represented by a cross) which causes disruption of the protein product. The second hit is often loss of the whole gene by deletion of part or all of the chromosome on which the gene resides.

an individual was heterozygous for esterase D their tumour was hemizygous, suggesting loss of material by deletion or monosomy (Godbout *et al*., 1983). The introduction of restriction fragment length polymorphisms (RFLPs) led to further studies showing loss of constitutional heterozygosity. RFLPs rely on the differences in large portions of DNA between individuals and therefore between the two equivalent chromosomal regions in any one individual. This difference means that enzymes (endonucleases) which cut at specific gene codes will cut at different sites on the chromosome. There is therefore a good chance that a gene or genetic marker will end up on two different lengths of DNA when a particular endonuclease is used. If an individual is shown to have two lengths (heterozygous) on constitutional testing, but only one in their tumour, then loss of constitutional heterozygosity has occurred. These RFLPs were used to localize the gene further until it was eventually cloned (Friend *et al*., 1986). Since the isolation of the *RB* gene, many groups have tried to isolate the underlying defects that cause retinoblastoma and that make the gene important in cell regulation. These studies have confirmed that *RB* acts as a typical tumour suppressor gene with an initial mutational event consisting of small changes in nucleotides resulting in truncation of the resultant protein product. These are usually nonsense mutations leading to an early stop codon or small frameshift deletions or insertions with a similar downstream effect. Nonetheless, pathogenic missense mutations do occur although they are more difficult to prove. Simple cosegregation of a missense mutation in a small family is not enough, although the presence of an amino acid change in a functional domain with cross-species conservation does add credence. In the final estimation it is only with functional assays that a pathogenic effect can be proven. Initial studies showed a relatively low rate of mutation identification in the *RB* gene (Liu *et al*., 1995), but with a combination of strategies including a coding sequence analysis taking in intron/exon boundaries (for splicing mutations) and a deletion strategy such as fluorescent *in situ* hybridization (FISH) and Southern blotting, the great majority of aberrations of the *RB* gene can be identified (Lohmann *et al*., 1996).

The retinoblastoma gene also predisposes to osteosarcoma. An individual who has had an enucleation and irradiation for retinoblastoma is 500 times more likely to develop the bone tumour. As the gene is involved in many common carcinomas such as that of the breast, it is likely that survivors will be at risk of these tumours also.

GENETIC SYNDROMES

These are usually readily identifiable by a clinical phenotype or by laboratory tests. The syndromes may be autosomal dominant or recessive or X linked (**Tables 1** and **2**). Of these it is the dominant conditions which are of most interest as they are likely to represent the inheritance of a faulty copy of a tumour suppressor gene, which predisposes the individual to common cancers. Although the conditions are generally uncommon, tumour suppressor genes in general are likely to play a fundamental role in the genesis of tumours, which affect a third of all humans in their lifetime. The identification of those causing genetic syndromes is likely to lead to more specific treatment using gene therapy, as well as earlier identification, monitoring and, most hopeful of all, prevention of common cancers.

Familial Adenomatous Polyposis (FAP)

FAP is the model condition by which researchers have hoped to transpose knowledge of a rare genetic disease to a commonly occurring cancer. FAP is an autosomal dominant condition characterized by the development of hundreds to thousands of adenomatous polyps in the colon and rectum, usually by 30 years of age (**Figure 3; see colour**

Table 1 Examples of autosomal dominant syndromes predisposing to cancer and their chromosomal location

Name of disease	Location (chromosomal)	Protein
FAP	5q	APC
NF1	17q	Neurofibromin
NF2	22q	Merlin/ schwannomin
Von Hippel–Lindau	3p	pVHL
MEN1	11q	Menin
MEN2	10q	RET
Gorlin	9q	PTCH
Tuberous sclerosis (TSC1)	11q	Hamartin
(TSC2)	16q	Tuberin
Juvenile polyposis	18q and other(s)	pDPC/SMAD4
Peutz–Jeghers	19p and other(s)	pSTK11/LKB1
Cowden	10q	PTEN
Tylosis	17q	Not found

Table 2 Autosomal recessive and X-linked conditions predisposing to malignancy

Name of condition	Chromosomal location	Protein
Fanconi anaemia	8 loci	4 found
Bloom syndrome	15q	pBLM
Ataxia telangiectasia	11q	pATM
Xeroderma pigmentosa	7 types	2 types
Chediak Higashi	1q	pLYST
Albinism	11q	OCA1, OCA2
Bruton	Xq	BLk
Wiscott Aldrich	Xp	CD43

plate section). If untreated this leads to the almost inevitable development of a colorectal cancer by 60 years of age. The condition may be associated with osteomas and epidermal cysts and this subdivision was designated Gardner syndrome (Gardner, 1951). However, most FAP families show these extraintestinal features to some extent. FAP has, in common with many other conditions, been mapped to a chromosomal region as a result of the finding of a constitutional (present in every cell) chromosomal anomaly in a manifesting case (Herrera *et al*., 1986). This patient had a small interstitial deletion on the long arm of chromosome 5 and manifested extracolonic features as well as multiple polyps. Following this discovery, Bodmer *et al*. (1987) localized the gene for FAP to 5q21–q22 by genetic linkage, using families mainly from the well-established St. Marks Polyposis Register in London. Of great interest was that the same region of chromosome 5 was implicated in sporadic colorectal cancer (Solomon *et al*., 1987). This meant that the gene could be localized more accurately by using tumour material from sporadic cases as well as from FAP cases. If the tumour is analysed for loss of genetic material using probes mapped to the implicated region of chromosome 5, a deletion map can be drawn. When the normal or 'wild-type' allele is lost from an FAP patient's tumours, the assumption is that a second hit has occurred, removing the only functioning copy of a tumour suppressor gene. The nature of this loss is known as loss of constitutional heterozygosity (LOH).

Once the gene had been localized to a relatively small chromosomal region, several research groups embarked on a project to isolate contiguous sequences of genes spanning the area. Subclones were then used to identify the position of candidate genes, which were expressed in normal colonic mucosa. Two of these genes, *APC* and *MCC* (Kinzler *et al*., 1991), were thought likely to be involved in tumorigenesis, because of the structure of the proteins for which they encoded. All that remained was to identify mutations in one of these genes, in the germ-line of patients with FAP. This was duly achieved when 10 *APC* germ line mutations were described (Nishisho *et al*., 1991). These mutations were not only likely to disrupt the protein structure, but were also found only in the *APC* gene. The idea that Gardner syndrome was a separate entity was refuted since the mutations occurred in patients with or without extraintestinal manifestations. However, FAP was one of the first conditions in which a clear correlation between genotype (the genetic change in *APC*) and phenotype (the clinical picture) emerged. Patients with mutations in the early part of the gene (5′: exons 2–5) had a very mild clinical picture with late onset of polyps (Spirio *et al*., 1993), whereas those with mutations from exon 9 through to codon 1450 of exon 15 had a classical disease course with nearly all patients manifesting the typical congenital retinal pigmentation. However, those with mutations beyond codon 1450 showed typical Gardner syndrome features (osteomas, cysts and desmoid disease) without retinal signs (Davies *et al*., 1995). There are even families who with extreme mutations in exon 15 show little else other than desmoid disease (Scott *et al*., 1996).

Currently it is possible to offer predictive genetic tests before symptoms to the majority of at-risk individuals. Although looking for germ-line mutations is laborious and not guaranteed to find the mutation, it is the most reliable. Nonetheless, testing using linkage analysis in families with more than one affected member is still very useful, especially when a germ-line mutation cannot be identified. This, combined with opthalmological screening for congenital hypertrophy of the retinal pigment epithelium (CHRPE), may reduce initial risks of 50% to well below 1% (Burn *et al*., 1991).

Von Hippel–Lindau

Von Hippel–Lindau (VHL) is another dominantly inherited familial cancer syndrome. The most frequent complications

are benign tumours of blood vessels, particularly in the eye (retinal angiomas), and haemangioblastomas of the cerebellum. Other features include renal cell carcinoma, phaeochromocytoma and renal, pancreatic and epididymal cysts. The syndrome is very variable but most individuals present before 40 years of age (Maher *et al*., 1990). The first clue to the location of the gene for VHL was the finding of a reciprocal translocation involving chromosomes 3 and 8 in a family with hereditary renal carcinoma (Cohen *et al*., 1979). Later, Teyssier *et al*. (1986) were able to show deletions of the short arm of chromosome 3 in other renal cell carcinomas. Linkage in families with VHL was confirmed on 3p in 1988 (Seizinger *et al*., 1988). The gene was finally cloned in 1993 (Latif *et al*., 1993) and codes for a relatively small protein.

Type 1 Neurofibromatosis (NF1)

NF1 is more common than NF2, but the disease may be so mild that an affected individual may never present to their doctor. The main manifestations are in the skin, with the appearance of café au lait patches and cutaneous neurofibromas in the first and second decades, respectively. The most famous potential misdiagnosis of NF1 was Joseph Merrick, the 'elephant man,' who in reality probably had Proteus syndrome (Clark, 1994). One potential serious complication of NF1 is optic gliomas, which may occur in up to 15% of cases (Listernick *et al*., 1989). These are usually very low grade and asymptomatic and if they are not specifically sought levels of around 1.5% are found. Other CNS gliomas do occur but their frequency is probably well below 5%. Meningiomas and vestibular schwannomas do not occur in excess frequency in NF1 (McGaughran *et al*., 1999). Phaeochromocytoma and spinal neurofibromas may develop as well as rhabdomyosarcomas, but these are relatively rare. Malignant change in neurofibromas can result in a malignant peripheral nerve sheath tumour (MPNST) in about 10% of NF1 patients in their lifetime (McGaughran *et al*., 1999).

The *NF1* gene was cloned in 1990, although it took over a year fully to characterize the gene from the first discovery of deletions in the germ-line of some familial cases (Viskochil *et al*., 1990). It is a massive gene containing over 300 kilobases of DNA divided into 50 exons (Collins, 1991). The gene transcribes a 327 kDa GAP protein containing 2818 amino acids. The protein, which binds to the oncogenic protein Ras, is found in all tissues. It is expressed at the cellular level in the perinuclear vesicles and microtubules. As p120 GAP is expressed more in the neural crest this may explain why *NF1* specifically affects neural tissue. Although diagnosis is possible by looking for germ-line mutations, this approach has not found any particular hot spots of mutation, although extensive analysis using a number of approaches as for retinoblastoma does detect 95% of mutations (Messiaen *et al*., 2000). Predictive diagnosis therefore still depends mainly on linkage in existing families.

There is now good evidence that at least two variant conditions of NF1 are caused by mutations in the *NF1* gene. Watson syndrome was shown to be linked to the *NF1* locus (Allanson *et al*., 1991) and NF–Noonan syndrome to be due to mutations in the gene itself (Colley *et al*., 1996).

Type 2 Neurofibromatosis (NF2)

NF2 is an autosomal dominant genetic disease characterized by the development of bilateral vestibular schwannomas (acoustic neuromas) in the second and third decades. Only recently has it been formally separated from the more common NF1 (von Recklinghausen disease), after the National Institutes of Health Consensus Development Conference Statement on Neurofibromatosis (1987). The first probable reported case of NF2 was that of Wishart (1822). Bilateral vestibular schwannoma had been thought to be part of von Recklinghausen neurofibromatosis (NF1) after reports of similarities in cases to those with the peripheral form (Cushing, 1917). Several reports emphasized the paucity of skin findings in families with bilateral vestibular schwannoma (Gardner and Frazier, 1930), and suggested that bilateral vestibular schwannoma represented a separate central form of von Recklinghausen neurofibromatosis. It was not until the separate assignment of NF1 to chromosome 17 (Seizinger *et al*., 1987) and NF2 to chromosome 22 (Rouleau *et al*., 1987) that the diseases were finally shown to be two distinct disorders. NF2, although less common than the type 1 form (incidence 1 in 35 000 compared with 1 in 3000) (Evans *et al*., 1992a), is more likely to present clinically at some time. All cases will develop a CNS tumour by 55 years of age. Although most of these tumours are benign (meningiomas, schwannomas; see **Figure 4**), 6% will develop a malignant glioma or ependymoma (Evans *et al*., 1992b).

The clue to the location of the *NF2* gene was not a constitutional chromosomal anomaly, but rather cytogenetic abnormalities found on chromosome 22 in human meningiomas and later in vestibular schwannomas and other tumours from NF2 patients. This candidate region was then confirmed as the likely location for the *NF2* gene by linkage analysis in a large US family (Rouleau *et al*., 1987). The gene was isolated simultaneously by two groups (Rouleau *et al*., 1993; Trofatter *et al*., 1993), and genotype phenotype correlations have been identified (Evans *et al*., 1998a). Mutations which give rise to a truncated protein are associated with a severe, multitumour, early-onset disease course, whereas those that give a nearly normal protein product (missense mutations) or no product (large deletions) give mild disease. Another feature of NF2 that is likely to be an important factor in other tumour-prone disorders is mosaicism (Evans *et al*., 1998b). If a mutation occurs after conception, say at the

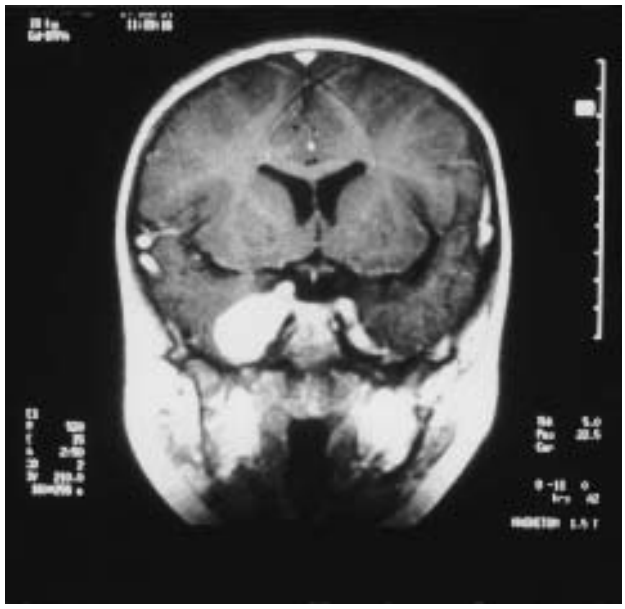


Figure 4 MRI scan of a 26-year-old man with type 2 neurofibromatosis. The scan shows bilateral enhancing tumours in the cerebello-pontine angles and meningiomas around the brain. The risk of developing bilateral tumours by chance is 1 in 2×10^6, yet 95% of individuals with mutations in the *NF2* gene develop bilateral vestibular schwannomas (acoustic neuromas).

eight-cell stage, roughly one eighth of all the cells will have an *NF2* mutation, which means that there are two different cell populations, one of which predisposes to tumours and could be transmitted to any offspring.

Gorlin Syndrome

Gorlin or naevoid basal cell carcinoma syndrome is another autosomal dominant condition which predisposes to malignancy. The condition is characterized by the development of multiple jaw keratocysts in the second decade and basal cell carcinomas in the third decade onwards. Gene mutation carriers also have a recognizable appearance or morphology. They have macrocephaly with bossing of the forehead and the face is usually covered with white milia. The facial features are often coarse and the shoulders slope downwards. Most individuals have a skeletal anomaly such as a bifid rib or wedge-shaped vertebra and ectopic calcification, particularly in the falx, is almost certain by 20 years of age.

Individuals with Gorlin syndrome are also at risk of developing the childhood brain malignancy medulloblastoma, which occurs in 5% of cases (Cowan *et al*., 1997) and cardiac and ovarian fibromas (Gorlin, 1987; Evans *et al*., 1993). Malignant transformation has been described in the ovarian fibromas (Strong, 1977), but they usually remain benign, although they can reach a large size and are often calcified (**Figure 5**). The clue to the location

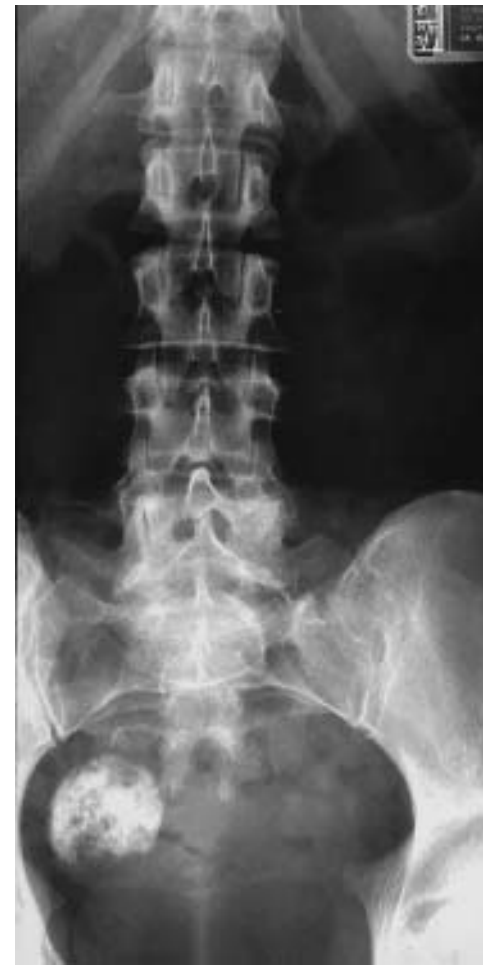


Figure 5 Large calcified ovarian fibroma on abdominal X-ray in a patient with Gorlin syndrome.

of the Gorlin gene again came from tumour deletion studies. Gailani *et al*. (1991) found that 40% of basal cell carcinomas that they studied had deletions of the proximal region on the long arm of chromosome 9. The condition has now been shown to be linked to that region using affected families and there is no locus heterogeneity (Farndon *et al*., 1992). The gene itself was subsequently identified as a homologue of the drosophila gene *PTCH* (Hahn *et al*., 1996). As the mean age at onset of medulloblastoma in Gorlin patients is 2 years compared with over 7 years in the general population, and there is loss of the normal copy of the gene in tumours (Cowan *et al*., 1997), this confirms *PATCHED* as a tumour suppressor in both medulloblastoma and basal cell carcinoma. Basal cell carcinomas occur at great frequency in the periphery of the radiation field 5–10 years after irradiation in Gorlin syndrome (**Figure 6; see colour plate section**).

Tuberous Sclerosis

Tuberous sclerosis is a condition in which hamartomas are a primary feature. These may occur in the brain (the

'tubers' of the name), or in the kidney, heart and elsewhere. Patients have a number of external features, which make early or presymptomatic identification possible. Depigmented patches of skin or 'ash leaf patches' can be seen with a Wood's light and a characteristic facial skin eruption known as adenoma sebaceum is often present. Subungual (under the finger nails) fibromas are another feature peculiar to tuberous sclerosis. Tumours may occur in the heart during foetal life (rhabdomyomas) and the malignancy to which the condition particularly predisposes is the brain tumour glioma. Unusually for a condition which is likely to be caused by one gene, two separate genetic loci have eventually been implicated. In 1987, Fryer *et al*. (1987) showed linkage to 9q in several families. However, many reports following this had contradictory results. This has now been shown to be due to locus heterogeneity, for although linkage to 9q has been confirmed in some families, the other major locus was actually cloned first in 1993 on chromosome 16 (European Chromosome 16 Tuberous Sclerosis Consortium, 1993). The later cloning of *TSC1* was partly due to the fewer individuals affected with the disease at this locus, although the linkage came first as actual families are more numerous (DeVries and Bolton, 2000).

Multiple Endocrine Neoplasias

The multiple endocrine neoplasias are further conditions which predispose to benign tumours and at least one malignancy. In MEN1 the organs affected are the parathyroid glands, pituitary and pancreas. The most serious problem is with islet cell tumours of the pancreas, which secrete gastrin. These cause the Zollinger–Ellison syndrome of which MEN1 makes up a large proportion of cases. The gastrin-secreting tumours may become malignant, seeding to the liver and other organs. However, many cases do not manifest the condition overtly even late in life. The serum calcium level is raised in 90–97% of cases, but laborious testing and repeated screening may be necessary. The availability of genetic tests has, therefore, greatly simplified screening of at-risk individuals. The location of the *MEN1* gene was confirmed at 11q13 by linkage analysis in families (Larsson *et al*., 1988). This and other studies have also shown loss of 11q alleles in the MEN1 tumours. The *MEN1* gene was eventually cloned in 1997 (Chandrasekharappa *et al*., 1997) and the protein product was termed menin. Presymptomatic testing is now available by mutation testing of an affected individual, or by linkage analysis in families with more than one affected individual.

MEN2a or Sipple syndrome is an autosomal dominant disease with high penetrance and variable expression. The association of medullary carcinoma of the thyroid and phaeochromocytoma are the hallmarks of the condition. Parathyroid tumours are less commonly found. MEN2 makes up 25% of all cases of medullary thyroid carcinoma, with nearly all MEN2 cases developing this tumour some time in life. The tumours in MEN2 are often bilateral and are preceded by C cell hyperplasia. Phaeochromocytoma occurs in 50% of individuals and is often multifocal. Screening of at-risk cases involves serum calcitonin estimation and monitoring of blood pressure, and has been greatly enhanced by the development of genetic testing, which removes over 50% of individuals from screening programmes. The test has targeted those in which thyroid cancer can be prevented by early or prophylactic thyroidectomy.

The *MEN2a* gene was localized to chromosome 10 by linkage analysis using RFLPs (Simpson *et al*., 1987), and later localized to 10q21.1 by *in situ* hybridization. Although researchers concentrated on trying to identify a tumour suppressor gene, it was eventually found that MEN2 was due to activating mutations in an oncogene called *RET* (Mulligan *et al*., 1993), although MEN2b differs from MEN2a in that the primary feature is the development of mucosal neuromas especially of the tongue. Medullary thyroid cancer is also a major feature and phaeochromocytoma also occurs but, in contrast to MEN1 and MEN2a, there is no parathyroid disease. Both conditions are caused by activating mutations in *RET*, although most of MEN2b is caused by a single mutation and MEN2a by five different substitutions at cysteine residues.

Other Dominant Syndromes

Tylosis, juvenile polyposis, Peutz–Jeghers syndrome, multiple exostosis and multiple lipomatosis are other dominantly inherited disorders which may predispose to malignancy.

Wilms' Tumour

Wilms' tumour, like retinoblastoma, has also been known for some time to have a hereditary element. However, the genetic basis is far more complex and the familial element much smaller. The first step to identifying a gene came with the discovery of a cytogenetically visible deletion in chromosome 11 in families with autosomal dominant aniridia who appeared to be predisposed to Wilms' tumour (Riccardi *et al*., 1978). Deletions in this area (11p13) also lead to genital and renal anomalies and mutations within the *WT1* gene itself lead to abnormal genital development (Pelletier *et al*., 1991). At first it was thought that this locus would be similar to retinoblastoma, but *WT1* has now been shown to be one of at least three genes involved in Wilms' tumour development. In 1989, Koufos *et al*. (1989) demonstrated tight linkage to 11p15.5 in a family with

Beckwith–Wiedemann syndrome (a mainly sporadic growth disorder with neonatal hyperinsulinism and features such as exompholos) and Wilms'. Beckwith–Wiedemann syndrome is now known to be due to complex mechanisms involving a number of genes including *CDKN1C* and *IGF2* where there is either loss of maternal copy or gain of paternal copy in an imprinted area (Lam *et al.*, 1999). In addition, Grundy *et al*. (1988) excluded both 11p13 and 11p15 in linkage analysis of a large family with dominant Wilms' tumour. A third and fourth locus has now been confirmed in families manifesting primarily Wilms' tumour alone, but a further locus probably exists.

Autosomal Recessive and X Linked Conditions

A list of these conditions and the chromosomal locations of the predisposing genes can be found in **Table 2**. These are less likely to present to the clinician as they are generally less common and mainly predispose to haematological malignancy.

COMMON CANCER PREDISPOSITION

Recent years have seen an enormous improvement in our understanding of the mechanisms of carcinogenesis. Most cancers require a number of genetic changes in a cell before an invasive tumour results. Few are likely to be caused purely by the loss of two copies of a single tumour suppressor gene as in retinoblastoma and the number of changes probably varies between four and 10. A combination of loss of function of tumour suppressor genes and activation of oncogenes is usually involved. The particular combination and order may alter both the histological and invasive nature of the cancer. There is now evidence that a minority of people who develop common cancers have inherited a faulty gene which puts them at high risk of malignancy, but this is not recognized as a syndrome apart from in the family history. Adenocarcinomas are more likely than carcinomas of squamous epithelium to have a strong hereditary component with 4–10% of all breast, ovarian and colon cancer resulting from an inherited gene defect. The discovery of germ-line (inherited) mutations in the *TP53* gene on the short arm of chromosome 17 in families with a peculiar combination of early and multiple tumours was the first proven example of this. Otherwise, predisposition can be relatively site specific with genes being isolated in recent years for melanoma, prostate cancer, pancreatic cancer and paraganglioma, but this chapter will focus on the three common cancers which have attracted the greatest attention.

Li–Fraumeni Syndrome (LFS)

This cancer predisposition syndrome was first outlined in 1969 by Li and Fraumeni (1969). They reported four families with autosomal dominant predisposition to soft tissue sarcoma, breast cancer and other tumours in children and adults. Many reports have followed, either describing further families or reporting an increased risk of cancers in first degree relatives of cases with soft tissue sarcoma. In 1988, Li *et al*. (1988) analysed 24 kindreds with an aggregation of tumours typical of the syndrome. They showed a predominance of soft tissue sarcoma, osteosarcoma and breast cancer, with an excess of adrenocortical carcinoma, brain tumours and leukaemia. Williams and Strong (1985) applied a segregation analysis to test the hypothesis that the disease was due to an autosomal dominant gene. They not only confirmed this, but also were able to predict that 50% of gene carriers would develop an invasive cancer by 30 years of age and 90% by 70 years. Although the syndrome (also known as SBLA) is rare, its importance lies in the unusual range of cancer predisposition and that it is caused by the first of the major predisposition genes to be identified.

Linkage studies using markers on many chromosomes was not an easy approach owing to the paucity of suitable families, because the condition is so lethal that often no living affected members are available. The genetic fitness of cases is also likely to be reduced owing to development of malignancies in childhood and the faulty gene may therefore die out in families, making fewer available for analysis. Also, until an individual has developed a syndromal cancer, their genetic status is unknown. Penetrance is high but not complete, so one cannot be completely sure of the status of an unaffected individual even in their sixties. The lack of an identifiable phenotype present in many of the conditions described earlier is also a major hurdle. Therefore, Malkin *et al*. (1990) used a candidate gene approach. They argued that the condition was probably due to a mutation in a tumour was an unlikely choice as retinoblastoma had not been observed in any LFS families. They chose the *TP53* gene, which was the second to be recognized, but the first to be cloned. This gene had been implicated in at least half the typical cancers featuring in LFS by analysis of the tumours by mutation and deletion studies. Malkin *et al*. examined the *TP53* gene in normal somatic cells of affected and unaffected members from five families. This was achieved by amplifying the genomic region encoding exons 5–8, which contain most of the conserved domains and are frequently mutated or deleted in tumours. They then sequenced the region using multiple primers. Affected members in all five families showed mutations in this region, with two occurring at codon 248, which is a hot spot for tumour mutations. All were in the conserved region IV in which no polymorphisms had been found in the germ-line before. They were also able to show loss of

the 'wild-type' allele in a tumour in one affected family member.

The involvement of *TP53* was then confirmed in a further family by Shrivastava *et al*. (1990). However, the initial impression that the process would be simple does not appear to be the case. A Manchester study (Santibanez-Koref *et al*., 1991) found that only two out of eight families had *TP53* mutations in exons 5–8, although they did confirm the hotspot at codon 248. Toguchida *et al*. (1991) found eight germ line mutations in 196 osteosarcoma patients but none in 200 controls. In contrast to previous studies, a family history was not present in some of the cases and the mutations were not all in the conserved regions of *TP53*. The absence of family history was not surprising, as one would predict a high new mutation rate in LFS to maintain the frequency of the condition in the population. Subsequent work has shown that with complete gene sequencing it is possible to identify mutations in over 70% of classical families. There is also evidence of genotype phenotype correlations, with much higher penetrance for mutations in the DNA core-binding domain (Birch *et al*., 1998). Recent evidence has also shown that up to 80% of childhood adrenocortical tumours having germ-line mutations in *TP53* (Varley *et al*., 1999). This makes this childhood tumour the most hereditary of all malignancies including retinoblastoma. While it is still likely that nearly all LFS families have *TP53* involvement, a mutation in the *hCHK2* gene has now been identified in at least one family (Bell *et al*., 1999). A typical family with a *TP53* mutation at codon 191 is shown in **Figure 7**. The early age of the breast cancers is particularly noticeable, with one woman having bilateral disease as well as other primary tumours. Clearly, predictive tests in this and other families are now possible. However, until a mutation is found in an affected family member, reassurance of at-risk members is not possible even if the conserved domains are screened.

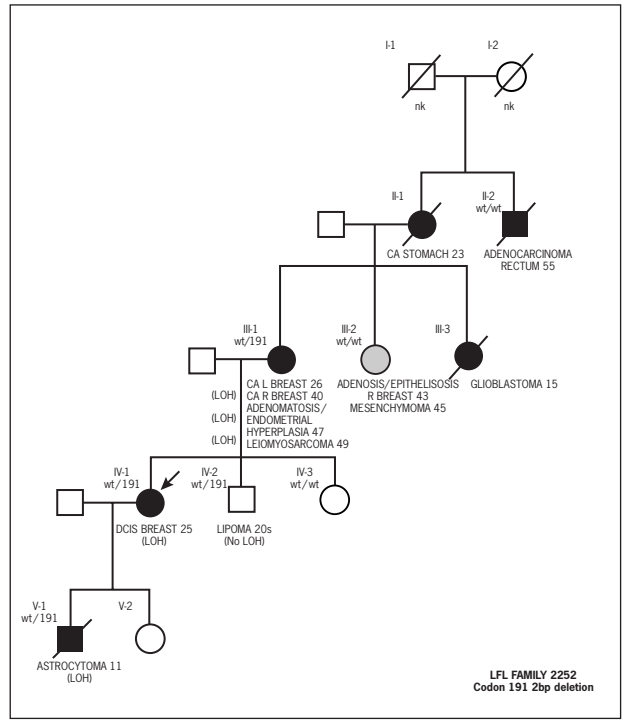


Figure 7 A family tree showing a dominant inheritance pattern of a TP53 codon 191 two base pair deletion in a family with LFS.

Breast Cancer

Breast cancer is now known to occur as part of a high penetrance predisposition such as in LFS, and in BRCA1/2 families, but may also be caused by mutations in genes such as *ATM* and *PTEN* which confer a risk of <50%. Breast cancer has long been known to have a familial tendency, as discussed earlier, and there is a profusion of supporting literature. Evidence from meticulous epidemiological studies shows that 4–5% of breast cancer is due to a dominant cancer gene with high penetrance and a population frequency in the USA of 0.003 (Claus *et al*., 1990). Studies in the UK have confirmed this population frequency and gave useful data on which risk estimation can be based. Important factors which point towards a possible familial predisposition are the number of relatives, particularly first degree, who have been affected, the age at which they developed the disease (early onset more significant) and whether bilateral or associated with other tumours.

The search for the gene or genes responsible for dominantly inherited breast cancer was dogged by some of the problems found in LFS. Ascribing status is difficult in unaffected cases even late in life and many of the known affected cases have died. Obtaining samples in some cases may depend on the use of stored paraffin block material, which may be unavailable and relies on having polymerase chain reaction technology and suitable probes, which are of course the norm nowadays. Additionally, as breast cancer is so common, affecting one in 11–12 women in their lifetime in the UK (HMSO, 1998), chance aggregations are likely to occur and non-gene mutation carriers in dominant families may be affected. The other problem, which could only be found by trial and error, was that of locus heterogeneity. Many chromosomal locations had been implicated by cytogenetic and LOH studies on tumour material. Chromosomal regions known to show involvement in more than 20% of tumours by molecular studies are 1p, 1q, 3p, 11p, 13q, 14q, 15q, 17p, 17q and 18q. Many of these regions were already known to contain tumour suppressor genes, e.g. *RB* on 13q, *TP53* on 17p and *DCC* on 18q, and these genes are likely to be involved in a multistage process towards malignancy. In a major breakthrough, Hall *et al*. (1990) were able to

show linkage in some breast cancer families to 17q12–21. They looked at over 20 families from whom they had collected samples over many years, including many cases that had subsequently died. Nevertheless, they still had to use a PCR probe to work with paraffin block tumour samples in some cases. When all families were included in the linkage analysis the region on 17q was excluded. However, when the families were stratified in terms of their average age at onset, the first seven families showed a significant linkage to 17q12 (lod score approaching 6). They argued that a large proportion of early-onset breast cancer families (<46 years) were caused by a mutated gene on 17q. Without the work of Hall *et al*., and subsequently by the Breast Cancer Linkage Consortium, it could have been many years before research was focused on this region. The problem of genetic locus heterogeneity was only overcome by a combination of meticulous collection of samples, innovative ideas and luck. Another possible gene that was implicated at about this time was the oestrogen receptor gene on chromosome 6. However, this has not since been confirmed. Following the discovery of linkage of breast cancer to 17q (Hall *et al*., 1990), Narod *et al*. (1991) undertook linkage on five families with breast/ovarian aggregation. They found that three of the families were linked to a locus at 17q12–q23 and their additive lod scores reached statistical significance. Subsequent work by the Breast Cancer Linkage Consortium showed that 80% of breast/ovarian families with four or more affected patients were linked the 17q locus (Easton *et al*., 1993). The following year heralded the identification of the first major breast cancer predisposing gene *BRCA1* (Miki *et al*., 1994). Surprisingly, *BRCA1* does not appear to be involved as a significant acquired mutation (somatic) in non-hereditary breast cancer. In the same month that the *BRCA1* gene was identified, the location of a second gene dubbed *BRCA2* was announced on chromosome 13. A year later *BRCA2* was cloned and again there was little evidence of involvement in sporadic disease (Wooster *et al*., 1995). It is now clear that although mutations in *BRCA1* and *BRCA2* account for the majority of high-risk breast cancer families (85%) and nearly all breast/ovarian families, in smaller aggregations they account for <50% of the hereditary element (Ford *et al*., 1998). While there is no doubt that *BRCA1/2* are highly penetrant genes, initial estimates of the lifetime risk of 85% (Easton *et al*., 1993) appear slightly high. Population studies do detect *BRCA1* and *BRCA2* mutations in blood samples from apparently sporadic breast cancer patients (Peto *et al*., 1999). Furthermore, founder mutations in the Jewish and Icelandic populations where *BRCA1/2* mutation frequencies can be as high as 2–2.5%, are associated with lifetime risks of breast cancer of 40–60% (Struewing *et al*., 1997). Outside populations with significant founder effects the frequencies of *BRCA1/2* mutations combined is probably no higher than 0.2%.

Having identified the most important high penetrance genes, the search is on for lower penetrance genes. Aggregation of breast cancer has been shown to occur in ataxia telangiectasia heterozygotes (Swift *et al*., 1987, 1991), who are the carriers of the recessive gene which causes a disease which predisposes especially to haematological malignancy in childhood. A mother of an affected child is at 3–5-fold risk of breast cancer, which would fulfil a dominant gene model with 25–40% penetrance and a population frequency of about 0.01. Since the isolation of the *ATM* gene (Savitsky *et al*., 1995) there have been conflicting studies as to whether this gene is a significant cause of breast cancer. Breast cancer is also thought to occur in 30% of women with Cowden's disease (a condition predisposing to multiple hamartomas), but since the discovery of the underlying gene defects in the *PTEN* gene, no studies have found the gene to be involved in familial aggregations of breast cancer.

Colon Cancer

It has been estimated that about 8% of colorectal cancer is due to the inheritance of a dominant predisposing gene (Solomon, 1990). Only a small proportion of this subset is due to FAP and there are perhaps 10 times as many individuals born with a gene for so-called hereditary nonpolyposis colorectal cancer (HNPCC). This latter inherited form of colorectal cancer has been outlined by Lynch and has in the past often been further subdivided into Lynch syndrome I and II. The type I form was considered to be site specific and to particularly predispose to proximal tumours, which has a major bearing on screening (Lynch *et al*., 1988). Type II was considered to predispose gene carriers to endometrial, ovarian, pancreatic, upper urinary tract and stomach cancers as well as colorectal and multiple tumours (Lynch *et al*., 1985). The peak age for these cancers to occur is the fifth decade with proximal colorectal tumours in two-thirds of cases manifesting this complication. The cancers are probably preceded by polyp development, but the mucosa is not lined with hundreds of them as in FAP.

It had been assumed that the gene or genes predisposing to HNPCC were tumour suppressor genes. In 1988, Vogelstein *et al*. (1988) proposed the classic model of progression to cancer in which several genes were involved starting with loss of tumour suppressor genes. Activation of oncogenes at a later stage such as the *ras* genes is also important, but it had not been considered that it was the rate of mutation that was the key factor in HNPCC. The most important tumour suppressor in colorectal cancer, the *APC* gene on chromosome 5 (that causes FAP), had been cloned in 1991 (Nishisho *et al*., 1991). Clues from tumour studies led to the isolation of several other tumour suppressor genes which are important in colorectal cancer development. The earliest of these was the *TP53* gene, whose

position on chromosome 17 was first implicated in 1981. The *DCC* gene on chromosome 18 was shown to be deleted in some colorectal cancers in 1990 and the *MCC* gene on chromosome 5 in 1991. None of these genes was found to be mutated in the germ-line of familial cases of colorectal cancer. The only report of positive linkage had been that of several Lynch type II families to the Kidd blood group on chromosome 18. The major breakthrough came from yeast genetics in which genes involved in repair of DNA suddenly became major candidates for human disease. For some time it had been noticed that tumours in HNPCC and sporadic patients showed instability in the DNA, which was manifested as a different size of microsatellite repeat in the tumour DNA compared with the blood. The discovery that a human version of a yeast DNA mismatch repair gene *MSH2* mapped to recently linked locus on chromosome 2 quickly led to the identification of the first HNPCC gene (Fishel *et al*., 1993). This then allowed a candidate gene approach, which proved that another mismatch repair gene, *MLH1*, was also an important contributor to HNPCC (Bronner *et al*., 1994). Since that time, two further genes, *PMS1* and *PMS2*, have been implicated in a tiny proportion of families and more recently *MSH6* has been found to cause HNPCC and also families with endometrial cancer. Although the mismatch repair genes are inactivated in both copies like a tumour suppressor gene, the mechanism to cancer development is different. Although it had been thought that as much as 13% of colorectal cancer could be due to HNPCC, it is now clear that the mismatch repair genes and in particular *MLH1* become inactivated somatically sometimes by methylation. More realistic estimates of HNPCC are therefore that it accounts for 1–2% of colorectal cancer. This means there are significant unidentified genes yet to be discovered.

Ovarian Cancer

Ovarian cancer, like breast and colon cancer, has had many reports of familial aggregation dating back at least to 1950. Increased risk of ovarian malignancy may be inherited as part of several genetic conditions. Gorlin syndrome (Strong, 1977) Peutz–Jeghers syndrome and XY females are all at heightened risk. In addition to this, ovarian cancer is part of HNPCC and breast/ovarian aggregation (now known to be due to *BRCA1* and *BRCA2*). There have been several reports of familial site-specific ovarian cancer (Fraumeni *et al*., 1975), but many contain cases of breast and other malignancies. The association of breast and ovarian cancer in both family reports and epidemiological studies of breast (Ridolfi *et al*., 1977) and ovarian cancer (Schildkraut *et al*., 1988) had suggested the presence of an autosomal dominant predisposing gene.

Tumour studies are less numerous than in either the breast or colon, but loss of constitutional heterozygosity has been found on chromosome regions: 3p, 6q, 11p, 13q, 17p, 17q and Xp. Following the isolation of *BRCA1* and *BRCA2*, it became clear that two of these loci (17q and 13q) were significant for hereditary disease and that apparently site-specific ovarian cancer was mainly caused by *BRCA1* (Steichen-Gersdorf *et al*., 1994). Whether this predisposition is mainly to ovary or breast may depend on the position of the mutation in each gene (Gayther *et al*., 1995). It is now thought unlikely that there is a significant other ovarian cancer gene (Ford *et al*., 1998). Predictive tests are now possible in many families once the underlying mutation has been identified. Current evidence suggests that about 40–60% of women at risk in these families will opt for testing and >50% will opt for prophylactic surgery for the ovaries and or breasts (Meijers-Heijboer *et al*., 2000).

CONCLUSIONS

The last 10 years has seen an enormous advance in our understanding of cancer and its familial elements. A great deal of this knowledge derives from the study of rare cancer-predisposing syndromes. This research is not esoteric because the cloning of these genes will benefit not only the small proportion of people who suffer from these conditions but also those who suffer from the common cancers occurring in these syndromes. Gene therapy, which could be directed at replacing the function of a deleted tumour suppressor gene, may be available in the early part of the twenty-first century. From the diseases in the first section this could be applied to cancer of the colon, skin, kidney and thyroid as well as virtually all common brain tumours.

The cloning of further tumour suppressor genes for breast, colon and ovarian cancer will have major implications in the treatment of these common cancers. The possibility of preventive treatment in high-risk families is also a real hope.

Currently, predictive genetic tests are available at some specialist genetic centers for FAP, NF1, NF2, von Hippel–Lindau disease, MEN1, MEN2 and tuberous sclerosis, HNPCC, BRCA1 and BRCA2 and can be combined with clinical screening protocols (**Table 3**). These may depend on a suitable family structure (blood samples for linkage are needed on the extended family including two affected cases) as even when the genes are cloned no guarantee can be made of identifying the underlying mutation (see NF1/BRCA1/2/HNPCC). Faster and more sensitive methods of gene mutation identification and sequencing will open the way for more readily available mutation studies in these cloned genes. Even while this chapter is being published further discoveries will be made and much of what is written here will be superseded in 5 years or so. The preceding sections should, however, give the reader a good grasp of the current state of the art and the discoveries that have brought it about.

Table 3 Chromosomal location and implications of various dominant cancer syndromes

Disease	Location	Tumours	Probable earliest tumour (years)	Risk in lifetime (%)	Start of screening (years)
FAP	5q	Adenomas	1st	100	10–16
		Bowel cancer	?4, 7	99	
NF1	17q	Neurofibroma	1st	100	Birth
		Glioma, sarcoma	1st	10	
NF2	22q	Schwannomas	1st	100	Birth
		Meningiomas	1st	60	
		Gliomas		10	
vHL	3p	Haemangioblastoma	1–2	90	5
		Renal carcinoma	20	70	15
MEN1	11q	Parathyroid, insulinoma, gastrinoma	5	95	5
MEN2a	10q	Medullary thyroid cancer, parathyroid, phaeochromocytoma	3	80	3–4
MEN2b	10q	As in MEN2a, except parathyroid	1	100	Birth
Gorlin	9q	Basal cell carcinoma	5	90	Birth
		Medulloblastoma	1	5	
Cowden	10q	Breast	30	30	35
		Thyroid			
LFS	17p	Sarcoma (bone/soft tissue), adrenal, breast cancer, gliomas.	1st	95	1st
BRCA1	17q	Breast, ovary, colon, prostate carcinoma	>16	80–90	30
BRCA2	13q	Breast, ovary, colon, prostate carcinoma	>16	80–90	30
		Male breast		10	
HNPCC	2p,3p 2q,7p	Colorectum, ovary, endometrium, ureter, gastric, pancreas	>16	80	25

REFERENCES

Allanson, J. E., *et al*. (1991). Watson syndrome: is it a subtype of type 1 neurofibromatosis. *Journal of Medical Genetics*, **28**, 752–756.

Bell, D. W., *et al*. (1999). Heterozygous germ line hCHK2 mutations in Li–Fraumeni syndrome. *Science*, **286**, 2528–2531.

Birch, J. M., *et al*. (1998). Cancer phenotype correlates with constitutional TP53 genotype in families with Li–Fraumeni syndrome. *Oncogene*, **17**, 1061–1068.

Bodmer, N. F., *et al*., (1987). Localisation of the gene for familial adenomatous polyposis on chromosome 5. *Nature*, **328**, 614–616.

Bronner, C. E., *et al*. (1994). Mutation in the DNA mismatch repair gene homolog hMLH1 is associated with hereditary non-polyposis colorectal cancer. *Nature*, **368**, 258–261.

Burn, J., *et al*. (1991). The UK Northern Region genetic register for familial adenomatous polyposis coli: use of age of onset, congenital hypertrophy of the retinal pigment epithelium, and DNA markers in risk calculation. *Journal of Medical Genetics*, **28**, 289–296.

Chandrasekharappa, S. C., *et al*. (1997). Positional cloning of the gene for multiple endocrine neoplasia type 1. *Science*, **276**, 404–407.

Clark, R. D. (1994). Proteus syndrome. In: Huson, S. M. and Hughes, R. A. C. (eds). *The Neurofibromatoses* 402–413 (Chapman and Hall, London).

Claus, E. B., *et al*. (1990). Age of onset as an indicator of familial risk of breast cancer. *American Journal Epidemiology*, **131**, 961–972.

Cohen, A. J., *et al*. (1979). Hereditary renal cell carcinoma associated with a chromosomal translocation. *New England Journal of Medicine*, **301**, 592–595.

Colley, A., *et al*. (1996). Neurofibromatosis/Noonan phenotype: a variable feature of Type 1 neurofibromatosis. *Clinical Genetics*, **49**, 59–64.

Cowan, R., *et al*. (1997). The gene for the Naevoid Basal Cell Carcinoma (Gorlin) syndrome acts as a tumour suppressor gene in medulloblastoma. *British Journal of Cancer*, **76**, 141–145.

Cushing, H. (1917). *Tumours of the Nervus Acusticus and the Syndrome of the Cerebello-pontile Angle*. (W.B. Saunders, Philadelphia).

Davies, D. R., *et al*. (1995). Severe Gardner's syndrome in families with mutations restricted to a specific region of the APC gene. *American Journal of Human Genetics*, **57**, 1151–1158.

DeVries, P. J. and Bolton, P. F. (2000). Genotype–phenotype correlations in tuberous sclerosis. *Journal of Medical Genetics*, **37**, E3.

Easton, D., *et al*. (1993). Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. *American Journal of Human Genetics*, **53**, 305–313.

Eeles, R. A., *et al*. (1996). *Genetic Predisposition to Cancer*. (Chapman and Hall, London).

European Chromosome 16 Tuberous Sclerosis Consortium (1993). Identification and characterisation of the tuberous sclerosis gene on chromosome 16. *Cell*, **75**, 805–808.

Evans, D. G. R., *et al*. (1992a). A genetic study of type 2 neurofibromatosis in the United Kingdom: I prevalence, mutation rate, fitness and confirmation of maternal transmission effect on severity. *Journal of Medical Genetics*, **29**, 841–846.

Evans, D. G. R., *et al*. (1992b). A clinical study of type 2 neurofibromatosis. *Quarterly Journal of Medicine*, **84**, 603–618.

Evans, D. G. R., *et al*. (1993). Complications of the Naevoid Basal Cell Carcinoma Syndrome: results of a population based study. *Journal of Medical Genetics*, **30**, 460–464.

Evans, D. G. R., *et al*. (1998a). Genotype–phenotype correlations in type 2 neurofibromatosis (NF2): Evidence for more severe disease with truncating mutations. *Journal of Medical Genetics*, **35**, 450–455.

Evans, D. G. R., *et al*. (1998b). Somatic mosaicism: a common mechanism for sporadic disease in tumour prone syndromes? Lessons from type 2 neurofibromatosis. *American Journal of Human Genetics*, **63**, 727–736.

Farndon, P., *et al*. (1992). Localisation of the gene for Gorlin (Naevoid basal cell carcinoma) syndrome on chromosome 9. *Lancet*, **i**, 581–582.

Fishel, F. S., *et al*. (1993). A mutator gene homolog MSH2 and its association with hereditary non-polyposis colorectal cancer. *Cell*, **260**, 1027–1038.

Ford, D., *et al*. (1998). Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. *American Journal of Human Genetics*, **62**, 676–689.

Francke, U. and King, F. (1976). Sporadic bilateral retinoblastoma and $13q^-$ chromosomal deletion. *Medical Pediatrics and Oncology*, **2**, 379–380.

Fraumeni, J. F., *et al*. (1975). Six families prone to ovarian cancer. *Cancer*, **36**, 364–369.

Friend, S. H., *et al*. (1986). A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. *Nature*, **323**, 643–646.

Fryer, A. E., *et al*. (1987). Evidence that the gene for tuberous sclerosis is on chromosome 9. *Lancet*, **i**, 659–661.

Gailani, M., *et al*. (1991). Evidence for a tumour suppressor gene on chromosome 9 in basal cell carcinomas of the skin. *American Journal of Human Genetics*, **49**, Supplement, 454.

Gardner, E. J. (1951). A genetic and clinical study of intestinal polyposis, a predisposing factor for carcinoma of the colon and rectum. *American Journal of Human Genetics*, **3**, 167–176.

Gardner, W. J. and Frazier, C. H. (1930). Bilateral acoustic neurofibromas: a clinical study and field survey of a family of five generations with bilateral deafness in thirty eight members. *Archives of Neurology Psychiatry*, **23**, 266–302.

Gayther, S. A., *et al*. (1995). Germline mutations of the BRCA1 gene in families: evidence for a genotype/phenotype correlation. *Nature Genetics*, **11**, 428–433.

Godbout, R., *et al*. (1983). Somatic inactivation of genes on chromosome 13 is a common event in retinoblastoma. *Nature*, **304**, 451–453.

Gorlin, R. J. (1987). Naevoid basal cell carcinoma syndrome. *Medicine*, **66**, 98–113.

Grundy, P., *et al*. (1988). Familial predisposition to Wilms tumour does not map to the short arm of chromosome 11. *Nature*, **336**, 374–376.

Hahn, H., *et al*. (1996). Mutations of the human homolog of *Drosophila* patched in the naevoid basal cell carcinoma syndrome. *Cell*, **85**, 841–851.

Hall, J. M., *et al*. (1990). Linkage of early onset familial breast cancer to 17q21. *Science*, **250**, 1684–1689.

Herrera, L., *et al*. (1986). Gardner syndrome in a man with an interstitial deletion of 5q. *American Journal of Medical Genetics*, **25**, 473–476.

HMSO (1998). *Cancer Statistics Registration England and Wales 1992*. (London, HM Stationery Office).

Knudson, A. G. (1971). Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences of the USA*, **68**, 820–823.

Koufos, A., *et al*. (1989). Familial Wiedemann–Beckwith syndrome and a second Wilms tumour locus both, map to 11p15.5. *American Journal of Human Genetics*, **44**, 711–719.

Lam, W. W., *et al*. (1999). Analysis of CDKN1C (p57KIP2) mutations in familial and sporadic Beckwith–Wiedemann syndrome (BWS) provides a novel genotype–phenotype correlation. *Journal of Medical Genetics*, **36**, 518–523.

Larsson, C., *et al*. (1988). Multiple endocrine neoplasia type 1 gene maps to chromosome 11 and is lost in insulinoma. *Nature*, **332**, 85–87.

Latif, F., *et al*. (1993). Identification of the von Hippel–Lindau disease tumour suppressor gene. *Science*, **260**, 1317–1320.

Li, F. P. and Fraumeni, J. F., Jr (1969). Soft tissue sarcomas, breast cancer, and other neoplasms: a familial syndrome? *Annals of Internal Medicine*, **71**, 747–752.

Li, F. P., *et al*. (1988). A cancer family syndrome in 24 kindreds. *Cancer Research*, **48**, 5358–5362.

Listernick, R., *et al*. (1989). Optic gliomas in children with neurofibromatosis type 1. *Journal of Pediatrics*, **114**, 788–792.

Liu, Z., *et al*. (1995). Germline mutations in the RB1 gene in patients with hereditary retinoblastoma. *Genes Chromosomes Cancer*, **14**, 277–284.

Lohmann, D. R., *et al*. (1996). Spectrum of RB1 germ-line mutations in patients with hereditary retinoblastoma. *American Journal of Human Genetics*, **58**, 940–949.

Lynch, H. T., *et al*. (1985). Hereditary nonpolyposis colorectal cancer (Lynch syndromes I and II): Clinical description of resource. *Cancer*, **56**, 934–938.

Lynch, H. T., *et al*. (1988). Differential diagnosis of hereditary nonpolyposis colorectal cancer (Lynch syndrome I and

Lynch syndrome II). *Diseases of the Colon and Rectum*, **31**, 372–377.

Maher, E. R., *et al*. (1990). Clinical features and natural history of von Hippel–Lindau disease. *Quarterly Journal of Medicine*, **77**, 1151–1163.

Malkin, D., *et al*. (1990). Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas and other neoplasms. *Science*, **250**, 1233–1238.

McGaughran, J. M., *et al*. (1999). A clinical study of type 1 neurofibromatosis in North West England. *Journal of Medical Genetics*, **36**, 197–203

Meijers-Heijboer, E. J., *et al*. (2000). Presymtomatic DNA testing and prophylactic surgery in families with a BRCA1 or BRCA2 mutation. *Lancet*, **355**, 2015–2020.

Messiaen, L. M., *et al*. (2000). Exhaustive mutation analysis of the NF1 gene allows identification of 95% of mutations and reveals a high frequency of unusual splicing defects. *Human Mutation*, **15**, 541–545.

Miki, Y., *et al*. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science*, **266**, 120–122.

Mulligan, L. M., *et al*. (1993). Germline mutations of the RET proto-oncogene in multiple endocrine neoplasia type 2a. *Nature*, **363**, 458–460.

Narod, S. A., *et al*. (1991). A familial breast–ovarian cancer locus on chromosome 17q12–q23. *Lancet*, **ii**, 82–83.

National Institutes of Health Consensus Development Conference Statement on Neurofibromatosis. (1987). *Archives of Neurology*, **45**, 575–579.

Nishisho, I., *et al*. (1991). Mutations of chromosome 5q21 genes in FAP and colorectal cancer patients. *Science*, **253**, 665–669.

Pelletier, J., *et al*. (1991). WT1 mutations contribute to abnormal genital system development and hereditary Wilms tumour. *Nature*, **353**, 431–434.

Peto, J., *et al*. (1999). Prevalence of BRCA1 and BRCA2 gene mutations in patients with early-onset breast cancer. *Journal of National Cancer Institute*, **91**, 943–949.

Ridolfi, R. L, *et al*. (1977). Medullary carcinoma of the breast. *Cancer*, **40**, 1365–1385.

Riccardi, V. M., *et al*. (1978). Chromosomal imbalance in aniridia–Wilms tumour association: 11p interstitial deletion. *Pediatrics*, **61**, 604–610.

Rouleau, G., *et al*. (1987). Genetic linkage analysis of bilateral acoustic neurofibromatosis to a DNA marker on chromosome 22. *Nature*, **329**, 246–248.

Rouleau, G. A., *et al*. (1993). Alteration in a new gene encoding a putative membrane-organizing protein causes neuro-fibromatosis type 2. *Nature*, **363**, 515–521.

Santibanez-Koref, M. F., *et al*. (1991). P53 Germline mutations in 2 out of 8 families with the Li Fraumeni syndrome. *Lancet*, **ii**, 1490–1491.

Savitsky, K., *et al*. (1995). A single ataxia telangectasia gene with a product similar to PI3-kinase. *Science*, **268**, 1749–1753.

Schildkraut, J. M. and Thompson, W. D. (1988). Relationship of epithelial ovarian cancer to other malignancies within families. *Genetics and Epidemiology*, **5**, 355–367.

Scott, R., *et al*. (1996). Familial infiltrative fibromatosis (Desmoid tumours) caused by a recurrent 3′ APC mutation. *Human Molecular Genetics*, **5**, 1921–1924.

Seizinger, B. R., *et al*. (1987). Genetic linkage of von Recklinghausen neurofibromatosis to the nerve growth factor receptor gene. *Cell*, **49**, 589–594.

Seizinger, B. R., *et al*. (1988). Von Hippel–Lindau disease maps to the region of chromosome 3 associated with renal cell carcinoma. *Nature*, **332**, 268–269.

Shrivastava, S., *et al*. (1990). Germ-line transmission of a mutated p53 gene in a cancer prone family with Li–Fraumeni syndrome. *Nature*, **348**, 747–749.

Simpson, N. E., *et al*. (1987). Assignment of multiple endocrine neoplasia type 2A to chromosome 10 by linkage. *Nature*, **328**, 528–530.

Solomon, E. (1990). Colorectal cancer genes. *Nature*, **343**, 412–414.

Solomon, E., *et al*. (1987). Chromosome 5 allele loss in human colorectal carcinomas. *Nature*, **328**, 616–619.

Spirio, L., *et al*. (1993). Alleles of the APC gene: an attenuated form of familial polyposis. *Cell*, **75**, 951–957.

Steichen-Gersdorf, E., *et al*. (1994). Familial site specific ovarian cancer is linked to BRCA1 on 17q12–21. *American Journal of Human Genetics*, **55**, 870–875

Strong, L. C. (1977). Genetic and environmental interactions. *Cancer*, **40**, 1861–1866.

Struewing, J. P., *et al*. (1997). The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews. *New England Journal of Medicine*, **336**, 1401–1408.

Swift, M., *et al*. (1987). Breast and other cancers in families with ataxia telangiectasia. *New England Journal of Medicine*, **316**, 1289–1294.

Swift, M., *et al*. (1991). Incidence of cancer in 161 families affected by ataxia telangiectasia. *New England Journal of Medicine*, **325**, 1831–1836.

Teyssier, J. R., *et al*. (1986). Recurrent deletion of the short arm of chromosome 3 in human renal cell carcinomas: shift of the c-raf1 locus. *Journal of the National Cancer Institute*, **77**, 1187–1191.

Toguchida, J., *et al*. (1991). A survey of germ-line and somatic p53 mutations in patients with bone and soft tissue sarcomas. *American Journal of Human Genetics*, **49**, Supplement, 458.

Trofatter, J. A., *et al*. (1993). A novel Moesin-, Ezrin-, Radixin-like gene is a candidate for the neurofibromatosis 2 tumour suppressor. *Cell*, **72**, 1–20.

Varley, J. M., *et al*. (1999). Analysis of a panel of patients with childhood adrenocortical tumours for germline TP53 mutations. *American Journal of Human Genetics*, **65**, 995–1006.

Viskochil, D., *et al*. (1990). Deletions and a translocation interrupt a cloned gene at the neurofibromatosis type 1 locus. *Cell*, **62**, 187–192.

Vogelstein, B., *et al*. (1988). Genetic alterations during colorectal tumour development. *New England Journal of Medicine*, **319**, 525–532.

Williams, W. R. and Strong, L. C. (1985). Genetic epidemiology of soft tissue sarcoma in children. In: Muller H. R. and

Weber, W. (eds), *Familial Cancer, First International Research Conference*. 151–153 (Karger, Basle).

Wishart, J. H. (1822). Case of tumours in the skull, dura mater, and brain. *Edinburgh Medical and Surgical Journal*, **18**, 393–397.

Wooster, R., *et al*. (1994). Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12–13. *Science*, **265**, 2088–2090.

Wooster, R., *et al*. (1995). Identification of the breast cancer susceptibility gene BRCA2. *Nature*, **378**, 789–792.

FURTHER READING

Bishop, D. T. and Thomas, H. J. W. (1990). The genetics of colorectal cancer. *Cancer Surveys*, **9**, 585–604.

Claus, E. B., *et al*. (1994). Autosomal dominant inheritance of early onset breast cancer. *Cancer*, **73**, 643–651.

Devilee, P. J. and Cornelisse, C. J. (1990). Genetics of breast cancer. *Cancer Surveys*, **9**, 605–630.

Eeles, R., *et al*. (eds) (1996). *Genetic Predisposition to Cancer*. (Chapman and Hall, London).

Evans, D. G. R. (1995). Practical implications of the new cancer genetics. In: Peckham, M., *et al*. (eds), *Oxford Textbook of Oncology*. (Oxford University Press, Oxford).

Hodgson, S. and Foulkes W. D. (1998). *Inherited Susceptibility to Cancer*. 3–19. (Cambridge University Press, Cambridge).

Huson, S. M., *et al*. (1989). A genetic study of von Recklinghausen neurofibromatosis in south east Wales. II. Guidelines for genetic counselling. *Journal of Medical Genetics*, **26**, 712–721.

Lemoine N. R., *et al*. (1994). *Cancer: A Molecular Approach*. (Blackwell Scientific, Oxford).

Varley, J., *et al*. (1997). Li–Fraumeni syndrome – A molecular and clinical review. *British Journal of Cancer*, **76**, 1–14.

Chapter 5

Human DNA Tumour Viruses

Beverly E. Griffin and Wilson Caparrós-Wanderley
Imperial College School of Medicine at St. Mary's, London, UK

CONTENTS

INTRODUCTION

Almost all vertebrate species, whether man, monkey or marsupial, have their own cohort of viruses. These are usually species specific, although they may share sequence homologies, genes and gene functions with similar viruses from other species. In practice, this has meant that a detailed study of a virus from one species may have profound impact on predicting the properties of a similar virus in another species, particularly where the viral sequences are available for comparison. However, this is not always the case, as best illustrated for two small, highly related DNA tumour viruses, the primate virus, simian virus 40 (SV40), and the mouse virus, polyomavirus (Py). For SV40, all activities for regulating cell growth, and transformation to a tumorigenic phenotype, are carried out by one viral protein, the large T-antigen (LT). However, with Py, differential splicing of the RNA, made from a single region of the viral genome, gives rise to distinct messages and two proteins – the so-called large and middle T-antigens (LT and MT) – where the function for altering cell growth is carried by one protein, LT, and that for inducing cellular transformation by the other, MT. Thus, cross-species speculations about related viruses need to be made with care. Similarly, the general notions about host range specificity of a virus cannot be deemed absolute, since mutations can occur that result in host alteration. One of the most interesting cases in point, in the DNA virus field, comes from the identification in Africa of a pox virus that appears recently to have crossed the species barrier from monkey to man. The current interest in 'emerging viruses' focuses on the origins of new species and evolutionary and environmental factors that contribute to their birth (Morse, 1993).

This chapter deals with a subset of DNA viruses, those associated with tumour formation, and is restricted to discussing in detail only those viruses where a good case can be made for an association with human cancer. In animal models, such as those employing SV40 and Py, the experimental evidence showing tumour causality in appropriate models is unambiguous. For human disease, the situation is by definition more complicated and causality or association depends on drawing together studies from many fields, including not only virology, but also epidemiology, oncology and molecular biology. Notably, with the four viruses that are discussed in detail, that is, papilloma viruses, two herpesviruses (Epstein–Barr and Kaposi sarcoma-related virus) and hepatitis B virus, all have the ability to persist in infected cells and evade host immune systems. For none of these viruses has a very good *in vitro* lytic system been identified, impeding progress in answering crucial questions about them. Even without this, much progress has been made, mainly owing to the judicious use of molecular methods, and model systems, where identified.

Over the last few years, the International Agency for Research on Cancer (IARC), Lyon, as part of their programme for evaluating carcinogenic risks to humans, has considered four of the human DNA viruses most likely to play a causal, predisposing, or auxiliary role in the development of cancer in humans. These are various members of the papillomavirus (PV) family (related to SV40 and mouse Py, and in former times considered to belong to the same papovavirus family), hepatitis B virus (HBV), and the two members, Epstein–Barr virus (EBV) and Kaposi sarcoma-associated virus (KSHV), of the herpesvirus family. The general conclusion reached by a consortium of individuals contributing to reports on three of these viruses (HPV, HBV and EBV) is that 'there is sufficient evidence' for their carcinogenicity, as it relates to defined forms of cancer (IARC, 1994, 1995, 1997). For KSHV, the most recently identified of these viruses, the conclusion reached in 1998 is that 'KSHV is probably carcinogenic to humans.' The ultimate proof of viral causality of malignancy will be concomitant eradication of the virus and of the disease, and is a target for the future. In

the case of HBV, where there are effective antiviral vaccines, this could be realized in the foreseeable future.

PAPILLOMAVIRUSES (PVs)

General Definition and Classification

Papillomaviruses (PVs) are a family of DNA viruses that cause hyperproliferative lesions of the mucosal and cutaneous epithelia (papillomas, warts and condylomas) in a wide variety of higher vertebrates, including humans. Most of these lesions are benign, self-limiting and regress with time, but some of them tend to progress towards malignancy and invasive carcinoma (e.g. carcinoma of the uterine cervix).

All PVs belong to the subfamily Papillomavirus, which constitutes one of the two members of the family Papovaviridae. The other member of this family, the subfamily Polyomavirus, is discussed later in this chapter. PVs and Polyomaviruses were initially grouped together because they share properties of small-sized, nonenveloped virions, icosahedral capsids, superhelical double-stranded DNA genomes, and use the nucleus as site of multiplication. Subsequent research has shown that, despite these similarities, the two genera are not evolutionarily related. They have different genomic organisations, their DNAs do not hybridize and there is no immunological cross-reactivity between them. Furthermore, in contrast to Polyomaviruses, PVs multiply only in differentiating epithelium and cannot be propagated *in vitro* (Howley, 1996).

PVs are highly species specific, hence their classification is based on their host range and DNA relatedness. Each virus is first named after its natural host followed by a number, and sometimes a letter, which indicates, respectively, its type and subtype (e.g. bovine (B)PV-4, Human (H)PV-6b, etc.) (**Table 1**). Classification of different isolates from one species into types and subtypes is based, at present, on their degree of sequence homology. On the basis of the site of infection, HPVs have also been classified into two main groups: cutaneous and mucosal. Each group can, in turn, be subdivided into 'high-' or 'low'-risk types according to the probability of malignant progression associated with the type of lesions they cause. Both this approach and the sequence homology method give rise to equivalent phylogenetic trees (**Figure 1**) (Shah, 1990).

Table 1 PV-associated pathological conditions

Species	Pathology	Virus Type[a]
Deer	Cutaneous fibropapillomas	DPV
Cattle	Alimentary tract carcinoma	**BPV-4**
	Cutaneous fibropapillomas	BPV-2
Cottontail rabbit	Skin carcinomas	**CRPV**
Humans	Skin warts	HPV-1, -2, -3, -7 and -10
	Epidermodysplasia verruciformis	HPV-**5**, **-8**, -17 and -20
	Anogenital warts (condylomas):	
	Exophytic condylomas	HPV-6 and -11
	Flat condylomas	HPV-**16**, **-18**, **-31**, **-33**, -42, **-43**, etc.
	Respiratory tract papillomas	HPV-6 and -11
	Conjunctival papillomatosis	HPV-6 and -11
	Focal epithelial hyperplasia (FEH)	HPV-13 and -32

[a]Virus types predominantly recovered from malignant lesions are indicated in bold.

Virion Structure

Nonenveloped icosahedral PVs replicate in the nucleus of squamous epithelial cells. PV particles are about 50 nm in diameter and encapsulate a single copy of the circular 8 kbp double-stranded DNA genome in the form of a chromatin-like complex with cellular histones. They have a density in caesium chloride of 1.34–1.36 g L^{-1} and, owing to the lack of lipids, are resistant to ether and other solvents (Pfister and Fuchs, 1994).

The viral capsid is composed of 72 capsomers with a star-shaped morphology and displaying a cylindrical channel along their axis (**Figure 2**). All capsomers are pentamers of the L1 protein, a 55-kDa protein which represents about 80% of the total capsid protein. L1 protein is required for virus attachment to the cell surface receptor and constitutes the basic structural component of the capsid. The remaining 20% of the capsid is composed of a 70-kDa protein known as L2. The exact function(s) of this protein is still unclear, but it may be involved in the efficient self-assembly of the viral capsid and attraction and/or proper positioning of the viral genome during viral assembly (Howley, 1996).

Genomic Organisation

The Papillomavirus genome is divided into an 'early' region (about 4.5 kbp in size), a 'late' region (about 2.5 kbp) and a long regulatory region (LCR) (about 1 kbp). There are two open reading frames (ORFs) in the late region (L1 and L2) and up to eight ORFs (E1 to E8) in the early region. There are no ORFs in the LCR, but this region contains the viral origin of replication and control elements for transcription and replication. In contrast to Polyomaviruses, all ORFs in the Papillomavirus

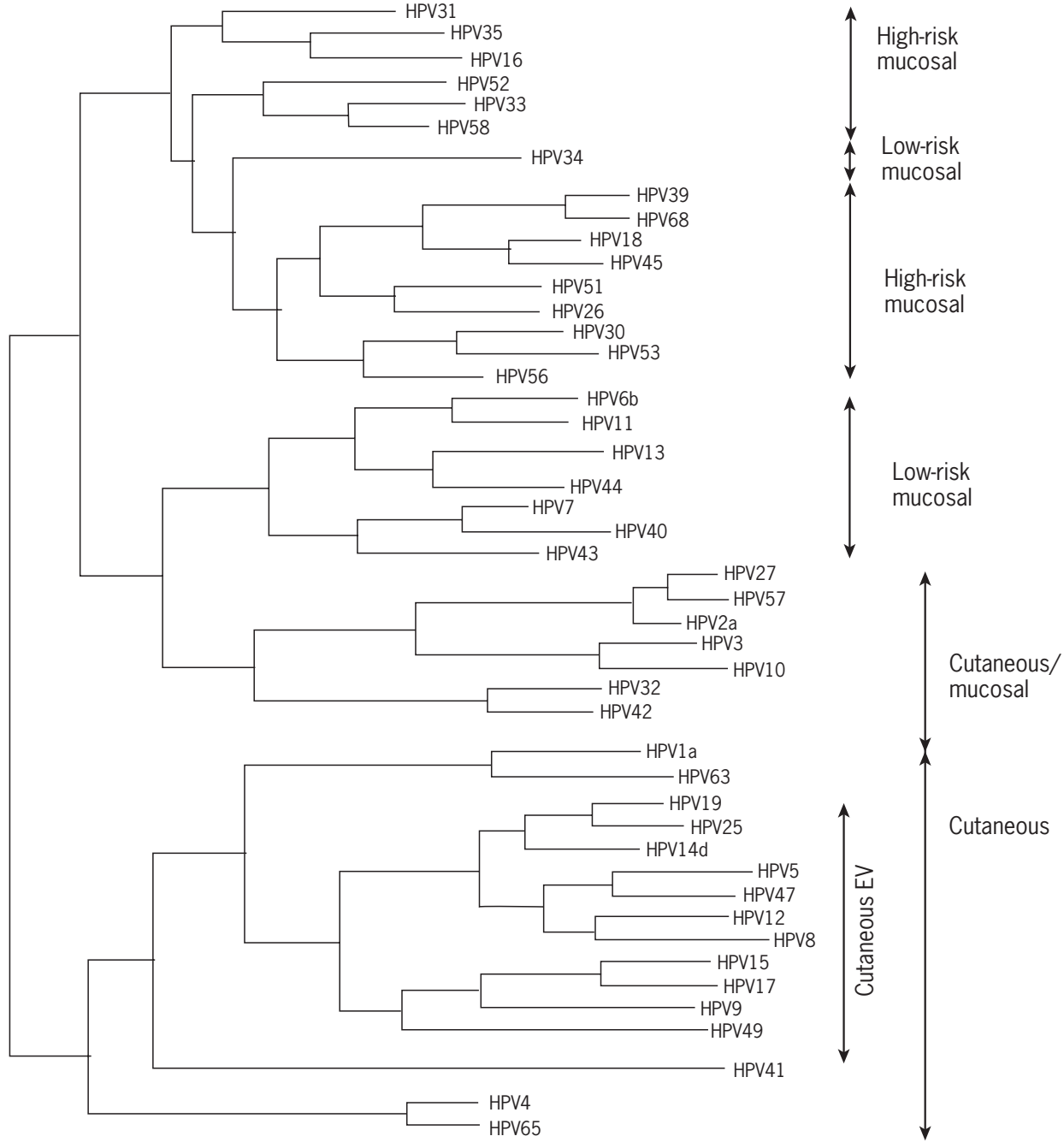


Figure 1 Phylogenetic tree of HPVs, based on DNA sequence homology of a 384-bp fragment of the E6 ORF. The clinical classification of HPV types, according to the site of infection, is indicated on the right. (Adapted from Van Ranst *et al.*, 1992.)

genome are located on one DNA strand (**Figure 3**) (Shah, 1990).

The properties of the proteins encoded by the two late ORFs, L1 and L2, have been described in the previous section.

The **E1** proteins, molecular weights (MW) 68–85 kDa, are essential for viral replication and in this role they are similar to the SV40 large T-antigen. They are phosphoproteins with DNA-dependent ATPase and ATP-dependent helicase activities. The 5′ portion of the E1 ORF can sometimes be translated as a smaller protein involved in modulation of viral replication (Chow and Broker, 1994).

The **E2** ORF codes for a family of proteins of which only the full-length member, MW 43–48 kDa, can support viral replication. Full-length E2 is also a transcriptional activator whilst truncated forms of E2 (also known as E2C

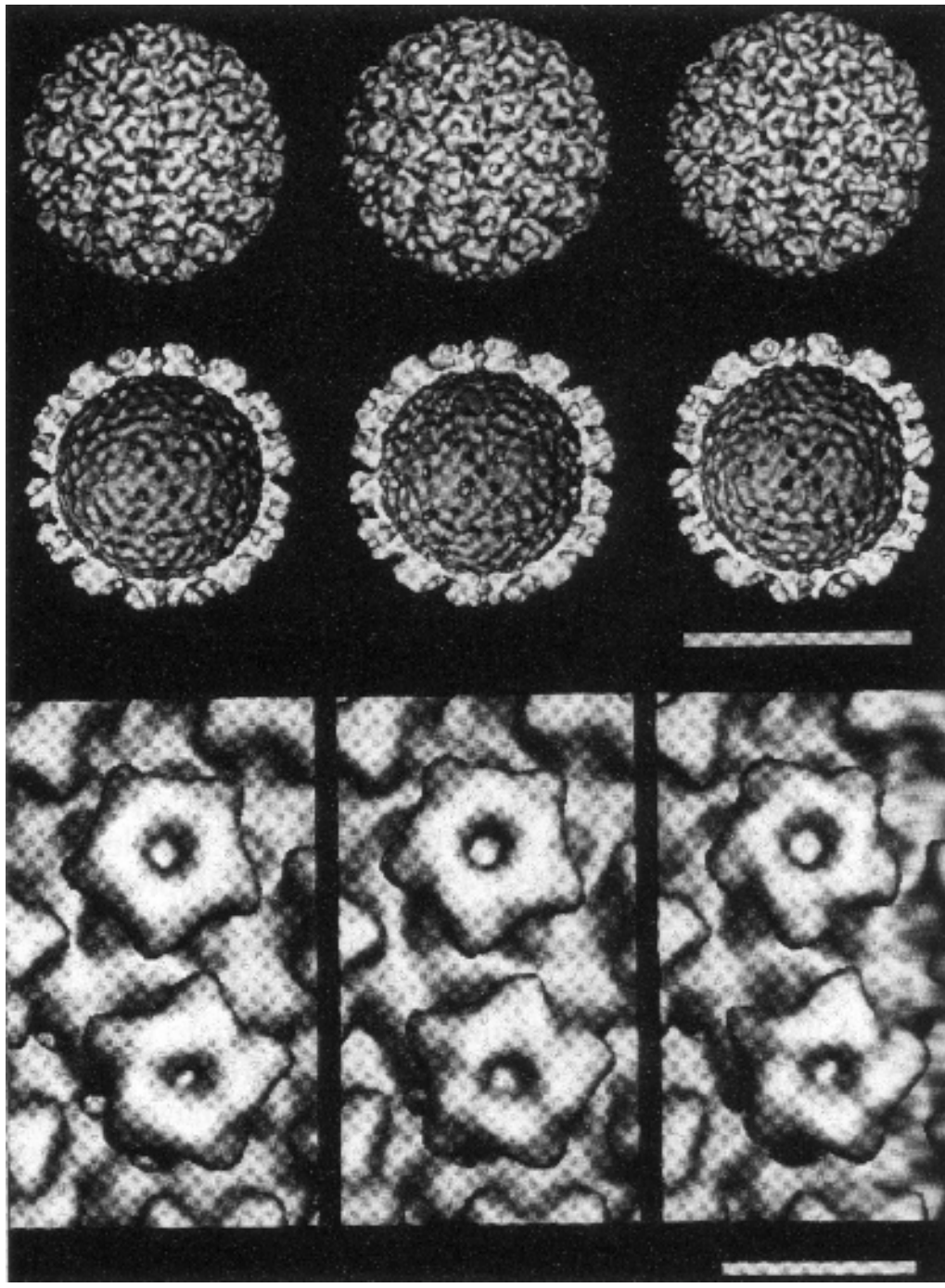


Figure 2 Structure of HPV. Surface-shaded representations of reconstructions of HPV-1 from warts (left columns), L1 capsids (middle columns) and L1L2 capsids (right columns). Outside view of capsids (top row), inside view (middle row) and a close-up view of pentavalent and hexavalent capsomeres (bottom row). Internal density to a radius of 20 nm was computationally removed to show internal features of the capsid. No differences are apparent. Bars = 50 nm (top and middle rows) and 10 nm (bottom row). (From Hagensee *et al.*, 1994, *Journal of Virology*, **68**, 4503–4505.)

and E2M), derived from different promoter usage and alternative mRNA splicing, are transcriptional repressors (Arrand, 1994; Chow and Broker, 1994).

The **E3** ORF is only present in some PVs and its function is not yet known.

The **E4** protein is expressed from a spliced mRNA (E1^E4) as a doublet of MW 17 and 16 kDa. E4 appears to be involved in the disruption of the cytokeratin network during the late phase of the viral life cycle. In HPV1 induced warts, E4 accumulates abundantly (up to 30% of

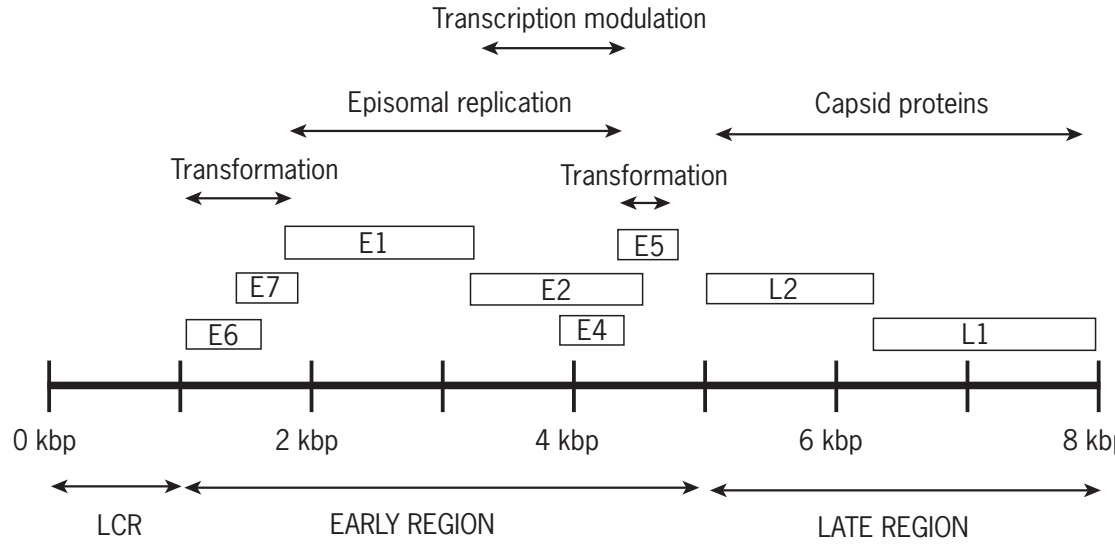


Figure 3 Organisation of a PV consensus genome (in kilobase pairs). The open boxes represent ORFs labelled E1–E6, or L1 and L2, according to their position in the 'Early' or 'Late' region of the genome. Locations of gene functions are listed above the genome. (Adapted from Pfister and Fuchs, 1994.)

the total protein mass) with the capsid proteins, but this does not occur in lesions caused by other PVs (Arrand, 1994).

The **E5** proteins are small (44–90 amino acids) and extremely hydrophobic polypeptides, which are present within intracellular membrane compartments, including the Golgi apparatus. In bovine (B)PVs, E5 is the major transforming protein and appears to stimulate mitogenesis by interfering with growth factor receptor signal transduction pathways (Stöppler *et al*., 1994).

The **E6** proteins, MW 16–18 kDa, have transforming and transcriptional transactivating activities and are localized in the nuclear matrix and in non-nuclear membrane fractions. E6 and E7 proteins are the major transforming proteins of HPVs. Both proteins appear to have arisen from duplication events involving a Cys-X-X-Cys motif. E6, similarly to adenovirus E1B and SV40 LT, associates with the tumour suppressor p53, resulting in its ubiquitin-dependent degradation. In HPVs, this degradation is mediated only by the E6 proteins of 'high-risk' but not 'low-risk' types, suggesting an important role for this process in the development of malignancy (Stöppler *et al*., 1994).

The **E7** proteins, MW 14–21 kDa, have transforming and transcriptional transactivating activities and are localized in the cytoplasm and the nucleolus. E7, similarly to the adenovirus E1A and SV40 LT, binds to the retinoblastoma tumour-suppressor gene product Rb-p105. This prevents Rb-p105 from interacting with the E2F transcription factor, thereby allowing initiation of the S-phase of the cell cycle. The E7 proteins of 'high-risk' HPVs are more effective in binding Rb-p105 than that of 'low-risk' types. E7 also has the ability to bind cyclins and cyclin-dependent kinases (cdks) and hence further disrupt the regulation of the cell cycle (Stöppler *et al*., 1994).

The **E8** ORF is only present in some PVs. In BPV4, the E8 protein is a small hydrophobic polypeptide localized in the cell membrane. E8 contributes to cell transformation by conferring anchorage-independent growth.

There are marked differences in the state and functional activity of PV genomes in benign tumours and in cancers of different species (Shah, 1990; Chow and Broker, 1994; Howley, 1996). In benign lesions the viral genome replicates as multicopy extrachromosomal plasmids. In carcinomas, however, the situation can be completely different. In cattle, the BPV type 4 genome is detected in alimentary tract papillomas but is apparently lost in carcinomas. In cottontail rabbit carcinomas the viral genome is present in the episomal form. In contrast, in some human genital tract cancers the viral genome, generally accompanied by deletions and mutations, is integrated into the cellular DNA. Integration with respect to the cellular DNA is not site-specific, but there is some specificity with respect to the site in the circular viral genome where the break for integration occurs. Viral genomes of HPV 16 and HPV 18 are found to be almost always interrupted in the E1–E2 region, producing a break that disrupts transcription of the E2 ORF, but not transcription of the E6, E7 and part of the E1 ORFs.

Transcription and Replication of the Viral Genome

The subdivision of the viral genome into 'early' and 'late' regions is based on the close association existing between viral replication and squamous epithelial differentiation. Upon entry into the basal stem cells of the epithelium, viral early gene expression is activated at a very low level,

leading to the temporary amplification and establishment of the DNA plasmids. Early gene expression in the basal cell layer also stimulates cell growth and, as the epithelial cells move upwards in tissues progressing through their differentiation programme, this pattern of expression is maintained. Finally, in the uppermost layers of the epithelium, viral replication and late gene expression are activated in the now fully differentiated epithelial cells, resulting in the production of infectious viral progeny (Chow and Broker, 1994; Howley, 1996).

Several factors contribute to the transcriptional complexity of PV genomes (Arrand, 1994; Pfister and Fuchs, 1994; Howley, 1996): first, the presence of multiple promoters (for example, the LCR of BPV-1 contains at least seven promoters); second, complex and multiple splice patterns which, associated with the activity of the different promoters and the use of different polyadenylation signals, give rise to an extensive variety of viral mRNAs, many of which are polycistronic; furthermore, some ORFs (e.g. E2 and E6) appear to be represented to different extents in different messages; finally, complex control by proteins and factors produced by the virus (E2, E6 and E7) or the host at both the intracellular (i.e. retinoic acid, NF-IL6, Oct-1, etc.) and extracellular (glucocorticoid hormones, TGF-β1 and -β2, EGF, etc.) levels occurs.

Detection of HPV Infection

Although detection and diagnosis of HPV-associated lesions can normally be achieved by colposcopy, histology and cytology, these methods are unable to identify specific HPV type(s) present in lesions. Serological responses against almost all HPV-derived antigens have been detected in infected individuals. However, the diagnostic utility of these serological responses is questionable because (1) they appear to be, for most antigens, non-type-specific, and (2) they persist for longer than the actual infection. Detection of HPV DNA is, therefore, the only reliable diagnostic tool available to establish current infection by specific HPV types. This approach involves the detection of the viral genome either directly (by Southern blot hybridization) or by polymerase chain reaction (PCR) protocols (Shah and Howley, 1996).

Pathogenesis of Infections

In contrast to some animal PV infections in which fibroblastic proliferation is prominent, the pathological effect of HPV infection is confined to the epithelium. All the layers of the normal epithelium are represented in the lesion, accompanied by certain characteristic histological features. The increased division rate of the basal cell layer leads to an irregularly thickened prickle cell layer, with abnormal mitoses also observed in all suprabasal layers. The granular layer contains foci of cells showing koilocytosis (cytoplasmic vacuolization) and nuclear changes (enlargement, hyperchromasia, degeneration and pyknosis). Koilocytosis is also a feature of the cornified layer of non-nucleated dead cells, which may also display hyperkeratosis (Shah, 1990).

HPV infection can be acquired in a variety of ways such as abrasions of the skin, sexual intercourse and passage through an infected birth canal, and results in a variety of clinical conditions (**Table 1**) (Shah and Howley, 1996). Most of these lesions have benign prognoses, but they may be associated with high levels of morbidity. For example, exophytic anogenital warts, one of the most common sexually transmitted diseases, usually cause itching, burning and pain, and have a significant negative effect on the psychosexual wellbeing of the individual. Other lesions, however, may undergo malignant transformation. In the case of flat anogenital warts, lesions in the uterine cervix may progress towards invasive carcinoma and 50% of the diagnosed population will ultimately die of the disease, accounting for about 15% of cancer-related deaths worldwide. The progression of benign papillomas to invasive cancers has certain characteristics which are shared in different species (Shah, 1990; Shah and Howley, 1996). First, only some of the virus types that infect a species have oncogenic potential. In epidermodysplasia verruciformis (EV), more than 20 different HPV types are recovered from the macular plaques characteristic of this disease. However, only two types, HPV5 and HPV8, predominate in the carcinomas that arise from these lesions. Similarly, over a dozen HPV types infect the human genital tract, but a majority of genital tract carcinomas are associated with only a few viral types, the so-called 'high-risk' types (predominantly HPV 16 and 18). Second, there is a long period between the initial infection and the development of invasive cancers. In humans this period may be between 5 and 40 years. Finally, cofactors are often involved in malignant progression. For example, carcinomas in EV patients arise preferentially in lesions that are exposed to sunlight.

Immunology of Infection

Viral infections are controlled by a combination of non-antigen-specific and antigen-specific immune responses. Most viruses induce these immune responses by causing lytic cell death which, in turn, causes inflammation and stimulates the production of cytokines. PV infection, in contrast, is non-lytic and, consequently little or no local inflammation is induced. This situation probably reflects the reduced ability of PVs to invoke effective immune responses that are capable of eliminating established lesions (Frazer, 1996). Nonetheless, there is evidence of involvement of the immune system in the control of PV infections.

Humoural (antibody) immune responses directed against almost all PV proteins have been detected in

infected individuals. However, only antibodies directed against L1 or L2 have been found to be neutralizing and to protect against infection. Antibodies against the E6 and E7 proteins of high risk HPVs, although not effective at the prophylactic level, are commonly associated with carcinoma of the uterine cervix. Antibodies against the E2 and E4 proteins have also been associated with carcinoma of the cervix, but not universally (Frazer, 1996; Shah and Howley, 1996).

The persistence of PV-induced lesions suggests that the development of an effective cellular immune response against PVs following infection is neither immediate nor universal. Nonetheless, several observations suggest that the host's cell-mediated immune response is responsible for limiting the growth and promoting the regression of PV-induced lesions. First, there is a high prevalence of PV-induced lesions and malignant tumours in hosts with suppressed cellular immunity. Second, the regression of anogenital and skin warts in humans is associated with a pronounced local infiltration of mononuclear cells (activated T lymphocytes, macrophages and, to lesser extents, NK cells and B lymphocytes) invading the epidermis and destroying the neoplastic tissue. On this basis, the cellular immune response in spontaneously regressing warts appears to be consistent with a delayed type hypersensitivity (DTH) reaction to foreign antigen. Nonetheless, the presence of specific cytotoxic CD8+ T lymphocytes (CTLs), which are normally involved in the resolution of viral infections, has been notoriously difficult to demonstrate in HPV infections. Indeed, only a handful of studies (i.e. Tarpey *et al*., 1994; Nakagawa *et al*., 1997) have shown E6- or E7- specific CTLs in humans. Finally, vaccination with PV antigens has been found to induce a specific T cell proliferative or CTL responses against L1, L2, E2, E6 and E7 in animals and humans (Frazer, 1996; Shah and Howley, 1996).

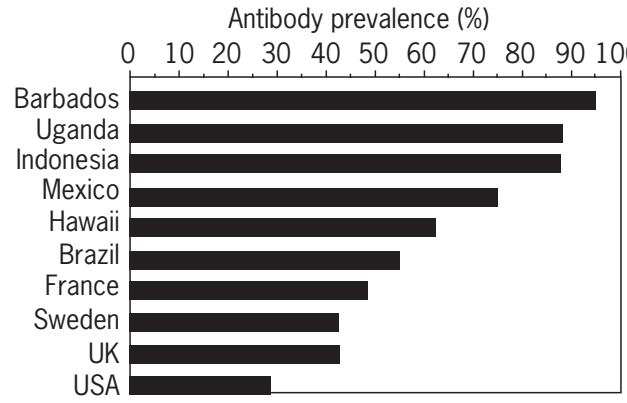


Figure 4 EBV antibody prevalence, age 4-6 years, in different parts of the world. (From IARC, 1997, p. 83.)

EPSTEIN–BARR VIRUS (EBV)

History, Definition and Classification

A relevant point to note about EBV (or HHV-4) is that it is almost ubiquitous in the adult human population, with the great majority of individuals carrying antibodies to the virus. Infection of B lymphocytes by EBV is mediated through interaction of a viral envelope glycoprotein with the receptor for the C3d complement component, CD21 (CR2), although, notably, EBV can infect epithelial cells that lack this receptor. The average age of seroconversion to this virus differs markedly in various parts of the world, usually being considerably later in the more socio-economic privileged parts of the world than in crowded, or poorer, populations. Viral antibody prevalence in terms of age is given in **Figure 4**. EBV has a particularly interesting international history. It was first observed by Epstein and colleagues in London in the 1960s on electron microscopic examination of a cell line (EB), established with extracts from an African tumour, called Burkitt's lymphoma (BL). From its physical appearance, the virus was defined as a member of the herpesvirus family. BL itself had been identified about 10 years earlier during travels by the Irish-born surgeon, Denis Burkitt in sub-Saharan Africa. Carvings from earlier periods showed this tumour of B lymphocytes to be a disease long prevalent in certain parts of Africa. Later, a genuinely serendipitous finding showed the virus to be the causative agent of infectious mononucleosis. This came about when a laboratory technician in Philadelphia, with no antibodies to EBV, developed mononucleosis and on subsequent testing was found to be EBV-antibody positive. A causal effect for EBV in mononucleosis was thus confirmed (reviewed by Griffin, 1998). An association between EBV and a tumour of epithelial cells, nasopharyngeal carcinoma (NPC), was discovered when sera from these patients were included in a general antibody screening programme in New York, also in the 1960s.

The natural reservoir of EBV, whether in the B lymphocyte or epithelial or other cell population, is still a matter of controversy. EBV is known to exist in circulating lymphocytes in the body (about 1 in 10^5 – 10^6 in a normal individual). In culture, the virus is capable of extending the lifetime of B lymphocytes for an unlimited time period, a phenomenon called immortalization. EBV is sub-classified as γ-herpesvirus, having a restricted host range with its site of latency residing in lymphocytes, compared for example with α-herpesviruses, such as the simplex viruses, which have broad host ranges, and are latent in sensory ganglia. Herpesvirus classifications are given in **Table 2**. Although many other mammalian herpesviruses, notably those from Old World primates, belong to the γ-herpesvirus subfamily, only one other human herpesvirus identified to date, that of KSHV (HHV-8), belongs to this subclass. A full list of

Table 2 Biological characteristics of herpesvirus subfamilies

Characteristic	Alpha	Beta	Gamma
Genus	Simplexvirus Varicella-Zoster virus	Cytomegalovirus Muromegalovirus	Lymphocryptovirus Rhadinovirus
Host range	Broad	Restricted	Restricted
Prevalent genomic organisation*	D, E	Variable	B, C
Productive cycle	Short	Long	Long
Spread in culture	Efficient	Moderate	Poor
Site of latency	Sensory ganglia	Lymphoreticular tissues	Lymphocytes
Proliferation of latently infected cells	No	No	Yes

(From IARC, 1997, p. 36.)
*See **Figure 5**.

herpesviruses and their taxonomies is given elsewhere (IARC, 1997).

Unlike the Papillomaviruses, there is no well-defined classification of different strains of EBV. The viral genome is composed of double-stranded DNA, with sizes that range from about 175 to >200 kbp pairs in cells from different sources. In lytically infected (virus producer) cells, the viral DNA is linear and carries small repetitive sequences at each end. In latently infected (non-virus-producing) cells, it is circular, having undergone recombination via its repetitive terminal sequences. With regard to its structure, EBV differs from other human herpesviruses, as illustrated in **Figure 5**, and in organisation more resembles that of its host cell DNA than do the other viruses. Its size variation is not a property that has been used in classification since the viral genome is composed of unique sequences interspersed with repetitive elements, and size is largely dictated by copy numbers of the repeats. The largest of these, called IR1, or *Bam*HI W after the restriction enzyme that cleaves it intact from the DNA (**Figure 6**), is >3000 bp in size. An obvious classification sought, but not found, has been one that would allow for association of specific viral types with different EBV-associated malignancies, two of which are noted above. To date, pathology-specific strains of virus have not been identified. Rather, restricted viral gene expression, in part controlled by the host cell, may play a role in the genesis of a pathological lesion. Genetic polymorphisms, designated 1 and 2, have, however, been identified which differ in sequences of some viral nuclear antigen (EBNA) genes, and to some extent in their biological properties, and their global localization. Unlike the better-defined distinctions of herpes-simplex viruses 1 and 2, which are separately classified, the functional differences between EBV 1 and 2 are not sufficiently distinct to allow for unique classification. Indeed, the polymorphisms may merely reflect 'hotspots' for mutational recombination events in the genome. Such a hypothesis is not totally fanciful. At least one viral isolate, Jijoye (from a primary African BL), on propagation in culture, has given rise to a novel isolate, P3HR-1, with a deletion that maps within one of the key latent EBV genes (that for EBNA-2, see below) affected by the polymorphism in EBV 1 and 2.

Virion Structure

Whereas smaller DNA viruses, such as SV40, Polyoma, Papilloma and Adenoviruses, are nonenveloped, all the herpesviruses have an outer envelope and within this, a capsid that contains the viral DNA. By electron microscopy (EM), in composition and appearance EBV resembles a typical herpesvirus with a toroid-shaped protein core wrapped with genomic DNA, as shown in **Figure 7**. Its nucleocapsid is composed of 162 capsomeres and its outer envelope is made up of glycoprotein (gp) spikes, many of which are composed of a 220/350-kDa protein, the principal target of a virus-neutralizing antibody response. Size variation of this protein reflects the number of glycosylated amino acid residues it contains. To date, gp220/350 is still the prime candidate for producing an EBV vaccine that might prevent, or delay, infection *in vivo*. The high lipid content of the envelope results in relative instability of EB virions at room temperature, and their rapid inactivation by lipid solvents, such as ether and chloroform, or by detergents. This is another difference between the herpesviruses and the small DNA viruses, the latter being generally stable under these conditions. Between the nucleocapsid and the envelope is a region called the tegument, which is frequently distributed asymmetrically, and by EM shows no distinctive features. The composition of the tegument in EBV has been much less carefully studied than in some other herpesviruses, notably herpessimplex viruses.

Genomic Organisation and Key Viral Latent Functions

EBV was the first herpesvirus to have its complete DNA sequenced, as presented simplistically in **Figure 6**, determined (Baer *et al*., 1984). In its overall structure, with unique sequences interspersed with repetitive elements, the viral genome appears to be a mini-version of its human host, with one notable exception, that is, every repetitive region (IR1–IR4 and TR) includes ORFs, occurs within a gene and also encodes a protein. There is no 'junk' DNA, the role often assigned to repetitive sequences in cellular

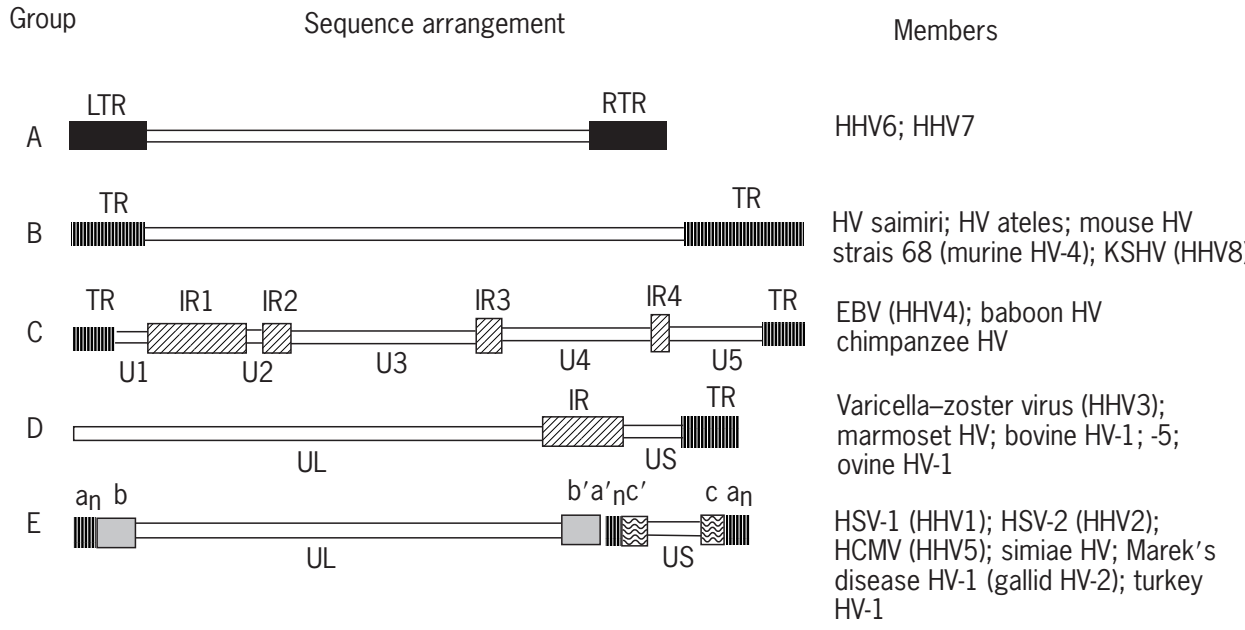


Figure 5 Architecture of herpesvirus genomes, showing unique (U) and repetitive (R) regions. Viruses have been grouped in categories (A–E) and their designations are given on the right. According to the nomenclature used by different groups, LTR = left terminal repeat; RTR = right terminal repeat; TR = terminal repeat; IR = internal repeat; UL = long terminal repeat; US = short teminal repeat. Symbols used in group E viruses represent sequence arrangements within repeats. (Adapted from IARC, 1997, p. 35.)

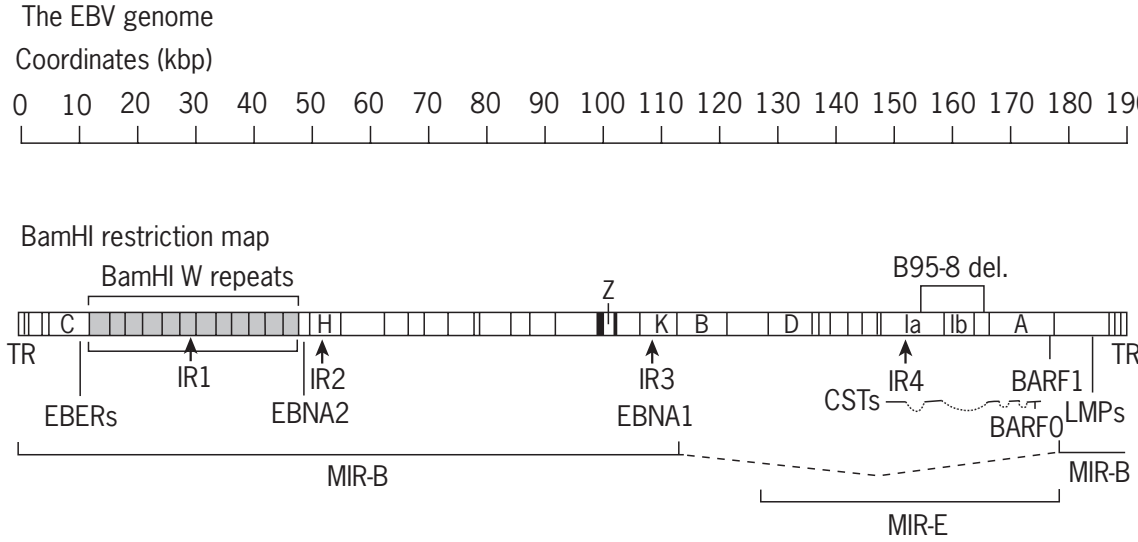


Figure 6 Physical map of the EBV genome, and location of some key gene and repetitive (IR) elements (see **Figure 5**). Coordinates for a 'typical' genome are given and genes allocated to the *Bam*HI restriction enzyme fragment in which they are located (see **Figure 10**). The CSTs encompass several restriction fragments. Its transcript is shown. MIR-B contains the minimum viral sequence required for immortalization of B lymphocytes and MIR-E the minimum for epithelial cells. (From Griffin and Xue, 1998, *Annals of Medicine*, **30**, 249–259.)

DNA, in EBV. Whereas genomes of the smaller tumour viruses, depending upon the stage in the cell cycle in which they are expressed, are classically divided into 'early' and 'late' ORFs, herpesviruses are divided into 'immediate early' (before DNA replication is initiated), 'early' and 'late' genes. An alternative classification divides their genes into 'latent' and 'lytic' functions. The latter, probably simplistic, classification is useful for discussion purposes

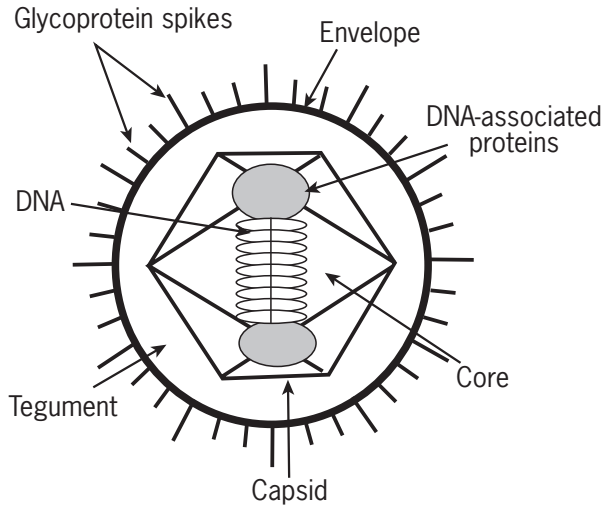


Figure 7 Schematic structure of a herpesvirus. (From IARC, 1997, p. 34.)

Table 3 Nomenclature of latent EBV gene products

Adopted terminology[a]	Alternative nomenclature[a]	
EBNA-1	EBNA-1	EBNA-1
EBNA-2	EBNA-2	EBNA-2
EBNA-3A	EBNA-3	EBNA-3A
EBNA-3B	EBNA-4	EBNA-3B
EBNA-3C	EBNA-6	EBNA-3C
EBNA-LP	EBNA-5	EBNA-4
LMP-1		
LMP-2A	TP-1	
LMP-2B	TP-2	
EBER-1		
EBER-2		

[a]EBNA, EBV nuclear antigen; LMP, latent membrane protein; EBER, EBV-encoded RNA; TP, terminal protein.
(From IARC, 1997, p. 53.)

when dealing with a large and complicated genome. Out of the 100 or so genes encoded by EBV, many of which are still poorly characterized, latent functions, as derived from data mainly drawn from EBV gene expression in B lymphocytes, consist of a small number of species only. These include six discrete EBV nuclear antigens (EBNAs), three discrete membrane antigens (LMPs) and two small RNAs (EBERs), of as yet unknown function. Interestingly, among other viruses studied in detail, only adenovirus encodes similar small RNAs (VA I and II) that structurally resemble EBERS, and although themselves not fully functionally characterized, are thought to modulate translation of viral proteins. For EBV, the EBERs are mainly localized in the nucleus and thus they may play alternative roles. Because of their very high levels of expression, EBERs have proved useful for detecting the presence of EBV in cells although, notably, they are apparently not expressed in all cells. For example, they are not found in a nonmalignant pathology associated with immunosuppression, oral hairy leucoplakia (OHL), where infected cells are frequently undergoing lytic replication. The nomenclatures used for the latent antigens in the EBV field are given in **Table 3**. The rest of the viral genes have been categorically designated as lytic, or lytically related. This distinct dichotomy into latent and lytic gene expression may be reassessed with time, since many EBV-associated tumours have recently been shown to express genes now designated as immediate early (or lytically related). Some of these, which may play initiating roles in the viral lytic cycle, may have other roles in tumours (discussed by Griffin and Xue, 1998). Alternatively, as proposed for KSHV (Ganem, 1998), a small amount of replication and thereby re-infection may be relevant to, and essential for, tumour growth.

These latent functions, on the assumption that most or all may play roles in the alteration of cell growth induced by EBV, at least for B lymphocytes in culture, are briefly defined as follows (IARC, 1997).

EBNA-1: a DNA-binding protein identified in all EBV-infected cells and responsible for EBV genome replication in latently infected cells. EBNA-1 is not recognized by the host cellular immune system, probably as a consequence of the glycine–alanine-rich repetitive (IR3) sequence within the protein. In transgenic mice, it is tumorigenic. This antigen and its pivotal function in EBV latency has recently been reviewed (Leight and Sugden, 2000)

EBNA-2: a transactivator both of other viral and cellular functions, and a key protein in B cell immortalization in culture. It is not generally expressed in EBV-associated tumours, although this does not rule out an early role in tumour induction. It is expressed in post-transplant lymphoproliferative disorders and in infectious mononucleosis.

EBNA-LP: appears to be important for the stimulation of B cell growth in culture and, like EBNA-2, be a contributing factor in post-transplant lymphoproliferative disorders and infectious mononucleosis. EBNA-2 and EBNA-LP are the first two proteins to be identified following cellular infection with the virus.

EBNA-3A, 3B and 3C: often considered together because they are derived from adjacent regions of the viral genome. EBNA-3A and 3C, but possibly not 3B, are involved in growth stimulation of B cells, but all three may have regulatory roles in the transcriptional control of other key viral functions. EBNA-3C has been compared in its properties to HPV E7 and adenovirus E1A proteins, both associated with cell growth alterations induced by their respective viruses.

LMP-1: often found expressed in EBV-associated tumours. In *in vitro* assays using heterologous promoters like SV40 LT, it is capable of inducing tumorigenic transformation of rodent fibroblasts in culture. It alters cytokeratin expression and inhibits cell differentiation.

This transmembrane antigen may recruit signalling antibodies and is absolutely required for both the initiation and maintenance of B cell growth in culture. In transgenic mice, LMP-1 produces a pathological response in keratinocytes, which has not been fully characterized.

LMP-2A and 2B: map across the terminal junctions of the viral DNA and therefore can only be expressed in latently infected cells, where the genome is circular. They do not appear to be directly involved in the *in vitro* growth stimulation of B cells, but may be important for the maintenance of latency. LMP-2A is a phosphoprotein, stably phosphorylated on tyrosine, and thus may have other unidentified functions.

Two other genes, *BARF1* and *BARF0*: more recently identified, both of which may play key roles, particularly in epithelial cell growth regulation. Their importance to B lymphocyte growth stimulation *in vivo* is less clear. The ***BARF1*** gene, like LMP-1, is fully competent for inducing tumorigenic cellular transformation of rodent cells, and even B lymphocytes in culture, when expressed under a strong, heterologous promoter. It has some homology with the human intercellular cell adhesion molecule 1 (ICAM-1). In limited studies carried out to date, BARF1 has been found expressed in most EBV-associated nasopharyngeal carcinomas (NPCs) examined. Its activities remain to be fully characterized. The second gene comes from ***Bam*HI I/A transcript**, also called complementary strand transcripts (**CSTs**), or ***BARF0*** gene. Primary CSTs extend over about 25 kbp of the viral genome (**Figure 6**) and spliced variants of it make up the major transcripts in NPCs. They were first identified in 1989 (Hitt *et al*., 1989) as a family of processed, multiply spliced polyadenylated RNAs and were subsequenty designated as 'complementary' in recognition of the fact that they were generated from the DNA strand with opposite polarity to that specifying numerous previously known viral genes. Each of the ORFs in the polycistronic CSTs, created by splicing events, overlap genes on the opposite strand, most of which are associated with lytic replication, which has led to the speculation that they may be involved in the maintenance of viral latency. CSTs are expressed also in BLs and other EBV-associated tumours, but at lower levels. They are often designated as latent functions as a consequence of their ubiquitous expression in tumours, but have also been found in lytically infected cells. A protein first described as a product of *BARF0*, the largest and terminal (3′ end of the gene, with its termination codon in the polyadenylation signal of the message) of the CST ORFs, was later identified in uninfected cells, casting doubt on its authenticity. BARF1 and CST expression and function(s) in EBV infected cells are key targets for future research.

The locations of some of the genes described on the physical map of EBV are given in **Figure 6**, and their designations and functions, where known, are summarized in **Table 4**. A unique working nomenclature has been established for EBV genes, where B stands for the *Bam*HI restriction DNA fragment containing a particular gene, a letter represents fragment size relative to the other *Bam*HI products (A being the largest and g the smallest in the sequenced B95-8 EBV genome; Baer *et al*., 1984), R (right) or L (left) denotes its direction (and polarity) on the conventional physical map of the genome, and a number denotes which reading frame is represented within a particular fragment. Thus, ***BARF1***, above, is the first rightwardly expressed ORF in the *Bam*HI A fragment. ***BARF0*** was not predicted by the DNA sequence, so it carries an aberrant designation. The differential expression of these genes in various EBV-associated tumours, or in lymphoblastoid cell lines (LCLs) generated by infecting B lymphocytes with the virus, have now led to subclassifications of viral latency, as simplistically illustrated for EBNAs and LMPs in **Figure 8**, and given in detail in **Table 5**.

The EBV genome also includes two other genes with interesting homologies to human genes: ***BCRF1*** and ***BHRF1,*** IL-10 and Bcl-2 homologs, respectively. Their roles in the virus have not been defined.

Cellular Immortalization *In Vitro*

Following a procedure first described for the small DNA tumour viruses, the minimal region of the EBV genome required for growth stimulation of cells in culture has been determined, using transfection protocols and fragments of the viral genome. The results of studies carried out on B lymphocytes (B) and epithelial (E) cells, representing the main tumour cell types associated with EBV tumours, are shown in **Figure 6**. Notably, there is no overlap between the minimal immortalizing regions (MIR-B and MIR-E) in these cells, supporting the argument that cell-type-specific functions may exist within the viral genome. For MIR-E, the data are consistent with findings on the transcription of EBV in NPCs, as determined by analysis of a comprehensive cDNA library made from the tumour (Hitt *et al*., 1989).

Pathogenesis

EBV is the causal agent for infectious mononucleosis, usually a self-limiting B cell proliferative disease, mainly a problem for economically privileged parts of the world where seroconversion and the development of antibodies to the virus occur late (**Figure 4**). With the hereditary immunodeficiency disorder, X-linked lymphoproliferative disease (XLP), or Duncan syndrome, fortunately rare, infection with EBV is usually fatal. Children that survive are at high risk of developing fatal lymphomas. With the so-called endemic form of Burkitt's lymphoma (BL), an acute problem for sub-Saharan Africa where it is the most prevalent cancer of children, there is a nearly 100% association with EBV. Again, in nasopharyngeal carcinoma (NPC), a head and neck tumour of poorly differentiated epithelial cells found with high frequency among the southern

Table 4 EBV gene products and proposed functions[a]

Open reading frame	Common name	Proposed function
Latent genes		
BKRF1	*EBNA-1*	Plasmid maintenance
BYRF1	*EBNA-2*	*trans*-Activation, transformation
BERF1	*EBNA-3A*	*trans*-Activation, transformation
BERF2	*EBNA-3B*	Unknown
BERF3/4	*EBNA-3C*	*trans*-Activation, transformation
BWRF1	*EBNA-LP*	*trans*-Activation, transformation
BNLF1	*LMP-1*	Transformation
BNRF1	*LMP-2A/2B*	Maintenance of latency
BARF0		*Unknown*
Immediate early genes		
BZLF1	*ZEBRA*	*trans*-Activation, initiation of lytic cycle
BRLF1		*trans*-Activation, initiation of lytic cycle
BlʹLF4		*trans*-Activation, initiation of lytic cycle
Early genes		
BMRF1		*trans*-Activation
BARF1		Limited homology to *ICAM-1*
BALF2		*DNA binding*
BALF5		*DNA polymerase*
BORF2		Ribonucleotide reductase subunit
BaRF1		Ribonucleotide reductase subunit
BXLF1		Thymidine kinase
BGLF5		Alkaline exonuclease
BSLF1		Primase
BBLF4		Helicase
BKRF3		Uracil DNA glycosylase
Late genes		
BLLF1	*gp350*	Major envelope glycoprotein
BXLF2	*gp85 (gH)*	Virus–host envelope fusion
BKRF2	*gp25 (gL)*	Virus–host envelope fusion
BZLF2	*gp42*	Virus–host envelope fusion, binds MHC class II
BALF4	*gp110 (gB)*	Unknown
BDLF3	*gp100–150*	Unknown
BILF2	*gp55–80*	Unknown
BCRF1		Viral interleukin-10
BHRF1		Viral *bcl-2* analogue

[a]See ***Table 3***. ZEBRA, EBV replication activation; gp, glycoprotein; MHC, major histocompatibility complex; BARFO, major ORF in CSTs, function unknown. (Adapted from IARC, 1997, p. 50.)

Chinese and in some other parts of Asia, the viral association is 100%. These associations (see Introduction), largely based on clinical, epidemiological and serological approaches, have now been known for nearly half a century. What still is not known, however, is the precise contribution of EBV to these diseases, whether causal or merely contributory. If contributory only, in no case has the corresponding co-factor(s) been definitively identified, although there are candidates such as malaria for BL and smoked, salted fish consumption for NPC. What is firmly established, however, is the fact that the geographical, racial and age incidence of individuals that develop these EBV-related malignancies are totally distinct (**Figure 9**).

During the last 20 years, following the cloning and sequencing of the viral genome (Baer *et al*., 1984), which has allowed for the development of alternative, sensitive assays for identifying EBV and its gene projects, other tumours have been associated to varying degrees with the presence of this virus. These include a variety of tumours of different histopathological types, including subsets of lymphoepitheliomas, Hodgkin disease, stomach and breast cancers and T cell lymphomas. Notably, in none of the cases does the frequency of association approach that seen for endemic BL and NPC. However, the Working Group set up to explore the risk of EBV to humans, (IARC, 1997) concluded that there is sufficient evidence for the

carcinogenicity of this virus, in the causation of BL and other non-Hodgkin lymphomas, immunosuppression-related lymphomas, Hodgkin's disease (HD), sinonasal angiocentric T cell lymphoma and NPC, to allow them to conclude that EBV is a human carcinogen. Subsequent to this document, more information on the expression of EBV in breast cancer has been published, and the viral genome has also been identified on two occasions in carcinomas of the liver, previously a preserve of the hepatitis viruses. The future will undoubtedly bring more 'associations' for this ubiquitous virus, and hopefully, if suitable animal models are identified, notions about its actual role in disease. The sole argument that this virus alone could be sufficient for inducing malignancies under appropriate circumstance comes from the fact that many of the polyclonal lymphomas that develop as a consequence of immunosuppression (natural or induced) have a high frequency of association with EBV.

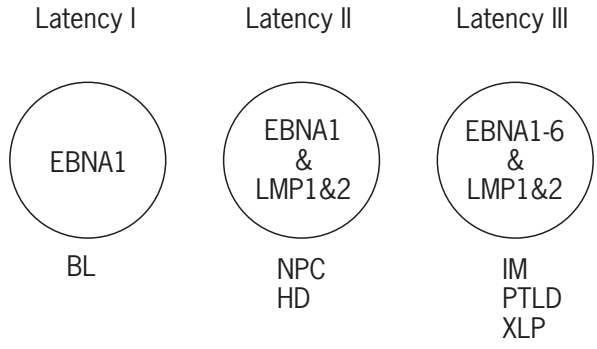


Figure 8 Patterns of latency gene expression in categories designated Latency I–III. BL = Burkitt's lymphoma; NPC = nasopharyngeal carcinoma; HD = Hodgkin's disease; IM = infectious mononucleosis; PTLD = post-transplant lymphoproliferative disease; XLP = X-linked lymphoproliferative disease (from Khanna *et al.*, 1999).

Immunological Considerations

One of the dominant characteristics about EBV is its adaptation to allow for persistence in its host(s), and gene expression, even in the presence of a functional immune system. EBV co-replicates with host DNA, and EBNA-1, required for latent replication, is tolerated, not eliminated, although there are epitopes for class I and class II HLAs in the viral antigen (Khanna *et al.*, 1999). The dominant feature in this protein that allows for its tolerance appears to be the repetitive (IR3) sequence that it harbours. In some cases of BL, where anti-EBNA-1 may be the sole antibody produced, this would allow for viral persistence. In situations where other antigens are expressed, for example in infectious mononucleosis or other EBV-associated malignancies, memory/activated T-cells appear to be important in limiting cell expansion and in targeting productively infected cells that express lytically related antigens. Immunological data suggest that vaccines designed to control primary EBV infection, a desirable objective in view of its carcinogenic role in humans, may profit by

Table 5 Patterns of latent EBV gene expression[a]

Type of latency	Gene product	Co-stimulatory molecules	Examples
IA	*EBERs*, *EBNAl*, *CSTs*		Burkitt's lymphoma
IB	*EBERs*, *EBNA1*, *CSTs* *LMP2A*		Gastric carcinoma
II	*EBERs*, *EBNA1*, *CSTs*, *LMP1*, *2A*, *2B*, *BARF1*	CD30 CD23 CD40 B7.1 LFA-1, -3 1CAM-1	Hodgkin disease Nasopharyngeal carcinoma T cell lymphoma
III	*EBERs*, *EBNA 1–4*, *6*	CD30	Post-transplant lympho proliferative disorder
	LP, *LMP-1*, *2A*, *2B*	CD23 CD40 CD44 B7.1 LFA-1, -3 1CAM-1	Infectious mononucleosis
Other	*EBERs*, *EBNA1*, *2*		Smooth-muscle tumours

[a]See **Tables 3** and **4** and **Figure 8**.
(From Griffin, 2000, *Mutation Research*, **462**, 395–405.)

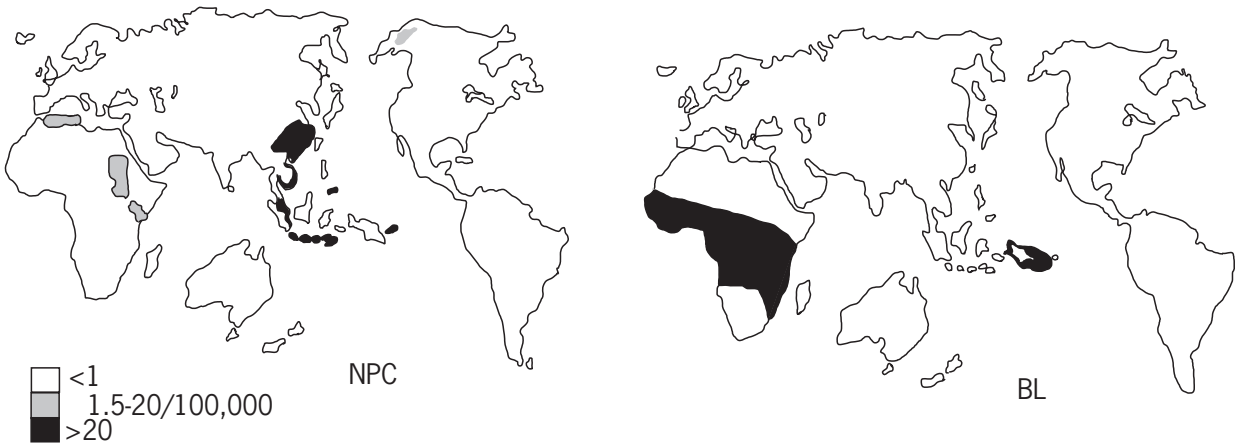


Figure 9 Comparative sites of highest frequencies of nasopharyngeal carcinoma (NPC) and Burkitt's lymphoma (BL), showing their disparate global locations. Black regions are sites of greatest frequencies and grey regions those of intermediate frequencies (for NPCs).

including dominant determinants of antigens associated with the viral life cycle (Khanna *et al*., 1999). This may be particularly relevant since animal models show that the development of neutralizing antibodies does not always correlate with protection from EBV infection. To this end, many of the dominant epitopes, including those found in latent and some lytically related proteins, have been mapped (**Figure 10**). Some of these might prove of value in the development of cytotoxic T lymphocyte (CTL) epitope-based vaccines, the aim of which would be to reduce morbidity and possibly clear infection. Since evidence suggests that many individuals having EBV-associated tumours retain detectable levels of EBV-specific T cells, needed for surveillance, this may be a reasonable approach. Even for BL, the tumour in which viral gene expression appears most tightly regulated, subpopulations of cells expressing lytically related antigens have been identified in some individuals (Labrecque *et al.*, 1999), making them also candidates for immunotherapeutic approaches. The recognition of the important contribution of EBV to diseases of humans has greatly stimulated efforts over the past few years to control this virus.

KAPOSI SARCOMA-ASSOCIATED HERPESVIRUS (KSHV)

History

In 1872, M. Kaposi, a Hungarian dermatologist, described a pigmented angiosarcoma, now called 'classic' or sporadic Kaposi sarcoma (KS), that mainly affected skin on the lower limbs, and was most prominent in elderly men of Mediterranean and eastern European origin. KS was also an African problem. In the 1960s and 1970s, the frequency and distribution of KS altered, and in many cases could be related to transplant therapies in other parts of the world. Whereas modest increases in KS were being reported in various countries prior to the onset of the syndromes now covered under the generic name AIDS, its frequency and epidemiology were drastically influenced by the spread of this virus. Over the past decade or so, although the histopathological presentations of all types of KS are identical, this malignancy has been generally subclassified into classic (sporadic), endemic (African), epidemic (AIDS related) and immunosuppression-associated (transplant) types, to reflect its origin. From being a comparatively rare form of cancer, KS is now fairly common in certain parts of the world. Exactly how common, however, is a controversial topic. The epidemiology of this cancer, and particularly the fact that in the early days it was the most common tumour in AIDS patients, with 15–20% of them developing KS, suggested that this disease might have an infectious aetiology (IARC, 1997). Thus, an active search to find such an agent was initiated.

The history of the discovery of KSHV is different from that of EBV, the human virus it most resembles, and owes much to the development of molecular biological methodologies. One of these in particular, called representational difference analysis (RDA), was used by a group in the USA, working with the husband and wife team Moore and Chang (Chang *et al*., 1994), in their search for a KS infectious agent. RDA consists of generating genomic representative entities from diseased and normal tissues, preferably from the same individual, using PCR amplification. These are stably associated with priming PCR sequences and hybridized to an excess of representative, nonligated amplified sequence, with no attached primers, from normal tissue. Following this procedure, only unique sequences found in the diseased tissues will contain priming sequences on both strands, which allows them to be substrates for subsequent PCR reactions. Repeating such a process enriches the sample for unique sequences.

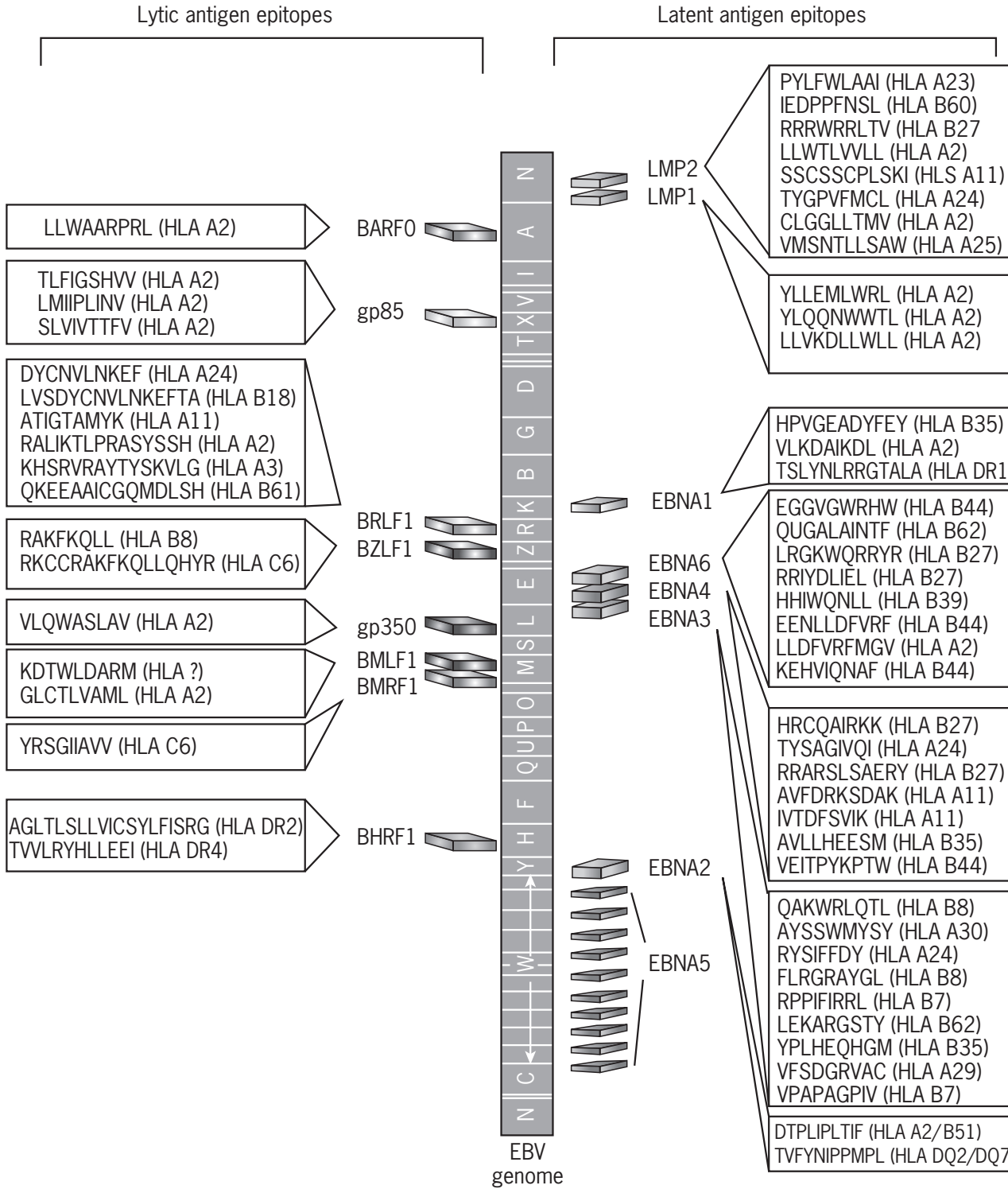


Figure 10 Schematic distribution of HLA class I and class II restricted cytotoxic T lymphocyte epitopes with EBV latent and lytic antigens, at the peptide level. BARF0, the largest ORF in the CST transcripts is given here as a lytic function, although it is also expressed during the latent cycle. (Adapted from Khanna *et al.*, 1999.)

These can then be purified and their sequences determined. By RDA, using tissues from AIDS-associated KS, Chang *et al*. (1994) identified sequences that were homologous with, but distinct from, other members of the γ-herpesvirus family, most notably EBV and the oncogenic primate virus, herpesvirus saimiri. They correctly concluded, as was subsequently shown, that this work was consistent with the presence of a new human

herpesvirus in KS lesions. Interestingly, the homologies they identified were with EBV late viral genes (in *BDLF1* and *BcLF1*, see EBV section).

Virion and Genome Structure

KSHV, or HHV8, has morphological features typical of herpesviruses (**Figure 7**), consisting of 100–150 nm particles surrounded by a lipid envelope, with an internal electron-rich central core. Its DNA was assessed by pulsed-field electrophoresis as 160–170 kb, consistent with that of other γ-herpesviruses, but more uniform than that observed with EBV. Both circular and linear forms of KSHV have been identified. Larger genomes reported to exist in some tumour-derived cell lines have been shown to result from DNA duplications, possibly associated with propagation in culture. In the same year, 1996, as the physical studies were reported, the complete sequence of the viral genome was published and an open reading frame map generated. This aptly illustrates the rapidity with which this field was being and has progressed. The sequence study (Russo, 1996) showed that the structure of KSHV was essentially similar to that of herpesvirus saimiri (HVS) (**Figure 5**). For KSHV, the genome has a single 140.5-kb long unique region, containing about 80 ORFs, flanked on either side by variable-length terminal repeats, about 800 bp in size. Within the genome, there were small repeat regions, some but not all of which appeared to be within ORFs, but overall there was little organisational similarity to EBV. In addition to numerous homologies with HVS, the sequence of KSHV also showed homologies with EBV genes, mainly those coding for late viral antigens where similarities that ranged from 44% to greater than 70% were observed. With EBV late genes, *BDLF1* and *BcLF1*, identified in the initial studies of Chang *et al*. (1994), the degree of homology at the DNA level was about 75%. Interestingly, the important viral DNA polymerases of these two viruses have 72% homology, although whether the enzymes themselves can be functionally interchanged is not known.

Although homologies between KSHV and EBV immediate early functions were observed, there were no homologues to EBV latent genes found in the ORFs of KSHV. Different isolates of KSHV appear to have highly conserved genomes. The phylogenetic tree of KSHV, based on aligned amino acid sequences as they relate to other herpesviruses, is shown in **Figure 11**. A close relationship with HVS, from squirrel monkeys (not apparently oncogenic in its natural host but tumorigenic to other nonhuman primates), and with equine herpesvirus 2 (with a more distant relationship with EBV) is seen (IARC, 1997).

Putative Key Tumour Genes

Studies on KSHV genes, to designate them as latent, immediate early, early and late genes, and identify those

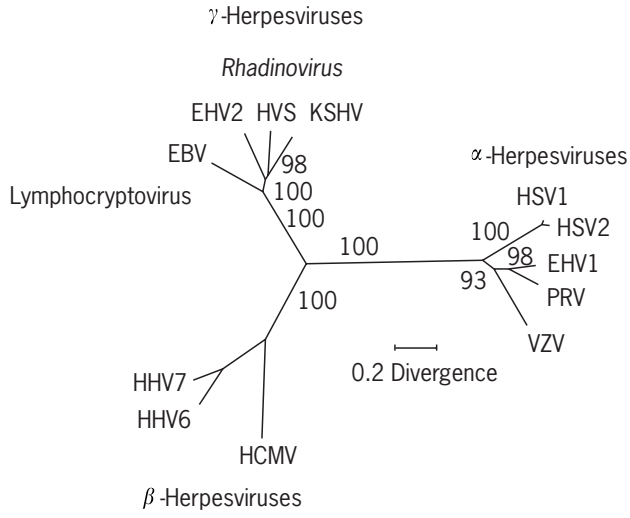


Figure 11 Phylogenetic tree of KSHV (HHV8) in relation to other herpesviruses. The comparison shows KSHV to be most closely related to the γ-herpesviruses, EBV, equine herpesvirus 2 (EHV2) and herpesvirus saimiri (HVS), its nearest relative (see **Figure 5**). (From IARC, 1997, p. 385.)

that may play key roles in the oncogenic activity of this virus, have been initiated. Here, as with other herpesviruses, latent transcripts are defined as constitutively expressed mRNAs which are not susceptible to chemical induction (e.g. with agents such as the promoter-stimulating phorbol ester (TPA) or n-butyrate, which affects chromatin structure) but are susceptible to inhibition by cycloheximide, an inhibitor of protein synthesis. Immediate early genes, on the other hand, are those whose transcripts are inducible, but resistant to the action of cycloheximide. Early gene expression is blocked by cycloheximide, but not by phosphonoacetic acid (PAA), an inhibitor of the virus-encoded DNA polymerase, whereas late lytic cycle gene expression is not blocked by the latter (Sun *et al*., 1999). The same definitions are used when considering EBV genes. In the case of KSHV, many apparently nonlatent genes have been found expressed in the virally associated tumours (see below), and the same is becoming apparent for EBV. Their roles in these settings are still undefined. Lines derived from primary effusion lymphoma (PEL) cells infected with KSHV have proved useful in identifying some of the genes in KSHV that act as possible tumour-inducing agents. Several of these (ORF 71–73, see below) classified as latent genes since their transcription products are constitutively expressed, are clustered in the viral genome. Notably, a completely different gene designation system from that for EBV has been adopted for KSHV based on gene numbers from the sequence.

LANA (latency associated antigen (ORF 73): a large (226/234-kDa) protein that reacts with sera from AIDS patients, characterized by a typical speckled nuclear pattern. Antibodies to LANA have been postulated to have

prognostic value for the likelihood of an individual developing KS. They do not cross-react with EBV-specific antigens.

v-FLIP (ORF 71): so named because of its homology with a cellular anti-apoptotic factor, c-FLIP, which regulates the apoptosis triggered by Fas and other members of this tumour necrosis factor receptor family. In KSHV, these genes are overlapping and probes for v-FLIP also recognize LANA, but not vice versa (Sun *et al*., 1999). Notably, HVS also expresses a v-FLIP, which appears to be a late function. Interestingly, probes for v-FLIP also recognize a lytic cycle transcript.

ORF K72: tentatively identified as a latent function, with about 30% amino acid homology to the human cellular cyclin, D2, and expressed in persistently infected cells alongside LANA and v-FLIP. In culture, K72 (alternatively, KSHV v-cyclin), phosphorylates the retinoblastoma protein, Rb. K72 is a small ORF (60 amino acids), and may also be transcribed during the lytic cycle thus, like v-FLIP, possibly playing more than one role related to its associated malignancies, in the virus.

In addition to these three apparently latent genes, numerous viral genes associated with later stages of the viral life cycle have also been identified in the sarcomas. One of the immediate early KHVS genes is structurally and functionally related to an EBV-transactivator gene. In functional assays in culture, this viral gene has been found competent to initiate reactivation of a cascade of genes associated with the virus lytic cycle. KSHV also encodes a number of homologues of proinflammatory cytokines, such as IL-6 and macrophage inflammatory protein (MIP), as well as *bcl-2*, a homologue of another anti-apoptotic gene (also found in EBV, in *BHRF1*) and v-GCR, a G protein-coupled receptor. By definition, several of these appear to act as late genes, but to be expressed in tumours.

At this stage in the understanding of KSHV and its role in malignancy, temporal expression of tumour-associated genes thus seems to differ in large part from that of EBV, where the dominant components are latent genes. Whether latent gene expression in tumours may also dominate in the case with KSHV remains to be seen. Data suggesting that KSHV may express proteins related to the membrane-associated oncogenes of EBV, LMP1 and 2A have been obtained (Glenn *et al*., 1999). For KSHV, on the other hand, it has been proposed that tumour growth may be enhanced by viral chemokines or cytokines expressed by adjacent infected cells that have undergone a switch from latency to a lytic cycle type, giving to such genes an important role in malignancy (Ganem, 1998). Notably, a similar situation exists for some EBV-associated BLs (see above), and may account for the proliferation of this tumour with its remarkable doubling times of 28–60 h.

Many of the studies on gene expression in KSHV are currently being made on PEL cell lines, since growth of cells in culture from KS often results in apparent loss of the viral genome. Whether PEL lines are suitable models for KSHV expression in tumour settings remains to be seen. Like lymphoblastoid cell lines as a model for the role of EBV in BL, they may be imperfect, but nonetheless of value for studying the function of the viral genes.

Pathogenesis

Attempts to detect KSHV in peripheral blood mononuclear cells (PBMCs) from healthy individuals in countries with a low prevalence of KS, even by very sensitive PCR approaches, has not been generally successful. In KS-risk countries, variable associations of the virus with PBMCs have been reported. On the assumption that KSHV may be sexually transmitted, studies on semen specimens have also been carried out, with controversial results. However, sexual behaviour does seem to be a risk factor in transmission of this virus and in the development of KS, the risk running parallel to that of HIV infection. The notion of a KS-associated infectious agent, independent of HIV infection, appears to be real, in that the virus has been identified in all four epidemiological forms of KS, with no significant differences in detection rates (IARC, 1997). In addition to an association with KS, another neoplastic condition, primary effusion lymphoma, a rare, distinct type of non-Hodgkin lymphoma, has also been associated with KSHV infection. The cells in this malignancy are usually large and irregularly shaped, with abundant cytoplasm, and prominent nucleoli and mitotic features, the latter properties also found in BL. Notably, both KSHV and EBV can often be identified in these tumours. In AIDS patients, this is a fulminant lymphoproliferation and the median survival time of the individual is short. Other B and T cell lymphomas, explored for its presence, have not revealed any KSHV. On the other hand, there is evidence for a role for this virus in Castelman disease, at least in AIDS patients. This is a rare, usually polyclonal, non-neoplastic disorder of unknown aetiology, first reported in 1956.

Several scenarios could account for the association of KSHV with these tumours, particularly with KS. First, the virus may be the aetiological agent, and one or more of the functions noted above, or other as yet undiscovered viral gene products, may play a critical role in disease. On the other hand, as often suggested for EBV, the virus may be a contributory factor to the malignancy, e.g. by stimulating cytokines which enhance cell growth. Alternatively, the virus may be a mere passenger with the capacity to infect the cell types now associated with tumours that harbour it. Again, as with EBV, it is difficult to distinguish among these possible scenarios. With both viruses, the fact that viral homologues exist in other primates, which in model studies are shown to produce tumours, can be viewed as supportive evidence for these herpesviruses as acting both risk factors and tumour-inducing agents, under the appropriate conditions.

HEPATITIS B VIRUS (HBV)

History

Jaundice (from *jaune*) is a disease of the liver that has been known for centuries. Its most notable characteristic is an orange–yellow discoloration of the skin and conjunctivae, caused by deposition of elevated levels of bilirubin produced from damaged hepatocytes. Viral hepatitis, a general term for infections of the liver, can be caused by a number of hepatitis viruses, only one of which, HBV, is classified as a DNA virus. HBV is an unusual DNA virus in that it has, as a component of its life cycle, an RNA reverse transcriptase activity, a function normally associated with RNA retroviruses. The mechanism of HBV replication is unique for a DNA virus, in that it involves an RNA intermediate. HBV has a striking tropism for hepatocytes, but also can be detected in PBMCs.

Although viral hepatitis is a major public health problem, the identification of its viral association(s) and of HBV as one of the infectious agents of the disease was long in coming. Prior to its ultimate discovery, epidemiological differences among the diseases had suggested the possible existence of more than one infectious agent. In the early 1960s, Blumberg, looking for inherited polymorphic traits in blood from different parts of the world, identified an antigen, subsequently designated as 'Australia antigen', in sera from an Australian aborigine which reacted specifically with an antibody found in serum from an American haemophilia patient. This antigen proved to be geographically restricted, being relatively rare in American and Western European individuals, but more common in African or Asian populations, and in patients with certain distinct pathologies, including leukaemia, leprosy and Down syndrome. The association of Australia antigen, now known as hepatitis B surface antigen (HBsAg), with viral hepatitis was made several years later. This seminal finding led others to undertake studies aimed at identifying the infectious agent. In 1970, Dane and colleagues first identified, by EM, virus-like particles in the sera of patients with Australia antigen-associated hepatitis (Dane *et al*., 1970). At this time, it was estimated that of the 3.5 billion people in the world, as many as 175 million might be carriers of HBV. In spite of the millions of viral carriers, the cancer now associated with HBV, hepatocellular carcinoma (HCC), is relatively uncommon, but nonetheless correlates in frequency with those parts of the world with the highest percentages of carriers of the virus, and possibly reflects a progressive disease, initiated by hepatocyte infection and proceeding through the development of chronic hepatitis, to tumour formation.

HBV is the prototype for a family of viruses, now called the Hepadnaviridae, found in woodchucks, ground and tree squirrels and Peking ducks, all sharing distinctive morphologies and genomic characteristics. Studies on these viruses lend support for a causal relationship between HBV and some liver cancers. In particular, the duck hepatitis virus has been associated with the development of hepatoma in its host. The Working Party dealing with the association between HBV and cancer (IARC, 1994) concluded that chronic infection with hepatitis B is carcinogenic to humans. It reached the same conclusion with regard to the RNA virus, hepatitis C (HCV). In making this judgment, several criteria for causality were used, one being that a strong association, as found with the hepatitis viruses, is a better indicator of causality than a weak association.

Virion Structure and the Virus Life Cycle

The HB infectious virion is a 42-nm double-shelled spherical particle (originally called the Dane particle) that consists of an outer envelope composed of HBsAg and an inner core, or nucleocapsid, with its own antigens, hepatitis B core (HBcAg) and e (HBeAg), antigens, together with HBsAg, acting as markers for the presence of intact virions and infectivity. Infectious virions also contain a small (3.2-kb) circular, partially single-stranded DNA, and an endogenous DNA polymerase that can produce a fully double-stranded genome. Electron microsopic (EM) analyses show that in patient's sera, however, the concentrations of incomplete HBsAg structures may exceed those of complete virions. One of these forms is a small (20–25 nm) spherical particle and the other(s) is a tubular or filamentous form varying in size from 20×20–200 nm. Structures of these particles are schematically illustrated in **Figure 12**. Interestingly, similar shaped tubular (or filamentous particles) are also seen in early EM pictures of the small DNA papovaviruses, where they have been postulated to represent precursor forms of the mature, virion spherical particles, although this has not been proved. Interestingly, the major capsid protein, VP1, of papova viruses, expressed on its own *in vitro*, spontaneously reassembles to form viral-like capsid particles, composed of icosohedral (major product) and tubular (minor product) structures.

The site of primary replication of HBV is the liver, although it has been postulated, based on abnormalities observed in patients with acute hepatitis, that haematopoietic stem cells many also support viral replication. HBV infection probably involves viral attachment to specific receptors, although these have not been identified, nor have mechanisms for attachment and penetration been elucidated. Recent data suggest that HBV nucleocapsids do not enter the nucleus. Rather, they are arrested at the membrane and release the partially single-stranded genomic DNA into the nucleus where it is converted to covalently closed-circular (CCC) double-stranded DNA, which in turn serves as the viral transcription template.

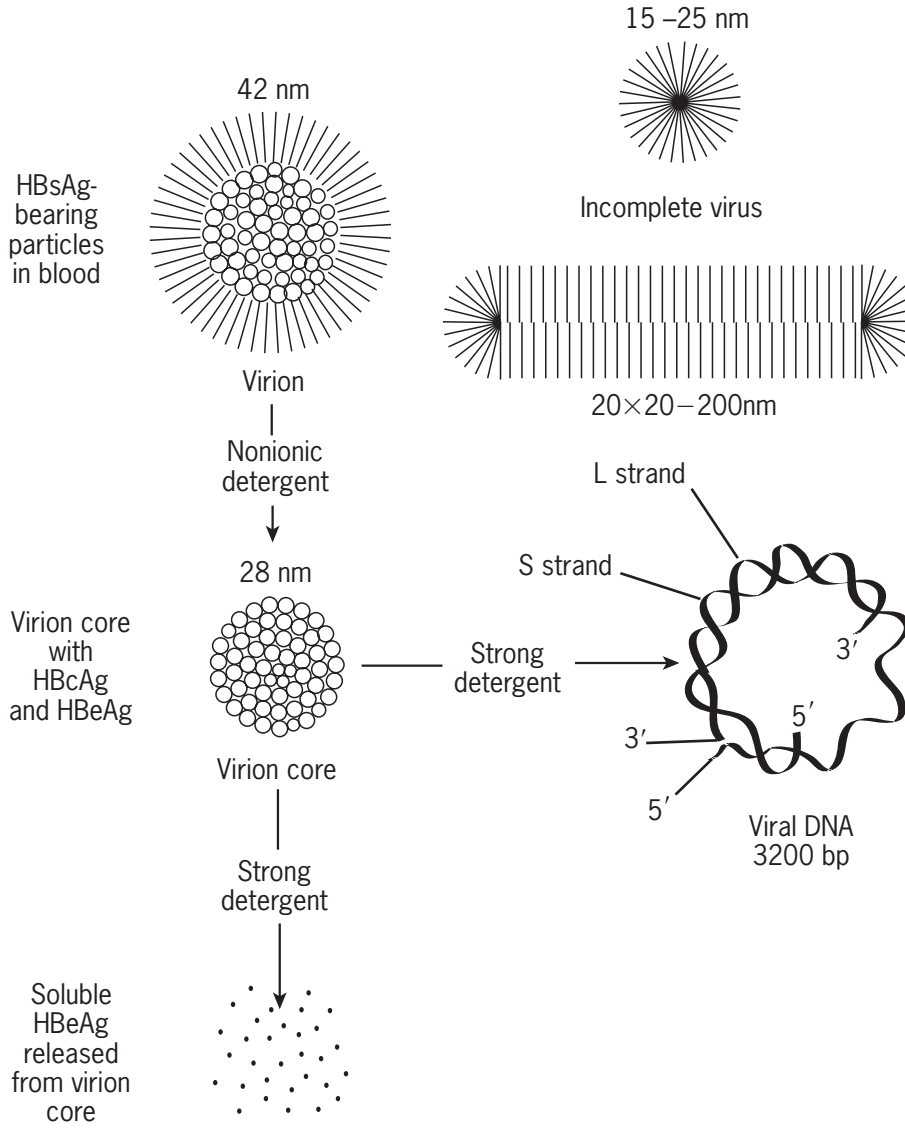


Figure 12 Diagram of the different forms adopted by hepatitis B virus (HBV).

Genome Structure, Replication and Antigens

The viral genome is organized into four transciption units, each with its own independent promoter. They share, however, a common polyadenylation site. Transcription yields four extensively overlapping viral RNAs, with sizes of 3.5, 2.4, 2.1 and 0.7 kb. These are exported into the cytoplasm where viral proteins are translated, and viral particle assembly and genome replication occurs (Chisari, 2000). HBV replication involves reverse transcription of the RNA pregenome (pRNA) to produce minus-strand DNA, the template for plus-strand DNA. Replication results in an encapsidated double-stranded open circular genome that is either recycled back to the the nucleus (to amplify the pool of CCC-DNA) or becomes enveloped by the viral protein and proceeds to complete the life cycle, as illustrated in **Figure 13**. The transcriptional products themselves are complex.

The 3.5-kb transcript: specifies viral genes reversibly transcribed as a first step in genome replication. The POL protein has numerous activities, acting as reverse transcriptase, DNA polymerase and RNase H, all essential functions for viral replication. The core protein (HBcAg) can form homodimers that self-assemble into capsid particles. In the cytoplasm, these also contain pregenomic viral RNA and POL, and the whole particle acts as the site for viral replication. The precore protein has a sequence that directs it to the endoplasmic reticulum (ER), where it undergoes limited proteolysis to produce the e antigen (HBeAg) which is secreted into the plasma membrane. Its role in the viral life cycle, in spite of its obvious importance, is still poorly understood.

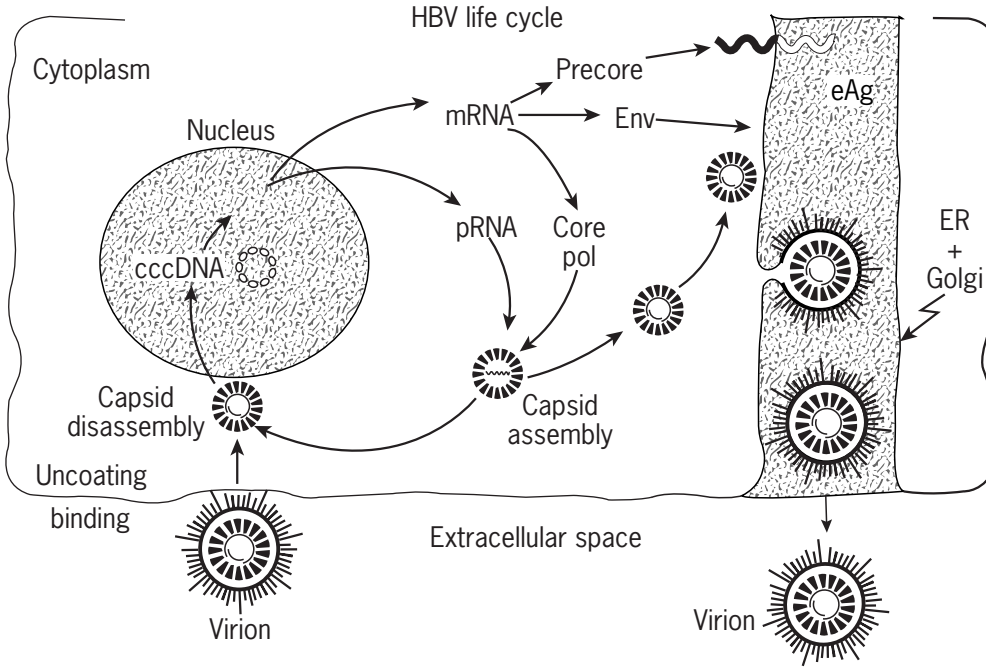


Figure 13 The HBV life cycle. Entry of the virus is still poorly defined. While the RNA containing capsid is maturing into a DNA-containing capsid, it migrates bidirectionally within the cytoplasm. One pathway terminates at the endopasmic reticulum (ER), where it interacts with envelope proteins which trigger an internal budding reaction, resulting in the formation of virions that are transported out of the cell by the default secretory pathway. The second pathway transports the maturing capsid to the nucleus to amplify the pool of covalently closed circular (CCC) DNA. (Adapted from Chisari, 2000.)

The 2.4- and 2.1-kb transcripts: produce the large, middle and small envelope proteins, which share common carboxy termini. The small, but major, transcript encodes the hepatitis B surface antigen (HBsAg). The middle and large transcripts which encode preS-2 and preS-1 antigens, respectively have amino acid extensions. The envelope proteins are cotranslationally inserted into ER membranes, where they aggregate, bud into the lumen and then are secreted by the cell as 22 nm subviral particles (**Figures 12** and **13**) or, if they have enveloped nucleocapsids before budding, become 42 nm infectious virions. The filamentous particles are generated when the large envelope protein is overexpressed. These are not usually secreted, but rather give an histologically distinct appearance (like 'ground glass') to cells, and hypersensitize them to the cytopathic effects of interferon-γ.

The 0.7-kb transcript: encodes the transactivator X protein. In the woodchuck model, the X protein is required to initiate infection. By virtue of its ability to transactivate expression of other genes, X is generally considered to be an important contributor to the pathologies induced by HBV. It was earlier thought to have properties associated with cell transformation and, on overexpression in transgenic mice, it can induce a high baseline incidence of HCC. However, in spite of considerable interest in X, its precise role *in vivo* still seems to be far from thoroughly characterized and its function as an oncogene has been called into question.

A transcriptional map of HBV, showing antigen locations, is given in **Figure 14**.

Pathogenesis

The natural history, clinical manifestations and geographical variation (as illustrated in **Figure 15**) of HBV infection, are highly variable. Chronic infection with HBV as related to endemicity, geography and mode and time of infection is given in **Table 6** and a brief resumé of the pathological consequences of infection in **Figure 16**. Serological patterns that accompany acute and chronic infections are given in **Figure 17**.

In spite of the wealth of knowledge that has been generated on this virus, the mechanism(s) by which HBV induces cellular transformation remains largely obscure. No viral oncogenes have been identified, and most or all of the viral antigens appear to be primarily involved in one or other aspect of the virus life cycle. In the woodchuck model, the X protein (see above) has been associated with infection. One promising lead on pathogenesis lies in the fact that HBV has a high mutation rate, which, although 100–1000 times lower than that observed with RNA

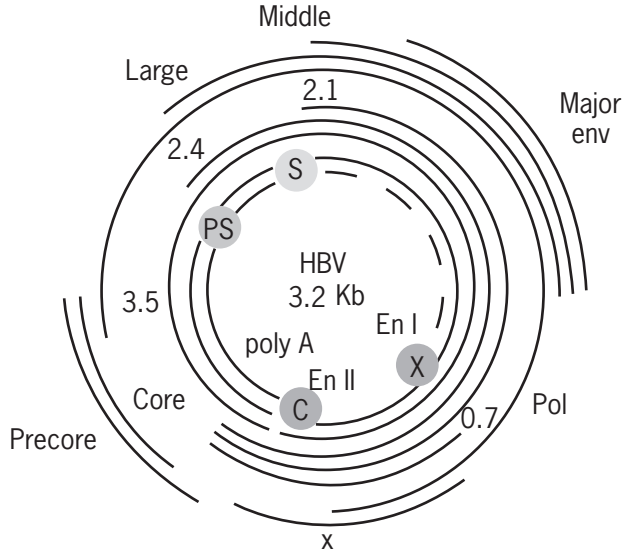


Figure 14 Transcriptional map of HBV, with the partially double-stranded 3.2-kb open circular genome present in virions shown in the centre. The core (C) pre-S (PS), HBs (S) and HBx (X) promoters are given inside round icons. (Adapted from Chisari, 2000.)

viruses, is nonetheless many times greater than normally seen with DNA viruses (IARC, 1994). However, although many HBV gene mutants have been identified and investigated, none to date has been proved beyond reasonable doubt to confer oncogenicity on the gene in question. Alternatively, in hepatocytes, HBV integrates into the host chromosome, giving rise to the possibility of insertional mutagenesis. Viral genome insertions in HCC have not, however, proved specific, nor have they been instrumental in pointing to insertional mutagenesis as explaining a viral role in HCC. The ability of X protein to affect expression of other genes allows for the possibility of either enhancement of expression of cellular genes associated with transformation, or down-regulation of tumour-suppressor genes, but has led to no definitive answers. Although approaches which consist of both direct and indirect influences on aberrant cell growth have been explored, and have generated interesting but frequently conflicting, data, the molecular route(s) by which hepadnaviruses predispose their host to malignancy remains an open question. There will probably be no simple answers, and one or more of the pathways that have been explored to date may play a role in diseases associated with this virus.

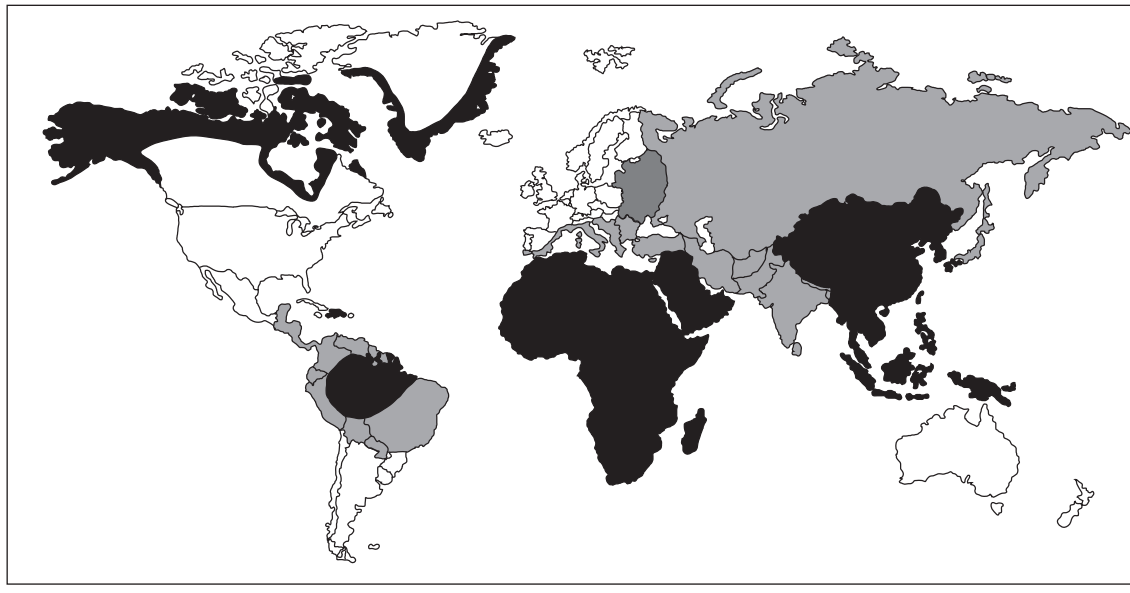


Figure 15 Geographical distribution of HBV. (From IARC, 1994, p. 56.) Black areas: high; grey areas: intermediate; white areas: low. (See **Table 6**.)

Table 6 Chronic infection with HBV : geography, mode and time

Endemicity	Geographical area	Predominant time of infection
High, $\geq$8%	China, Southeast Asia, Pacific Basin, sub-Saharan Africa, Amazon Basin	Perinatal, childhood
Intermediate, 2–7%	East, central and southern Europe, Middle East, South Asia, Japan	Perinatal, childhood, adulthood
Low, $<$ 2%	North America, western Europe, Australia, southern Latin America	Adulthood

(Taken from IARC, 1994, p. 59.)

Hepatitis B virus

- Hepatotropic, noncytopathic, 3.2-kb circular DNA
- Acute, chronic hepatitis, hepatocellular carcinoma (HCC)
- Over 2 billion people infected
- Over 350 million people chronically infected
- 100-fold increased risk (40% lifetime risk) of HCC
- Over 1 million deaths each year

Figure 16 HBV, disease association and frequencies. (Adapted from Chisari, 2000.)

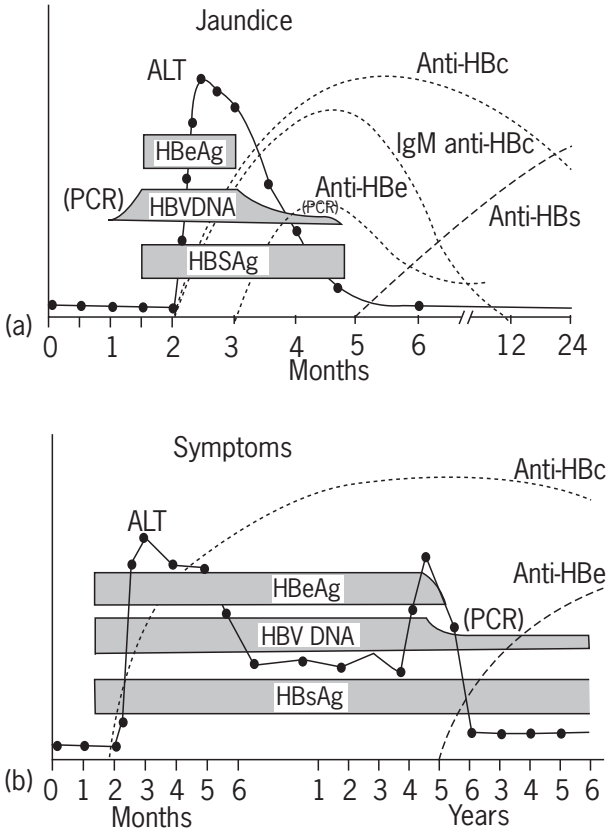


Figure 17 Serological and molecular patterns of HBV expression in (a) acute and (b) chronic HBV infections. Hatched bars show patterns of antigen expression; broken lines indicate periods in which HBV DNA (as noted) is detectable by PCR. Expression of alanine aminotransferase (ALT) used as control. (From IARC, 1994, p. 60.)

A hypothesis, following on from infection through to chronic illness and HCC, as outlined in **Figure 18** (Chisari *et al*., 2000), draws on all the factors noted above, and others, in predicting a course for disease. Such a complex scenario would emphasize the need for more than viral surveillance for the eradication of HCC.

Immune Biology of HBV and Immunoprophylaxis

The host–cell interactions that allow for the persistence of a virus, and the failure of the immune system to eliminate it in an immunocompetent individual, is a topic of considerable relevance for the DNA tumour virus field. For largely noncytopathic viruses, such as HPV, EBV, KSHV and HBV, they must either overwhelm an effective immune response or adopt mechanisms that allow for avoidance, as suggested by one or more of the hypothesized routes for progression from infection to generation of hepatocellular carcinoma (**Figure 18**). One approach for EBV therapy, as discussed elsewhere, assumes that the immune system may need to, and can, be stimulated specifically to recognize viral genes that might be expressed in its associated tumours, with beneficial effects. As noted, however, realistically such an approach is aimed at reducing morbidity, rather than effecting cure (Khanna *et al*., 1999). Such an approach may be even more valid for HBV, which can infect virtually all the hepatocytes in the liver, suggesting that the number of infected cells might actually outnumber relevant antigen-specific T-cells by several orders of magnitude in acute hepatitis infections and HCC. Thus, in individual situations where there are not sufficient cytotoxic T cells to contain infection, stimulation of the immune system might prove effective. On the other hand, there are studies to suggest that in some individuals with acute hepatitis, even in the presence of a vigorous T cell response, not all virus may be cleared. This has been explored using sensitive PCR assays where, several decades after complete clinical and serological recovery from this disease, low levels of viral DNA were detected in sera and PBMCs in some cases. Whether this result represents individuals at risk of reinfection and/or developing HCC for other, possibly genetic reasons, remains to be assessed. Notably, in HBV vaccine studies in Taiwan on children of different ethnic origins, unidentified host factors were postulated to explain the hyporesponsiveness seen among some populations (Hsu, 1996).

For HBV, it was earlier demonstrated that serum containing HBsAg retained its immunogenic properties even after heat inactivation. This observation proved the basis for plasma-derived, and later recombinant, s-antigen vaccines against HBV. In the 1980s, large-scale vaccination studies on children were initiated in both Taiwan and China, countries with the highest rates of HBV endemicity in the world, with the reasonable expectations that universal childhood immunization would allow HBV infections to be controlled in these areas within a few generations. In some high-risk areas, vaccination programmes to immunize every newborn child have been initiated and subsequent HCC incidence in these areas is being carefully monitored. As this malignancy, as with

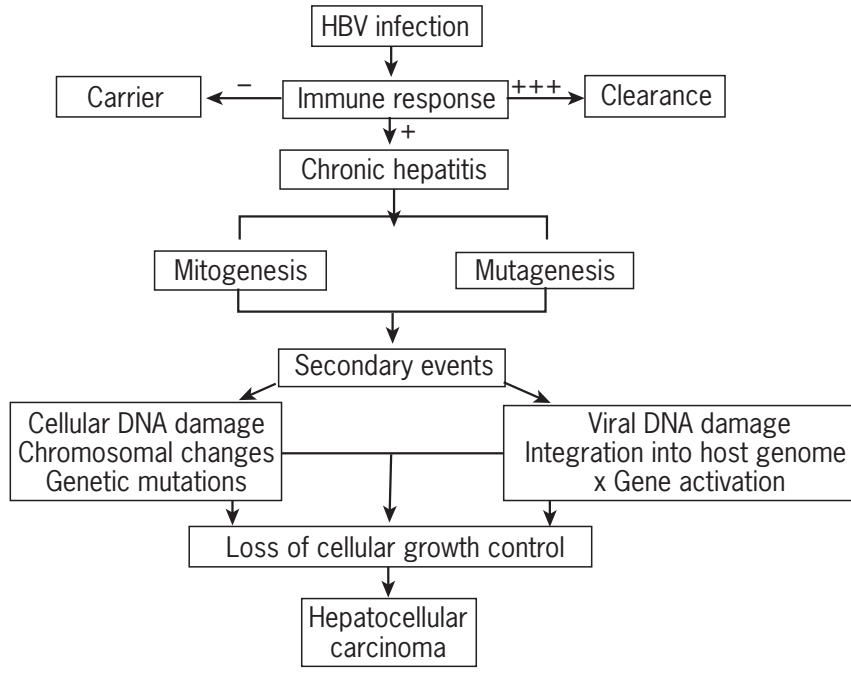


Figure 18 Hypothesis. Route from HBV infection to the genesis of hepatocellular carcinoma (HCC). (Adapted from Chisari, 2000.)

EBV-associated NPC, is mainly confined to adult populations, the data on this topic are not yet available. However, in Taiwan 10 years on, for children who completed a complete course of vaccination (four doses), anti-HBV antigen antibodies were detectable in a high proportion (82%) of them. Interestingly, in China it has now been found that tree shrews (*Tupaia belangeri chinenesis*) can be infected with human HBV and the infection passed to offspring in a high proportion of cases. Thus, a useful model may evolve for studying many of the unanswered questions about the relationship of HBV infection, the development of HCC and the variable responses to vaccination that have been observed.

OTHER DNA VIRUSES

Studies on growth changes in cells in culture or tumours produced by other DNA viruses in model (frequently immunoincompetent) animals, or their isolation from human tumours, have led to their tentative assignment as oncogenic, or potentially oncogenic, viruses. These include the ubiquitous human Polyomavirus, BK, that has been detected in brain tumours of different histological types, and also in KS, osteosarcomas and kidney carcinomas. Another human Polyomavirus, JC, causally associated with a pathological condition, PML (progressive multifocal leucoencephalopathy), has also been explored with regard to tumour formation. Both BK and JC encode large T antigens (LTs) that are related to that of SV40, and *in vitro* their genes stimulate cell growth (Barbanti-Brodano *et al.*, 1998). However, any significance in malignant growth in humans has not yet been established, although BK remains a candidate human oncogenic virus. The Adenoviruses, of which there are many distinct strains, have genes (*E1A* and *E1B*) that act as oncogenes in culture, interacting with tumour-suppressor genes. They, particularly, Ad12, can induce tumour formation, at least in animal models. Likewise, the human herpesvirus 6 (HHV-6) can transform mouse and human epidermal keratinocytes in culture, generating cell lines that are tumorigenic in athymic mice. Herpes-simplex virus (HSV) sequences have been found in human cervical cancer, but research on this topic was curtailed with the more definitive discovery of papillomaviruses in this malignancy. To date, roles for neither HSV nor HHV-6 in human malignancies have been established. The best evidence for viral causation of human malignancies appears to lie with the four DNA viruses dealt with in detail above, and with some of the RNA viruses, considered elsewhere, although it would be unwise to assume that other, undiscovered, candidate human oncogenic viruses do not exist. Newer, more sensitive methods of analysis make the discovery of novel cancer-related viruses a challenging research field.

REFERENCES

Arrand, J. R. (1994). Molecular genetics of human papillomaviruses. In: Stem, P. L. and Stanley, M. A. (eds), *Human Papillomavirus and Cervical Cancer: Biology and Immunology.* Chap. 2, 28–40. Oxford Medical Publications, Oxford.

Baer, R., *et al*. (1984). DNA sequence and expression of the B95-8 Epstein–Barr virus genome. *Nature*, **310**, 207–211.

Barbanti-Brodano, G., *et al*. (1998). BK and JC human polyoma viruses and simian virus 40: natural history of infection in humans, experimental oncogenicity and association with human tumors. *Advances in Virus Research*, **50**, 69–99.

Chang, Y., *et al*. (1994). Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma. *Science*, **266**, 1865–1869.

Chisari, F. V. (2000). Viruses, immunity and cancer: lessons from hepatitis B. *American Journal of Pathology*, **156**, 1118–1132.

Chow, L. T. and Broker, T. R. (1994). Papillomavirus DNA replication. *Intervirology*, **37**, 150–158.

Dane, D. S., *et al*. (1970). Virus-like particles in serum of patients with Australia-antigen-associated hepatitis. *Lancet*, **1**, 695–698.

Frazer, I. H. (1996). Immunology of papillomavirus infection. *Current Opinion in Immunology*, **8**, 484–491.

Ganem, D. (1998). Human herpesvirus 8 and its role in the genesis of Kaposi's sarcoma. *Current Clinical Topics in Infectious Diseases*, **18**, 237–251.

Glenn, M., *et al*. (1999). Identification of a spliced gene from Kaposi's sarcoma-associated herpesvirus encoding a protein with similarities to latent membrane proteins 1 and 2A of Epstein–Barr virus. *Journal of Virology*, **73**, 6953–6963.

Griffin, B. E. (1998). Relation of Burkitt's tumor-associated herpes-type virus to infectious mononucleosis. *Reviews in Medical Virology*, **8**, 61–66.

Griffin, B. E. (2000). Epstein–Barr virus (EBV) and human disease: facts, opinions and problems. *Mutation Research*, **462**, 395–405.

Griffin, B. E. and Xue, S.-A. (1998). Epstein–Barr virus infections and their association with human malignancies: some key questions. *Annals of Medicine*, **30**, 249–259.

Hagensee, M. E., *et al.* (1994). Three-dimensional structure of vaccinia virus produced human papillomavirus type 1 capsids. *Journal of Virology*, **68**, 4503–4505.

Hitt, M. M., *et al*. (1989). EBV gene expression in an NPC tumor. *EMBO Journal*, **8**, 2639–2651.

Howley, P. M. (1996). Papillomavirinae and their replication. In: Fields, B. N., *et al*. (eds), *Virology*, 3rd edn. Chap. 65, 2045–2076 (Lippincott-Raven, Philadelphia).

Hsu, L. C. (1996). Ethnic differences in immune responses to hepatitis B vaccine. *American Journal of Epidemiology*, **143**, 718–724.

IARC (1994). *Hepatitis Viruses*, Monograph **59** (IARC, Lyon).

IARC (1995). *Human Papillomaviruses*, Monograph **64** (IARC, Lyon).

IARC (1997). *Epstein–Barr Virus and Kaposi's Sarcoma Herpesvirus/Human Herpesvirus 8*, Monograph **70** (IARC, Lyon).

Khanna, R., *et al*. (1999). Vaccine strategies against Epstein–Barr virus-associated diseases: lessons from studies on cytotoxic T-cell-mediated immune regulation. *Immunological Review*, **170**, 49–64.

Labrecque, L. G., *et al*. (1999). Expression of Epstein–Barr virus lytically related genes in African Burkitt's lymphoma: correlation with patient response to therapy. *International Journal of Cancer*, **81**, 6–11.

Leight, E. R. and Sugden, B. (2000). EBNA-1: a protein pivotal to latent infection by Epstein–Barr virus. *Reviews in Medical Virology*, **10**, 83–100.

Morse, S. S. (1993). *Emerging Viruses* (Oxford University Press, Oxford).

Nakagawa, M., *et al*. (1997). Cytotoxic T lymphocyte responses to E6 and E7 proteins of human papillomavirus type 16: relationship to cervical intraepithelial neoplasia. *Journal of Infectious Diseases*, **175**, 927–931.

Pfister, H. and Fuchs, P. G. (1994). Anatomy, taxonomy and evolution of Papillomaviruses. *Intervirology*, **37**, 143–149.

Russo, J. J. (1996). Nucleotide sequence of the Kaposi's sarcoma-associated herpesvirus (HHV 8). *Proceedings of the National Academy of Sciences of the USA*, **93**, 14862–14867.

Shah, K. V. (1990). Biology of human genital tract papillomaviruses. In: Holmes, K. G., *et al*. (eds), *Sexually Transmitted Diseases*, 2nd edn. Chap. 37, 425–431 (McGraw-Hill, New York).

Shah, K. V. and Howley, P. M. (1996). Papillomaviruses. In: Fields, B. N., *et al*. (eds), *Virology*, 3rd edn. Chap. 66, 2077–2109 (Lippincott-Raven, Philadelphia).

Stöppler, H., *et al*. (1994). Transforming proteins of the papillomaviruses. *Intervirology*, **37**, 168–179.

Sun, R., *et al*. (1999). Kinetics of Kaposi's sarcoma-associated herpesvirus gene expression. *Journal of Virology*, **73**, 2232–2242.

Tarpey, I., *et al*. (1994). Human cytotoxic T lymphocytes stimulated by endogenously processed human papillomavirus type 11 E7 recognize a peptide containing a HLA-A2 ($A^{*}0201$) motif. *Immunology*, **81**, 222–227.

Van Ranst, M., *et al*. (1992). Phylogenetic classification of human papillomavirus: Correlation with clinical manifestations. *Journal of General Virology*, **68**, 4503–4505.

FURTHER READING

Arrand, J. R. and Harper, D. R. (1998). *Viruses and Human Cancer.* (BIOS Scientific Publishers, Oxford).

Boshoff, C. (1998). Kaposi's sarcoma associated herpesvirus. In: Newton, R., *et al*. (eds), *Cancer Surveys* 33. Infections and Human Cancer.157–190. (Imperial Cancer Research Fund, London).

Evans, A. S. and Kaslow, R. A. (eds) (1997). *Viral Infections of Humans. Epidemiology and Control*, 4th edn. (Plenum Medical Books, New York).

Fields, B. N., *et al*. (eds) (1996). *Fields Virology*, 3rd edn, Vol. 2. (Lippincott-Raven, Philadelpia).

Reitz, M. S., *et al*. (1999). Perspectives on Kaposi's sarcoma: facts, concepts and conjectures. *Journal of the National Cancer Institute*, **91**, 1453–1458.

Rezza, G., *et al*. (1999). Human herpesvirus 8 seropositivity and risk of Kaposi's sarcoma and other acquired immunodeficiency syndrome-related diseases. *Journal of the National Cancer Institute*, **91**, 1468–1474.

Yan, R. Q., *et al*. (1996). Human hepatitis B virus and hepatocellular carcinoma. I. Experimental infection of tree shrews with hepatitis virus. *Journal of Cancer Research and Clinical Oncology*, **122**, 283–288.

Websites

The human papillomavirus database: http://hpv-web.lanl.gov.

Virus databases on-line: http://www.ncbi.nlm.nih.gov/ICTV; http://www.ncbi.nlm.nih.gov/htbin-post/Taxonomy; http://www.virology.net.

Chapter 6

RNA Viruses

Mitsuaki Yoshida
Banyu Tsukuba Research Institute, Tsukuba, Ibaraki, Japan

CONTENTS

GENERAL DESCRIPTION OF RNA TUMOUR VIRUSES

Introduction

It is well established that tumour formation proceeds through multiple steps, each of which consists of mutation and selection of the mutated cells. Various signals for induction of cellular proliferation and fixation of the abnormal phenotypes by genetic mutation promote cellular conversion into a more malignant state. External cellular signals are transduced into the nucleus through multiple pathways and finally induce specific gene expression and cellular responses. Genetic alterations of these signalling pathways, transcriptional machinery or the target genes themselves are the origin of cancers.

Infection by some specific viruses represents one of these genetic alterations and triggers these multiple-step processes for the final induction of specific cancers in both animals and humans. In human cancers, the real causes of the cancers are mostly unknown, and therefore there is no effective way to diagnose healthy individuals who will develop cancers in their near future. This is possible, however, when the cancers result from specific viral infection since the virus carriers are easily identified by their antibodies against specific viruses. Such situations provide an opportunity not only to prevent the cancers, but also to investigate the mechanism of cancers directly in humans.

As human tumour viruses, human T cell leukaemia virus (HTLV) type 1 (HTLV-1), hepatitis C-type virus (HCV), hepatitis B-type virus (HBV), human papillomavirus (HPV) and Epstein–Barr virus (EBV) are well established. Among these, the first two, HTLV-1 and HCV, are the RNA viruses, i.e. they have RNA as their genomes. These two viruses, however, are classified into unrelated groups, HTLV-1 to the retroviridae and HCV to the flabiviridae, which are different in many respects. In this chapter, retroviruses are the main subject.

Viral Replication

The term RNA tumour viruses generally represents retroviruses that contain reverse transcriptase (RT), which transcribes genomic RNA into DNA upon infection. Cores of the viral particles contain two copies of a single-stranded, positive RNA, Gag protein and reverse transcriptase. The core is enveloped with membrane similar to the plasma membrane of host cells, on which the viral Env (envelope) glycoprotein is exposed. The interaction of the Env protein with a receptor on a target cell membrane is required for infection. Interaction of the Env with the receptor induces membrane fusion between the viral particle and cell, allowing the core to be incorporated into the cell, then the genomic RNA is reverse transcribed into complementary DNA (cDNA) by the particle RT (**Figure 1**). The cDNA is then converted to double-stranded DNA and integrated into the host chromosomal DNA forming a 'provirus.' During these processes, a long terminal repeating sequence (LTR) is formed at both ends which contains many elements essential to viral gene

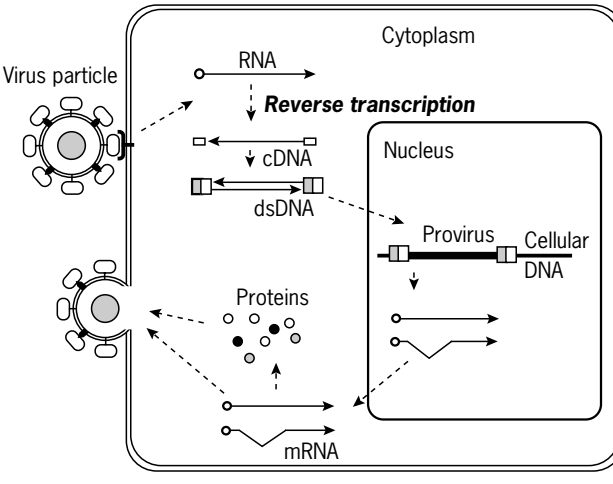


Figure 1 Life cycle of retroviruses.

expression and replication. Details of the mechanism of replication have been reviewed elsewhere (Weiss *et al*., 1985). Once the proviruses are integrated into germ cells, the retroviruses can be transmitted vertically to successive generations of hosts by transmission of proviral genomes (endogenous viruses), in addition to the standard viral replication and infection (exogenous viruses).

The proviruses are transcribed into RNA by the cellular transcription machinery and a subpopulation of the viral transcripts is spliced into subgenomic mRNA. Both spliced and unspliced viral RNA species are translated into viral proteins; genomic RNA (unspliced RNA) into Gag and Pol proteins and subgenomic RNA into Env protein. The viral proteins and genomic RNA are assembled at specific sites under the plasma membrane into the particles and released by budding. Generally, retroviral replication is not harmful to the host cells except in some cases such as human immunodeficiency virus (HIV). The receptors for a few retroviruses have been identified as membrane proteins, and their expression in many types of cells is consistent with the array of cells at risk of retroviral infection.

Genome Structures and Tumorigenesis

There are two types of retroviral genome; one type carries and the other does not carry host-derived oncogenes (**Figure 2**). Replication-competent retroviruses generally have *gag*, *pol* and *env* genes but do not carry oncogenes. Some of these viruses induce leukaemia or lymphoma after a long latency and are thus called chronic leukaemia viruses. This type of retrovirus induces leukaemia through activation of expression of the adjacent cellular genes by the integrated LTRs. When a provirus is, by chance, integrated in the vicinity of a proto-oncogene and induces abnormal expression of the proto-oncogene it may lead to tumorigenesis. This 'promoter-insertion mechanism of viral carcinogenesis' operates in a variety of tumour systems; *Myc* activation by avian leukaemia virus (ALV) in B cell lymphoma, *erbB* activation by ALV in chicken erythroblastosis and *int1* and *int2* activation by mouse mammary tumour virus (MMTV) in mouse mammary tumours. Because proviral integration is not site specific, repeated integration through viral replication is usually required before the provirus appears in a tumorigenic site. This explains the long latency after infection for tumorigenesis; however, once it integrates into the right place of the host genome, viral replication is no longer required for tumorigenesis.

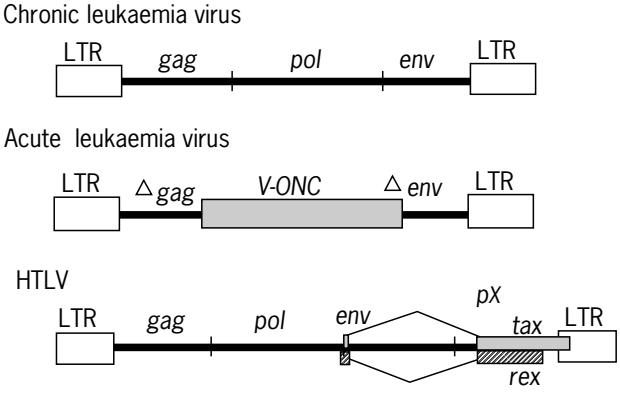


Figure 2 Genome structure of proviruses.

Endogenous viruses, which are proviruses vertically transmitted through germ cells, mostly have this type of genome. Many copies of the endogenous viruses are maintained in human cells, as many as thousands in some cases, but they are not very replicative even though they have complete genomes. Various indications of possible participation of endogenous viruses in human cancers have been reported, but, it is still a subject for further study.

The other type of retroviruses carry an oncogene derived from a cellular proto-oncogene (Weiss *et al*., 1985). Capture of a oncogene in the viral genome, which is derived from host cell DNA, in turn results in a deletion of some portion of the viral genome. Consequently, acute sarcoma/leukaemia viruses are generally replication defective. This type of virus has to be infected together with a chronic leukaemia virus as a 'helper' for their replication. One exception is Rous sarcoma virus, which has the oncogene *src* between *env* and the *3′LTR* and is competent in replication. The viral oncogene of the acute leukaemia viruses is responsible for transformation of infected cells and does not require viral replication or site-specific integration, inducing specific tumours with short latency. In contrast to the broad specificity of retroviral infection *in vitro*, these viruses are able to induce tumours in relatively few tissues *in vivo*. Specificity of retroviral tumorigenesis is restricted by either the class of viral oncogene or the tissue-specific promoter activity of the integrated *LTR*, in addition to the nature of the viral receptors on the host cells.

HUMAN TUMORIGENIC RETROVIRUSES: HTLVs

Human retroviruses include HIV, human endogenous viruses and human foamy viruses. The replication and pathology of HTLV-1 are considered throughout this chapter, but with some exceptions, observations on HTLV-1 are applicable to the other members of the HTLV group (see the following section).

HTLVs and Disease

HTLV-1 is an aetiological factor in adult T cell leukaemia (ATL) (Poeisz *et al*., 1980; Hinuma *et al*., 1981; Yoshida *et al*., 1982). ATL is a unique T cell malignancy with a CD4-positive phenotype. HTLV-1 prevalence and ATL

are clustered in southwestern Japan. The sera of ATL patients contain antibodies that react specifically with cell lines established from the ATL patient. The antigens in these cell lines were later proved to be HTLV-1 proteins by molecular characterization of the viral genome. The extensive epidemiological studies on antibodies and ATL clearly established that HTLV-1 is closely associated with ATL. The nationwide and worldwide epidemiology indicated that HTLV-1 infection is also associated with a neurological disease, HAM/TSP (HTLV-1-associated myelopathy/tropical spastic paraparesis).

The genome of HTLV-1 provirus consists of *LTR–gag–pol–env–pX–LTR* (Seiki *et al*., 1983) and is distinct from the standard retroviral genome, *LTR–gag–pol–env–LTR* (**Figure 2**). The existence of a functional *pX* sequence is unique to this virus, thus HTLV-1 forms an independent retroviral group, HTLV. After characterization of HTLV-1, another virus, HTLV-2, was isolated from a patient with hairy T-cell leukaemia (Chen *et al*., 1983). The genome is about 60% homologous to HTLV-1, but its pathogenicity is not yet established.

Viruses similar to HTLV-1, simian T cell leukaemia viruses (STLV), were isolated from various species of nonhuman primates, including chimpanzee. The STLVs share a 90–95% identity of genomic sequence with HTLV-1 and some isolates show more homology than those among human isolates. Although STLV is widely distributed in monkeys, no typically leukaemic animals have been observed. A few cases of a leukaemia-like disease have been noted in STLV-infected monkeys in zoos, but an aetiological connection between STLV and the disease remains to be established. Another member of the HTLV group is bovine leukaemia virus (BLV). BLV infects and replicates in B cells of cows and induces B cell lymphoma. BLV also infects lymphocytes of sheep and induces leukaemia after a short latent period.

Epidemiology

Nearly all ATL patients carry antibodies against HTLV-1 proteins. These antibodies are easily detected by indirect immunostaining of cells infected with HTLV-1, by an enzyme-linked immunosorbent assay (ELISA), by a particle-agglutination assay or by Western blotting. Some populations of healthy adults also carry HTLV-1 antibodies and these sero-positive persons are defined as the viral carriers. In fact, HTLV-1 can be detected in such individuals using the polymerase chain reaction (PCR). In considering the epidemiology of HTLV-1, geographic clustering, age-dependent prevalence, genomic variation and geographic origin of HTLV-2 are informative.

Geographic Clustering

HTLV-1 antibodies are recognized in 5–15% of adults clustered in southwestern Japan, the Caribbean islands and South America, Central Africa and Papua New Guinea and the Solomon Islands in Melanesia. The prevalence of sero-positive adults varies significantly from one district to another within these areas of endemicity. For example, in a particular isolated island in Kyushu, Japan, 30–40% of people over 40 years of age might be infected, whereas on a neighbouring island, the prevalence may be far lower. Significantly, ATL and HAM/TSP are also clustered, overlapping HTLV-1 in distribution. ATL patients and healthy carriers found sporadically in nonendemic areas mostly originate from one of the endemic areas.

HTLV-2 is frequently isolated in the United States and other countries from intravenous drug abusers and persons infected with HIV. HTLV-2 is endemic in South America, and also in Pygmy populations in Africa.These regions are likely to be the natural reservoir of HTLV-2 sporadically observed in the other places.

Age-dependent Prevalence

The prevalence of virus carriers increases with age after 20 years increasing sharply around 40–50 years of age and reaching a maximum in people aged between 50 and 60 years. The prevalence is significantly (1.6 times) higher in females than in males, but the incidence of ATL is similar in both sexes. The increase in prevalence among females can be attributed to sexual transmission of HTLV-1 from husbands to wives and also suggests that such infection is not leukaemogenic.

Although global epidemiology identified the age-dependent increase of antibody prevalence, cohort studies over 10 years in Japan revealed that sero-conversion of adults from antibody negative to positive is very rare. These observations are unable to explain the sharp increase of sero-prevalence in the 40–50 years age group. After extensive epidemiological studies of antibody prevalence, it is now accepted that the age-dependent increase is a reflection of the reduction of the infection risk at the early stage of life. Artificial milk became popular around 40 years ago in these areas in Japan, and thus reduced the incidence of breast milk-born infection of HTLV-1 (discussed below).

Genomic Stability

The viral genome is well conserved (over 96%) in Japan and the Caribbean area. Viral isolates from Africa and Papua New Guinea may vary somewhat more, but the variations are very limited. Retroviral genomes are thought to be unstable relative to those of other viruses because reverse transcription has no proofreading mechanism. Such genomic stability stands in sharp contrast to the highly labile genome of HIV. This may be associated with the very low competency of HTLV-1 replication *in vivo*.

Infection of HTLV-1

Infection In Vitro

Viral particles of HTLV-1 show extremely low infectivity *in vitro*, but co-cultivation with virus-producing cells can transmit HTLV-1 to a variety of human cells, including T and B lymphocytes, fibroblasts and epithelial cells, and also cells from monkeys, rats, rabbits and hamsters but, curiously, not mice. In these infected cells, the provirus is integrated into random sites in the chromosomal DNA, and most of the viral genes are successfully expressed. However, in non-T cell lines, the integrated proviruses become latent in expression, otherwise usually inducing the fusion of infected cells forming syncytia that ultimately die.

Only CD4-positive T cells are frequently immortalized upon infection with HTLV-1 (Miyoshi *et al*., 1981) and the immortalized cells express high levels of IL-2Rα, proliferating in an IL-2-dependent fashion. Animal retroviruses that do not carry an oncogene generally do not immortalize cells *in vitro*, and therefore immortalization by HTLV-1 appears to be unique and suggests that the virus may have a particular function. Accordingly, a contribution of *pX* to this effect has been proposed.

In contrast to *in vitro* infection, the cells of both ATL patients and asymptomatic viral carriers infected *in vivo* are almost exclusively T cells with the CD4 + phenotype. Furthermore, infected cells *in vivo* do not express viral information at significant levels. Reverse transcriptase-mediated PCR (RT-PCR) on mRNA indicates that over 95% of infected cells fail to express viral genes *in vivo* irrespective of whether they are in a transformed or non-transformed state.

Natural Transmission

HTLV-1 can be transmitted *in vivo* through (1) blood transfusion, (2) nursing with breast milk, and (3) sexual relations.

Blood Transfusion

Retrospective studies of blood transfusions showed that 60–70% of recipients of fresh, sero-positive blood were infected with HTLV-1. Transfer of infected cells from donor to recipient is required for viral transmission, and therefore fresh, sero-positive plasma does not support the infection. Blood transfusion-mediated transmission of HTLV-1 seems not to induce ATL (see Sexual Transmission, below), but does induce HAM/TSP. Therefore, rejection of HTLV-1-positive blood can protect recipients against both HTLV-1 infection and development of HAM/TSP. The blood-mediated transmission of HTLV-1 explains a high prevalence of up to 20% of abusers of intravenous drugs by the sharing of unsterilized needles.

Mother to Child

Viral transmission from mother to child was originally suggested by epidemiological evidence: most mothers of sero-positive children were carriers of the virus and about 30% of the children of sero-positive mothers were themselves sero-positive. Neonatal infection was initially suspected, but surveys of lymphocytes in cord blood from a large number of children born to sero-positive mothers have virtually, but not completely, excluded this possibility. Instead, breast milk was found to be a likely source of transmissible virus. Supporting this, milk taken from sero-positive mothers and given to adult marmosets leads to the appearance of antibodies in these monkeys. More direct evidence stems from a practical trial demonstrating that cessation of breast-feeding by sero-positive mothers drastically reduced the sero-conversion rates of their children (see the last section).

Sexual Transmission

Wives with sero-positive husbands are very frequently sero-positive. Conversely, the husbands of sero-positive wives show the same frequency of sero-positivity as do men of the region under study. On these grounds, it seems that the virus can be transmitted from husband to wife but not vice versa. Infected T cells have been found in semen from men infected with HTLV-1 and these cells are considered to transmit the virus from male to female. The higher rate of sero-positivity in female (1.6 times) is explained by this transmission. The sex-specific incidence of ATL does not mirror this difference, suggesting that HTLV-1 infections sexually transmitted to females are not leukaemogenic to ATL.

Viral Gene Expression

Proviral Genome Unique to HTLV-1

The HTLV-1 proviral genome cloned from leukaemic-cell DNA from an ATL patient is 9032 bp long and consisted of *LTR–gag–pol–env–pX–LTR* (Seiki *et al*., 1983). The presence of a *pX* region on the 3′ side of the *env* gene distinguishes the HTLV-1 genome from those of other retroviruses (**Figure 2**).

In general in retroviruses, the LTRs function as units regulating viral gene expression and replication. Furthermore, the *pX* region of HTLV contains additional, overlapping, regulatory genes (**Figure 3**): *tax*, *rex* and a gene whose function remains unknown, which encode p40tax, p27rex, and p21x. The Tax protein, p40tax, is a potent *trans*-activator of proviral transcription (Sodroski *et al*., 1984; Fujisawa *et al*., 1985) and thus is essential to viral gene expression (Yoshida, 1995). The second protein, p27rex, is a *trans*-acting modulator of RNA processing, which allows expression of the unspliced *gag* and *env* mRNAs in the cytoplasm. Expression of these unspliced

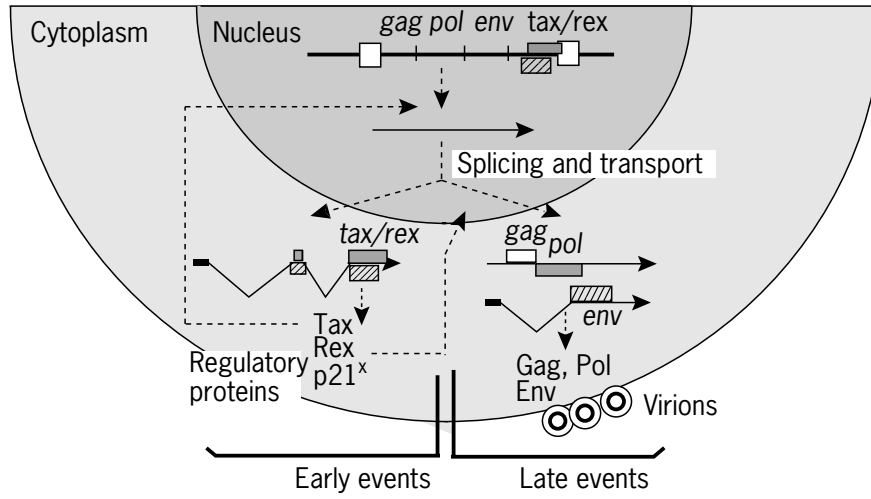


Figure 3 Transcription and viral RNA processing.

mRNAs is essential for expression of the viral structural proteins. Thus, Rex is also essential for HTLV-1 gene expression and replication. These systems are unique among retroviral regulations and have similarities to Tat and Rev systems of HIV, which are also essential for the replication of HIV.

The *pX* sequence is also able to encode various proteins when the sequence is alternatively spliced. The function of some of these products was characterized using expression vectors, but the physiology of these proteins remains to be analysed.

Transcriptional Activation

Retroviral LTR elements contain a TATA box, a transcriptional enhancer and a poly(A) signal, all of which are essential for viral gene expression and replication of RNA tumour viruses. These elements are recognized by cellular transcriptional factors for RNA polymerase II and retroviral gene expression depends upon the cellular machinery of the host. In addition to LTR-mediated regulation, HTLV-1 contains the *tax* gene, which acts in trans to stimulate viral transcription (**Figure 3**). Tax function depends on three direct repeats of the 21-bp sequence transcriptional enhancer, in the LTR. The interaction of Tax with the enhancers has been proposed to be indirect, that is, Tax interacts with a cellular protein that binds to the enhancer DNA to mediate trans-activation.

In addition to the viral genome, Tax activates cellular genes. Since the transcription of the gene for the α-chain of the IL-2 receptor (IL-Rα) was reported to be activated by Tax, many other genes were identified also to be activated but through different mechanisms (see later). These include the gene for GM-colony-stimulating factor (GM-CSF), the proto-oncogenes c-*fos* and c-*jun*, the genes for parathyroid hormone-related protein (PTHrP), MHC class I antigen and many others.

RNA Processing

The primary transcript, the viral genome, of HTLV-1 contains genes for Gag, Pol, Env, Tax, Rex and a few others in this order with some overlapping. Only the first coding frame on an mRNA is translated into protein in eukaryotic cells, and therefore the retrovirus needs to splice the primary transcript into various species of viral mRNA to encode other proteins. This splicing is regulated by Rex, a pX protein of HTLV-1 (**Figure 3**). Viral replication requires three species of mRNA: 8.5-kb genomic (unspliced) RNA as *gag* and *pol* mRNA, a 4.2-kb singly spliced sub-genomic RNA as *env* mRNA and a 2.1-kb doubly spliced mRNA for the expression of Tax and Rex proteins. The viral transcripts early after infection or induction are all spliced into completely spliced *tax*/*rex* mRNA. Tax and Rex are then expressed, and viral transcription is enormously enhanced and at the same time, unspliced RNA for Gag, Pol and Env are expressed in the cytoplasm. This regulation is essential for viral protein expression, since host-cell mRNAs are generally all spliced into mRNA to be transported to the cytoplasm.

Rex function in turn reduces the level of spliced mRNA that encodes regulatory proteins, including Tax protein, and eventually reduces viral transcription. Thus, Rex enhances the expression of viral structural proteins, but suppresses total viral gene expression. In short, Rex exerts feedback control of the viral gene expression, thus resulting in a transient expression of HTLV-1.

For this regulation, target RNA requires to have a cis-acting element (RxRE) consisting of 205 nucleotides located in the 3′ region of the viral RNA. The unique secondary structure of this element allows Rex protein to bind to it. A nuclear export signal in the Rex protein suggests transport of a Rex–RNA complex into the cytoplasm without processing. HIV has a *rev* gene, strikingly similar in function to *rex* of HTLV-1. In the HIV system, Rev protein binds to RvRE in the *env* coding sequence.

Adult T Cell Leukaemia (ATL)

It is now established that HTLV-1 is associated with ATL, HTLV-1-associated myelopathy or tropical spastic paraparesis (HAM/TSP) and uveitis. Other diseases such as chronic lung disease, monoclonal gammopathy, chronic renal failure, strongyloidiasis and nonspecific dermatomycosis are also suggested to associate with the viral infection; however, further systematic studies are required to establish the exact relationships. Here, only ATL is described.

The Leukaemic Cells

ATL cells are T cells with the CD4+ phenotype and, usually, a highly lobulated nucleus. These cells always carry HTLV-1 proviruse(s) and the site of integration is monoclonal in a given ATL patient. In 70–80% of cases of ATL patients examined, one copy of the complete provirus was integrated into each leukaemic cell. Occasionally, one or two copies of defective provirus are integrated into the DNA of a single cell. Even in the defective genomes, preservation of the pX region in defective proviruses suggests its importance in tumorigenesis.

The leukaemic cells express a high level of IL-2Rα on their surfaces. Production of PTHrP, IL-1β or GM-CSF by tumour cells has also been described. In almost all cases, leukaemic cells carry aberrant chromosomes, and there are frequently multiple abnormalities, such as trisomy of chromosome 7 and 14q11, 14q32 and 6/q15 translocations. The abnormality involving 14q32 was found in 25% of ATL patients, but the others appeared less frequently.

Clinical Features

ATL (Uchiyama *et al*., 1977) is classified into three phases, smouldering, chronic and acute phases, depending on clinical features. In smouldering ATL, patients commonly have from one to several per cent of morphologically abnormal T cells in their peripheral blood, but do not show other signs of severe illness and are therefore thought to be in an early stage of ATL development. In smouldering ATL, the abnormal cells are not aggressively malignant, but are HTLV-1 infected and expanded clonally. The onset of ATL is observed between 20 and 70 years of age, the peak rate of onset being in the 40s and 50s. The male-to-female ratio of ATL is 1.4:1.0. Symptoms of ATL vary from patient to patient, but are frequently complicated by skin lesions, enlargement of lymph nodes, liver and/or spleen and infiltration of leukaemic cells into the lungs and other organs. Patients usually have antibodies to HTLV-1 proteins, show an increased level of serum LDH and suffer from hypercalcaemia. The acute form, or phase, of ATL is aggressive and resistant to treatment; consequently, most patients in this phase die within 6 months of its onset.

Molecular Mechanism of Pathogenesis

Viral Function in Leukaemogenesis

The aetiological role of HTLV-1 in ATL has been demonstrated by sero-epidemiology and molecular biology of HTLV-1 and ATL. The bases of this are as follows: (1) ATL and HTLV-1 geographically overlap (population level); (2) most ATL patients are infected with HTLV-1 (individual level); (3) leukaemic cells from ATL patients are infected with HTLV-1 (cell level); (4) more importantly, the leukaemic cells show monoclonal integration of proviral DNA (molecular level); and (5) HTLV-1 has the capacity to immortalize human T cells *in vitro* (biochemical level). The evidence in (4) indicates that the leukaemic cells originated from a single HTLV-1-infected cell and, thus, that HTLV-1 plays a causative role in leukaemogenesis. There are estimated to be approximately one million carriers of HTLV-1 in Japan, and about 500 new cases of ATL are reported each year. About 2–5% of all carriers of HTLV-1 are thought to develop ATL during their life span (Tajima, 1990).

The site for provirus integration is monoclonal in ATL cells, but not the same among ATL patients (Seiki *et al*., 1984). Therefore, the promoter insertion model is unlikely since it requires a common integration adjacent to a proto-oncogene. Consequently, a 'trans-acting function' of HTLV-1 is postulated in leukaemogenesis. Molecular biology studies of HTLV-1 showed that the Tax protein functions as a 'trans-acting factor.' Consistent with these observations, Tax was found to immortalize T cells in an IL-2-dependent fashion, to transform rat embryonic cells in cooperation with c-*ras*, and to induce mesenchymal tumours in Tax transgenic mice. The central role of Tax in leukaemogenesis is thus proposed.

Trans-activation of Transcription

Activation of Enhancer Binding Protein

Tax trans-activates transcription via specific enhancers such as the 21-bp enhancer in the LTR, the NF-κB binding site in the gene for interleukin-2 receptor α and serum responsive element (SRE) in the c-*fos* gene . However, Tax is unable to bind directly to the enhancer DNA sequence. Instead, Tax binds to enhancer binding proteins; the first group includes CREB (cyclic AMP-responsive element (CRE) binding protein, CREM (CRE modulator protein), ATF-1 and ATF-2, which bind to the 21-bp enhancer in the HTLV-1 LTR. The second group is the family of NF-κB such as p50, p65, c-Rel and p52, which bind to the NF-κB binding site in IL-2 receptor α gene, and the third group is SRF (serum response factor) which binds to SRE (serum response element) in the c-*fos* or c-*egr* gene. These transcription factors are regulated by signal-dependent phosphorylation in normal cells; however, Tax binding permits

activation of these factors without specific phosphorylation, thus establishing constitutive activation of these genes.

It is now demonstrated that Tax binding to these transcription factors recruits CBP (CREB binding protein) on to the enhancer–CREB complexes (Kwok *et al*., 1994), because of Tax's affinity for CBP (**Figure 4**). CBP has a histone-acetylating activity and is normally unable to bind to CREB protein unless it is phosphorylated by protein kinase A. Therefore, the ternary complex of the enhancer–CREB–Tax–CBP without phosphorylation would acetylate histones bound to DNA nearby and thus activate transcription initiation. This simple binding hypothesis, however, may need further careful characterization.

Inactivation of Transcriptional Inhibitors

NF-κB proteins are activated by Tax binding in the nucleus (see the previous section). However, in resting cells, NF-κB proteins are retained in the cytoplasm by forming complexes with IκB. Stimulation of cells induces phosphorylation of IκB protein and its degradation. The released NF-κB then migrates into the nucleus because of its nuclear translocation signal and binds to its specific DNA sequences.

Tax also binds to IκBalpha which results in destabilization of IκB–NF-κB complexes. It is uncertain whether Tax induces phosphorylation of IκB and /or degradation of IκB. Activation of a protein kinase activity by Tax suggests the former mechanism, but the binding of Tax to a component of the proteasome may suggest the latter mechanism. Whatever the mechanism, Tax is able to down-regulate the transcriptional inhibitor, IκB, in the cytoplasm and thus activates NF-κB-dependent transcription. Tax appears to exert its effects via two independent mechanisms targeting an activator, NF-κB, and its inhibitor, IκB.

Trans-repression of Transcription

When the trans-activating function of Tax was extensively investigated, Tax was reported to trans-repress transcription of DNA polymerase beta (Jeang *et al*., 1990), which is required for repair of damaged DNA. Tax was then demonstrated to trans-repress the transcription of a set of growth-inhibitory genes such as *p18INK4c*, *NF-1*, *lck* and a apoptotic gene *bax*. Furthermore, Tax was also shown to trans-repress p53-dependent transcription, which affects the tumour-suppressor function of p53. These effects, therefore, suggest an abnormally enhanced proliferation of HTLV-1-infected cells. It is of interest to know how a trans-activator, Tax, is able to trans-repress other sets of genes.

The underlying mechanism was in fact shown to be rather simple, that is, inhibition of a transcriptional coactivator family, CBP/P300 (**Figure 4**). An E-box binding protein E47 in *p18INK4c* expression and p53 in p53-dependent transcription are essential to interact with CBP/p300 to achieve efficient expression. In these systems, Tax binds to CBP/p300 and interferes with the interaction between CBP/p300 and enhancer binding proteins, E47 or p53, resulting in trans-repression of specific transcription. This implies that Tax would be able to suppress many other genes since the CBP/p300 protein serves as a coactivator for a huge number of genes. It is therefore suggested that the cascade of Tax activity in transcriptional regulation

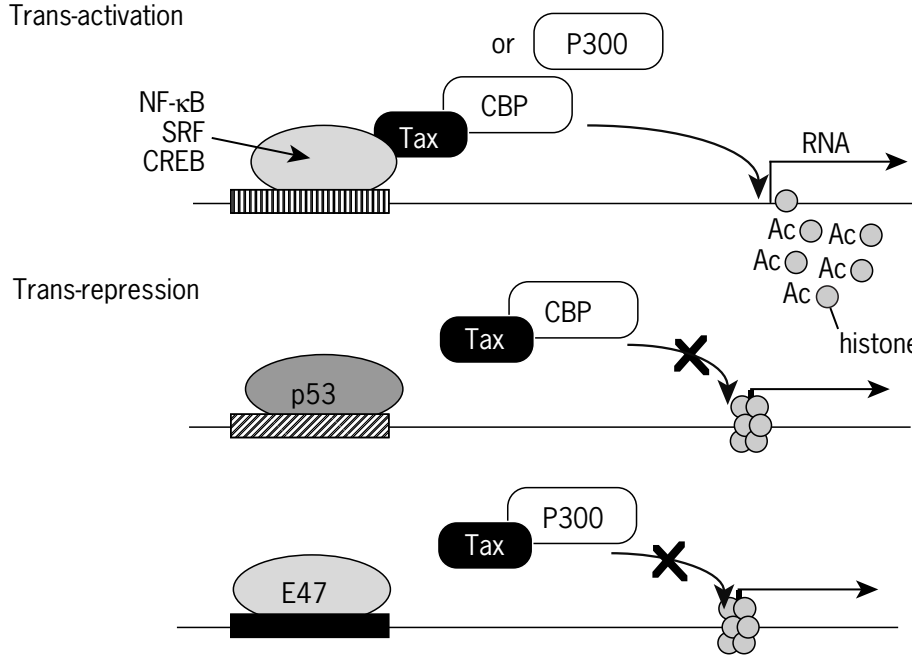


Figure 4 Tax binding to CBP or p300 to activate and repress the specific transcription.

would be unexpectedly wide in its targets and highly variable in its effect, depending on the level of expression.

Inhibition of Tumour-Suppressor Proteins

Independently of transcriptional regulation, the Tax protein was also found to interact with an inhibitor of cyclin-dependent kinase 4 (CDK4), p16INK4a (Suzuki *et al*., 1996) and a *Drosophila* large disc tumour-suppressor protein, Dlg. Furthermore, the direct binding inhibits the function of tumour-suppressor proteins. These suppressive effects on tumour-suppressor functions strongly suggest that Tax protein contributes to the development of ATL.

Cell Cycle Inhibitor

p16INK4a and p15INK4b are inhibitors of CDK4/6 and their inactivation results in activation of the kinases. Upon activation of CDK4/6, Rb is phosphorylated and is no longer able to bind E2F, thus releasing the active form of E2F, which then binds to target DNA sequences and initiates expression of various genes important for DNA synthesis. Tax binding to and inhibition of both p16INK4a and p15INK4b is able to activate CDK4 and promote cell entry into S phase (**Figure 5**). p16INK4a is frequently deleted in many human tumour cells, particularly in melanoma and haematopoietic tumours, and the deletion has been suspected to play a critical role in tumour induction and progression. Therefore, the functional knockout of the tumour-suppressor protein mimics the effect of gene deletion and may contribute to development of ATL.

With respect to the inactivation of Rb signalling pathway by p16INK4a, it is of interest to point out that DNA tumour viral proteins such as SV40 T-antigen, adenovirus E1A and papillomavirus E6/7 target Rb protein in their transformation (Nevins, 1992). The observation that developmentally unrelated tumour viruses share the common signalling pathway by targeting different molecules suggests that the Rb pathway is critical for the normal regulation of cell proliferation.

hDlg that Associates with APC, Another Tumour-suppressor Protein

Another target of Tax is hDlg, which is a signalling molecule downstream of Wnt/Frizzled and upstream of β-catenin in their signalling pathways. hDlg binds to the C-terminus of the tumour-suppressor protein APC involved in transducing cytostatic signals. Tax binds to the same domain of hDlg *in vitro* and *in vivo* and competes with APC, thus abrogating the growth-retarding signalling. In addition to the competitive displacement of APC from hDlg, Tax further induces hyperphosphorylation of hDlg as demonstrated by its slower migration in gel electrophoresis. It is noteworthy that APC and hDlg are significantly expressed in normal T cells (T. Suzuki and M. Yoshida, unpublished observation), hence these interactions might have roles in T cells, although these are not well understood.

Cell Cycle Check Point Protein, HsMAD1

Cell cycle processes are inspected at checkpoints to determine whether the scheduled processes are verified. Thus, once the checkpoint system is compromised, damaged cells can go through their cell cycle and proliferate, fixing genetic abnormality in the daughter cells. Tax of HTLV-1 can bind to the human homologue (HsMAD1) of yeast mitotic checkpoint protein MAD1 (Jin *et al*., 1998). HsMAD1 is a component of the mitotic

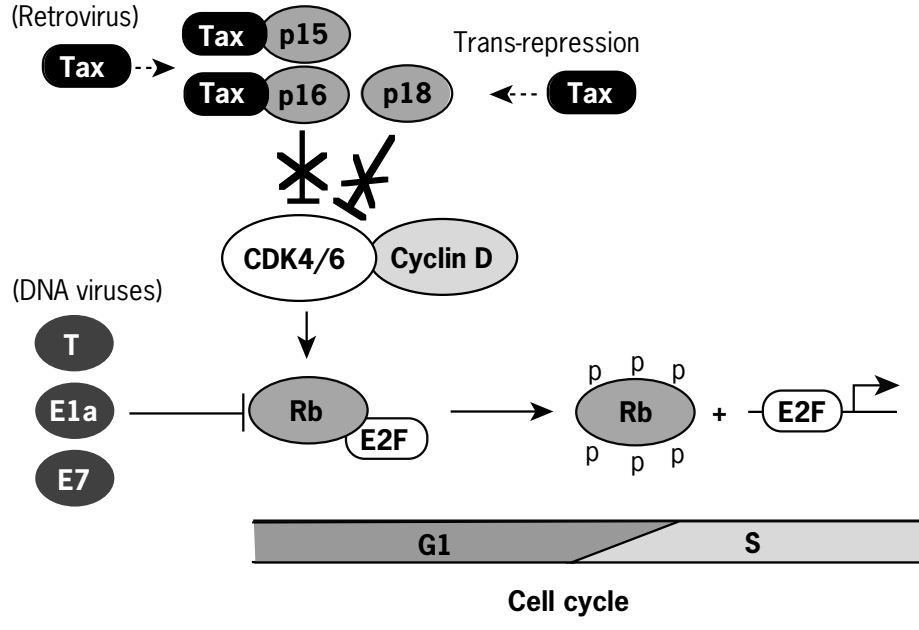


Figure 5 Tax binding to and inactivation of tumour-suppressor proteins, p16INK4a and p15INK4b.

checkpoint system which prevents anaphase and commitment to cellular division until chromosomal alignment is properly completed. Therefore, abrogation of the mitotic checkpoint function of HsMAD1 may be linked to chromosomal abnormalities which are observed at unusually high frequency in ATL cells.

Common Targets with DNA Tumour Viruses

The Tax protein appears to be pleiotropic through interacting with so many cellular regulators: activation and repression of transcription of different sets of genes, inhibition of CDK4 inhibitors and inhibition of tumour-suppressor proteins. Some of the target molecules and target pathways of Tax protein are shared by transforming proteins of DNA tumour viruses, T antigens of SV40, E1A/B of adenoviruses and E6/7 of papillomaviruses. The most striking shared target is the Rb signalling pathway: Tax targets p16INK4a and p15INK4b and the transforming proteins of DNA tumour viruses target Rb, both resulting in activation of E2F family and promotion of cells into the S phase of the cell cycle (**Figure 5**). Other examples shared are the transcriptional target coactivators CBP/p300 and hDlg; CBP/p300 is affected by T antigen and E1A, and hDlg by E6 of high-risk HPV and E4 9ORF1 of adenoviruses.

The genes for p16INK4a and Rb proteins are frequently mutated or deleted in spontaneous human cancer cells and are believed to play critical roles in tumorigenesis. Therefore, functional knockout of these gene products by viral proteins is similarly expected to play crucial roles in specific tumour induction. It is also interesting that developmentally unrelated viruses and spontaneous genetic mutations target the same signalling pathway for malignant transformation. By analogy with these spontaneous mutations or deletions, the virally induced functional inactivation of either p16INK4a or Rb protein would be primarily responsible for the induction of specific tumours in humans.

Possible Mechanism of Pathogenesis

Cooperation of the Pleiotropic Function of Tax Protein

As summarized in the previous section, Tax protein effects are pleiotropic and function at different levels, including transcriptional regulation, cell cycle promotion, tumour-suppressor function and genomic stability. It is reasonable to speculate that the pleiotropic functions of Tax would be mostly cooperative in ATL development (**Figure 6**). This interesting possibility is discussed below.

Cell proliferation. It has been demonstrated that various genes activated by Tax are in fact able to promote cell proliferation *in vitro*; transcriptional activation of lymphokines such as IL-6, GM-CSF, TGF-β and some others, lymphokine receptors such as IL-2Rα and some nuclear oncogenes such as c-*fos*, c-*egr* and c-*jun*, are all growth promoting. These transcriptional trans-activations are mediated independently through NF-κB or SRF and some others through CREB. Furthermore, trans-repression of *p18INK4c*, *lck* and *NF-1* would result in promotion of cell growth, since these genes are mostly growth-retarding. Direct inhibition of tumour-suppressor proteins such as p16INK4a, p15INK4b, cyclin D3 and hDlg all result in abnormal cell proliferation of infected cells. Indeed, most of the genes and proteins targetted directly or indirectly by Tax are likely to cooperate in promotion of cell proliferation upon infection by HTLV-1. This consideration is consistent with the fact that normal cell regulation is protected by redundant mechanisms and also that tumorigenesis proceeds through accumulation of multiple gene abnormalities.

Fixation of genetic abnormality. In addition to the genes discussed in the previous section, Tax also affects

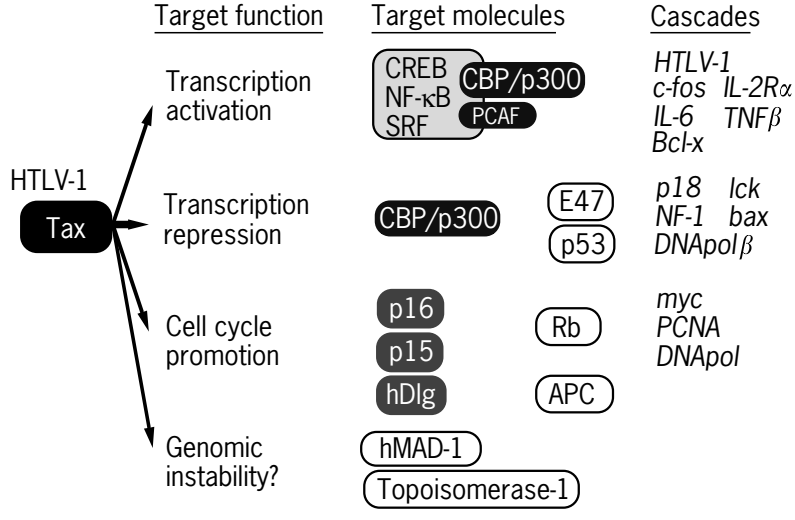


Figure 6 Summary of the pleiotropic function of Tax.

genes which are not directly linked to cell proliferation. These include trans-repression of DNA polymerase-β and Bax, trans-activation of Bcl-X_L and inhibition of topoisomerase I, PCNA and HsMAD. DNA polymerase-β DNA topoisomerase I and PCNA are involved in repair of damaged DNA and therefore reduction of these activities through any mechanism would result in higher mutation rate fixing genetic abnormality in infected cells. In fact, a mutagenic effect of Tax on host cell chromosomes has been directly demonstrated. In relation to DNA damage and its repair, p53 function is also affected by Tax; p53-dependent transcription is trans-repressed through interfering with p53 binding to CBP (see the previous section). Furthermore, p53 function is impaired by phosphorylation at Ser15. Therefore, Tax is able to abolish the cell cycle checkpoint function of p53, thus leading cells to accumulate DNA mutations. This would also be bypassed through Tax binding to HsMAD, a component of the mitotic checkpoint, suggesting promotion of abnormal cell division.

Escape from apoptosis. It is widely accepted that unbalanced activation of some signalling pathway for cell growth or undesired mutation of critical genes would induce apoptosis and eliminate these abnormal cells. Therefore, tumorigenic mutation or viral function would be counteracted were apoptosis induction to be effective. Surprisingly, the Tax protein was shown to trans-activate Bcl-X_L and trans-repress Bax, an inhibitor and mediator of apoptosis, respectively, thus preventing apoptosis of these cells. These effects of Tax seem to be critical for finalizing its effects *in vivo*, otherwise potent activities of Tax would be cancelled by elimination of the Tax-expressing cells by apoptosis.

In total, Tax induces abnormal cell growth, genomic instability and fixation of the abnormality through its pleiotropic functions. Tax is therefore concluded to be a tumour initiator and promoter in the development of ATL.

Low Expression of Tax In Vivo

The pleiotropic functions of Tax thus far identified *in vitro* are all consistent with its implication as an aetiological factor for ATL. However, the extremely low expression of Tax *in vivo* offers a different impression from those expected from its properties *in vitro*. Infected T cells in peripheral blood do not produce significant amounts of Tax mRNA and protein, irrespective of whether they are transformed or not. Expression is detected only by sensitive PCR, but not by any other techniques. Semiquantitative PCR indicated that over 95% of infected cells are absolutely negative for expression of the viral message. Such extremely inefficient expression of the viral genes might be beneficial for the virus to escape from host immune responses, but raises a question concerning the role of Tax in tumorigenesis. It should be emphasized that continuous expression of the viral proteins in infected individuals are suggested by the persistent prevalence of the antibodies.

These somewhat mysterious observations may be explained in several ways: first, Tax plays essential roles in the early stages of transformation, but it is no longer required for maintenance of the transformed phenotype. Second, very low levels of Tax may be sufficient for the maintenance of abnormal phenotypes of infected cells. However, this is unlikely because most infected cells are absolutely negative for Tax expression. Third, Tax is transiently expressed in a small population of infected cells at one time and in another cell population at other times. This would be possible in T cells, some of which are stimulated by antigens or other signals, but its activation would be soon terminated unless the stimulation was continuous. Different specificities of stimulation would induce Tax expression in different populations of T cells. However, it is not easy to distinguish these possibilities experimentally.

The other possibility for low expression of HTLV-1 genes is defective proviruses *in vivo* in cells. This is, however, not the case, since the expression of viral genes is rapidly induced when primary cells from infected individuals are cultured. The mechanism for such restriction of viral expression *in vivo* is not well understood. The extremely low expression of Tax in primary tumour cells is in contrast to transforming genes of DNA tumour viruses, which are significantly expressed and responsible for maintenance of the abnormal growth of transformed cells. For example, *E6/E7* genes of human papillomavirus are expressed in the HeLa cell line maintained for a long period in culture and their repression arrests cell growth.

Clonal Expansion of Infected T cells

Another point in question is the clonal burst of infected T cells. Growth stimulation through the pleiotropic functions of Tax would result in a random population of proliferating cells, since it stimulates growth of most T cells infected with HTLV-1. However, leukaemic cells are always monoclonal. Therefore, an additional genetic event is postulated to trigger the clonal selection of infected T cells. However, not much is known about the mechanism of clonal selection for leukaemic cells. It might not be associated with the viral function. The notion that further alteration is necessary is consistent with the long delay in ATL development after HTLV-1 infection.

Prevention of HTLV-1 Infection

Infection by HTLV-1 is easily detected by antibodies against HTLV-1 proteins. Assay kits for ELISA, Western blotting and particle agglutination have been produced for the diagnosis of HTLV-1 infection and are commercially available. Worldwide screening of HTLV-1 prevalence has been carried out using these systems. Transfusion of sero-positive blood results in transmission of HTLV-1 to two-thirds of the recipients. With the introduction of HTLV-1 screening systems in blood banks, sero-positive blood is now rejected in Japan and viral transmission through transfusion has been greatly reduced.

This has also resulted in an effective reduction of transfusion-related HAM/TSP. The application of these systems to populations in all endemic areas is now clearly shown to prevent HTLV-1 infection.

The major, natural route of viral transmission is from mother to child through infected T cells in breast milk (Hino *et al*., 1985). Curiously, mothers with high levels of antibodies to Tax protein transmit the virus to their offspring at a higher rate than do those with low titres of Tax antibodies. It is possible that efficient replication of HTLV-1 would stimulate antibody production at high levels but that the antibodies might not significantly inhibit viral replication. Non-breast-feeding has been examined among sero-positive mothers in Nagasaki City, Japan, to prevent the viral transmission into children. By consent, pregnant women are surveyed for HTLV-1 antibodies; those who are sero-positive are encouraged to avoid breast feeding. The trial indicated a drastic reduction in the incidence of sero-positive children, from about 30% to just a few per cent. The success of this trial provides direct evidence for viral transmission through milk and suggests the possibility of eliminating ATL in the next few generations.

Epidemiological studies have established that HTLV-1 infects individuals at an early age through breast milk. In fact, within 6–12 months of birth, one-quarter to one-third of children born to sero-positive mothers develop antibodies to HTLV-1. A dramatic increase in age-specific rates of sero-positivity is, however, observed only in those over 40 years old. Sexual transmission of HTLV-1 from husband to wife may account for a portion of this increase, but cannot explain the increase in males. One possible speculation to account for this phenomenon is that the age-dependent prevalence is a reflection of the infection in the young period of 0–1 years old. These epidemiological results suggest that prevention of milk and transfusion-mediated transmission might be adequate.

REFERENCES

Chen, I. S., *et al*. (1983). Human T-cell leukemia virus type II transforms normal human lymphocytes. *Proceedings of the National Academy of Sciences of the USA*, **80**, 7006–7009.

Fujisawa, J., *et al*. (1985). Functional activation of the long terminal repeat of human T-cell leukemia virus type I by a trans-acting factor. *Proceedings of the National Academy of Sciences of the USA*, **82**, 2277–2281.

Hino, S., *et al*. (1985). Mother-to-child transmission of human T-cell leukemia virus type-I. *Japanese Journal of Cancer Research*, **76**, 474–480.

Hinuma, Y., *et al*. (1981). Adult T-cell leukemia: antigen in an ATL cell line and detection of antibodies to the antigen in human sera. *Proceedings of the National Academy of Sciences of the USA*, **78**, 6476–6480.

Jeang, K. T., *et al*. (1990). HTLV-I trans-activator protein, tax, is a trans-repressor of the human beta-polymerase gene. *Science*, **247**, 1082–1084.

Jin, D. Y., *et al*. (1998). Human T cell leukemia virus type 1 oncoprotein Tax targets the human mitotic checkpoint protein MAD1. *Cell*, **93**, 81–91.

Kwok, R. P., *et al*. (1994). Nuclear protein CBP is a coactivator for the transcription factor CREB. *Nature*, **370**, 223–226.

Miyoshi, I., *et al*. (1981). Type C virus particles in a cord T-cell line derived by co-cultivating normal human cord leukocytes and human leukaemic T cells. *Nature*, **294**, 770–771.

Nevins, J. R. (1992). E2F: a link between the Rb tumor suppressor protein and viraloncoproteins. *Science*, **258**, 424–429.

Poeisz, B. J., *et al*. (1980). Detection and isolation of the C retrovirus from fresh and cultured lymphocytes of a patient with T cell lymphoma. *Proceedings of the National Academy of Sciences of the USA*, **77**, 7415–7419.

Seiki, M., *et al*. (1983). Human adult T-cell leukemia virus: complete nucleotide sequence of the provirus genome integrated in leukemia cell DNA. *Proceedings of the National Academy of Sciences of the USA*, **80**, 3618–3622.

Seiki, M., *et al*. (1984). Nonspecific integration of the HTLV provirus genome into adult T-cell leukaemia cells. *Nature*, **309**, 640–642.

Sodroski, J. G., *et al*. (1984). Trans-acting transcriptional activation of the long terminal repeat of human T lymphotropic viruses in infected cells. *Science*, **225**, 381–385.

Suzuki, T., *et al*. (1996). HTLV-1 Tax protein interacts with cyclin-dependent kinase inhibitor p16INK4A and counteracts its inhibitory activity towards CDK4. *EMBO Journal*, **15**, 1607–1614.

Tajima, K. (1990). The 4th nation-wide study of adult T-cell leukemia/lymphoma (ATL) in Japan: estimates of risk of ATL and its geographical and clinical features. The T- and B-cell Malignancy Study Group. *International Journal of Cancer*, **45**, 237–243.

Uchiyama, T., *et al*. (1977). Adult T-cell leukemia: clinical and hematologic features of 16 cases. *Blood*, **50**, 481–492.

Weiss, R., *et al*. (1985). *RNA Tumor Viruses* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY).

Yoshida, M., *et al*. (1982). Isolation and characterization of retrovirus from cell lines of human adult T-cell leukemia and its implication in the disease. *Proceedings of the National Academy of Sciences of the USA*, **79**, 2031–2035.

Yoshida, M. (1995). HTLV-1 oncoprotein Tax deregulates transcription of cellular genes through multiple mechanisms. *Journal of Cancer Research and Clinical Oncology*, **121**, 521–528.

FURTHER READING

Yoshida, M. (2001). Multiple viral strategy of HTLV-1 for dysregulation of cell growth control. *Annual Review of Immunology*, **19**, 475–496.

Chapter 7

Genomic Instability and DNA Repair

Kurt W. Kohn
National Cancer Institute, National Institutes of Health, Bethesda, MD, USA

Vilhelm A. Bohr
National Institute on Aging, National Institutes of Health, Baltimore, MD, USA

CONTENTS

INTRODUCTION

Preservation of genome integrity is critical for the functional preservation of dividing cells. Genomic instability and DNA damage set the stage for carcinogenesis, both in the initiation stage and in the evolution toward malignancy. The frequency estimates for various types of DNA damage normally incurred in mammalian cells are summarized in **Table 1**. Multiple mechanisms have evolved to repair DNA damage and preserve genome integrity. Several mechanisms are conserved widely or even universally among organisms and thus play a fundamental role in maintaining the integrity of living species.

Table 1 Estimated frequencies of DNA lesions normally occurring in mammalian cells

Damage	Events per cell per day
Single-strand breaks	55 000
Depurinations	13 000
Depyrimidinations	650
Guanine-O6 methylation	3100
Cytosine deamination	200
Glucose-6-phosphate adduct	3
Thymine glycol	270
Thymidine glycol	70
Hydroxymethyluracil	620
Guanine-8 oxygenation	180
Interstrand cross-link	8
Double-strand break	9
DNA-protein cross-link	Unknown

Multistage carcinogenesis, the stepwise accumulation of genetic changes favouring malignant behaviour, is brought about jointly by chemical mutagenesis and genomic instability mechanisms. Spontaneous mutations, estimated at perhaps three per cell during a human lifetime, however, are insufficient by themselves to account for cancer incidence (Coleman and Tsongalis, 1999). The probability of accumulating the several (perhaps five or more) genetic changes thought necessary for malignancy is greatly enhanced by the development of genomic instability, which can function as a mutator phenotype. For most malignant tumours, therefore, genomic instability plays an important part in causality (see reviews by Breivik and Gaudernack, 1999).

Genomic instability and/or defective DNA repair are characteristic of several human genetic disorders, including ataxia telangiectasia, Fanconi anaemia, Bloom syndrome, Werner syndrome, xeroderma pigmentosum,

Cockayne syndrome and Nijmegen breakage syndrome. Most of these genetic defects predispose affected individuals to the development of malignant tumours (see the chapter *Inherited Predispositions to Cancer*).

GENOMIC INSTABILITY

Genomic instability implies an abnormally high rate of genomic alterations. Not only do tumours contain genome abnormalities, they also have increased genomic heterogeneity among their cells. Two types of genomic instability syndromes have been established in tumours: microsatellite instability and chromosome instability. Malignant tumours almost always have one or the other, but rarely both. Thus malignancy can develop by one route or the other, but does not need both (see review by Coleman and Tsongalis (1999)).

Both types of genomic instability have been observed early in tumour development while the lesions are still small and benign. Genomic abnormalities and variation increase as tumours progress toward malignancy. Genomic variation progresses by a combination of increased rate of variation (genomic instability) and selection of cell clones adapted to the malignant lifestyle (see the chapter *Cell Proliferation in Carcinogenesis*).

Genomic abnormalities can arise by several mechanisms having different regional characteristics in DNA. Local alterations include sequence amplifications, deletions, insertions and point mutations. Chromosomal translocations and rearrangements arise by breakage/rejoining or other recombination events. Mitotic abnormalities give rise to aneuploidy by unequal chromosome segregation and to multiploidy by failure of nuclear division. These processes lead to genomic heterogeneity, which becomes increasingly rampant during the progress of malignancy; the most autonomously replicating genotypes become selected as tumours progress. Thus genomic instability is implicated in both the origin and progression of most malignant tumours.

Microsatellite Instability

Microsatellite instability implies variation in the length of homopolymer regions (particularly poly(A) sequences) or of dinucleotide or trinucleotide repeat regions. Microsatellite instability is caused by a defect in mismatch repair (MMR). For a compilation of microsatellite instability reports for various tumours and a discussion of mechanism of production, see Coleman and Tsongalis (1999).

Microsatellite instability is discussed further in the section on MMR.

Chromosome Instability

Chromosome instability implies not only abnormal variation in gross chromosome number (aneuploidy), but also an increased rate of chromosomal alterations. Aneuploid tumour cells have been estimated to have a 10–100-fold increased rate of chromosome gain or loss and a similar increase in the rate of loss of heterozygosity at specific genomic sites (Lengauer *et al*., 1998). Sometimes the loss of a chromosome is balanced by the duplication of the remaining allelic chromosome. Loss of heterozygosity of a given genetic region can be detected by DNA electrophoresis and hybridization even when the corresponding chromosome is present in the euploid 2 copies.

For a compilation of frequencies of gain or loss of DNA from each chromosome arm in various tumours (as determined by comparative genomic hybridization), see Rooney *et al*. (1999). Interestingly, some chromosome arms show gains much more often than losses, some show the reverse and many show nearly equal frequencies of gain or loss. For a given tumour type, on the other hand, each chromosome arm may show frequent gain or loss, but rarely both. This may reflect distinctive preferred patterns of selection among different tumour types in the context of chromosome instability.

Although the association of microsatellite instability with defective MMR is well established, the molecular origin of chromosome instability is only beginning to be elucidated. Unlike microsatellite instability, which can be complemented by cell fusion with a chromosome-instability cell type, the converse is not the case: whereas microsatellite instability is a recessive character, chromosome instability is dominant and may be due to a gain-of-function mutation of a single protein (Lengauer *et al*., 1998). Chromosome instability is not a consequence of increased chromosome number *per se*, because tetraploid cells resulting from the fusion of two cells having microsatellite instability does not yield chromosome instability. Sometimes gross chromosome instability is accompanied by p53 mutation, but the early stages of chromosome instability often appear much earlier than p53 mutation. Moreover, mutation or inactivation of p53 does not by itself affect chromosome stability (Lengauer *et al*., 1998).

Centrosome Abnormalities

Abnormal centrosome function in cancer cells can cause chromosome instability. Sometimes the centrosomes become overduplicated, causing multifocal spindles and grossly abnormal mitoses. Centrosome dysfunction however is not the only route to aneuploidy. Improper behaviour of the mitotic spindle can be due to a variety of as yet poorly defined defects in mitotic checkpoints, which can cause unequal chromosome segregation between daughter cells.

The molecular basis of centrosome dysfunction is not yet clear. Although abnormal centrosome function is sometimes associated with p53 loss or mutation, this is not a strict association, because normal centrosome function together

with aneuploidy has been observed in the presence of mutated p53. Thus, p53 mutation does not by itself cause centrosome dysfunction, but may be implicated in other routes to aneuploidy.

THE NEED FOR DNA REPAIR

Living organisms are constantly exposed to stress from environmental agents and from endogenous metabolic processes. An important factor is exposure to oxidative reagents or oxidative stress, largely arising as a side effect of mitochondrial energy metabolism. The resulting reactive oxygen species (ROS) attack proteins, lipids and DNA. Since proteins and lipids are readily degraded and resynthesized, the most significant consequence of oxidative stress is thought to be DNA modifications, which can become permanent via the formation of mutations and other types of genomic damage.

Many different types of DNA base changes have been observed following oxidative stress, and these lesions are widely considered as instigators of cancer, development, ageing and neurological disorders (for review, see Wiseman and Halliwell, 1996). The endogenous attack on DNA by ROS generates a low steady-state level of DNA adducts that have been detected in the DNA of human cells (Dizdaroglu, 1991). Over 100 oxidative base modifications in DNA have been detected in human cells (Wiseman and Halliwell, 1996). The best known and most widely studied is 8-hydroxyguanosine (8-oxoG). Oxidative DNA damage accumulates in cancerous tissues and is thought to contribute to carcinogenesis. For example, higher levels of oxidative base damage were observed in lung cancer tissue compared with surrounding normal tissue and a ninefold increase in 8-oxoG, 8-hydroxyadenine and 2,6-diamino-4-hydroxy-5-formamidopyrimidine in DNA in breast cancer tissue compared with normal tissue has been reported (Wiseman and Halliwell, 1996). DNA damage can also occur after direct attack by external or exogenous sources. Radiation from various sources can directly damage bases in DNA. For example, ultraviolet (UV) irradiation from exposure to sunlight creates certain DNA lesions. The main ones are the cyclobutane pyrimidine dimers formed usually between two adjacent thymine bases in DNA and the pyrimidine–6,4–pyrimidine dimer photoproducts. Irradiation from γ-ray sources or X-rays creates many different kinds of lesions in DNA, including base modifications, sites with a loss of base, and breaks in a DNA strand. DNA breaks can be single- or double-stranded. Many food constituents can directly damage DNA. These include carcinogens or chemicals that react directly with DNA or do so after metabolic modification. Some of these agents alkylate DNA bases, some forming bulky adducts. For example, aromatic amines are found in a variety of foods and are known to cause DNA damage and to be highly mutagenic. A number of poisons attack DNA directly. An example is mustard gas or nitrogen mustard which chemically modifies DNA bases and produces cross-links between bases on the same or on opposite DNA strands (Kohn, 1996). Interstrand cross-links cause havoc in the cell by completely blocking the progress of polymerases. DNA repair pathways have evolved to deal with all these lesions in DNA. Some are listed in **Figure 1**, which shows the general pathways of DNA repair, including those that are thought to exist in mitochondria.

Some DNA lesions can be repaired directly back to the original DNA structure, as in the case of the alkyltransferase

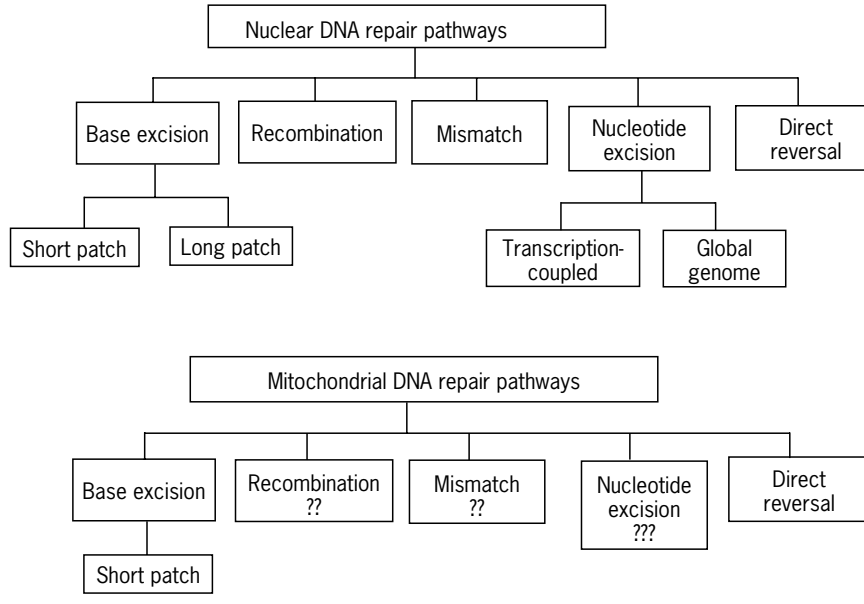


Figure 1 Repair mechanisms for nuclear and mitochondrial DNA in mammalian cells.

and photolyase reactions. Most lesions, however, require more complex repair mechanisms, such as base excision repair, nucleotide excision repair, mismatch repair and recombinational repair. In general, the more bulky DNA base modifications are removed by nucleotide excision repair, while less bulky ones are repaired by base excision repair pathways. These are not firm distinctions, however, since there is much more overlap between these pathways than previously thought. (See the chapters on *The Formation of DNA Adducts* and *Physical Causes of Cancer*.)

REPAIR BY DNA ALKYLTRANSFERASE

Of all DNA repair mechanisms, the simplest and most perfect is mediated by O^6-alkylguanine-DNA alkyltransferase. It is also the most specialized, applying only to a particular type of chemical damage to DNA bases, namely chemical adducts at the guanine-O6 or cytosine-O4 positions. These types of adducts are produced by chemical carcinogens such as nitrosamines and by chemotherapeutic agents such as nitrosoureas (see the chapter *The Formation of DNA Adducts*). Adducts at the guanine-O6 or cytosine-O4 positions alter the base-pairing preferences and therefore are highly mutagenic. Cells normally contain an alkyltransferase, which efficiently removes alkylations from these positions and restores the original chemistry of the base. Thus the repair is perfect in that the DNA is brought back precisely to its original state.

The alkyltransferase that accomplishes this feat is an unusual enzyme, because it is, in part, its own substrate: it transfers an adduct from a DNA base to a sulfhydryl group at the enzyme's active site. The alkylated sulfhydryl group on the enzyme is so stable that it cannot be removed to regenerate the active enzyme molecule: each enzyme molecule can act only once, whereupon it becomes permanently inactivated. The reaction therefore is stoichiometric rather than catalytic, and fails to meet a classical criterion for an enzyme. Nevertheless, in terms of structure and mechanism, this is clearly an enzyme in that it lowers the energy barrier of a reaction. Its unusual feature is that one of the stable products of the reaction happens to be part of the enzyme protein itself. This same feature, however, makes the repair very fast and efficient. It is one of the few DNA repair processes that can be accomplished in essentially a single step. However, the number of DNA adducts that can be removed is limited by the number of active enzyme molecules present in the cell.

REPAIR OF SINGLE-BASE DAMAGE

Base excision repair (BER) is thought to be the major way in which the cell deals with most types of damage to single bases in DNA, although nucleotide excision repair (NER) (see later) may also play a part. An exceptional case, however, is presented by adducts at the O6 position of guanine; as described above, these adducts can be removed by an alkyltransferase while leaving the normal guanine base in place. The alkyltransferase-mediated repair of guanine-O6 adducts is unusual and remarkable in that it is error free. On the other hand, adducts at the guanine-N7 position, which is the most common alkylation site on DNA, lead to loss of the guanine base. This can occur either through the action of a repair glycosylase or by spontaneous hydrolysis of the glycosidic bond (alkylation at N7 facilitates spontaneous release of guanine from DNA). Either way, the loss of the base leaves the DNA with an unsubstituted deoxyribose unit, known as a base-free site or AP site (AP stands for apurinic/apyrimidinic). A particularly important BER mechanism removes uracil residues that normally arise spontaneously due to the occasional hydrolysis of the 4-amino group of cytosine. This is carried out by a uracil glycosylase.

Hydrolysis of 5-methylcytosines yields a normal base, thymine, which would not be recognized as abnormal by a repair glycosylase. 5-Methylcytosine in mammalian DNA occurs only at 5′-CpG-3′ dinucleotides, which consequently are strongly disfavoured in most regions of the genome. Some DNA regions, usually outside of coding sequences, however contain islands rich in CpG sequences which, when methylated, are subject to GC → AT transitions.

REPAIR OF BASE-FREE SITES

Two independent mechanisms exist for repair of base-free (or 'abasic') sites: single-nucleotide gap-filling and long-patch repair. Either process may be preceded by the action of a glycosylase which cleaves the glycosidic bond between the base and sugar moieties of DNA. An AP-endonuclease then cuts the DNA strand containing the base-free site immediately on the 5′ side of the lesion, and yields a 5′-sugar-phosphate terminus and a 3′-OH terminus. A repair polymerase (typically Pol β) then extends the 3′ end and displaces the base-free sugar residue. In the long-patch mechanism, the displacement of the damaged strand extends to include between 2 and about 10 nucleotide residues. The displaced DNA segment (sometimes referred to as a 'flap') is removed by a flap endonuclease (FEN1). The DNA strand can then be made whole by the action a DNA ligase. Long-patch repair may be carried out via Pol β, which does not require proliferating cell nuclear antigen (PCNA), or via Pol δ which is PCNA dependent (discussed by Prasad *et al*., 2000). These authors recently found in a reconstituted system of long-patch BER that FEN1 and Polβ can cooperate in the linked processes of strand displacement (a Pol β function) and displaced-strand cleavage (a FEN1 function). The two enzymes mutually stimulated each other in this system.

A PCNA-dependent reconstituted long-patch BER system was found to require replication protein A (RPA) for optimum activity; PCNA and RPA seemed to function coordinately (Dianov *et al*., 1999).

It is not yet clear how the two base excision repair pathways (involving short- or long-patch repair of the consequent base-free sites) are regulated. It may depend on the type of glycosylase involved or on the type of DNA polymerase. The BER complex situated at the AP site in DNA involves DNA polymerase, FEN1, AP-endonuclease and PCNA, and the organisation of this complex is very important. In addition to the short- and long-patch BER pathways, there appears to be another BER pathway of transcription coupled repair which deals with repair of active genes. As mentioned above, there is much emerging evidence for overlap and interaction between the repair pathways. For example, the xeroderma group G (XPG) protein participates in both BER and NER.

NUCLEOTIDE EXCISION REPAIR (NER)

NER is the most versatile of the DNA repair mechanisms. It repairs a variety of bulky adducts that distort the DNA helix, but only if both chemical damage and helix distortion are present. One of the most important functions of NER is to repair photoproducts due to sunlight UV exposure. One of these products, pyrimidine–6,4–pyrimidine dimers, (representing about 20% of photoproducts), causes large DNA helix distortions and is efficiently repaired by NER. Another photoproduct, cyclobutane–pyrimidine dimers, is more abundant (about 80% of the total), but causes less helix distortion and therefore is less efficiently repaired. The repair of the latter type of lesion may be aided specifically by the XPE protein, but the exact mechanism is not clear. NER also repairs bulky carcinogen adducts, such as DNA adducts of polycyclic aromatic hydrocarbons, and DNA cross-links produced by anticancer drugs such as cisplatin. In addition, NER can function (albeit inefficiently) as a backup for base damage that evades BER or the alkyltransferase repair mechanism (see reviews by Balajee and Bohr (2000), Batty and Wood (2000) and de Boer and Hoeijmakers (2000)).

The proteins associated with the seven complementation groups of xeroderma pigmentosum (XP) (XPA to XPG) play parts in the NER mechanism. XP is a highly cancer-prone disease in which the patients suffer from high incidence of skin and internal cancers, due to defective DNA repair. The NER process is depicted in an interaction diagram in **Figure 2**. The diagram shows how the proteins involved in NER assemble at the site of a DNA lesion and cooperate to excise a DNA single-strand segment containing the lesion.

The global form of NER operates throughout the genome. A modified form of NER, known as transcription-coupled repair (TCR), however, operates preferentially at sites of transcription. TCR differs from global NER in the mechanism by which the DNA helix at the lesion site is opened to permit access to the repair machinery. In global NER, the lesion is recognized by the XPC:HR23B heterodimer which then recruits the transcription factor IIH (TFIIH) complex. TFIIH contains two DNA helicases (XPB and XPD) that unwind the DNA locally in opposite directions from the lesion site. At about the same time, XPA comes into play and serves as a nucleus for the assembly of other repair components at the lesion site. XPA may also participate in the recognition of lesions. The DNA unwinding may be assisted by the single-strand binding protein RPA. In addition, RPA can bind several key proteins, including XPA, and thus may assist in assembling the repair complex.

In the case of TCR, a DNA lesion is detected in the course of transcription by RNA polymerase II and its associated proteins. The polymerase stalls at the site of the lesion, which somehow leads to the assembly of the repair complex. Although the details of how this happens are not clear, the polymerase presumably dissociates and is replaced by TFIIH and XPA; it is thought that the Cockayne syndrome group B (CSB) protein plays a part in this process.

TFIIH has at least two modes of action. First, it functions in transcription initiation through the action of its associated helicases, as well as through the action of cyclinH:cdk7 (another component of TFIIH) which phosphorylates the C-terminal tail of RNA polymerase II and thereby allows transcription to start. (Another function of cyclinH:cdk7 is to phosphorylate cyclin-dependent kinases in the course of cell cycle regulation (Kohn, 1999).) Second, the helicases XPB and XPD (which are also part of the TFIIH complex) play an important part in NER and TCR (the helicases bind more tightly in the TFIIH complex than does cyclinH:cdk7, which may not be present in the repair form of TFIIH).

The subsequent steps of NER or TCR, leading to the excision of an oligonucleotide containing the lesion, are summarized in **Figure 2** and its caption. After excision, the resulting gap in the damaged DNA strand must be filled. This is accomplished by DNA polymerse δ or ε, followed by DNA ligase I. The assembly of the polymerase at the 3′ terminus of the stand gap first requires PCNA, which clamps on to the DNA and RFC, which loads PCNA on to the 3′ terminus.

When an NER complex has assembled at a DNA lesion, the complex may send a signal indicating the presence of DNA damage. A linchpin in this communication is p53. The ability of p53 to bind RPA could be part of this communication link.

Transcription-coupled NER

As already mentioned, NER has two pathways: (1) transcription-coupled repair (TCR) and (2) global genome repair (GGR). The TCR pathway repairs lesions in the

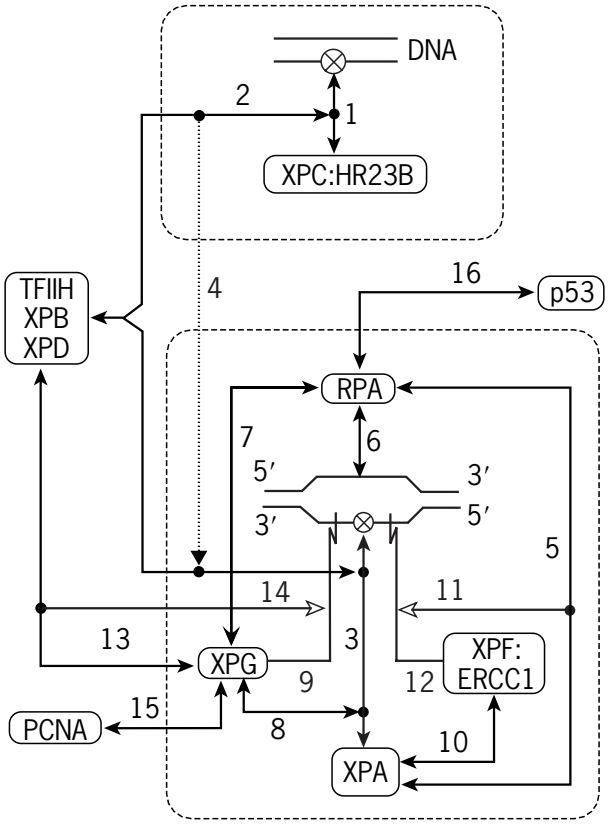


Figure 2 Molecular interactions in nucleotide excision repair (NER). The symbols used in this and subsequent diagrams are defined in (Kohn, 1999) (or see http://discover.nci.nih.gov/kohnk/interaction_maps.html). The upper dashed box depicts a normal double-stranded DNA helix bearning a lesion which is recognized and bound by a heterodimer consisting of XPC and HR23B. A DNA segment surrounding the lesion is then unwound by the XPB and XPD helicases (components of the TFIIH transcription factor complex). The lower dashed box depicts the region of unwound helix and the excision of a DNA segment containing the lesion. The transition from the closed to the unwound state of the DNA (with its associated proteins) is indicated by the hatched arrow with the solid triangle arrowhead. During this transition, the XPC:HR23B complex is replaced by XPA. XPA serves to assemble several proteins that participate in the excision of the lesion. In the case of transcription-coupled repair (TCR), the DNA helix around the lesion is already unwound due to the transcription process, and therefore repair can begin with XPA (lower box) and does not require XPC or HR23B. The numbered steps are as follows. (1) A lesion in one strand of an intact DNA helix becomes bound by a dimer consisting of the XPC and HR23B proteins. (2) This dimer (XPC:HR23B) binds the TFIIH complex which contains the DNA helicases XPB and XPD (Yokoi *et al.*, 2000). All of these proteins are needed for the initial opening of the DNA helix at the site of the lesion. XPB and XPD function with opposite polarity: (3′ → 5′ and 5′ → 3′, respectively) to unwind the DNA for a short distance on both sides of the lesion (no diagram symbol is available for unwinding). (3) XPA is then recruited to the lesion (however, in the case of TCR, the helix is already open owing to the presence of RNA polymerase II, and XPA can bind to the lesion without the aid of XPC:HR23B). (4) The transition arrow (solid triangle arrowhead, shown here hatched) indicates that the DNA helix opens and XPC:HR23B is replaced by XPA, and the TFIIH complex now is bound to XPA instead of to XPC:HR23B. (5) XPA binds the DNA single-strand-binding protein RPA. (6) RPA binds the undamaged strand where the helix has been opened. Thus RPA helps to stabilize the XPA complex at the site of the lesion (the length of the unwound region in NER is similar to the 30 nucleotides required for optimum binding of RPA to DNA single strand). (7) RPA recruits endonuclease XPG. (8) XPG binds XPA while XPA is bound to the lesion (XPG may be required for XPA to replace XPC:HR23B at the site of the lesion; this is not shown in the diagram). (9) XPG incises the lesion-containing strand on the 3′ side of the lesion (approximately 6–14 nucleotide residues away from the lesion). (10) XPA recruits the XPF:ERCC1 heterodimer to the lesion site. (11) RPA interacts with XPF and directs the endonuclease activity of XPF to the 5′ side of the lesion. (12) XPF incises the lesion-containing strand on the 5′ side of the lesion (approximately 16–25 nucleotides away from the lesion) (a single-strand segment containing the lesion is thereby released whose modal length is 27 (24–32) nucleotides, independent of the type of lesion. The sites of incision by XPF and XPG may be at or close to the transitions between unwound and helical DNA, since these two enzymes are structure specific endonucleases). (13) TFIIH interacts strongly with XPG. (14) TFIIH may help to position XPG on the 3′ side of the lesion. (15) XPG recruits PCNA which is required for the subsequent DNA repair synthesis that fills the gap left by the excised strand segment. (16) RPA can bind p53 and thereby perhaps serve to signal the presence of DNA damage. For references, see reviews by Balajee and Bohr (2000); Batty and Wood (2000); and de Boer and Hoeijmakers (2000) and the annotation list for DNA repair in Kohn (1999).

transcribed strand of transcriptionally active genes and is dependent on RNA polymerase II (RNA pol II) (Balajee and Bohr, 2000). The GGR pathway removes lesions from genes regardless of whether they are transcriptionally active or inactive.

Cells of a rare genetic disease, Cockayne syndrome (CS), are defective in TCR, but proficient in GGR of UV-induced DNA damage. Affected individuals suffer from postnatal growth failure resulting in cachectic dwarfism, photosensitivity, skeletal abnormalities, mental retardation, progressive neurological degeneration, retinopathy, cataracts and sensorineural hearing loss. Two complementation groups, CS-A and CS-B, have been identified and the corresponding genes have been cloned. The cellular phenotype of CS includes increased sensitivity to a number of DNA-damaging agents including UV radiation, ionizing radiation and hydrogen peroxide (Friedberg, 1996).

A characteristic feature of CS cells is that they do not recover the ability to synthesize RNA efficiently after UV damage; this phenotype is consistent with a defect in TCR. The *CSA* and *CSB* genes have been cloned and their products characterized. The *CSA* gene product is a 44-kDa protein that belongs to the 'WD repeat' family. Members of this protein family are structural and regulatory proteins,

but usually lack enzymatic activity. The *CSB* gene product is a 168-kDa protein that belongs to the SWI/SNF family, which are DNA and RNA helicases with seven conserved sequence motifs. CSB has an acidic amino acid stretch, a glycine-rich region and two putative NLS sequences. CSB is a DNA-stimulated ATPase, but is not able to unwind DNA in a conventional strand displacement assay (Selby and Sancar, 1997).

The precise molecular role of CSB is not clear at present. CSB may facilitate repair of active genes by recruiting DNA repair proteins to actively transcribed regions. *In vitro,* CSB forms a complex with RNA polymerase II, DNA and the RNA transcript in a manner that requires ATP hydrolysis (Tantin *et al*., 1997). This quaternary complex recruits another molecular complex including the TFIIH core subunits p62 and XPB. TFIIH is a complex factor thought to promote local DNA unwinding during transcription initiation by RNA pol II and promoter escape, as well as in NER (Balajee and Bohr, 2000).

It is also possible that CSB indirectly stimulates TCR by facilitating transcription. Members of the SWI/SNF family are involved in regulating transcription, chromatin remodelling and DNA repair, including such actions as disruption of protein–protein and protein–DNA interactions. The *CSB* gene product could have a similar function. In fact, it is still a matter of debate whether CS is due to a primary defect in transcription or DNA repair (Friedberg, 1996). Some evidence suggests that CSB may indirectly stimulate TCR by facilitating the process of transcription (Balajee *et al*., 1997; Selby and Sancar, 1997). Thus, CSB may be a transcription elongation factor and a repair-coupling factor acting at the site of RNA pol II-blocking lesions, and the CS phenotype may arise from deficiencies in both transcription and DNA repair. The biological function of CSB in these different pathways may be mediated by distinct functional domains of the protein.

It is well established that the *CSB* phenotype involves a defect in TCR of UV-induced DNA damage, although

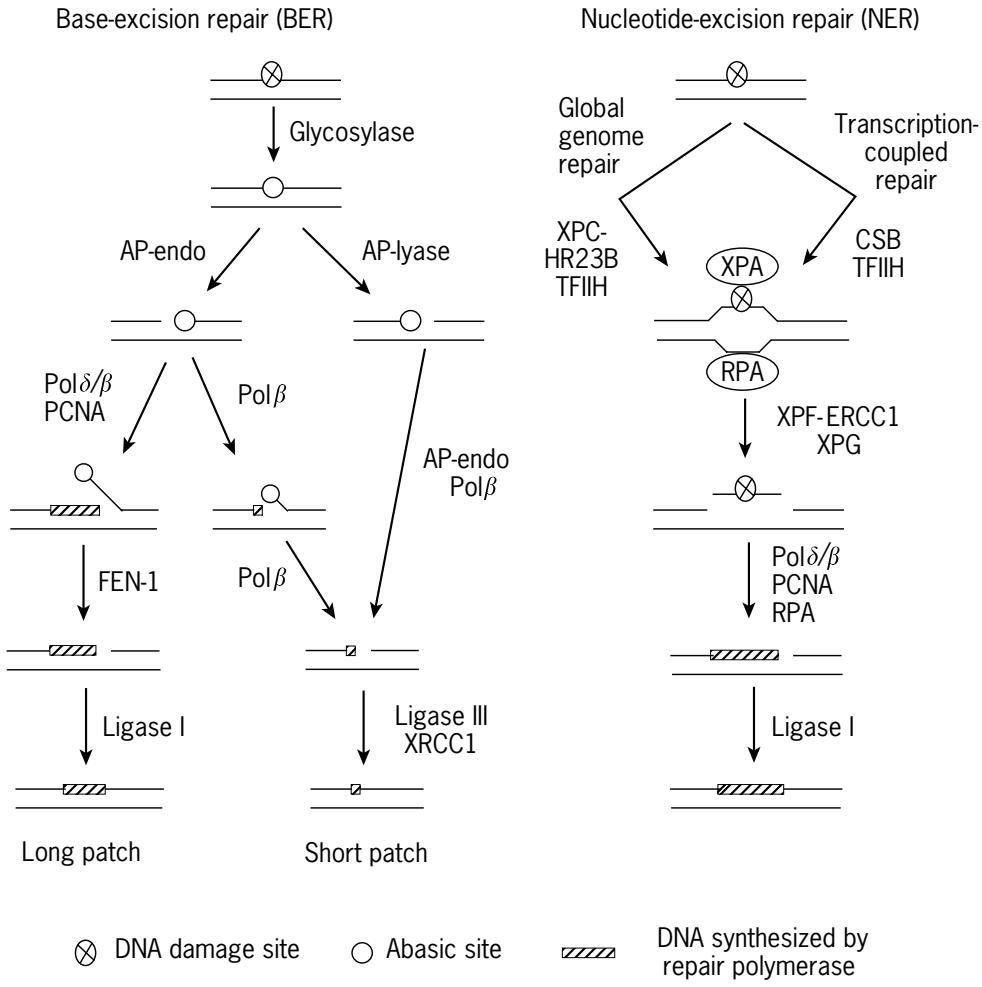


Figure 3 The pathways of base excision repair (BER) and nucleotide excision repair (NER). Some of the proteins involved are shown. The general steps are recognition, incision, replication and ligation. NER has a subpathway called transcription-coupled repair (TCR) and within BER there are two pathways, long-patch and short-patch BER. The individual pathways are described in the text.

CSB may also function in TCR of oxidative damage (Le Page *et al*., 2000).

The various types of BER and NER are summarized in **Figure 3**.

MITOCHONDRIAL DNA REPAIR IN MAMMALIAN CELLS

Oxidative phosphorylation is an essential metabolic pathway that takes place in the mitochondrion and produces reactive oxygen species (ROS), such as superoxide, hydrogen peroxide, and hydroxyl radicals. Mitochondrial DNA (mtDNA), because of its close proximity to the electron transport chain, is at risk of damage from the ROS produced by oxidative phosphorylation. The most common oxidative DNA lesions in mtDNA are uracil, 8-oxoG and thymine glycol. If unrepaired, uracil can cause GC–AT transition mutations. Thus, it is predicted that insufficient DNA repair capacity in the mitochondrion could lead to mitochondrial dysfunction and degenerative disease including altered energy balance and other pathophysiological states associated with ageing and cancer.

For many years it was thought that mitochondria had no DNA repair capacity, and that this could explain the accumulation of DNA lesions found here with ageing. Recently, however, it has become evident that there is, indeed, efficient DNA repair in these organelles, but that they do not appear to have the same variety of repair pathways that are found in the nuclear DNA. The known or suspected DNA repair mechanisms present in mitochondria are indicated in **Figure 1**. The many question marks at repair pathways indicate lack of concise knowledge; there is renewed interest in the exploration of mitochondrial DNA repair.

There have been reports of recombinational repair and mismatch repair pathways in mitochondria and the general repair pathways in mtDNA were recently reviewed (Croteau *et al*., 1999). In mtDNA, UV-induced lesions, but not oxidative lesions, are repaired. 8-Oxoguanine is efficiently removed from mtDNA, as it is from actively transcribed genes in the nucleus, and the repair efficiency is similar in all regions of the mitochondrial genome (Croteau *et al*., 1999).

The mitochondria contain enzymes that participate specifically in BER in the mitochondrion. An early indication of the existence of mitochondrial BER (mtBER) was the isolation of a mammalian mitochondrial endonuclease that recognizes and cleaves AP sites. Later, an *in vitro* reconstituted repair assay was performed using mitochondrial enzymes from *Xenopus laevis* and an abasic site-containing DNA substrate (Pinz and Bogenhagen, 1998). A DNA ligase was also purified from mitochondria that may be related to nuclear DNA ligase III. Recently, it was confirmed that the gene encoding human DNA ligase III produces two forms of the ligase, one nuclear and one mitochondrial (Lakshmipathy and Campbell, 1999). The mitochondrial DNA Pol β possesses a 5′-deoxyribose phosphate lyase activity via β-elimination, suggesting that Pol β may play a role in mitochondrial BER (Longley *et al*., 1998). PCNA has also been shown to stimulate Pol β-mediated DNA synthesis, suggesting that PCNA may be an auxiliary factor in mitochondrial mtDNA replication and repair.

DNA MISMATCH REPAIR (MMR)

MMR deals with at least two types of replication errors: (1) single base-pair mismatches and (2) insertion or deletion loops that arise by slippage between the template and replicating strands. Slippage tends to occur in sequence-repeat regions, such as segments of poly(A) or of dinucleotide repeats such as $(AC)_n$, where base pairing near the replication point can easily shift position. Repeated sequences sometimes occur incidentally in coding regions where slippage could cause frame-shift mutations.

A marker for defective repair of insertion/deletion loops is microsatellite instability. Since microsatellites generally occur outside of genes, abnormal length variation of these sequence repeats does not usually cause mutation. Microsatellite instability, however, serves as a sensitive indicator of the harder to detect changes brought about by the same mechanism within genes.

The significance of the MMR mutator phenotype in tumours with microsatellite instability is shown by the finding of frame-shift mutations in numerous tumour-suppressor genes (*APC*, *TGFβ-RII*, *IGF-IIR*, *BAX*, *BRCA1*, *BRCA2*) and some DNA repair genes (*hMSH3*, *hMSH6*, *BLM*) (Buermeyer *et al*., 1999). These frame shifts usually occur within mononucleotide tracts.

In human cells, MMR is carried out by a choreography of multiprotein complexes made up by hMSH2, hMSH3, hMSH6, hMLH1, hPMS2 and probably hMLH3 (**Figure 4**). Also involved are excision and replication proteins: Pol δ, PCNA, RFC, EXO1 and FEN1, which degrade and replace the error-containing segments of newly synthesized strand. The system must also include a DNA ligase which remains to be identified. How the system distinguishes the newly synthesized strand is unknown, although it may well be through recognition of a strand terminus.

Single base-pair mismatches and insertion/deletion loops are recognized by hMSH2:hMSH6 heterodimers (also known as hMutSα complex) which serve as the starting point for the assembly of a ‘repairosome’ typically containing hMLH1, hPMS2 and the excision and replication machinery proteins. An alternative route is through hMSH2:hMSH3 heterodimers (hMutSβ) which recruits a repairosome via hMLH1 and hMLH3; this complex can repair insertion/deletion loops, but not single base-pair mismatches. Loss of *hMSH2* or *hMLH1* thus abrogates the repair of both types of defects. As expected, cells defective

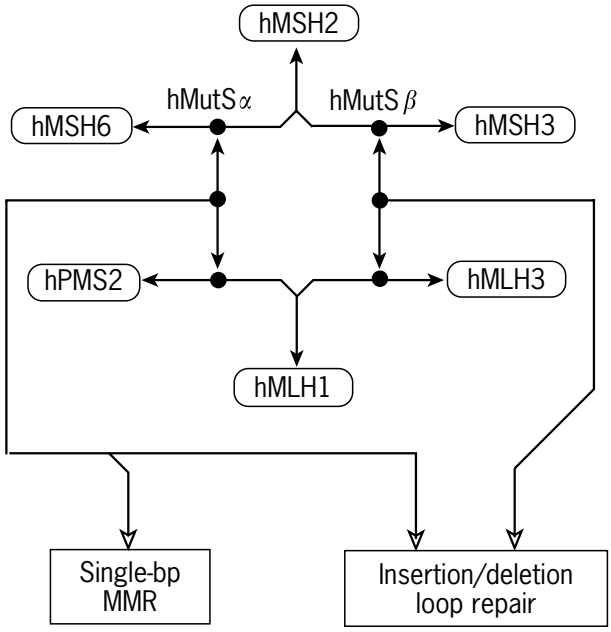


Figure 4 Molecular interactions and functional capabilities of proteins implicated in DNA mismatch repair (MMR). The hMutSα complex, consisting of hMSH2 and hMSH6, combines with hMLH1 and hPMS2 to form the nucleus for a repairosome complex that repairs both insertion/deletion loops and single base-pair mismatches. The hMutSβ complex, consisting of hMSH2 and hMSH3, combines with hMLH1 and hMLH3 and leads to a repairosome complex that can repair insertion/deletion loops, but not single base-pair mismatches.

in *hMSH6* display a mutator phenotype but exhibit little or no microsatellite instability (Jiricny and Nystrom-Lahti, 2000)(because insertion/deletion loops can still be repaired via the hMutSβ path).

It has been proposed that the hMSH2:hMSH6 complex functions akin to a molecular switch, owing to its ADP/ATP-binding ability and ATPase activity (Fishel, 1998). In its ADP-bound state, hMSH2:hMSH6 binds strongly to DNA mismatches. This binding then facilitates the exchange of ADP for ATP (which otherwise occurs only very slowly). In the ATP-bound state, hMSH2:hMSH6 can dissociate from DNA lesions. Its ATPase activity then recycles hMSH2:hMSH6 to the active ADP state. At some point in the cycle, the components of the repairosome are assembled.

MMR, Apoptosis and Chemotherapy Resistance

In addition to its role in DNA repair, the MMR system seems to signal the presence of DNA damage to the apoptosis-initiating system, which may be why MMR-defective tumour cells tend to have increased resistance to DNA-damaging drugs such as cisplatin (see reviews by Li (1999) and Jiricny and Nystrom-Lahti (2000)). Treatment of a mixture of MMR-proficient and deficient cells with cisplatin resulted in enrichment of the MMR-deficient population (Fink *et al*., 1998). Moreover, cisplatin resistance in ovarian cancer recently was reported to be linked to suppression of *hMLH1* due to hypermethylation in the gene's promoter region (Strathdee *et al*., 1999).

MMR-deficient cells also resist killing by alkylating agents that methylate DNA guanine-O6 positions. Such alkylations are mutagenic, because these alkylated guanines base-pair preferentially with T. MMR-deficient cells are alkylation tolerant: they retain the alkylations, but are not killed by them. The cost of survival, however, is mutagenesis. Treatment of *MSH2*-knockout mice with agents that methylate DNA guanine-O6 positions failed to induce apoptosis in the small intestine (a prominent response in wild-type animals) (Toft *et al*., 1999). This *MSH2*-dependent apoptosis was partially mediated by a p53-dependent pathway.

The apoptosis resistance also carries over to other DNA damaging agents including 6-thioguanine (which becomes incorporated into DNA as a bogus base), cisplatin and topoisomerase blockers. Resistance to these agents is conferred by loss of *MSH2, MSH6, MLH1* or *PMS2* functions (but not by loss of *MSH3* function) in several mammalian systems. In addition to loss of apoptotic response, the resistant cells do not exhibit the usual G2/M cell cycle arrest. The components of the MMR system thus appear to have an essential role in the transmission of DNA damage signals (see reviews by Buermeyer *et al*. (1999) and Li (1999)).

The role of MMR in apoptosis signalling may have relevance for chemotherapy with DNA-damaging agents, because drug resistance may develop by loss of MMR function in a single selection step (Aebi *et al*., 1996). Loss of MMR may also confer resistance to low doses of ionizing radiation (Fritzell *et al*., 1997; DeWeese *et al*., 1998) (see review by Li (1999)).

The route by which signals from the MMR system induce apoptosis remains to be elucidated; it may in part involve phosphorylation of p53 and/or the related p73 family proteins (Duckett *et al*., 1999; Li, 1999). The function of p73 in the induction of apoptosis in cisplatin-treated cells may be regulated by tyrosine kinase c-Abl (Gong *et al*., 1999). Since MMR is targeted exclusively to newly synthesized DNA strands (or to strand regions containing nearby strand breaks (Duckett *et al*., 1999)), base damage in the template strand could not be removed: the MMR system could sense the mismatch caused by the base damage, but would attempt to repair the wrong strand. This futile repair cycle is one model proposed as the initiator of the apoptosis signal. Alternatively, the MMR recognition complex might assemble at damage-induced mismatches near replication forks, block replication and thereby induce apoptosis (Li, 1999).

Thus the MMR system corrects DNA mismatches caused by base damage in newly synthesized DNA strands (or in

strands near break sites). However, when presented with damage that it cannot repair, the system sends out an apoptosis-inducing signal. Loss of components of the MMR system allows cells to survive and proliferate while retaining an accumulation of DNA damage. Treatment of MMR-defective tumours with drugs that alkylate DNA at guanine-O6 positions may therefore be ineffective or even detrimental (Li, 1999).

MMR and Colon Cancer

For reasons unknown, MMR defects are associated mainly with cancer of the colon (predominantly right colon), endometrium and ovary. MMR is most closely associated with HNPCC, the most common cancer predisposition syndrome; 70% of HNPCC kindred have germ-line mutations in one of the MMR-associated genes. About 60% of the mutations are in *hMLH1* and about 35% in *hMSH2* (Jiricny and Nystrom-Lahti, 2000; http://www.nfdht.nl). Tumours of HNPCC kindred with mutations in *hMSH2* or *hMLH1* have strong mutator phenotypes and high microsatellite instability (Buermeyer *et al*., 1999). One copy of the gene is mutated in the germline of HNPCC patients, and both copies are mutated in their tumours (which do not exhibit loss of heterozygosity). Loss of MMR requires inactivation of both copies of one of the critical genes, and occurs with increased probability if one of the copies is already mutated in the germ line. In contrast to other cancers, which usually have rampant aneuploidy and loss of chromosome arms, HNPCC tumours have few allelic losses and often have a diploid karyotype (Rosen, 1997). Although chromosome instability and aneuploidy are early events in the development of most other cancers, the initial tumorigenic event in HNPCC is loss of mismatch repair, which induces more localized genome alterations.

Microsatellite instability is also present in 15% of colon cancer patients who have no family history of colon cancer. As in HNPCC, these sporadic tumours occur predominantly on the right side of the colon. In most of these cases, MMR is inactivated, not by mutations, but by transcriptional silencing of the *hMLH1* gene. The *hMLH1* gene is subject to silencing by hypermethylation of its promoter, which is a primary factor in sporadic gastrointestinal tumours having microsatellite instability. In cell lines derived from such tumours, this methylation can be reversed by treating the cells with 5-azacytidine, which eventually restores hMLH1 expression. In most of the tumours, both copies of *hMLH1* are distinguishably suppressed by hypermethylation, although occasionally one of the alleles is inactivated by mutation. For further details and references, see Buermeyer *et al*. (1999) and Markowitz (2000).

MMR and TGFβ Receptor

TGFβ-RII, a receptor in the TGFβ tumour-suppressor pathway, has a special relationship with MMR-deficient colon and stomach cancers. This relationship is due to two circumstances. First, the coding sequence of the human *TGFβ-RII* gene contains a homopolymer tract of 10 adenines that is subject to frame-shift mutation in MMR-deficient cells. This causes premature transcript termination with loss of most of the cytoplasmic domain of the receptor. Among colon tumours, these *TGFβ-RII* frame shifts are found exclusively and almost universally in those tumours that exhibit microsatellite instability. Second, TGFβ functions in intestinal crypts to cause cell cycle arrest and apoptosis when these cells reach the luminal region of the crypts, which is where TGFβ is concentrated. If signalling through the TGFβ pathway is abrogated, crypt cells can continue to proliferate and can initiate the sequence of changes that eventually leads to malignancy. By contrast, MMR-deficient endometrial cancers have a much lower frequency of *TGFβ-RII* frame-shift mutations than do the MMR-deficient gastrointestinal cancers (reviewed by Markowitz (2000)).

The significance of the TGFβ pathway in colon cancer is confirmed by the occurrence, in 15% of microsatellite-stable colon cancers, of inactivating mutations of the *TGFβ-RII* gene outside the poly(A) tracts. Some microsatellite-stable colon cancers with normal *TGFβ-RII* genes bear mutations in downstream components of the TGFβ pathway, such as Smad2 and Smad4. For a recent review, see Markowitz (2000).

A molecular interaction diagram of the essentials of the TGFβ pathway leading to cell proliferation is shown in **Figure 5** and explained in the caption.

Frame-shift mutation in the poly(A) tract of the *TGFβ-RII* gene is often the initiating lesion in microsatellite instability-high non-familial colon cancers. This was shown by a tight correlation between the adenoma-to-carcinoma transition and poly(A) tract length alteration in retrospective tissue pathology samples (Grady *et al*., 1998). A more recent study suggested that the initial gene alteration could be in either the *TGFβ-RII* or the *BAX* gene, or in both (Calin *et al*., 2000). Since BAX is a proapoptotic factor, mutation of this gene may contribute to the apoptosis resistance of some microsatellite instability-high tumours. Nevertheless, TGFβ itself has a proapoptotic effect (pathway unknown) on normal intestinal epithelial cells in culture (references cited by Grady *et al*. (1998)). (See chapter *Signalling by TGF-β*.)

RECOMBINATIONAL REPAIR: REPAIR OF DNA DOUBLE-STRAND BREAKS (DSB) AND CROSS-LINKS

DSB constitute a common type of DNA damage, produced by ionizing radiation, replication blocks and certain DNA-reactive drugs such as bleomycin. The formation and repair of DSB are also part of the immune system's

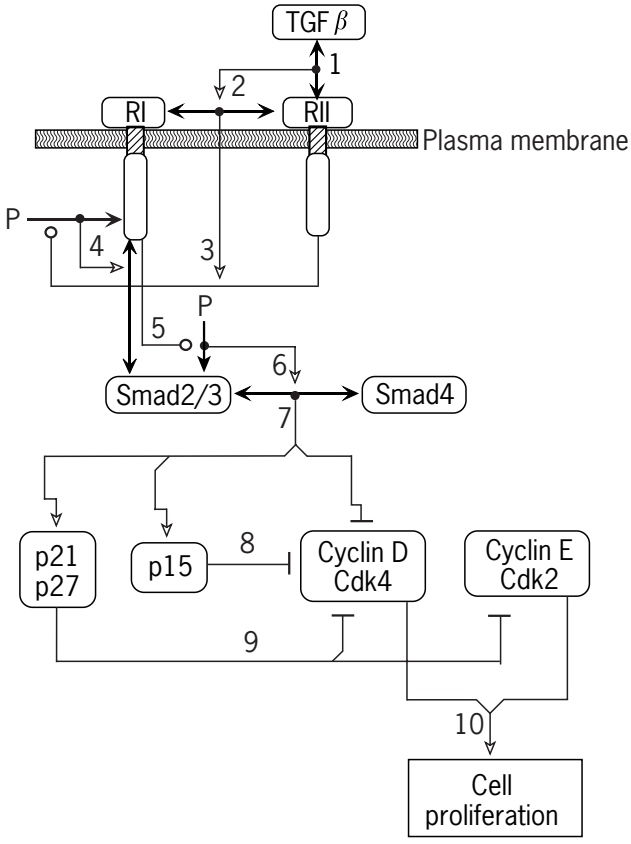


Figure 5 Molecular interaction diagram of the essentials of TGFβ signalling to cell proliferation. (1) TGFβ binds to the type II receptor, TGFβ-RII. (2) This stimulates heterodimer formation between TGFβ-RII and TGFβ-RI. (3) The RII subunit can then phosphorylate the cytoplasmic domain of RI. (4) A phosphorylated site on RI binds Smad2, thereby recruiting this protein to the plasma membrane. (5) This permits the kinase domain of RI to phosphorylate Smad2. (6) Phosphorylated Smad2 binds Smad4. (7) The Smad2:Smad4 heterodimer translocates to the nucleus, where it stimulates the expression of G1/S phase inhibitors p15, p21 and p27, and inhibits the G1/S phase stimulators cyclin D and cdk4 (Massague, 1998; Massague and Wotton, 2000). (8) p15 inhibits Cdk4. (9) p21 and p27 inhibit both Cdk4 and Cdk2 kinase activities. (10) Both cyclin D- and cyclin E-dependent kinases are required for entry of cells into S phase. (For further details, see the chapter *Signalling by TGF-β*.)

V(D)J recombination process. Defects in DSB repair can cause translocations and other DNA rearrangements (Flores-Rozas and Kolodner, 2000). There are two types of recombinational repair of DSB, differing in whether or not the DNA ends to be joined require extensive sequence homology.

Non-homologous End Joining (NHEJ): Repair of DNA Double-strand Breaks

The NHEJ mechanism repairs DNA double-strand breaks without the need for extensive sequence homology between the DNA ends to be joined, although a few complementary base pairs are needed to provide cohesive ends. NHEJ is an error-prone repair process, because it usually creates small deletions. NHEJ is responsible for the rejoining of DSB during V(D)J recombination in the processing of immunoglobulin genes. Defects in NHEJ in mice cause severe combined immune deficiency and radiation sensitivity.

Small deletions are a by-product of NHEJ, because of the need for cohesive DNA ends ('microhomology'), generated by resecting a few nucleotides from one of the DNA strands. The required microhomology may occasionally be as short as a single base pair, although two or three is more efficient. This is in marked contrast to the extensive homology needed by the 'single-strand annealing' repair mechanism, which produces large deletions (see below). NHEJ usually requires resection of only a short region to reach a sequence of microhomology by chance. The exonuclease that resects one of the DNA strands and stops when it detects microhomology is thought to be Mre11, a component of the Mre11:Rad50:Nbs1 module (see the heading The DNA Replication Checkpoint) (Paull and Gellert, 2000).

NHEJ proceeds with the binding of Ku (Ku70–Ku80 dimer) to the ends of the broken DNA (the physical relationship between Ku and Mre11 is unknown). The DNA-bound Ku protein recruits a tight dimer consisting of XRCC4 and Lig4. The DNA ligase activity of Lig4 is thereby activated and efficiently seals the broken strands (**Figure 6**). Ku can also recruit and activate the catalytic subunit of the DNA-dependent protein kinase (DNA-PKcs) to the break site. DNA-PKcs, however, is not as essential as Ku, and its exact role in NHEJ is not clear. DNA-PK may phosphorylate Ku70, or RPA (a single-strand binding heterotrimer that may play a role in the process) or DNA-PK itself (see Nick McElhinny *et al*. (2000) and a review by Karran (2000)).

Homologous Recombination, Another Mechanism of DSB Repair

In mammalian cells, homologous recombination was until recently thought to be much less common than NHEJ. Indeed, transfected DNA usually integrates nonhomologously. However, it now appears that homologous recombination does play a major role in DSB repair (see review by Jasin (2000)).

Repair by homologous recombination is not subject to the loss of nucleotides at the joining site that makes NHEJ error-prone. However, when double-strand breaks occur

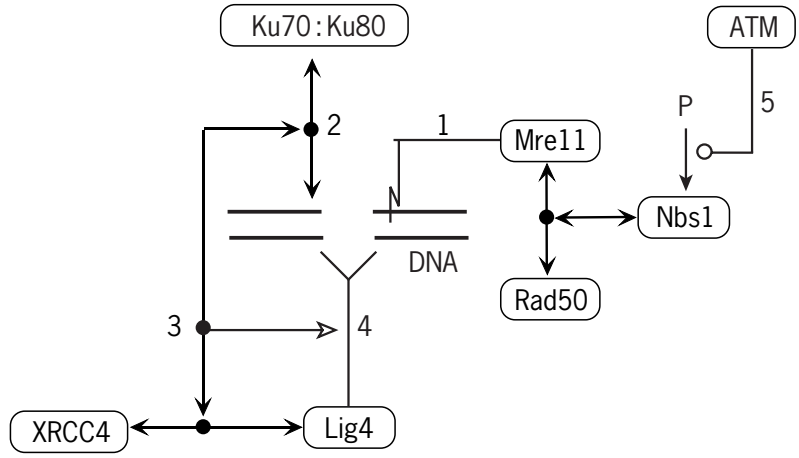


Figure 6 The NHEJ mechanism of DNA double-strand break repair, as currently understood. (1) One DNA strand is resected a short distance by the exonuclease Mre11 (probably functioning as a Mre11:Rad50:Nbs1 trimer) until a point of microhomology is reached, which provides cohesive ends. (2) The double-strand ends bind Ku70:Ku80 dimer. (3) The DNA-bound Ku70:Ku80 recruits Lig4:XRCC4, a tight dimer. (4) Lig4 is thereby activated and brought to the site where it can ligate the strands. Other molecules, such as DNA-PKcs and RPA, may have roles in this process, but their functions remain to be defined. (5) Nbs1 is phosphorylated by ATM in response to DNA damage, as a result of which DNA replication is inhibited (mechanism unknown).

within repetitive sequences (such as Alu-family elements, of which there are about 10^6 copies scattered in the human genome), recombination can occur between Alu elements in different parts of a chromosome or between different chromosomes. This can cause deletion or expansion of regions within a chromosome or translocation between chromosomes. Nevertheless, the preference for recombination between aligned sister chromatids presumably minimizes major chromosome aberrations.

When recombination occurs between homologous chromosomes one of which contains a defective critical gene, the normal copy may be lost. The resulting loss of heterozygosity often can be readily detected. A classical example is the retinoblastoma gene, *Rb*, a recessive tumour suppressor. Heterozygous carriers of a mutant *Rb* gene are susceptible to loss of the functional copy of the gene in an occasional cell, an event that starts the cell on the road to malignancy.

Repair of double-strand breaks or cross-links by homologous recombination requires complementary sequences between the damaged DNA and an undamaged homologue, such as a sister chromatid. First, the 5′-terminating strands of the double-strand break are resected by an exonuclease, so as to leave a 3′-terminated protruding single strand. Repair of DNA cross-links is thought also to begin with processing of the lesion to yield a 3′-terminated protruding strand, but the details of how this is accomplished are unclear.

The protruding 3′-terminus then binds Rad52, a large heptameric doughnut-shaped protein that protects the strand from degradation. Seven Rad52 molecules, assembled as a symmetrical ring, bind specifically to DNA single-strand ends and prevent further exonuclease attack. Although the Rad52 ring has a large central hole, there is no evidence of DNA within the channel. The length of single-strand tail associated with the Rad52 ring is estimated as 36 nucleotides. This terminal region may be exposed and configured to facilitate base pairing with a complementary strand (Parsons *et al*., 2000).

The single-stranded region behind the Rad52 ring is then covered by a contiguous array of Rad51 molecules to form a nucleoprotein filament, which is capable of strand exchange. Rad51 is a structural and functional homologue, of bacterial RecA that is conserved from yeast to humans. Unlike RecA, however, the eukaryotic Rad51 requires ATP hydrolysis to bind properly on DNA. One Rad51 monomer binds per three nucleotides of DNA single strand. Also, unlike RecA, Rad51 requires a 3′ or 5′ extension of DNA single strand to initiate strand exchange. (see review by Karran (2000) and references cited by Namsaraev and Berg (2000)).

The Rad51 nucleoprotein filament is a loose helix which facilitates invasion of the damaged donor strand into a homologous double-strand region that may be located on another chromatid. The strands of the undamaged recipient DNA become locally separated, while base-paired heteroduplex forms with the donor single strand from the damaged DNA. The pairing between donor and recipient strand (‘heteroduplex’ region) is extended by a process called ‘branch migration’ in which the recipient double strand opens, and its original base pairs are replaced by heteroduplex. DNA replication machinery assembles at the 3′-terminus of the donor strand and extends the strand while further displacing the original complementary strand.

Finally, the displaced strand is cleaved by an endonuclease and the donor strand is ligated to the 5′-terminus of the recipient strand.

The second strand of the damaged DNA may be processed by way of a second recombination event. Alternatively, a new complementary strand is synthesized from the template provided by the displaced recipient strand. A distinction between the two mechanisms is that the former involves only leading-strand synthesis, whereas the latter involves both leading- and lagging-strand syntheses.

The rad51–DNA complex may include additional components that provide, as yet undefined functions, conceivably involved in the processing of chromatin structures. Five sequence relatives ('paralogues') of Rad51 have been demonstrated to engage in a pattern of mutual interactions: Rad51B, Rad51C, Rad51D, XRCC2 and XRCC3. As will be described later, these molecules may function together as multiprotein complexes (Schild *et al*., 2000). Knockout of any of the components conferred a high degree of genomic instability and enhanced sensitivity to the DNA cross-linking agents mitomycin and cisplatin (Schild *et al*., 2000). Tumours that might have acquired genomic instability by loss of function of one of these paralogues would be predicted to be sensitive to DNA cross-linking and double-strand break-inducing agents.

The Rad51 System for Homologous Recombination Repair

As already mentioned, Rad51 and its relatives (Rad51B, Rad51C, Rad51D, XRCC2 and XRCC3) engage in a pattern of mutual interactions and may function together as multiprotein complexes (Schild *et al*., 2000). The demonstrated pattern of interactions is summarized in **Figure 7**.

In cells subjected to ionizing radiation or mitomycin, Rad51 aggregates in foci in the nucleus, made visible by immunostaining. The Rad51 foci occur at DNA damage sites in S phase cells, particularly in regions of post-replicative chromatin. This preference conforms with the preferential double-strand break repair when cells are in late S phase or G_2. The foci may represent repair assemblies that function when sister chromatids are available for homologous recombination (Karran, 2000).

The formation of these foci requires Rad51B (Takata *et al*., 2000), consistent with the ability of Rad51B to enhance interactions between members of the Rad51 complex (**Figure 7**). Another family member, Rad54, is also required (Karran, 2000). Rad54 binds Rad51 (Tan *et al*. (1999), cited by Karran (2000)) and promotes DNA double-strand break repair carried out by homologous recombination with sister chromatids (Dronkert *et al*., 2000).

More extended complexes are possible, since Rad51, directly or indirectly, can bind Rad52, Rad54, p53, BRCA1, BRCA2 and c-Abl. Rad52 can bind the single-strand binding protein RPA. Larger assemblies are possible if other protein–protein interactions can exist simultaneously. For example, p53 has been reported to have binding sites (in order from N- to C-terminus) for MDM2, DP1, PARP, c-Abl, RPA, XPB/D, p19ARF, p300/CBP, BRCA1 and 14-3-3. These, and potential chains of further binding interactions, have been summarized in a molecular interaction map (Kohn, 1999) (see also http://discover.nci.nih.gov/

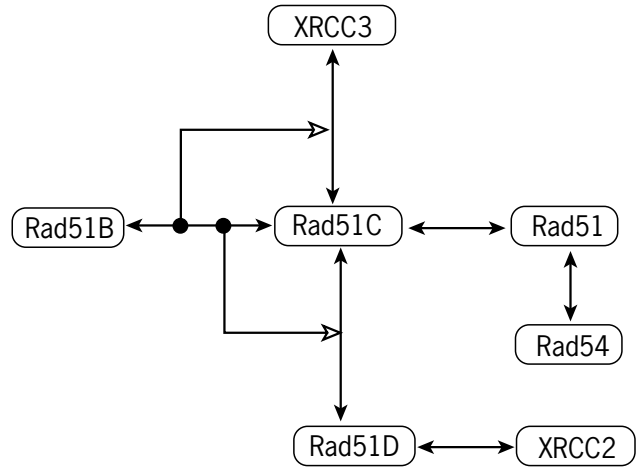


Figure 7 Interaction pattern among Rad51 family members, as reported by Schild *et al*. (2000) and Tan *et al*. (1999) (cited by Karran (2000)). The double-arrowed lines indicate demonstrated binary interactions. The interactions seem capable of building up multimolecular assemblies, as suggested by binding experiments using three components and by yeast three-hybrid experiments. The lines with open triangular arrowheads represent the finding that Rad51B enhances the Rad51C interactions with Rad51D or XRCC3. An additional finding (not included in this diagram) was that Rad51B could bind two molecules of Rad51C in the yeast three-hybrid system, suggesting that the complex could in effect become doubled (Schild *et al*., 2000).

kohnk/interaction_maps.html). Large multimolecular assemblies consisting of different subsets of binding interactions may be formed by remodelling for different functions. The interaction set includes cycles (e.g. Rad51-p53-BRCA1-Rad51 or p53-cAbl-Rad51-BRCA1-p53) which might form a network of molecular chains and perhaps account for the observably large nuclear foci within which DNA repair sites seem to be localized.

Single-strand Annealing

An alternative mechanism of homologous recombination repair, called single-strand annealing, is important when no sister chromatid is available. This mechanism involves resection or realignment of the broken DNA duplex to regions of homology that may be some distance from the break site. Completion of the repair then only requires DNA strand scission of ligation events. Consequently, the segment of DNA intervening between the realigned regions of homology is deleted, which makes this repair mechanism highly error prone.

DNA REPAIR INVOLVING RecQ-FAMILY HELICASES

The RecQ DNA-helicase of *Escherichia coli* is implicated in the suppression of illegitimate recombination and the repair of DNA double-strand breaks. Five human helicases homologous to RecQ are known: BLM, WRN, RecQ4 (also called RecQL4), RecQ5 and RecQ1 (**Figure 8**). Defects in the first three have been recognized as causes of rare genetic diseases: Bloom syndrome, Werner syndrome and Rothmund–Thomson syndrome, respectively (see the chapter *Inherited Predispositions to Cancer*). All three are associated with genetic instability which appears to be due to loss of a helicase activity. The function of these helicases in mammals is only beginning to be elucidated.

The DNA Helicase Defective in Bloom Syndrome (BLM)

The BLM helicase has been the most extensively studied. It is a large protein (1417 amino acids) which contains the motifs characteristic of DNA and RNA helicases. Bloom syndrome cells exhibit a high frequency of chromosome breaks and exchanges. A characteristic of the BLM defect, not shared by the other helicase defects, however, is an increase in reciprocal exchanges between sister chromatids. The cells have a prolonged S phase, thought to be due to difficulty in dealing with stalled replication forks or abnormal replication fork configurations which may occasionally form in the normal course of events. BLM can unwind relatively short DNA duplexes in an ATP-dependent reaction. It can unwind longer helices when aided by the single-strand binding protein, RPA, to which it binds tightly (via the 70-kDa subunit of the RPA trimer) (Brosh *et al*., 2000). BLM, however, does not efficiently unwind DNA from a blunt end, suggesting that a single-stranded tail is needed (to which RPA might anchor) in order to initiate unwinding of relatively long duplexes. RPA binding and cooperative function is a characteristic

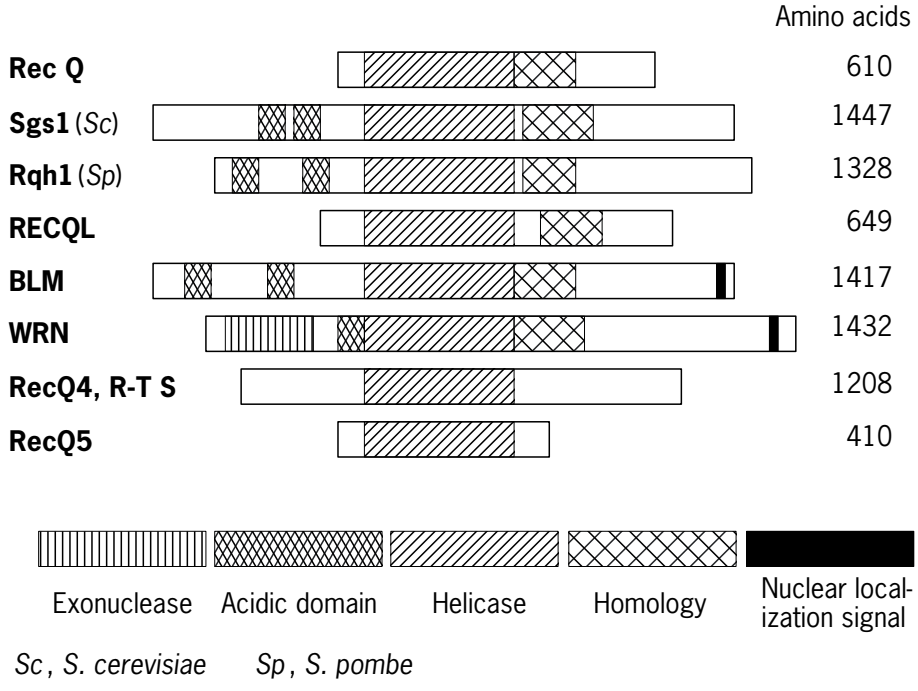


Figure 8 The RecQ family of helicases.

also of the WRN helicase. BLM preferentially unwinds four-stranded DNA helices consisting of Gs stabilized by Hoogsteen pairing; such structures sometimes occur in regulatory regions of genes, and related structures occur in telomeres (Brosh *et al*. (2000) and references cited therein).

BLM recognizes and binds to Holliday junctions (crossover structures between two double helices) and promotes branch migration (movement of the crossover junction along the helices). A possible configuration of a stalled replication fork may form by rewinding of the template strands with displacement of the two newly replicated strands which then pair with each other. The consequent structure is in fact a Holliday junction and could be acted upon by nucleases that normally 'resolve' Holliday junctions. The result would be a double-strand break (Karow *et al*., 2000) (**Figure 9**). The duplex formed by the pairing of the newly replicated strands could be recognized as a double-strand end by the homologous recombination repair system (e.g. involving Rad52 and Rad51) and lead to homologous recombination between sister chromatids. A role of the Rad52–Rad51–Rad54 repair system is supported by the recent finding that the enhanced sister chromatid exchange in *BLM*−/− cells requires Rad54 (Wang *et al*., 2000).

BLM is concentrated in part in nuclear foci and, during S phase, also in nucleoli. Nucleoli contain highly repetitive ribosomal DNA sequences which perhaps tend to form homologous crossovers during replication. BLM perhaps helps to prevent or reverse such crossovers, thereby reducing the chance of loss or expansion of ribosomal DNA regions. BLM has been found in a large multi-molecular protein contain BRCA1 and several DNA repair-related proteins (Wang *et al*., 2000).

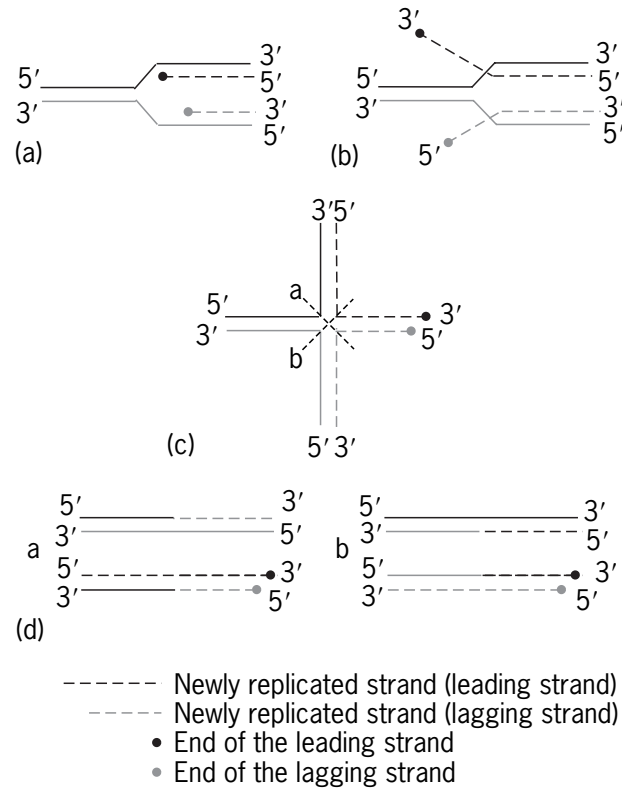


Figure 9 How a helicase defect can lead to a double-strand break involving a parental DNA strand in the vicinity of a stalled replication fork. (a) A normal replication fork. The newly replicated strands are shown by dashed lines with filled circles depicting strand termini. (b) After rewinding of the template helix. This may happen when there is a defect in a helicase that normally operates to unwind the template. The replication fork has backed up, displacing parts of the newly replicated strands. (c) After the displaced segments of newly replicated strands have base-paired with each other. The strands are shown re-configured in order to display the symmetry of a Holliday junction. The Holliday junction can be enzymatically 'resolved' by breakage and rejoining of strands. This can happen in two ways, indicated by the dashed diagonal lines labelled a and b which intersect the strands that are cut and recombined. (d) The final products after the Holliday junction has been resolved in the two possible ways, a and b. In both cases, one of the parental DNA strands becomes joined to newly replicated DNA ending in double-strand termini, equivalent to a double-strand break.

The Protein Defective in Werner Syndrome (WRN)

Werner syndrome (WS) is a homozygous recessive disease characterized by early onset of normal ageing including wrinkling of skin, greying of hair, cataracts, diabetes and osteoporosis. Neoplasms, particularly sarcomas, are observed at higher prevalence in WS patients than in normal individuals of the same age. The symptoms of WS begin to appear near puberty, and most patients die before reaching age 50. Because the clinical features of WS are similar to symptoms of ageing in normal individuals of more advanced age, WS is considered to be a segmental progeria. One of the motivations to study WS is its resemblance to ageing; thus, knowledge of the mechanism and molecular basis of WS might give insight into normal ageing and ageing-associated diseases such as cancer.

WS is caused by mutation in a single gene, *WRN*. The protein product of *WRN* is a member of the RecQ family of helicases, which also includes the yeast protein Sgs1 and the Bloom syndrome protein (BLM). Biochemical evidence shows that the WRN protein is an NTP-dependent DNA unwinding enzyme (and a DNA-dependent NTPase). It was recently shown that WRN is also a 3′–5′ exonuclease (reviewed in Bohr *et al*., 2000). Thus, WRN is a helicase, exonuclease and ATPase. WRN is the only member of the RecQ protein family that is a DNA exonuclease.

WRN and Genomic Instability

Despite extensive study of cells from WS patients, the precise molecular deficiencies involved in WS remain to be defined. Genomic instability of WS cells has been well documented and is consistent with a defect in replication, recombination or DNA repair. Some WS cells undergo premature replicative senescence and delayed progression of S phase, and some WS cells are hyper-recombinogenic. The *WRN* homologues *S. cerevisiae Sgs1*, *S. pombe rqh1* and *E. coli recQ* suppress illegitimate recombination. One possibility is that the WRN helicase is an anti-recombinase, but this hypothesis has not been tested directly.

It has been suggested that WS cells are genetically unstable because alternate DNA structures are not properly resolved. One such alternate structure is the DNA triplex, which can form in a DNA sequence-dependent manner. Sequences that can form triple helices are abundant in the human genome. WRN helicase unwinds a 3′ tailed triple helix DNA substrate in an NTP-dependent manner. Thus, it is possible that triplex structures persist in WS cells and contribute to the variegated translocation mosaicism observed in WS cells.

WRN and DNA Repair

Some evidence suggests a role for the WRN protein in DNA repair. For example, WS cells are sensitive to the carcinogen 4-nitroquinoline 1-oxide (4-NQO). However, WS cells are not hypersensitive to UV light or several other DNA-damaging agents. WS cells may also be partially defective in transcription. Several observations suggest that WRN may be a general activator of transcription by RNA polymerase II (Balajee *et al*., 1999).

WRN can proficiently unwind short DNA duplexes ($\sim$30 bp) in a reaction dependent on nucleoside triphosphate hydrolysis. In the presence of RPA, however, WRN can unwind long DNA duplexes (up to 800 bp). Moreover, WRN can bind directly to RPA (Brosh *et al*., 1999).

WRN does not preferentially bind DNA damage. It could, however, play an important role as a molecular sensor of DNA damage. It may interact with the DNA minor groove in its action as a helicase, as suggested by studies using the minor groove-binding drugs netropsin and distamycin (Brosh *et al*., 2000a,b).

The progress of the 3′–5′ exonuclease reaction catalysed by WRN is blocked by bulky DNA adducts and oxidative lesions, such as 8-oxoguanine (Cooper *et al*., 2000; Machwe *et al*., 2000). Although WRN does not bind preferentially to DNA lesions, it may sense their presence in DNA via protein–protein interactions, perhaps as an early step in damage recognition, and may recruit DNA repair enzymes to the site of a lesion.

WRN can bind to a single-stranded site in a recombination intermediate Holliday structure, where also the RuvA protein binds. This unwinding is ATP dependent, suggesting that it is due to the helicase activity of WRN. WS cells are not deficient in *in vitro* DNA repair assays for nucleotide excision repair (NER) of various bulky adducts in DNA or in base excision repair (BER) of abasic sites, and they are not generally sensitive to tested DNA-damaging agents, with the exception of 4NQO. Thus, as mentioned above, the influence of WRN on DNA repair may be subtle.

Interaction of WRN with Ku, a Protein Needed for DNA Double-strand Break Repair

WRN interacts physically and functionally with the Ku heterodimer. Ku strongly stimulates the 3′–5′ exonuclease activity of WRN but does not affect its helicase or ATPase activities (Cooper *et al*., 2000). The Ku heterodimer and DNA-PK are key proteins in DSB repair (Featherstone and Jackson, 1999). It has been proposed that a helicase and an exonuclease are required in DSB repair to remove the single-stranded overhangs. Thus, it is tempting to speculate that WRN provides both of these enzymatic functions during DSB repair. However, this model predicts that WS cells would have a defect in DSB repair, and there is no evidence of such a defect at the cellular level at present. Alternatively, the interaction between WRN and Ku may have a role in replication, but the role of Ku in that process is still not clear. Ku-deficient transgenic mice have a distinct senescent phenotype, suggesting that there may be another pathway in which WRN and Ku may cooperate.

Ku, the Ku70 : Ku80 heterodimer, can bind DNA and the catalytic subunit of DNA-PK. In the presence of DNA, Ku activates the kinase activity of DNA-PKcs. Ku, however, may have other functions, independent of DNA-PKcs. Ku can bind at DNA double-strand breaks, or at junctions between DNA single- and double-strand regions. Once bound, Ku can move along the DNA in an ATP-independent manner. There have been some suggestions that Ku might have helicase activity, but recent studies suggest that it does not (Cooper *et al*., 2000). Since Ku strongly interacts with the WRN helicase it may attract that helicase activity to the site. Multiple Ku units can load on to the same DNA segment (Frit *et al*., 2000). In addition to DNA-PKcs, Ku may bind other repair-related proteins and perhaps serve as nucleus of a multiprotein repair focus at a point of DNA damage. Ku binds to telomeric ends and may participate in DNA repair or telomerase repair functions here.

What is the Role of WRN in DNA Repair?

In summary, several lines of evidence support the notion that WRN is involved in DNA damage recognition and processing. The WRN exonuclease may function in an early step of DNA repair to recognize DNA lesions. The enzyme recognizes and arrests at some oxidative DNA base lesions. This arrest may then attract DNA repair proteins to

the site, including RPA, p53 and PCNA, which have also been implicated in early steps of DNA damage recognition. WRN is likely to be in a protein complex involving various DNA replication proteins (Lebel *et al*., 1999) and this complex is also likely to contain proteins involved in BER.

DNA REPAIR POLYMERASES

There have been major developments in our understanding of the role of various mammalian DNA polymerases. There are now about 12 characterized DNA polymerases and more are constantly being discovered. In **Table 2** we list some of the properties of these. One of the interesting features under study is that they differ considerably in their fidelity or proofreading of the DNA template. Mutations in these polymerases have been directly associated with human cancer-prone diseases such as xeroderma pigmentosum complementation group V (Woodgate, 1999). This field is evolving very rapidly, and new polymerases are constantly being discovered.

THE Hus1:Rad1:Rad9 SLIDING CLAMP AND Rad17 CLAMP-LOADER MODEL

Hus1, Rad1, Rad9 and Rad 17 are components of a so-called 'Rad checkpoint' DNA damage response system that is conserved from yeast to humans. Although the role of these molecules in mammalian cells is not yet clear, recent evidence suggests a provocative model for their molecular mode of function (Rauen *et al*. (2000) and references cited therein). Hus1, Rad1 and Rad9 form a heterotrimer complex that resembles PCNA in structure and function. PCNA forms a trimeric clamp that encircles the DNA like a doughnut and binds DNA polymerase, with which it can slide along the DNA, keeping the polymerase from falling off. Like PCNA, the Hus1:Rad1:Rad9 heterotrimer may form a clamp around the DNA. It then perhaps slides along the DNA in search of points of damage, or perhaps it binds and keeps in place a repair polymerase. As in the case of PCNA, a 'clamp loader' is needed to open the trimeric doughnut and reassemble it around the DNA. The clamp loader for PCNA is RFC, a five-subunit protein. Recent evidence suggest that the clamp loader for Hus1:Rad1:Rad9 may be Rad17, which shows sequence homology with all five subunits of PCNA (Rauen *et al*., 2000).

Hus1:Rad1:Rad9 trimer normally is distributed throughout the nucleus. In response to DNA damage, however, it concentrates in foci and becomes difficult to extract. This may reflect the clamping of the trimer around the DNA. The association of Rad17 with the trimer, however, is transient, as would be expected for a clamp loader.

The biological functions of this system are still poorly understood. The homologous system in yeast seems to control the S and G2 cell cycle checkpoints. In mammals, Hus1 may do more than just monitor DNA damage. Hus1 is expressed in all examined tissues and throughout embryonic development. *Hus1*-knockout mouse embryos are able to complete gastrulation, but shortly afterwards develop severe abnormalities and die midway in gestation. Some of the embryos survived to the point of having a beating heart. Cells in the *Hus1*-null embryos proliferated at a normal rate, but died by apoptosis at an abnormally high rate (Weiss *et al*., 2000). Cells from *Hus1*-null mouse embryos exhibit increased spontaneous chromosomal abnormalities, suggesting that Hus1 function is needed to maintain chromosome stability (Weiss *et al*., 2000).

Table 2 Eukaryotic polymerases and their fidelity in replication of undamaged DNA. (Adapted from Wang, 1999.)

Polymerase	Error frequency (mutations per base pair)
Alpha (α)	10^{-4}
Delta (δ)	10^{-5}
Epsilon (ε)	10^{-6}
Gamma (γ)	10^{-6}
Beta (β)	5×10^{-4}
Zeta (ζ)	10^{-4}
Eta (η)	10^{-2}
Iota (ι)	10^{-1}
Theta (θ)	?
Kappa (κ)	10^{-4}
Lambda (λ)	?
Mu (μ)	10^{-1}

MULTIMOLECULAR ASSEMBLIES AND NUCLEAR FOCI

DNA repair functions may often be organized in large multimolecular structures. Often these structures are assembled in large nuclear foci that can be seen by means of fluorescent antibodies. Components of the repair systems may be localized in foci which function as repair factories where DNA lesions could be brought for processing. Exchangeable components might then be efficiently shared among several repair tasks simultaneously in progress. Nuclear foci sometimes undergo rearrangement or remodelling during the cell cycle or in response to DNA damage or other types of stress. Nuclear foci or molecular repair assemblies have been found involving the Rad51 system already discussed in the section on homologous recombination. Other repair systems that may

function as multimolecular assemblies include systems based on Ku (see earlier) and on BRCA1.

A BRCA1-associated Genome Surveillance Complex (BASC)

BRCA1 has several protein-binding domains, and may be associated with large multiprotein complexes in the nucleus, including the DNA repair-related proteins MSH2, MSH6, MLH1, ATM, BLM and the Rad50:Mre11:Nbs complex (Wang *et al*., 2000a,b). The exact structure and function of these complexes is still not clear, however, because it has been difficult to demonstrate specific direct binding in cells (Jun Qin, personal communication). BRCA1-based multimolecular foci perhaps assemble in alternative arrangements with different components, making it difficult to establish individual interactions *in vivo*. In addition, BRCA-1 directly participates in TCR of oxidative DNA damage.

THE DNA REPLICATION CHECKPOINT

DNA damage during S phase normally causes temporary arrest of DNA replication. Proteins required for this checkpoint include ATM, Mre11, Rad50 and NBS1. Genetic defects occur in ATM (ataxia telangiectasia), NBS1 (Nijmegen breakage syndrome) and Mre11 (ataxia telangiectasia-like disorder) (see the chapter *Inherited Predispositions to Cancer*). In all three syndromes, DNA damage (e.g. by ionizing radiation) fails to arrest replication and leads to extensive chromosome damage (Petrini, 2000). Although the mechanistic details are not yet in, we know a few steps in the process. DNA damage causes phosphorylation of NBS1. This is required for replication arrest. ATM senses the DNA damage and carries out the phosphorylation of NBS1. The activity of JNK (c-Jun N-terminal kinase), which is normally stimulated in response to DNA damage, fails to respond in cells derived from patients having any of the three syndromes (see review by Petrini (2000)). **Figure 6** includes the molecular interactions for which there is evidence.

REFERENCES

Aebi, S., *et al*. (1996). Loss of DNA mismatch repair in acquired resistance to cisplatin. *Cancer Research*, **56**, 3087–3090.

Balajee, A. S. and Bohr, V. A. (2000). Genomic heterogeneity of nucleotide excision repair. *Gene*, **250**, 15–30.

Balajee, A. S., *et al*. (1997). Reduced RNA polymerase II transcription in intact and permeabilized Cockayne syndrome group B cells. *Proceedings of the National Academy of Sciences of the USA*, **94**, 4306–4311.

Balajee, A. S., *et al*. (1999). The Werner syndrome protein is involved in RNA polymerase II transcription. *Molecular Biology of the Cell*, **10**, 2655–2668.

Batty, D. P. and Wood, R. D. (2000). Damage recognition in nucleotide excision repair of DNA. *Gene*, **241**, 193–204.

Bohr, V. A., *et al*. (2000). Werner syndrome protein: biochemical properties and functional interactions. *Journal of Experimental Gerontology*, **35**, 695–702.

Breivik, J. and Gaudernack, G. (1999). Genomic instability, DNA methylation, and natural selection in colorectal carcinogenesis. *Cancer Biology*, **9**, 245–254.

Brosh, R. M. Jr, *et al*. (1999). Functional and physical interaction between WRN helicase and human replication protein A. *Journal of Biological Chemsitry*, **274**, 18341–18350.

Brosh, R. M., Jr, *et al*. (2000a). Potent inhibition of Werner and Bloom helicases by DNA minor groove binding drugs. *Nucleic Acids Research*, **28**, 2420–2430.

Brosh, R. M., Jr, *et al*. (2000b). Replication protein A physically interacts with the Bloom's syndrome protein and stimulates its helicase activity. *Journal of Biological Chemistry*, **275**, 23500–23508.

Buermeyer, A. B., *et al*. (1999). Mammalian DNA mismatch repair. *Annual Review of Genetics*, **33**, 533–564.

Calin, G. A., *et al*. (2000). Genetic progression in microsatellite instability high (MSI-H) colon cancers correletes with clinico-pathological parameters: a study of the TGFβRII, BAX, hMSH3, hMSH6, IGFIIR and BLM Genes. *International Journal of Cancer (Pred. Oncology)*, **89**.

Coleman, W. B. and Tsongalis, G. J. (1999). The role of genomic instability in human carcinogenesis. *Anticancer Research*, **19**, 4645–4664.

Cooper, M. P., *et al*. (2000). Ku complex interacts with and stimulates the Werner protein. *Genes and Development*, **14**, 907–912.

Croteau, D. L., *et al*. (1999). Mitochondrial DNA repair pathways. *Mutation Research*, **434**, 137–148.

de Boer, J. and Hoeijmakers, H. J. (2000). Nucleotide excision repair and human syndromes. *Carcinogenesis*, **21**, 453–460.

DeWeese, T. L., *et al*. (1998). Mouse embryonic stem cells carrying one or two defective Msh2 alleles respond abnormally to oxicative stress inflicted by low-level radiation. *Proceedings of the National Academy of Sciences of the USA*, **95**, 11915–11920.

Dianov, G. L., *et al*. (1999). Replication protein A stimulates proliferating cell nuclear antigen-dependent repair of abasic sites in DNA by human cell extracts. *Biochemistry*, **38**, 11021–11025.

Dizdaroglu, M. (1991). Chemical determination of free radical-induced damage to DNA. *Free Radicals in Biology and Medicine*, **10**, 225–242.

Dronkert, M. L. G., *et al*. (2000). Mouse RAD54 affects DNA double-strand break repair and sister chromatid exchange. *Molecular Cellular Biology*, **20**, 3147–3156.

Drummond, J. T., *et al*. (1996). Cisplatin and adriamycin resistance are associated with MutLα and mismatch repair deficiency in an ovarian tumor cell line. *Journal of Biological Chemistry*, **271**, 19645–19648.

Duckett, D. R., *et al*. (1999). hMutSα- and hMutLα-dependent phosphorylation of p53 in response to DNA methylator damage. *Proceedings of the National Academy of Sciences of the USA*, **96**, 12384–12388.

Featherstone, C. and Jackson, S. P. (1999). Ku, a DNA repair protein with multiple cellular functions? *Mutation Research*, **434**, 3–15.

Fink, D., *et al*. (1998). Enrichment for DNA mismatch repair-deficient cells during treatment with cisplatin. *International Journal of Cancer*, **77**, 741–746.

Fishel, R. (1998). Mismatch repair, molecular switches, and signal transduction. *Genes and Development*, **12**, 2096–2101.

Flores-Rozas, H. and Kolodner, R. D. (2000). Links between replication, recombination and genome instability in eukaryotes. *Trends in Biochemical Science*, **25**, 196–200.

Friedberg, E. C. (1996). Cockayne syndrome – a primary defect in DNA repair, transcription, both or neither? *BioEssays*, **18**, 731–738.

Frit, P., *et al*. (2000). Ku entry into DNA inhibits inward DNA transactions *in vitro*. *Journal of Biological Chemistry*, **275**, 35684–35691.

Fritzell, J. A., *et al*. (1997). Role of DNA mismatch repair in the cytotoxicity of ionizing radiation. *Cancer Research*, **57**, 5143–5147.

Gong, J. G., *et al*. (1999). The tyrosine kinase c-Abl regulates p73 in apoptotic response to cisplatin-induced DNA damage. *Nature*, **399**, 806–809.

Grady, W. M., *et al*. (1998). Mutation of the type II transforming growth factor-beta receptor is coincident with the transformation of human colon adenomas to malignant carcinomas. *Cancer Research*, **58**, 3101–3104.

Jasin, M. (2000). Chromosome breaks and genomic instability. *Cancer Investigations*, **18**, 78–86.

Jiricny, J. and Nystrom-Lahti, M. (2000). Mismatch repair defects in cancer. *Current Opinions in Genetic Development*, **10**, 157–161.

Karow, J. K., *et al*. (2000). The Bloom's syndrome gene product promotes branch migration of Holliday junctions. *Proceedings of the National Academy of Sciences of the USA*, **97**, 6504–6508.

Karran, P. (2000). DNA double strand break repair in mammalian cells. *Current Opinions in Genetic Development*, **10**, 144–150.

Kohn, K. W. (1996). Beyond DNA crosslinking: history and prospects of DNA-targeted cancer treatment – 15th Bruce F. Cain Memorial Award Lecture. *Cancer Research*, **56**, 5533–5546.

Kohn, K. W. (1999). Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Molecular Biology of the Cell*, **10**, 2703–2734.

Lebel, M., *et al*. (1999). The Werner syndrome gene product co-purifies with the DNA replication complex and interacts with PCNA and topoisomerase I. *Journal of Biological Chemistry*, **274**, 37795–37799.

Lengauer, C., *et al*. (1998). Genetic instabilities in human cancers. *Nature*, **396**, 643–648.

Le Page, F., *et al*. (2000). Transcription coupled repair of 8-oxoguanine in murine cells: the ogg1 protein is required for repair in nontranscribed sequences but not in transcribed sequences. *Proceedings of the National Academy of Sciences of the USA*, **97**, 8397–8402.

Li, G.-M. (1999). The role of mismatch repair in DNA damage-induced apoptosis. *Oncology Research*, **11**, 393–400.

Longley, M. J., *et al*. (1998). Identification of 5′-deoxyribose phosphate lyase activity in human DNA polymerase gamma and its role in mitochondrial base excision repair *in vitro*. *Proceedings of the National Academy of Sciences of the USA*, **95**, 12244–12248.

Machwe, A., *et al*. (2000). Selective blockage of the $3' \rightarrow 5'$ exonuclease activity of WRN protein by certain oxidative modifications and bulky lesions in DNA. *Nucleic Acids Research*, **28**, 2762–2770.

Markowitz, S. (2000). TGFβ receptors and DNA repair genes, coupled targets in a pathway of human colon carcinogenesis. *Biochimica Biophysica Acta*, **1470**, M13–M20.

Massague, J. (1998). TGFb signal transduction. *Annual Review of Biochemistry*, **67**, 753–791.

Massague, J. and Wotton, D. (2000). Transcriptional control by the TGFβ/Smad signaling system. *EMBO Journal*, **19**, 1745–1754.

Namsaraev, E. A. and Berg, P. (2000). Rad51 uses one mechanism to drive DNA strand exchange in both directions. *Journal of Biological Chemistry*, **275**, 3970–3976.

Nick McElhinny, S. A., *et al*. (2000). Ku recruits the XRCC4-Ligase IV complex to DNA ends. *Molecular Cellular Biology*, **20**, 2996–3003.

Parsons, C. A., *et al*. (2000). Precise binding of single-stranded DNA termini by human RAD52 protein. *EMBO Journal*, **19**, 4175–4181.

Paull, T. T. and Gellert, M. (2000). A mechanistic basis for Mre11-directed DNA joining at microhomologies. *Proceedings of the National Academy of Sciences of the USA*, **97**, 6409–6414.

Petrini, J. H. J. (2000). The Mre11 complex and ATM: collaborating to navigate S phase. *Current Opinions in Cell Biology*, **12**, 293–296.

Pinz, K. G. and Bogenhagen, D. F. (1998). Efficient repair of abasic sites in DNA by mitochondrial enzymes. *Molecular Cellular Biology*, **18**, 1257–1265.

Prasad, R., *et al*. (2000). FEN1 stimulation of DNA polymerase β mediates an excision step in mammalian long patch base excision repair. *Journal of Biological Chemistry*, **275**, 4460–4466.

Rauen, M., *et al*. (2000). The human checkpoint protein hRad17 interacts with the PCNA-like proteins hRad1, hHus1, and hRad9. *Journal of Biological Chemistry*, **257**, 29767–29771.

Rooney, P. H., *et al*. (1999). Comparative genomic hybridization and chromosomal instability in solid tumours. *British Journal of Cancer*, **80**, 862–873.

Rosen, N. (1997). Cancers of the gastrointestinal tract. In: Devita, V. T., Jr, *et al*. (eds), In: *Cancer: Principles and Practice of Oncology*. 971–1054 (Lippincott-Raven, Philadelphia).

Schild, D., *et al*. (2000). Evidence for simultaneous protein interactions between human Rad51 parallogs. *Journal of Biological Chemistry*, **275**, 16443–16449.

Selby, C. P. and Sancar, A. (1997). Human transcription-repair coupling factor CSB/ERCC6 is a DNA-stimulated ATPase but is not a helicase and does not disrupt the ternary transcription complex of stalled RNA polymerase II. *Journal of Biological Chemistry*, **272**, 1885–1890.

Strathdee, G., *et al*. (1999). A role for methylation of the hMLH1 promoter in loss of hMLH1 expression and drug resistance in ovarian cancer. *Oncogene*, **18**, 2335–2341.

Takata, M., *et al*. (2000). The Rad51 paralog Rad51B promotes homologous recombinational repair. *Molecular Cellular Biology*, **20**, 6476–6482.

Tantin, D., *et al*. (1997). Recruitment of the putative transcription-repair coupling factor CSB/ERCC6 to RNA polymerase II elongation complexes. *Molecular Cellular Biological*, **17**, 6803–6814.

Toft, N. J., *et al*. (1999). Msh2 status modulates both apoptosis and mutation frequency in the murine small intestine. *Proceedings of the National Academy of Sciences of the USA*, **96**, 3911–3915.

Vogel, H., *et al*. (1999). Deletion of Ku86 causes early onset of senescence in mice. *Proceedings of the National Academy of Sciences of the USA*, **96**, 10770–10775.

Wang, T. S. F. (1999). Cellular DNA polymerases. In: DePamphilis, M. L. (ed.), In: *DNA Replication in Eukarytoic Cells*. 461–493 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY).

Wang, W., *et al*. (2000a). Possible association of BLM in decreasing DNA double strand breaks during DNA replication. *EMBO Journal*, **19**, 3428–3435.

Wang, Y., *et al*. (2000b). BASC, a super complex of BRCA1-associated proteins involved in the recognition and repair of aberrant DNA structures. *Genes and Development*, **14**, 927–939.

Weiss, R. S., *et al*. (2000). Inactivation of mouse *Hus1* results in genomic instability and impaired responses to genotoxic stress. *Genes and Development*, **14**, 1886–1898.

Wiseman, H. and Halliwell, B. (1996). Damage to DNA by reactive oxygen and nitrogen species: role in inflammatory disease and progression to cancer. *Biochemical Journal*, **313**, 17–29.

Woodgate, R. (1999). A plethora of lesion-replicating DNA polymerases. *Genes and Development*, **13**, 2191–2195.

FURTHER READING

Balajee, A. S. and Bohr, V. A. (2000). Genomic heterogeneity of nucleotide excision repair. *Gene*, **250**, 15–30.

Batty, D. P. and Wood, R. D. (2000). Damage recognition in nucleotide excision repair of DNA. *Gene*, **241**, 193–204.

Buermeyer, A. B., *et al*. (1999). Mammalian DNA mismatch repair. *Annual Review of Genetics*, **33**, 533–564.

Flores-Rozas, H. and Kolodner, R. D. (2000). Links between replication, recombination and genome instability in eukaryotes. *Trends in Biochemistry Science*, **25**, 196–200.

Haber, J. E. (2000). Partners and pathways: repairing a double-strand break. *Trends in Genetics*, **16**, 259–264.

Jasin, M. (2000). Chromosome breaks and genomic instability. *Cancer Investigation*, **18**, 78–86.

Jiricny, J. and Nystrom-Lahti, M. (2000). Mismatch repair defects in cancer. *Current Opinions in Genetics Development*, **10**, 157–161.

Lengauer, C., *et al*. (1998). Genetic instabilities in human cancers. *Nature*, **396**, 643–648.

Lindahl, T. and Wood, R. D. (1999). Quality control by DNA repair. *Science*, **286**, 1897–1905.

McCullough, A. K., *et al*. (1999). Initiation of base excision repair: glycosylase mechanisms and structures. *Annual Review of Biochemistry*, **68**, 255–285.

Oshima, J. (2000). The Werner syndrome protein: an update. *BioEssays*, **22**, 894–901.

Pegg, A. E. (2000). Repair of O^6-alkylguanine by alkyltransferases. *Mutation Research*, **462**, 83–100.

Vessey, C. J., *et al*. (2000). Genetic disorders associated with cancer predisposition and genomic instability. *Progress in Nucleic Acids Research*, **63**, 189–221.

Chapter 8

Telomerase

Carmela P. Morales, Woodring E. Wright and Jerry W. Shay
The University of Texas Southwestern Medical Center at Dallas, Dallas, TX, USA

CONTENTS

INTRODUCTION

Intense interest in the enzyme telomerase has occurred in the field of cancer biology over the past 5 years. Recent evidence suggests a role for telomerase during the multistep process of carcinogenesis. Reactivation or upregulation of telomerase is found in the majority of human cancers and appears to be responsible for the limitless replicative potential of malignant cells, a hallmark of cancer. In normal cells, the 'replicative lifespan' is tightly regulated by an internal divisional clock which limits the number of divisions that a cell can undergo during its lifetime. This divisional clock, known as the telomere, is located at each end of all linear chromosomes. By setting maximum limits on the number of times a cell can divide, telomeres may serve to prevent genetically aberrant cells from accumulating the additional mutations they need to become malignant. For the successful propagation and continued growth of malignant cells, therefore, telomere control on cell growth must be subverted. Current evidence indicates that in almost all human cancers this is achieved by reactivation of the enzyme telomerase.

TELOMERES AND THE 'END-REPLICATION' PROBLEM

Telomeres are long stretches of noncoding DNA located at the ends of all eukaryotic chromosomes (**Figure 1**). In vertebrates, telomeres are comprised of simple, repetitive noncoding DNA sequences. Human telomeres contain the six base pair sequence TTAGGG, repeated many thousands of times.

The length of telomeres varies from chromosome to chromosome. Evidence suggests that as chromosome 'caps,' telomeres have at least three critical functions: to protect chromosome ends from enzymatic degradation and abnormal fusion reactions; to serve as a buffer zone to protect against the 'end-replication' problem; and to serve as a gauge for mitotic age (the divisional clock).

The role of telomeres in maintaining chromosomal integrity was proposed by Barbara McClintock in 1941. Studying telomeres in maize chromosomes, McClintock observed that if not capped by telomeres, the ends of chromosomes had a tendency to fuse. Her observations

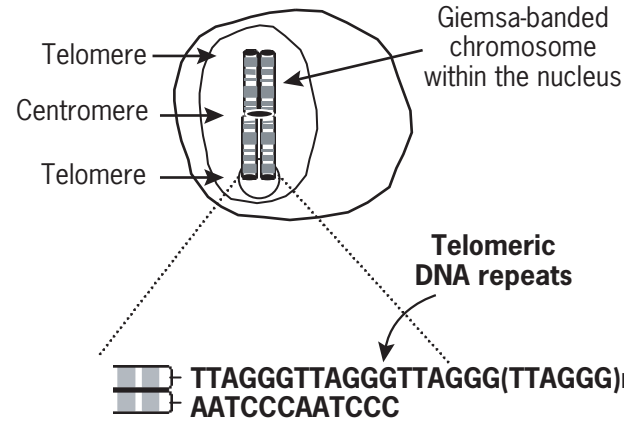


Figure 1 Location of telomeres at the ends of all eukaryotic chromosomes. Telomeres are comprised of the simple DNA sequence TTAGGG, repeated several thousand times (*n*).

were confirmed 50 years later in yeast and mice when it was demonstrated that without telomeric ends, chromosomes undergo aberrant end-to-end fusions, forming multicentric chromosomes with a propensity to break during mitosis, activating DNA damage checkpoints, and in some cases leading to widespread cell death (Zakian, 1989).

The second major function of telomeres relates to the process of semiconservative DNA replication. During each round of cell division, 50–200 base pairs are lost from the ends of linear human chromosomes (Hastie *et al*., 1990; Lindsey *et al*., 1991; Allsopp *et al*., 1992, 1995). This 'end-replication' problem (**Figure 2**) occurs because conventional DNA replication machinery is unable to replicate completely the 3′ ends of chromosomal DNA during the S phase of each cell cycle. The polymerases that copy parental DNA strands prior to cell division synthesize DNA only in the 5′ to 3′ direction and require a short RNA primer to begin. These primers are then degraded and filled in by DNA synthesis extending from the upstream primer. However, at the end of a linear chromosome there is no 'upstream' DNA synthesis to fill in the gap between the final RNA priming event and the end of the chromosome.

This replication strategy predicts that with each round of cell division, there will be progressive shortening of the 3′ end of chromosomal DNA. Telomeric DNA therefore provides a cushion of expendable noncoding sequences to protect against the potentially catastrophic attrition of important chromosomal material.

TELOMERES ARE THE DIVISIONAL CLOCK

The existence of an internal divisional clock was first suggested in 1965 by Leonard Hayflick, who demonstrated that cells maintained in culture have a finite capacity to proliferate. In 1972, Olovnikov suggested that erosion of the chromosome ends could lead to the loss of essential genes and an exit from the cell cycle. Harley *et al*. introduced a modification of the Olovnikov theory, proposing a telomere-based mechanism to account for the process of 'cellular ageing' (Harley *et al*., 1990). Specifically, it was proposed that after a certain number of divisions, telomeres are no longer sufficient to protect chromosome ends from degradation and aberrant fusion reactions. Through signalling mechanisms that are not entirely understood, a few short telomeres may trigger exit from the cell cycle at G1 and entry into senescence, a postmitotic state characterized not only by a lack of further cell division, but also

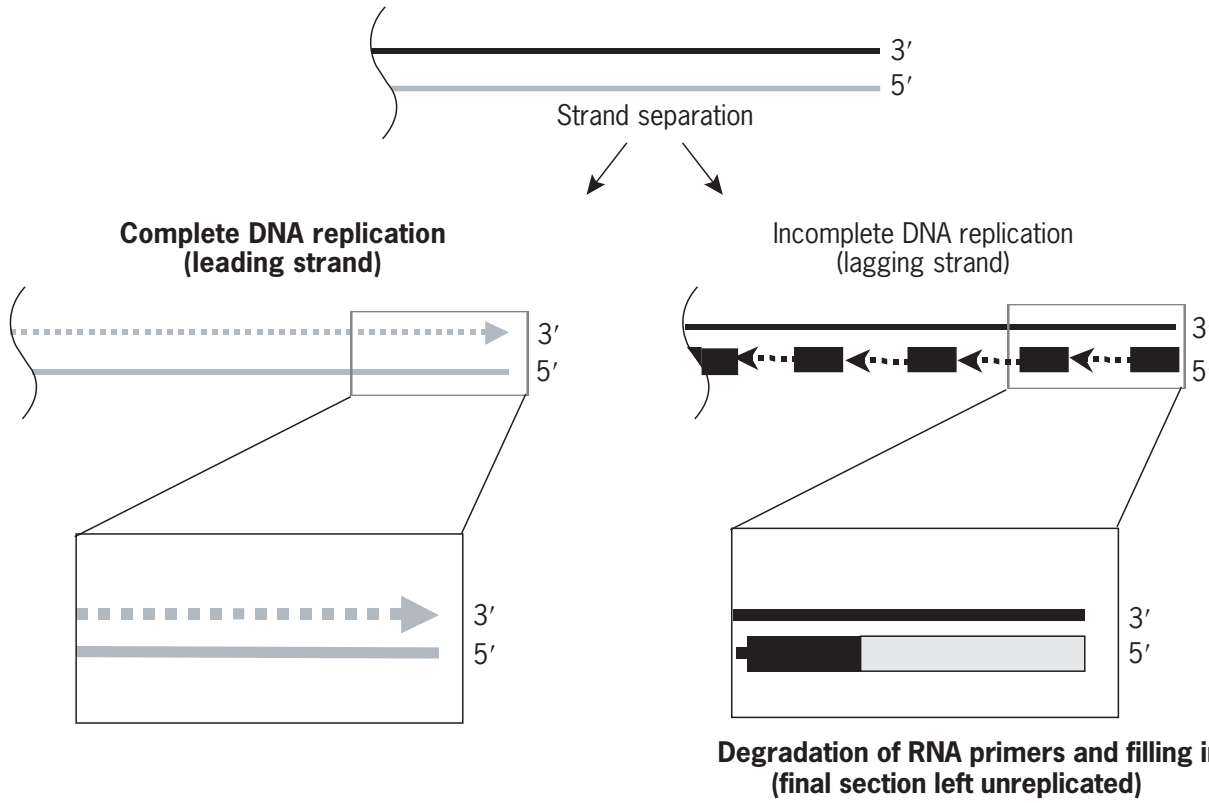


Figure 2 The end replication problem. DNA polymerases require a short RNA primer (black rectangles) to initiate DNA replication in the 3′ to 5′ direction. Since the extreme 3′ end of the chromosome cannot accommodate a primer, this part of the chromosome cannot be replicated, leading to a loss of telomeric DNA with each round of cell division.

by an altered pattern of gene expression and continued metabolic activity for long periods of time. In the case of normal human fibroblasts, a correlation was found between the number of divisions the cells could undergo in culture and initial telomere length, regardless of the age of the fibroblast donor. Additionally, average telomere length in blood and colonic mucosa was shown to decrease with biological age. In adults, sperm telomeres were found to be several kilobase pairs longer than in somatic tissues. Finally, significantly shorter telomere lengths have been demonstrated in some primary cells from patients with the premature ageing syndrome Hutchinson–Gilford progeria compared with normal age-matched controls. These cells also exhibited a reduced proliferative capacity compared with age-matched controls when maintained in culture. Although a large amount of correlative data supported the notion that telomere length determines the proliferative capacity of human cells, a direct test of this hypothesis (**Figure 3**) was lacking. The ability to elongate telomeres experimentally led to the observation that hybrid cells with artificially long telomeres had a longer lifespan than that of cell hybrids in which telomeres had not been elongated (Wright *et al*., 1996). These observations provided the first direct evidence that telomere length is the counting mechanism that limits the proliferative capacity of human cells.

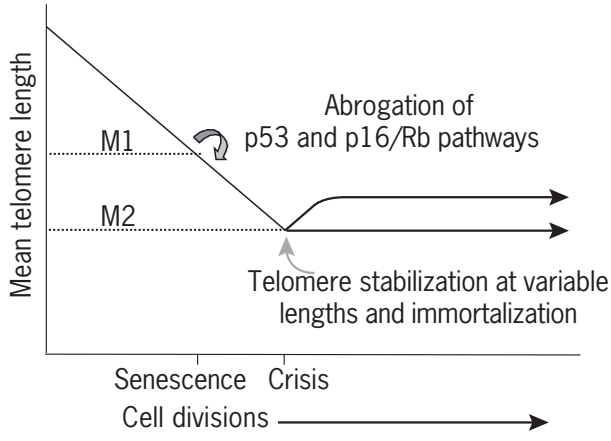


Figure 3 The telomere hypothesis. Telomeric repeats are lost with each round of cell division, leading to a decrease in mean telomere length with accumulated divisions. Upon reaching a certain length, short telomeric DNA sequences trigger entry into the senescence pathway, during which cells remain metabolically active but are no longer able to divide. This stage is also known as mortality stage 1 (M1). Cells can be forced to proliferate beyond this point by abrogating p53 and pRb or their respective pathways. During this period of extended life, cells continue to divide in the face of progressive telomere shortening. After reaching a critically short telomere length, most cells enter a crisis, undergoing widespread cell death (apoptosis). This second stage is also known as mortality stage 2 (M2). Only those rare cells which engage mechanisms that stabilize telomere length are able to continue proliferating for indefinite periods.

The ability to alter cellular proliferative capacity by manipulating telomere length provided a mechanistic basis for earlier observations of cellular lifespan *in vitro*. Normal human fibroblasts maintained in culture undergo a finite number of divisions as determined by their initial telomere length, after which they enter a state of growth arrest (senescence), also known as mortality stage 1 (M1).

Cells nearing the end of their lifespan can be forced to proliferate beyond this point by the introduction of certain viruses or oncogenes that abrogate the function of the tumour-suppressor genes *p53* and *pRb*. These observations suggest that p53 and pRb perhaps mediate cell cycle exit at G1 in response to telomere shortening. Bypass of M1 allows additional rounds of cell division until further, critical telomere shortening occurs, resulting in a state of 'crisis,' characterized by widespread cell death. This second stage is known as mortality stage 2 (M2). As a low-frequency ($\sim 10^{-7}$) event in human cells, a subpopulation of cells escapes from crisis, giving rise to cells which now have an unlimited proliferative capacity (immortalized). The characteristic feature of such immortal cells is the ability to maintain their telomeres.

The dual role of telomeric DNA as protector of chromosomal integrity and mitotic clock implicates cellular senescence as a natural and effective initial protection mechanism against the development of cancer. It is generally believed that tumours are initiated by multiple genetic events in cells which result in the inappropriate activation of growth stimulatory signals, an insensitivity to antigrowth signals and a resistance to apoptosis. However, transformation to fully malignant derivatives does not occur in most cases because the majority of these aberrant cells will have exhausted their endowment of allowed divisions.

TELOMERASE

Early studies demonstrated a significantly shorter telomere length in most cancers compared with noncancerous tissue from the same patient (Hastie *et al*., 1990). In culture, cancer cells generally have short but stable telomeres, suggesting that human cancers have developed strategies for the maintenance of telomeric DNA at a length above the critical threshold. In 85–95% of human cancers, this telomere stabilization is achieved by reactivation or upregulation of the ribonucleoprotein enzyme telomerase.

Telomerase is an RNA–protein complex which utilizes its RNA as a template for the addition of TTAGGG repeats to the 3′ ends of chromosomes, thereby compensating for losses due to the end-replication problem (**Figure 4**).

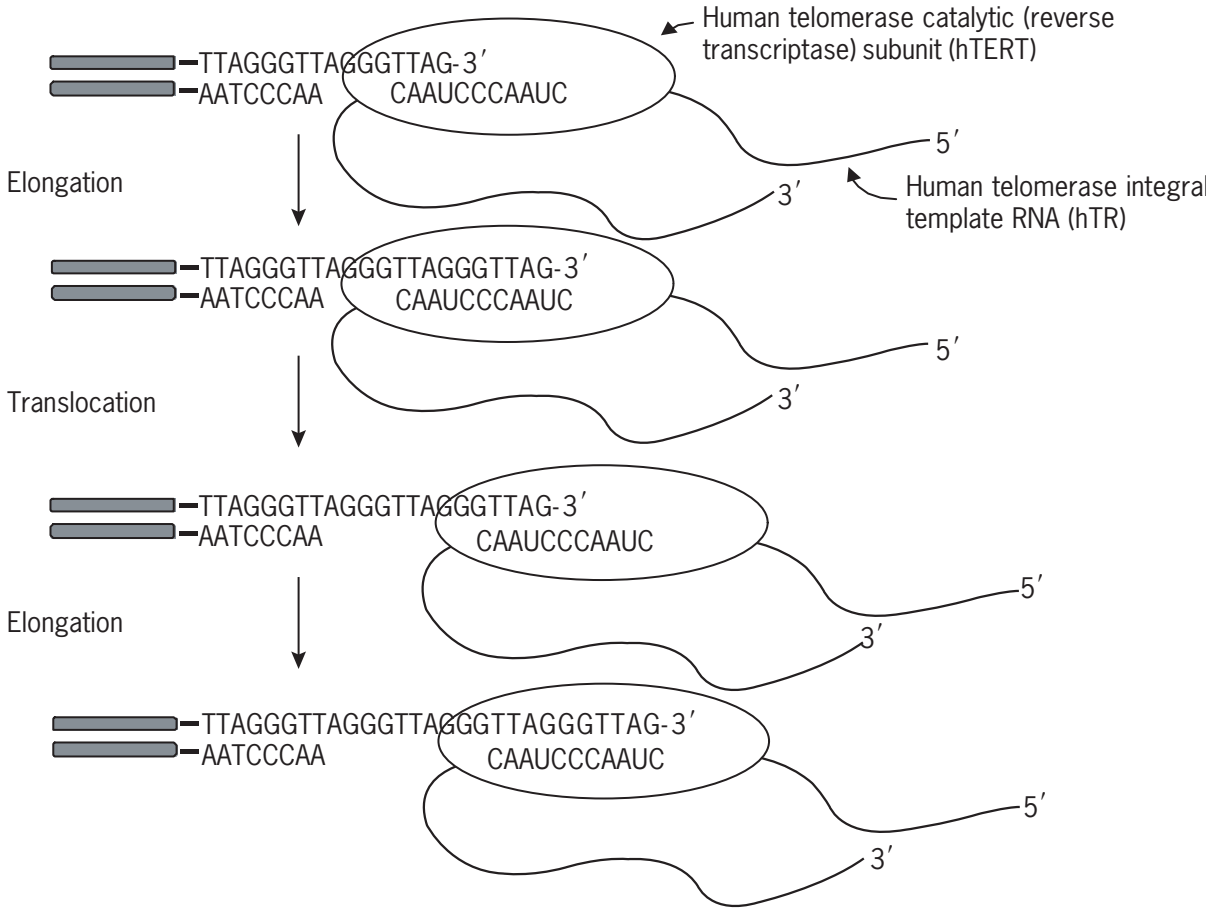


Figure 4 Telomerase is an enzyme minimally comprised of a catalytic protein component and an RNA subunit. The RNA serves as a template for the addition of TTAGGG repeats to the ends of chromosomes, leading to maintenance of telomeric DNA. This enzyme is processive, meaning that a variable number of repeats may be added to a chromosome prior to disengaging the chromosome end.

First discovered in *Tetrahymena* by Elizabeth Blackburn and Carol Greider, telomerase activity has now been detected in extracts from almost all organisms, with the exception of bacteria and viruses, which have circular genomes, and *Drosophila*, which have retrotransposons instead of telomeres. In humans, most adult somatic cells lack telomerase activity. However, telomerase is present at high levels in germ cells, early embryos, activated T and B cells and germinal centres of lymphoid organs. Telomerase activity is also detectable at lower amounts in the basal cells of renewal tissues (skin and intestine). In somatic tissues and cells (including T cells), however, the presence of detectable levels of telomerase activity is not sufficient to prevent long-term telomere attrition. The variable levels of telomerase in normal human tissues are illustrated in **Figure 5**.

In humans, telomerase is composed of two essential components: an integral RNA (hTR), which provides the template for the synthesis of telomere repeats, and a protein subunit (hTERT), which provides catalytic activity. The cloning of hTERT and hTR made it possible to test directly the hypothesis that telomere shortening regulates the entry into cellular senescence. Using an *in vitro* system, the combination of hTERT and hTR was first shown to reconstitute telomerase activity (Weinrich *et al*., 1997; Beattie *et al*., 1998). Second, the introduction of hTERT into telomerase-negative primary cells resulted in telomerase activity. Exogenous hTR was not needed because it is present even in cells that do not normally have telomerase activity. Finally, it was shown that telomere maintenance by exogenous telomerase was sufficient for the immortalization of human mammary epithelial cells, foreskin fibroblasts, retinal pigmented epithelial cells and umbilical vascular endothelial cells. Taken together, these experiments provided direct evidence that short telomere length directs entry into cellular senescence and that telomere maintenance by telomerase is sufficient to bypass this growth arrest under most circumstances. Importantly, the introduction of telomerase prior to either M1 or M2 is sufficient for immortalization, indicating that telomeres

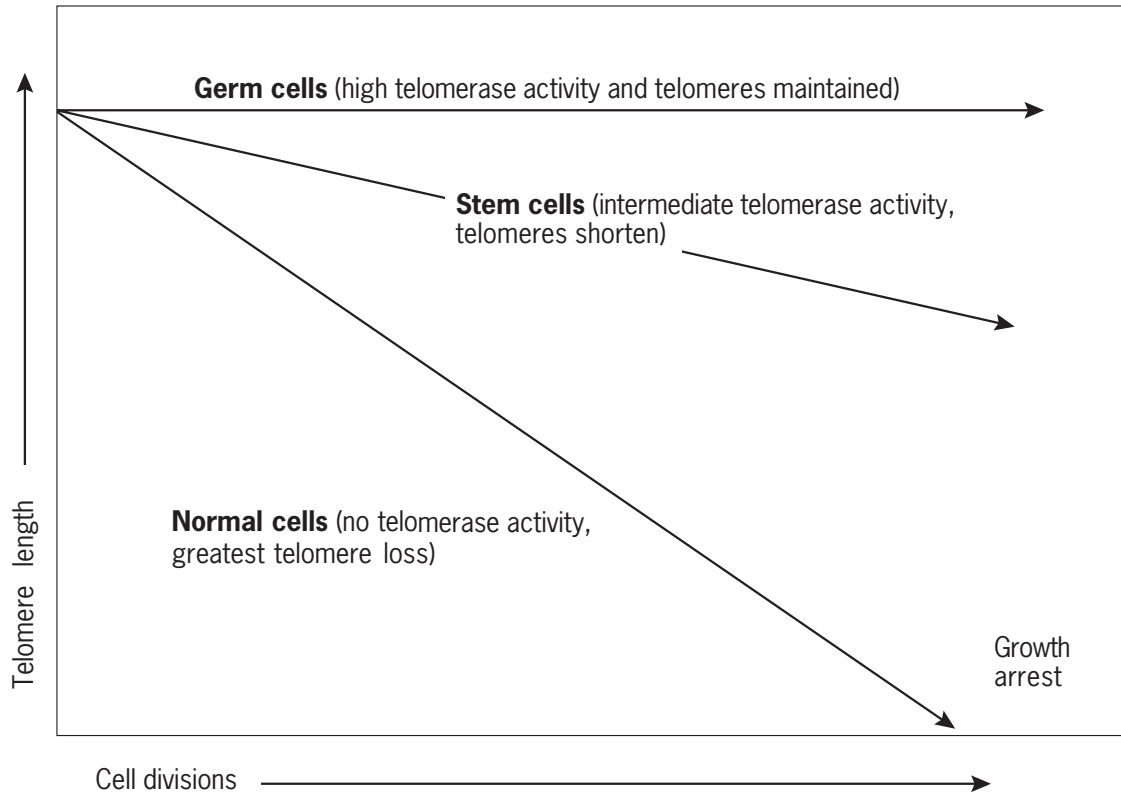


Figure 5 Variable levels of telomerase in normal human tissues. Germ cells of the reproductive system maintain high levels of telomerase activity throughout life and therefore do not sustain telomere shortening. Stem cells of renewal tissues express modest levels of telomerase, leading to a blunted rate of telomere shortening. Somatic cells and tissues lack detectable telomerase activity and sustain the greatest rates of telomere loss.

are associated with both the M1 and M2 stage of growth arrest.

Various factors such as oxidative stress, introduction of an activated Ha-*ras* oncogene and γ-irradiation have been shown to induce a senescent-like state in cells much younger than the Hayflick limit. Several reports have also suggested that in addition to telomerase, inactivation of the Rb/$p16^{INK4a}$ pathway is required for the immortalization of some human epithelial cells. The protein $p16^{INK4a}$ is an inhibitor of cyclin-dependent kinases, and its levels have been shown to increase after only a few passages in culture. During the establishment of human mammary epithelial cells under standard culture conditions, there appears to be a 'self-selection' process, such that only the cells which have lost $p16^{INK4a}$ expression (usually by methylation of the promoter) are able to survive the initial culture period. In these surviving cells, exogenous telomerase expression leads to immortalization. In most cases, no appreciable decrease in telomere length can be demonstrated to account for this 'self- selection' process, termed M0. These observations suggest that the involvement of $p16^{INK4a}$ at M0 is telomere-independent, a finding that stands in direct contrast to the role of p53 in mediating of growth arrest at the telomere-dependent M1 stage. Recent observations in keratinocytes suggest that the loss of $p16^{INK4a}$ expression is not required if the cells are maintained under optimized culture conditions, such as co-culturing with irradiated fibroblast feeder layers (Ramirez *et al.*, 2001). These findings suggest that artificial cell culture conditions may account for the premature growth arrest (M0) in epithelial cells, and that this response is mediated by induction of $p16^{INK4a}$.

ASSAYS FOR TELOMERASE

The standard method for measuring telomerase activity is a highly sensitive PCR-based assay termed the TRAP (telomere repeat amplification protocol) assay (**Figure 6**).

In this assay, extracts are first prepared from primary tissue or cultured cells by lysing the cells with a detergent, releasing telomerase into the extract solution. An aliquot of this solution is then added to a reaction mixture containing a short primer and deoxynucleotide triphosphates (dNTPs). If telomerase is present, it will elongate the primer with TTAGGG repeats. The products of this elongation step are

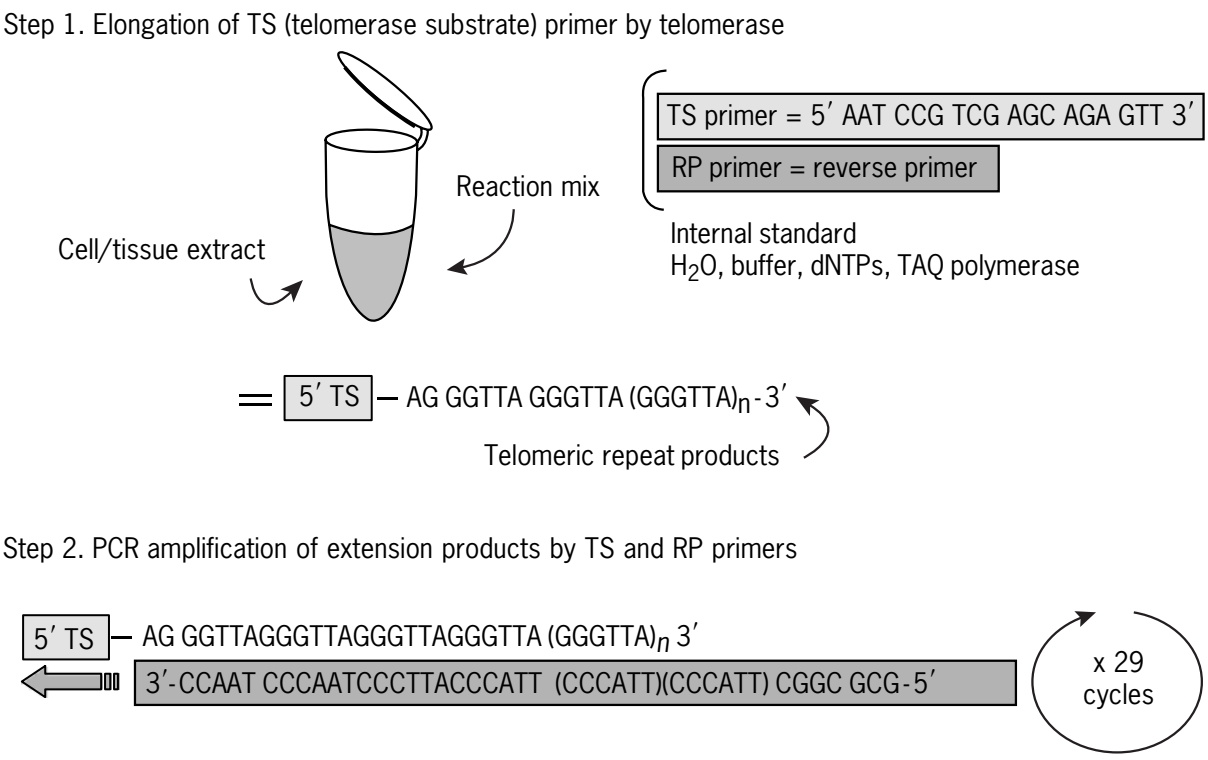


Figure 6 The telomerase activity assay. This sensitive PCR-based assay can be applied to extracts from a variety of cells and tissues. In the first step, extracts are incubated with labelled nucleotides and also a synthetic telomere end (TS primer) at room temperature. If the extract contains telomerase, the enzyme will synthesize the addition of TTAGGG repeats to the TS primer. The second step utilizes PCR to amplify the extended products in the presence of additional primers and polymerases. These amplified products are then separated by electrophoresis to generate the characteristic six-base-pair ladder indicating telomerase activity. ITAS is an internal standard included to control for the possibility of enzyme inhibitors.

heterogeneous in length, representing multiples of the six-base-pair TTAGGG sequence. Using a second primer which is complementary to the telomerase repeat and a DNA polymerase known as TAQ polymerase, each product is amplified using the polymerase chain reaction (PCR). The amplified products are then run on a polyacrylamide gel, creating a six-base-pair ladder. The laddering effect occurs because telomerase is a processive enzyme, adding telomeric repeats in multiples of TTAGGG (**Figure 7**). This amplification protocol increases the sensitivity of the assay such that telomerase activity can be detected in samples containing as few as 0.1% positive cells. An internal standard is also incorporated into the assay, since some tissue extracts contain molecules that inhibit PCR and give false-negative results. In addition, this internal standard permits semiquantitative analysis of relative telomerase activity levels.

Alternative approaches to the measurement of telomerase have recently been developed. Unlike TRAP, which is a functional assay carried out on extracts of cells or tissues, *in situ* techniques are designed to visualize the components of telomerase at a cellular level. *In situ* hybridization for the RNA component of human telomerase (hTR) can be applied to formalin-fixed, paraffin-embedded tissues as well cultured cells and cell smears. Several studies have demonstrated good concordance between telomerase activity as measured by TRAP and telomerase RNA by *in situ* hybridization. Although normal cells do contain hTR, the levels in normal tissues are sufficiently low that they do not complicate the observation of elevated hTR levels in tumours. Antibodies for the immunhistochemical detection of the telomerase protein component are now becoming commercially available, even though their utility remains to be established. Studies using other antitelomerase antibodies, however, been shown to correlate with telomerase activity by TRAP. These techniques may have some advantages over the PCR-based assay. The excellent morphological preservation of cellular detail provided by *in situ* hybridization or immunohistochemistry may be helpful in localizing telomerase to specific cell types. Furthermore, *in situ* telomerase assays could be readily adapted by clinical laboratories, many of which already utilize such techniques for the detection of other proteins.

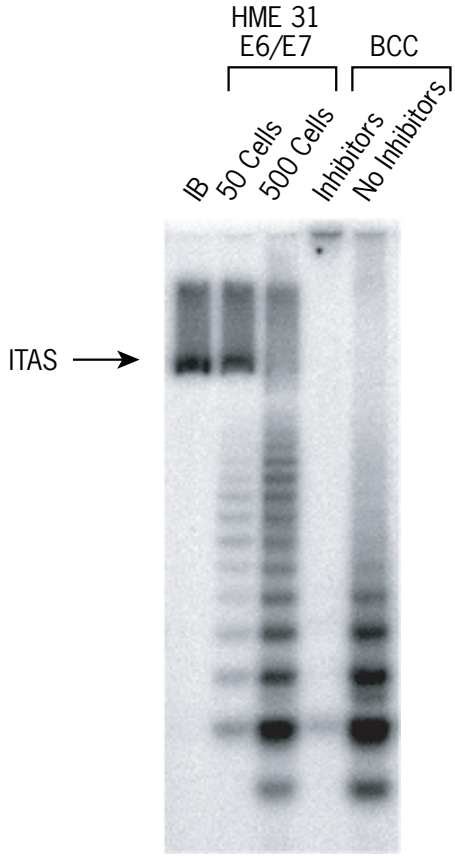


Figure 7 Telomerase activity in two different basal cell carcinomas (BCC). Human mammary epithelial cells expressing the E6 and E7 oncoproteins (HME 31 E6/E7) serve as a positive control for quantitation. The lane containing lysis buffer (LB) serves as a negative control. This representative telomerase assay gel reveals the characteristic 6 base-pair ladder indicative of enzymatic activity, and the internal standard (ITAS) that serves to normalize sample-to-sample variation. Absence of the ITAS signal, as demonstrated in the first BCC sample, indicates the presence of PCR inhibitors. Without this internal control, this sample may be misinterpreted as lacking telomerase activity. Quantitation of telomerase activity is done by determining the ratio of the internal standard to the telomerase ladder.

ASSOCIATION BETWEEN TELOMERASE ACTIVITY AND CANCER

The TRAP assay has made possible the large-scale testing for telomerase activity in a wide variety of human cancers and normal tissues. Using TRAP, telomerase activity has been detected in 85–95% of all human cancers and cancer cell lines, whereas adjacent normal tissue and mortal cells in culture are generally telomerase negative (**Table 1**).

Thousands of individual malignancies representing all of the major organ systems have been tested to date, including those originating from the head and neck, lung, gastrointestinal, pancreatic and biliary tract, liver, breast, male and female reproductive tract, kidney/urinary tract, central nervous system, skin and blood (Shay and Bacchetti, 1997). Preinvasive and preneoplastic lesions, such as colorectal adenomas, high-grade prostatic intraepithelial neoplasia, *in situ* breast carcinoma and those from the head/neck and lung tissue, are positive in 30% to almost 100% of cases. These observations provide strong evidence that most human malignancies are associated with the reactivation or upregulation of telomerase. Given that most normal human cells have the capacity to undergo 60–70 population doublings, it may at first glance seem difficult to invoke telomere shortening as a barrier to cancer formation, because after 60 doublings, a single cell would generate a tumour mass of approximately 10^{15} kg! However, not only do evolving malignancies exhibit high rates of turnover due to chronic, widespread apoptosis and differentiation within the tumour, but also many clonal expansions occur.

A schematic demonstrating the relationship between cell turnover and tumorigenesis is shown in **Figure 8**.

In this scenario, a single healthy cell would be expected to generate a population of 10^6 cells after 20 doublings. As a rare event, one cell in this population acquires a genetic mutation which confers to it a selective growth advantage over the remaining cells. Owing to its newly acquired growth advantage, this cell then generates a clone of similarly altered cells. This process is repeated several times, with each successive generation resulting from an event in a single cell from the previous generation. In some cases, an additional generation is required to convert a minimally functional recessive mutation into a strong phenotype through loss of the remaining wild-type allele (loss of heterozygosity (LOH)). Rarely, a single cell emerges which has acquired sufficient advantageous mutations to result in the development of a malignancy. Using such a scenario, it can be seen that the number of cells in the final tumour grossly underrepresents the number of cells required to produce it. Because so many doublings are required for the development of a cancer, most potentially tumorigenic cells probably senesce owing to critical telomere erosion. Thus, cellular senescence could be viewed as a powerful initial blockade against carcinogenesis.

The timing of telomerase reactivation in human tumours appears to vary considerably from organ to organ. In most cases, the mean telomere length in a variety of tumour types is substantially shorter than those in normal tissues from the same patient. Taken in the context of the telomere hypothesis, these observations suggest that telomerase reactivation or upregulation occurs only after dramatic telomere loss, and serves to stablilize shortened chromosome ends and permit continued cell proliferation. In most colorectal, oesophageal and pancreatic adenocarcinomas, telomerase reactivation appears to be a late event, occurring during the transition from low- to high-grade

Table 1 Telomerase activity in human cancers[a]

Pathology	% Positive	Pathology	% Positive
Head/neck and lung		*Breast*	
Normal oral mucosa	32	Fibrocystic disease/fibroadenoma	0
Head/neck squamous cell carcinoma	86	Carcinoma *in situ*	75
Non-small cell lung carcinoma	78	Carcinoma (ductal and lobular)	88
Small cell lung carcinoma	100	Adjacent tissue	5
Adjacent lung tissue	4	*Reproductive tract*	
Gastrointestinal tract		Normal adult ovary	33
Gastric metaplasia/adenoma[b]	27	Normal myometrium/endometrium	0
Gastric carcinoma	85	Leiomyoma	0
Adjacent gastric tissue	25	Leiomyosarcoma	100
Colorectal adenoma[b]	45	Cervical/vaginal/endometrial cancer	100
Colorectal carcinoma	89	Ovarian carcinoma	91
Adjacent and normal colon tissues[b]	25	Normal adult testis	100
Pancreas and liver		Normal prostate	0
Benign pancreatic lesions (all)	0	BPH without carcinoma[c]	5
Pancreatic carcinoma	95	BPH with carcinoma[c]	11
Adjacent pancreatic tissue	14	High-grade PIN[d]	60
Benign pancreatic brushings	0	Prostate carcinoma	90
Malignant pancreatic brushings	100	*Skin*	
Normal liver tissue	0	Normal epidermis[b]	44
Nonmalignant liver disease (all)[b]	29	Squamous cell carcinoma	83
Hepatocellular carcinoma	86	Basal cell carcinoma	95
Adjacent liver tissue	2	Melanoma	86
Kidney/urinary tract		*Haematological tissues*	
Normal urothelium	0	Myeloma	100
Dysplastic urothelium	43	Lymphoma, low grade	86
Bladder carcinoma (all stages)	92	Lymphoma, high grade	100
Bladder carcinoma (washings)	73	Tonsils, normal	100
Bladder carcinoma (voided urine)	29	Myelodysplastic syndrome	67
Renal cell carcinoma	83	CML, chronic[e]	71
Adjacent renal tissue	0	CML, early accelerated	33
Wilm tumour	100	CML, blast stage	100
Adjacent renal tissue (Wilm)	33	CLL, early[f]	14
Neural tissues		CLL, late	57
Normal retina	0	Acute promyelocytic leukaemia	100
Retinoblastoma	50	Acute lymphocytic leukaemia	80
Glioblastoma multiforme	75	Acute myelogenous leukaemia	73
Oligodendroglioma	100		
Anaplastic astrocytoma	10		
Meningioma, ordinary	17		
Meningioma, atypical	92		
Meningioma, malignant	100		
Ganglioneuroma	0		
Neuroblastoma	94		
Adjacent neural tissue	0		

[a]Adapted from Shay and Bacchetti, 1997. [b]Telomerase activity weak compared with carcinomas. [c]BPH = benign prostatic hypertrophy. [d]PIN = prostatic intraepithelial neoplasia. [e]CML = chronic myelogenous leukaemia. [f]CLL = chronic lymphocytic leukaemia.

dysplasia. Additionally, early-stage neuroblastomas lack or have low levels of telomerase activity, whereas late-stage disease has high levels of telomerase activity. In other malignancies, such as head and neck cancers, lung carcinomas and breast carcinomas, telomerase activity is present in early preneoplastic lesions, albeit at lower levels than frankly malignant tissues. Since most of these data were obtained by studying tissue extracts, it is difficult to determine whether telomerase activity in preneoplastic lesions is due to the infiltration of microscopic quantities of tumour cells or to low-level telomerase activity in preneoplastic cells. The development of *in situ* techniques, such as immunohistochemistry with telomerase antibodies or *in situ* hybridization for the telomerase template RNA

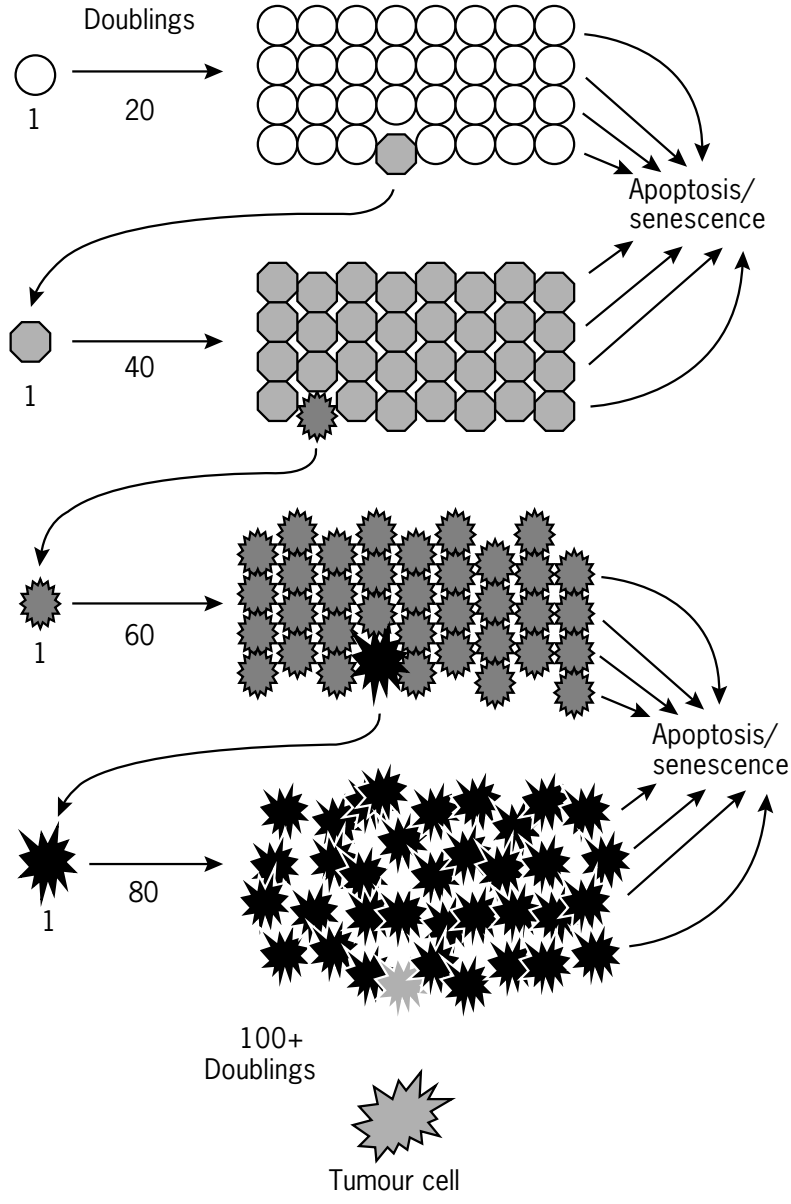


Figure 8 Relationship between cell divisions and tumorigenesis. Within a normal population of cells, a single cell acquires a mutation which endows the cell with a growth advantage. After 20 doublings, a clone of cells emerges, one of which undergoes an additional advantageous mutation. While the other cells senesce or die, this cell survives to generate a clone of similarly mutated cells. The cycle continues until one mutant emerges which has acquired all of the necessary mutations for tumorigenesis. According to this scheme, a fully tumorigenic cell can result in less than 100 doublings.

(hTR), will be important in clarifying these important issues. In the case of cervical cancer, *in situ* hybridization for hTR showed focal increases in hTR expression at the level of *in situ* carcinomas.

Although most human cancers express high levels of telomerase, a substantial portion (10–15%) are telomerase negative. There are several explanations for this observation. First, although the TRAP assay is capable of detecting telomerase activity with only 1–10 tumour cells, because tumours are heterogeneous, some sampled specimens may contain no or insufficient numbers of telomerase-positive tumour cells to be detected by the assay. Second, some cancers may not have reached a point where telomerase activity is required. Such tumours may in fact still be mortal and therefore truly telomerase negative. Third, as mentioned previously, it has been reported that some tissue extracts contain inhibitors of either the elongation or amplification steps of the TRAP assay, leading to false-negative results. Other reasons for false-negative results include technical errors, such as poor sample preservation, sampling error or misloading of the specimen into the reaction mixture. Finally, there is experimental evidence for one or more

alternative mechanisms (possibly based on recombination) for lengthening of telomeres (ALT). Some immortalized cell lines show evidence of ALT activity characterized by the absence of telomerase activity but the presence of very long and heterogeneous telomeres. Although the mechanism is not well known, the existence of an ALT pathway has important theoretical implications for telomerase inhibition as a treatment for cancer. To date, however, there is no experimental evidence suggesting that tumour cells with telomerase can be converted to the ALT pathway.

EXPLOITATION OF TELOMERASE IN CANCER DIAGNOSTICS AND PROGNOSTICS

The strong association between telomerase and most human malignancies has prompted a flurry of studies exploring the potential clinical utility of telomerase as a diagnostic cancer marker. In addition to measuring telomerase activity in tissue extracts, the TRAP assay has been successfully applied to a wide variety of samples including bladder washings, sedimented cells in voided urine and colonic effluent, oral rinses, brushes and washes, endoscopic brushings, biliary aspirates, ascitic fluid, blood, fine needle aspirates and frozen sections. Formalin-fixed, paraffin-embedded pathological material can also be tested for the presence of micrometastasis using recently developed *in situ* hybridization to the telomerase template RNA (hTR) and immunohistochemical detection of the catalytic component (hTERT). In oesophageal carcinomas, a marked increase in hTR occurs during the transition from low- to high-grade dysplasia, suggesting that telomerase is important for the development of advanced lesions in oesophageal carcinogenesis. Although initial data appear promising, these techniques need to be validated by comparison with the standard TRAP assay. However, current evidence demonstrates a good correlation between telomerase activity and *in situ* levels of hTR.

The utility of telomerase in predicting the outcome of cancer (prognosis) is based on the notion that without telomere maintenance, malignant cells will be unable to sustain long-term proliferation and eventually undergo cell death and tumour regression, contributing to a favourable outcome. The best evidence for such a scenario is in stage 4S neuroblastomas that lack detectable telomerase activity. These cases are associated with large rate of spontaneous tumour regression. This indicates that, at least in some cancers, telomerase is not absolutely required for malignancy but that tumours without telomerase may ultimately regress if they do not engage a mechanism for telomere stabilization. In ordinary meningiomas, a strong correlation was found between telomerase activity and disease relapse. Telomerase activity has also been observed to confer a worse prognosis in other malignancies such as neuroblastoma, acute myeloid leukaemia, breast cancer and some gastrointestinal cancers.

Another potential role for telomerase is in the detection of residual disease after surgical resection or adjuvant chemo- and/or radiation therapy. Telomerase activity has been detected in cells adjacent to tumours, suggesting the presence of small foci of residual malignant cells. Such information could be used to restage a lesion and identify a subset of patients who would benefit from additional therapy. Although there are reports that inflammatory cells express low levels of telomerase activity, analysis of malignant lymph nodes reveals levels of telomerase at least sixfold higher than their benign counterparts.

TELOMERASE INHIBITION IN CANCER

Several lines of evidence support the notion that inhibition of telomerase may be an effective anti-cancer strategy. As mentioned previously, telomerase is present in most human malignancies. Although the introduction of certain viral oncoproteins or the abrogation of tumour-suppressor genes may confer an extended lifespan, in the absence of a mechanism for telomere maintenance, these cells eventually reach a period of crisis and undergo widespread cell death.

Telomerase is a challenging molecule for drug development because of the long period required to reach sufficient telomere shortening. To be considered telomerase-specific, inhibitors should fulfil several criteria: (1) inhibitors should reduce telomerase activity without initially affecting proliferation; (2) treatment with inhibitors should result in telomere shortening with each round of cell division; (3) treated cells should eventually undergo growth arrest or apoptosis; (4) there should be a correlation between initial telomere length and time to growth arrest or cell death; (5) control of chemically related molecules or inhibitors lacking the ability to inhibit telomerase activity should not have an effect on cell proliferation and telomere length. Numerous conventional chemotherapeutic agents have been reported to inhibit telomerase. These reports are based on observations that treatment with such agents resulted in widespread cell death and loss of telomerase activity. However, the interpretation of the results is suspect since they do not fulfil the criteria expected for telomerase inhibitors.

There are several important theoretical considerations associated with telomerase inhibitor therapy. First, during the initial period of telomere shortening, continued cell proliferation could result in clinically significant tumour growth. Second, discontinuation of therapy for even short periods of time during the treatment period could result in the rapid induction of telomerase and telomere relengthening. Third, selective pressure on tumour cells being treated with telomerase inhibitors could lead to drug resistance due to the emergence of cells with alternative

mechanisms of telomere maintenance. Finally, there is a theoretical concern that normal cells with telomerase activity (germ cells and renewal tissues) would also be susceptible to telomerase inhibition. However, these cells generally have a longer than average telomere length and are much more slowly dividing than tumour cells, and thus would be expected to be relatively resistant to the consequences of telomerase inhibition.

The telomerase template RNA (hTR) and the catalytic core of the protein subunit (hTERT) are two obvious choices for drug design since both components are absolutely required for telomerase activity. Agents tested to date include antisense oligonucleotides and synthetic peptide nucleic acids targeted against the template region of telomerase (hTR). Using agents that fulfil these criteria, several groups demonstrated that telomerase inhibition in cultured cancer cells resulted in a marked (70–95%) decrease in telomerase activity, telomere shortening and widespread cell death after periods ranging from 2 to 3 months, depending on the initial telomere length. Discontinuation of the drug resulted in a rapid reactivation of telomerase and regrowth of telomeres to their initial length. In tumour cells that were inhibited until the point of cell death, there was no evidence for the development of resistant cells, suggesting that the alternative mechanism of telomere maintenance may not be readily adopted in these cells.

Overall, these initial observations suggest that sustained telomerase inhibition may be an effective and feasible anticancer strategy. However, given the delayed effect of such agents, the most appropriate setting for telomerase inhibition would appear to be the prevention of relapse due to small numbers of remaining cells or to cells resistant to initial conventional therapy. Other strategies would combine telomerase inhibitors with other agents, such as angiogenesis inhibitors to target tumour cells specifically and effectively. Additional novel and potentially effective approaches against telomerase are beginning to emerge, such as ribozymes directed against the template region of telomerase RNA, which cleave the RNA and render telomerase inactive. Molecules which couple the telomerase promoter to apoptotic genes are also in the early stages of development. Such approaches would be expected to have the dual benefit of optimizing tumour cell specificity while producing a more rapid biological effect. However, the effect of such agents on telomerase-competent stem cells and germ-line cells remains an important consideration.

TELOMERASE FOR THE DEVELOPMENT OF *IN VITRO* MODELS OF CANCER PROGRESSION

The introduction of telomerase into some normal human cells resulted in bypass of M1 and immortalization. After a doubling of their normal lifespan, these immortalized cells maintain a normal diploid karyotype and DNA damage and cell cycle checkpoints remained fully intact, suggesting that normal cells immortalized with telomerase do not develop additional cancer-associated changes. In cells expressing the Simian virus 40 Large T antigen for long periods of time followed by the introduction of an activated *ras* oncogene, the addition of telomerase appears to be sufficient for transformation into full tumorigenicity. Taken together, these results suggest that while telomerase expression *per se* does not result in genomic destabilization, in the context of underlying mutations, telomerase contributes to tumorigenicity by providing aberrant cells with an unlimited proliferative capacity.

The ability of telomerase to immortalize cells without altering the underlying genetic background has also been demonstrated in cells with inherited susceptibility syndromes, such as ataxia–telangiectasia, Bloom syndrome, xeroderma pigmentosum and premature ageing syndromes such as Werner syndrome and Hutchinson–Gilford progeria. Thus, telomerase may be an important tool for establishing premalignant cell lines which can be used for the development of *in vitro* models of cancer progression, for amassing large numbers of cells required for other assays or as standard cellular reagents for microarray analysis. Microarray is a novel technique which utilizes microchip technology to analyze cellular RNA for changes in patterns of genetic expression in mutated cells versus their normal counterparts.

The role of telomerase in cancer progression will undoubtedly represent a major continuing area of investigation in the field of cancer biology. The tight association between telomerase and cancer, the ability to generate immortalized human cell lines for studies of cancer progression and the development of telomerase inhibitors for use as anticancer agents all underscore the fundamental role of telomere maintenance as a major player in the development and continued unlimited growth of cancer cells.

REFERENCES

Allsopp, R. C., *et al*. (1992). Telomere length predicts the replicative capacity of human fibroblasts. *Proceedings of the National Academy of Sciences of the USA*, **85**, 10114–10118.

Allsopp, R. C., *et al*. (1995). Telomere shortening is associated with cell division *in vitro* and *in vivo*. *Experimental Cell Research*, **220**, 194–200.

Beattie, T. L., *et al*. (1998). Reconstitution of human telomerase activity *in vitro*. *Current Biology*, **8**, 177–180.

Harley, C. B., *et al*. (1990). Telomeres shorten during ageing of human fibroblasts. *Nature*, **345**, 458–460.

Hastie, N. D., *et al*. (1990). Telomere reduction in human colorectal carcinoma and with ageing. *Nature*, **346**, 866–868.

Lindsey, J., *et al*. (1991). *In vivo* loss of telomeric repeats with age in humans. *Mutation Research*, **256**, 45–48.

Ramirez, R. D., *et al*. (2001). Putative telomere-independent mechanisms of replicative aging reflect inadequate growth conditions. *Genes and Development*, **15**, 398–403.

Shay, J. W. and Bacchetti, S. (1997). A survey of telomerase activity in human cancer. *European Journal of Cancer*, **33**, 787–791.

Weinrich, S. L., *et al*. (1997). Reconstitution of human telomerase with the template RNA component hTR and the catalytic protein subunit hTRT. *Nature Genetics*, **17**, 498–502.

Wright, W. E., *et al*. (1996). Experimental elongation of telomeres extends the lifespan of immortal × normal cell hybrids. *EMBO Journal*, **15**, 1734–1741.

Zakian, V. A. (1989) The structure and function of telomeres. *Annual Reviews of Genetics*, **23**, 579–604.

FURTHER READING

Cech, T. R. (2000). Life at the end of the chromosome: telomeres and telomerase. *Angewandte Chemie International Edition in English*, **39**, 34–43.

Greider, C. W. (1999). Telomerase activation-one step on the road to cancer? *Trends in Genetics*, **15**, 109–112.

Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, **100**, 57–70.

Wynford-Thomas, D. (1999). Cellular senescence and cancer. *Journal of Pathology*, **187**, 100–111.

Chapter 9

Apoptosis

John C. Reed
The Burnham Institute, La Jolla, CA, USA

CONTENTS

INTRODUCTION

Cell death is a part of normal physiology for most metazoan species. During development, redundant or unwanted cells are removed through programmed cell death, making important contributions to morphogenesis, organogenesis and other processes (Vaux and Korsmeyer, 1999). Programmed cell death typically occurs through a ritualistic process known as 'apoptosis,' a term adapted from the Greek and which has analogies with the falling away of leaves from trees in the autumn. Among the features of cells undergoing apoptosis are chromatin condensation, nuclear fragmentation, plasma membrane blebbing, cell shrinkage and ultimately shedding of membrane-delimited cell fragments known as apoptotic bodies (Wyllie, 1997).

In adult mammals, programmed cell death plays an essential role in tissue homeostasis, offsetting new cell production with cell death in all self-renewing tissues. Roughly one million cells commit suicide every second in the adult human, thus making room for new cells produced in tissues such as the bone marrow, skin and gut on a daily basis. So massive is the flux of cell birth and death through our bodies that the average person will produce and in parallel eradicate a mass of cells equivalent to his or her entire body weight each year. Consequently, defects in the cell death machinery which prevent the programmed turnover of cells can result in cell accumulation, thereby imparting a selective growth advantage to neoplastic cells without necessarily involving concomitant defects in the cell division cycle (Reed, 1999).

Defects in the pathways responsible for programmed cell death also play important roles in multiple aspects of tumour cell biology, besides cell accumulation. For example, because cancer requires the accumulation within a single clone of multiple genetic lesions, enhanced cell longevity as a result of defective apoptosis may indirectly promote cancer. The genetic instability that characterizes many cancers is also indirectly assisted by defects in apoptosis, since errors in DNA management typically disrupt cell cycle checkpoints, triggering a cell suicide response. Growth factor and hormone independence, hallmarks of many advanced cancers, can also be attributed in part to alternations in the cell death machinery, which permit cancer cells to thrive in the absence of these factors that cells normally require for maintenance of their survival. Metastasis is also assisted by defects in apoptosis, permitting tumour cells to survive in a suspended state (such as during circulation through the blood or lymph), whereas normal epithelial cells undergo apoptosis when detached from extracellular matrix and thus are confined to predefined locations in the body. Immune surveillance mechanisms are also thwarted by defects in apoptosis, since cytolytic T-cells and natural killer (NK) cells depend on components of cell death machinery to kill target cells. Finally, defects in apoptosis pathways contribute to resistance of cancer cells to chemotherapy and radiation, raising the threshold of drug- or radiation-induced damage necessary to trigger a cell suicide response (**Table 1**).

The core cell death machinery consists of families of genes and their encoded proteins, many of which are conserved throughout metazoan evolution (Metzstein *et al*., 1998) . Several protein domains which are entirely or nearly unique to apoptosis pathways are found within apoptosis-suppressing and -inducing proteins, including caspase protease domains, caspase-recruitment domains (CARDs), death domains (DDs), death effector domains (DEDs), caspase-inducible DNA endonuclease (CIDE) domains,

Table 1 Relevance of defective apoptosis to cancer: *pathogenesis, progression, therapy resistance*

Cell accumulation (cell death < cell division)
Longevity (accumulation of genetic lesions)
Genomic instability (tolerate DNA mistakes)
Immune surveillance (resistance to immune attack)
Growth factor/hormone independence (survival without paracrine/endocrine growth factors)
Angiogenesis (resistance to hypoxia; hypoglycaemia)
Metastasis (survival without attachment)
Chemoresistance/radioresistance (increased threshold for cell death)

Table 2 Protein domains associated with apoptosis

Caspase (catalytic) domains
BIR domains (IAPs)
Caspase-recruitment domains (CARDs)
Death domains (DDs)
Death effector domains (DEDs)
Bcl-2 homology (BH) domains
CIDE domains

BIR domains and Bcl-2 homology (BH) domains (**Table 2**) (Reed, 2000). Three-dimensional structures have been obtained for at least one of each of these domains, providing the foundation of a clearer understanding of the molecular mechanisms of apoptosis regulation and, in some cases, ideas for how one might modulate apoptosis proteins with therapeutic intent (Fesik, 2000).

APOPTOSIS IS CAUSED BY PROTEASES

The biochemical events responsible for apoptosis can be linked to the activation in cells of a family of cysteine proteases, known as caspases, which cleave specific target proteins in cells at aspartic acid residues. It is these proteolytic cleavage events which directly or indirectly explain the morphological changes that we recognize as 'apoptosis.' As many as 14 caspase-encoding genes have been identified in humans and mice. Homologous genes are also found in lower organisms within the animal kingdom, such as the nematode *Caenorhabditis elegans* and the fruit fly *Drosophila melanogaster*.

Caspases are initially synthesized as inactive zymogens in cells. How, then, are caspases activated? The answer is by proteolytic cleavage, occurring at conserved aspartyl residues, thus generating the large ($\sim$20 kDa) and small ($\sim$10 kDa) subunits that comprise the catalytic component of these enzymes. Structures of several caspases have been solved by X-ray crystallography, revealing a hetero-tetrameric assembly of two large and two small subunits, with two active sites per molecule. The observations that caspase zymogen activation involves cleavage at aspartyl residues and that active caspases also cut proteins at aspartyl residues have obvious implications. Namely, pro-caspases can become cleaved and activated either as a result of cleavage by other active caspases or through 'autoproteolytic' mechanisms which will be discussed below. Thus, once some caspase activation has occurred, the process can spread to other caspases through amplification steps in which one active caspase molecule cleaves and activates multiple caspase zymogens, as well as cascades of proteolytic activation of caspases through sequential stepwise mechanisms (Salvesen and Dixit, 1997; Thornberry and Lazebnik, 1998).

At least three mechanisms have been proposed for explaining how the 'first' caspase becomes activated: (1) the induced-proximity model; (2) the 'safety-catch' mechanism; and (3) introduction of an exogenous protease (**Figure 1**). In the induced-proximity model (**Figure 1**), two or more caspase zymogens are clustered together as a result of protein interactions, thus bringing them into close association. Because the zymogen forms of caspases are not entirely without proteolytic activity (some possessing about 1% of the activity of the processed 'active' enzymes), the close proximity allows these weakly active zymogens to trans-process each other, thereby generating the autonomously active proteases. This clustering of caspase zymogens typically is made possible by the presence in certain pro-caspases of an N-terminal prodomain which is located proximal to the portions of the pro-protein that give rise to the p20 (large) and p10 (small) catalytic subunits. These N-terminal prodomains serve as protein-interaction motifs that permit self-association of zymogens or that mediate interactions with adapter proteins which assemble multiprotein zymogen-activation complexes, sometimes referred to as 'apoptosomes' (Salvesen and Dixit, 1999).

Not all caspases possess significant N-terminal prodomains, thus creating a hierarchy of proteolytic cascades in which upstream caspases that possess large prodomains become activated as a result of protein interactions, and these initiator caspases then cleave and activate downstream effector caspases which possess only short N-terminal pro-peptides that are incapable for mediating protein interactions. This hierarchy of caspase proteolytic cascades is also supported by comparisons of the substrate preferences of upstream initiator and downstream target caspases, using combinatorial peptide library screening and other techniques. These studies have revealed the presence of tetrapeptide substrate cleavage sites preferred by upstream caspases within the sequences of the proforms of the downstream caspases, at sites corresponding to the authentic *in vivo* cleavage sites that separate the large and small subunits. The tetrapeptide cleavage sites preferred by the downstream caspases, in turn, are often found in several of the target proteins which are known to become cleaved at aspartyl residues during apoptosis, including

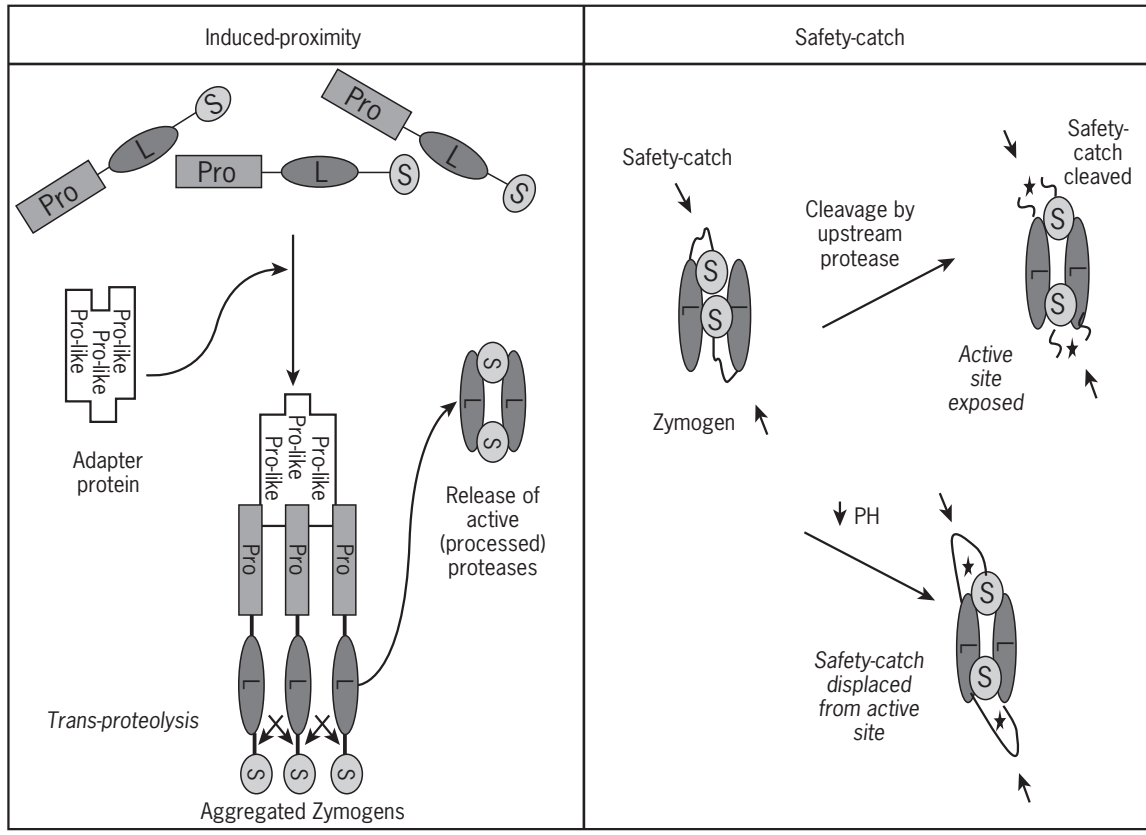


Figure 1 Mechanisms of caspase activation. Two of the proposed mechanisms of caspase activation are depicted, the induced-proximity model and the safety-catch model . In the induced-proximity model, pro-caspase zymogens are present as monomers that are single polypeptide chains with N-terminal prodomains (Pro) followed by the regions corresponding to the large (L) and small (S) catalytic subunits. Bringing these zymogens into close proximity, either by interactions with adapter proteins which typically share sequence and structural similarity with the prodomain regions of pro-caspases (shown) or by overexpression of pro-caspases resulting in self-aggregation (not-shown), allows these proenzymes to trans-process each other. This trans-processing is possible because the zymogens possess weak protease activity. Once proteolytically processed, separating the large and small subunit fragments and (often) removing the N-terminal prodomain, the active protease assembles, consisting of heterotetramer with two L and two S subunits. In the safety-catch mechanism, an autoinhibitory loop in the zymogen is envisioned as occupying or blocking the active site of the enzyme. Typically, cleavage at or near this autoinhibitory loop located between the L and S subunits exposes the active site (indicated by a star), creating the active protease (top). Alternatively, autorepression may be relieved by lowering pH, protonating acidic residues in the autoinhibitory loop and causing the loop to leave the active site of the enzyme (bottom).

the actin-regulatory protein gelsolvin, the chromatin-regulatory protein poly-ADP ribosyl polymerase (PARP), the endonuclease inhibitor ICAD and its relatives of the caspase-inducible DNA endonuclease (CIDE)-family proteins, nuclear lamins, certain protein kinases and various other substrate proteins (Thornberry and Lazebnik, 1998; Cryns and Yuan, 1999).

The safety-catch mechanism envisions an auto-repressing loop in the caspase zymogen which occupies the active site of the protease and maintains it in a highly constrained inactive state (**Figure 1**). Activation then is speculated to involve displacement of the auto-repressive loop from the active site, allowing the caspase possibly to auto-process and thereby auto-activate itself. This mechanism has been proposed for certain downstream effector caspases, such as caspase-3 and -7, which contain three adjacent acidic residues (aspartic or glutamic acid) in the candidate auto-repression loop. Neutralizing the negative charge of these acidic residues by lowering the pH or by site-directed mutagenesis (with conversion to uncharged alanine) lowers the barrier to caspase activation. It remains to be determined whether the proposed safety-catch mechanism can be extended to other caspases.

Finally, an example of caspase activation resulting from introduction of an exogenous protease has come from investigations of mechanisms of target cell killing by cytolytic T cells (Lowin *et al*., 1995). Here, perforin-mediated channels are used to inject granzyme B, a serine protease

with specificity (like caspases) for cleavage of substrates at aspartyl residues. Granzyme B is capable of directly cleaving and activating most caspase-family zymogens. Interestingly, however, intracellular trafficking of granzyme B appears to be under regulation by cellular factors that may limit its interactions with caspases in some scenarios.

Inactivation of caspases can occur through several mechanisms in cancers (**Table 3**). For example, structural alterations in caspase-family genes, such as deletions and loss of function mutations, have been documented in tumour specimens or cancer cell lines, including initiator and effector caspases, suggesting that inactivation or elimination of caspase-family genes represents one mechanism by which malignant cells may escape apoptotic elimination. In addition, some caspase-family genes, including caspase-2 and -9 in mammals, are capable of producing shorter protein isoforms through alternative mRNA splicing that function as trans-dominant inhibitors of their full-length counterparts, possibly by forming heterodimers composed of full-length and truncated procaspases or by competing for binding to upstream activators. Although little work has been performed to date, it is intriguing to speculate that tumours may overexpress the trans-dominant inhibitory isoforms of some caspases as an additional mechanism for subverting apoptosis. Protein phosphoryation is another mechanism capable of directly suppressing caspases in tumours. The only reported example of this so far is human caspase-9, which can become phosphorylated by the kinase Akt (protein kinase-B), thereby suppressing the active caspase-9 enzyme as well as suppressing activation of the pro-caspase-9 zymogen. Interestingly, elevated levels of Akt activity are observed in many tumours as a result of (1) amplification of the Akt-family genes, (2) increased signalling by upstream protein tyrosine kinases and Ras-family oncoproteins, (3) aberrant production of Akt co-activators and (4) inactivation of PTEN—the product of a tumour-suppressor gene which normally suppresses Akt activation by dephosphorylating second messenger polyphosphoinositol lipids (Datta *et al*., 1999). Theoretically, other post-translational modifications of caspases could also participate in their inactivation in cancers, including *S*-nitrosylation or glutathionylation of the active site cysteine.

Table 3 Mechanisms of caspase inactivation and suppression in cancers

Mechanism	Examples
Mutations in caspase genes	Caspase-3, -5, -8
Expression of dominant-negative caspase isoforms by alternative mRNA splicing	Caspase-2, -9
Phosphorylation	Caspase-9
Nitrosylation	Caspase-3
Overexpression of caspase-inhibiting IAPs	XIAP, cIAP-1, Livin

INHIBITOR OF APOPTOSIS (IAP) FAMILY PROTEINS FUNCTION AS CASPASE INHIBITORS

All protease networks studied to date include inhibitors which control flux through proteolytic cascades and which establish thresholds for protease activation which must be surpassed to trigger biological processes. The inhibitors of apoptosis proteins (IAPs) represent a family of anti-apoptotic proteins conserved throughout metazoan evolution that appear to serve this role. IAPs were first identified in the genomes of baculoviruses, where they suppress apoptosis induced by viral infection of host insect cells. Subsequently, cellular homologues of the baculovirus IAPs were discovered in humans, mice, flies and other animal species (Deveraux and Reed, 1999; Miller, 1999).

Membership in the IAP family requires two things: (1) an ability of the protein to suppress apoptosis, at least when overexpressed in cells, and (2) the presence of at least one copy of a conserved domain known as a BIR domain (‘baculovirus IAP repeat’). The BIR domain represents a zinc-binding fold. One to three copies of the BIR domain are found in IAP family proteins, sometimes in association with other domains such as RING fingers, putative nucleotide-binding domains and caspase recruitment domains (CARDs) (described below).

A single BIR domain can be necessary and sufficient for inhibition of certain caspases by human IAP-family proteins. For example, the second of three BIR domains found in the human XIAP protein (where X indicates that the gene maps to the human X-chromosome) directly binds to and potently inhibits ($K_i < 1\ \text{nmol}\ \text{L}^{-1}$) caspase-3 and -7. Thus, the concept has emerged that BIR domains represent caspase-inhibitory structures that bind active caspases and suppress them within cells. Interestingly, BIR domains do not inhibit all caspases, but rather exhibit clear selectivity. The second BIR domain of XIAP (in combination with adjacent residues), for instance, inhibits the downstream effector proteases, caspase-3 and -7, but does not bind or suppress the upstream initiator proteases, caspase-1, -8, -9 or -10. In contrast, the third BIR domain of XIAP suppresses caspase-9 but not other caspases tested. Although having caspase-inhibitory activity, at least some IAP family proteins may also participate in other processes, particularly signal transduction pathways involved in regulating kinases responsible ultimately for NF-κB induction and JNK activation. Contrasting the roles of IAPs as caspase inhibitors versus signal transduction modulators remains an active area of research.

Altered expression of IAP-family genes has been documented in cancers (LaCasse *et al*., 1998). For example, the *cIAP-2* gene becomes involved in chromosomal translocations in certain types of lymphomas, resulting in deregulation of this gene. The *cIAP-2* gene also is a target of the transcription factor NF-κB, a member of the

Rel family of oncoproteins which suppresses apoptosis. Moreover, a member of the IAP-family called survivin is inappropriately overexpressed in the majority of human cancers. Survivin is the smallest of the IAPs identified thus far, containing a single BIR domain in humans and mice. This particular IAP-family member is particularly intriguing because it appears to have a dual role in apoptosis suppression and in cell division. The survivin protein is physically associated with the mitotic spindle apparatus during M-phase and beyond into anaphase, evidently playing an essential role in cytokinesis. Homologues of survivin are found in yeast and in *C. elegans*, performing essential functions in cell division but lacking any clear role in cell death control. Thus, it is speculated that survivin's evolutionarily conserved core function is related to cytokinesis, and that an apoptosis checkpoint function for this protein may have evolved later in higher organisms as a way of linking defects in the late stages of cell division to an apoptotic response. Circumstantial evidence implicates survivin in caspase suppression in mammalian cells analogous to other IAPs, but a direct role for this BIR domain containing protein in caspase inhibition remains equivocal. Furthermore, not all BIR-containing proteins are involved in apoptosis regulation, indicating that this zinc-binding fold can serve alternative functions, despite a conserved arrangement of histidine and cysteine residues. Most likely, BIRs are protein-interaction domains that bind and inhibit caspases in some circumstances but not others.

Regardless of the mechanism, overexpression of survivin clearly reduces apoptosis in response to a variety of apoptogenic stimuli whereas antisense-mediated reductions in survivin expression or gene transfer-mediated expression of dominant-negative mutants of survivin sensitize tumour cells in culture to apoptosis induction by anticancer drugs. These findings, coupled with the observation that survivin is rarely expressed in normal adult tissues whereas survivin mRNA and protein levels are markedly increased in most cancers, have elevated the status of survivin as a potential drug-discovery target for cancer therapy (Altieri *et al*., 1999).

RECEPTOR-MEDIATED MECHANISMS OF CASPASE ACTIVATION

Several mechanisms of caspase activation have been elucidated, including the coupling of ligand binding at cell-surface receptors to caspase activation. Moreover, defects in at least some of these mechanisms have been uncovered in cancers.

The tumour necrosis factor (TNF) family of cytokine receptors includes at least six members that trigger apoptosis and which share a conserved protein-interaction domain in their cytosolic region which is essential for their cytotoxic activity (Ashkenazi and Dixit, 1998; Wallach *et al*., 1999). These TNF-family receptors include TNFR1, Fas (CD95), DR3 (weasle; tweak), DR4 (trail receptor-1; Apo2), DR5 (trail receptor-2) and DR6 (**Figure 2**). The conserved cytosolic domain is known as the 'death domain' (DD) and is comprised of a six α-helical bundle which forms trimers and possibly higher-order oligomers in response to ligation of the extracellular domains or simply when these receptors are overexpressed in the absence of their cognate ligands. Hereditary loss-of-function mutations or trans-dominant inhibitory mutations in the DD of Fas (CD95) have been associated with a lymphoproliferative disorder and autoimmunity in the *lpr*/*lpr* strain of mice and in humans with autoimmune lymphoproliferative syndrome (ALPS). Similarly, somatic mutations in the death domain of Fas have been detected in certain human neoplasms, including myelomas, lymphomas, leukaemias and carcinomas of the lung and bladder.

The DDs of TNF-family death receptors bind adaptor proteins which interact, in turn, with the N-terminal domains of specific initiator caspases (**Figure 2**). The assembly of these multiprotein complexes is thought then to activate the associated pro-caspases by the induced-proximity model described above. Fadd (Mort1) represents one such adaptor protein. This protein contains a DD which binds directly to the DD of Fas, as well as a protein interaction module known as a death effector domain (DED). The DED is similar to the DD in structure, comprised of six α-helices, but constitutes a separate domain family which can be differentiated by sequence homology. Two of the known human caspases contain DEDs in their N-terminal prodomain regions, pro-caspase-8 and -10. These pro-caspases each contain two DEDs upstream of the catalytic segments and are both capable of binding directly to the DED of Fadd. Studies of cells derived from caspase-8 knockout mice have revealed an obligatory role for this initiator caspase in apoptosis induction by Fas and TNFR1. Similarly, gene ablation studies indicate an essential role for Fadd in apoptosis induction by these TNF-family death receptors.

Another example of a caspase-activating adapter protein that associates with TNF-family receptor complexes is Raidd (Cradd). The Raidd protein contains a DD and another protein interaction module known as a CARD domain. The CARD domain of Raidd specifically binds a homologous domain found in the N-terminal prodomain region of pro-caspase-2. Structures of these CARD domains have been solved, revealing again a characteristic fold comprised of six α-helices and suggesting that complementary patches of acidic, basic and hydrophobic residues on the surfaces of these domains account for the selectivity of their interactions with each other but not with other CARD-family proteins. In most cases, interactions of Fadd and Raidd with TNF-family receptors is mediated by an intermediate DD-family protein, Tradd, via associations of their DDs.

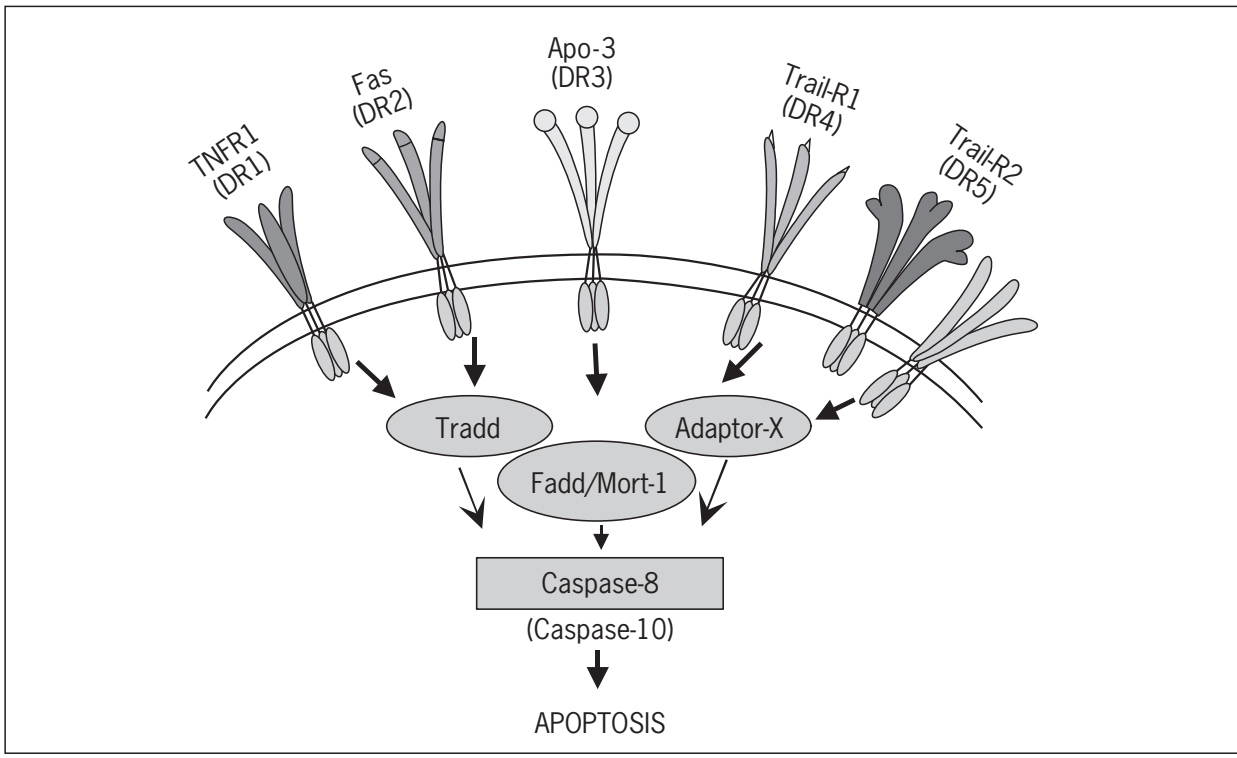


Figure 2 Caspase activation by TNF-family receptors. The currently known human members of the tumour necrosis factor (TNF) receptor family which contain cytosolic death domains (DDs) are depicted. The intracellular DDs of these receptors bind DD-containing adaptor proteins, such as Tradd, Fadd and others yet to be identified. These adaptor proteins then interact with initiator caspases which contain homologous protein interaction domains. The TNF family of receptors provided the first example of a receptor system which transduces signals into cells via proteolysis, thus representing a milestone in cell biology research.

Additional proteins which can participate indirectly in recruitment of pro-caspases to receptors include RIP and RIP-2 (Cardiac) which contain a conserved kinase domain of unknown functional significance in association with either a DD or CARD, respectively. Interestingly, whereas the CARD domain of RIP-2 binds the N-terminal CARD domain of pro-caspase-1, other portions of this protein reportedly interact with TRAF6, an adapter protein that associates with certain TNF-family receptors and Toll-family receptors (**Figure 3**). It should be noted, however, that caspase-1 and its close relatives, human caspase-4 and -5, may be only tangentially involved in apoptosis, and instead are principally involved in proteolytic processing of proinflammatory cytokine precursors, particularly pro-interleukin-1β and pro-IL-18. Caspase-1 knockout mice, for example, are grossly normal and exhibit only very modest defects in apoptosis induction, while manifesting extreme resistance to endotoxin-induced sepsis and displaying profound deficiencies in production of IL-1 and IL-18 *in vivo* (Reed, 1998a).

TNF-family death receptors play important roles in immune system interactions with tumours. One of the principal weapons used by cytolytic T cells for killing tumour targets, for example, is Fas-ligand (FasL). Consequently, defects in Fas-induced apoptosis can contribute to tumour avoidance of immune surveillance mechanisms. Moreover, upon achieving a Fas-resistant state, it has been shown that some tumours can then tolerate expressing FasL on their surfaces, thus using this death ligand as a weapon to kill neighbouring normal cells as well as activated lymphocytes.

Multiple mechanisms of tumour resistance to Fas and other TNF-family death ligands have been elucidated, including prereceptor, receptor and postreceptor defects (**Table 4**) (Tschopp *et al*., 1998, 1999; Ashkenazi and Dixit, 1999). Prereceptor defects include the production of soluble 'decoy' receptors or fragments of receptors that compete with the transmembrane receptors for ligand binding. For instance, a soluble form of Fas can be produced by alternative mRNA splicing in which the exon encoding the transmembrane anchoring domain is skipped, resulting in production of a secreted version of this receptor. In the case of Trail (Apo2L), three genes have been identified which produce 'decoy' receptors that can compete with the death receptors, DR4 (TrailR1) and DR5 (TrailR1), for binding to Trail, thereby sparing cells from the apoptotic effects of this death ligand. Proteolytic removal of the extracellular domain of TNF-family receptor is also a possibility. Another potential resistance mechanism is found in the SODD (silencer of death domains) protein. SODD (also

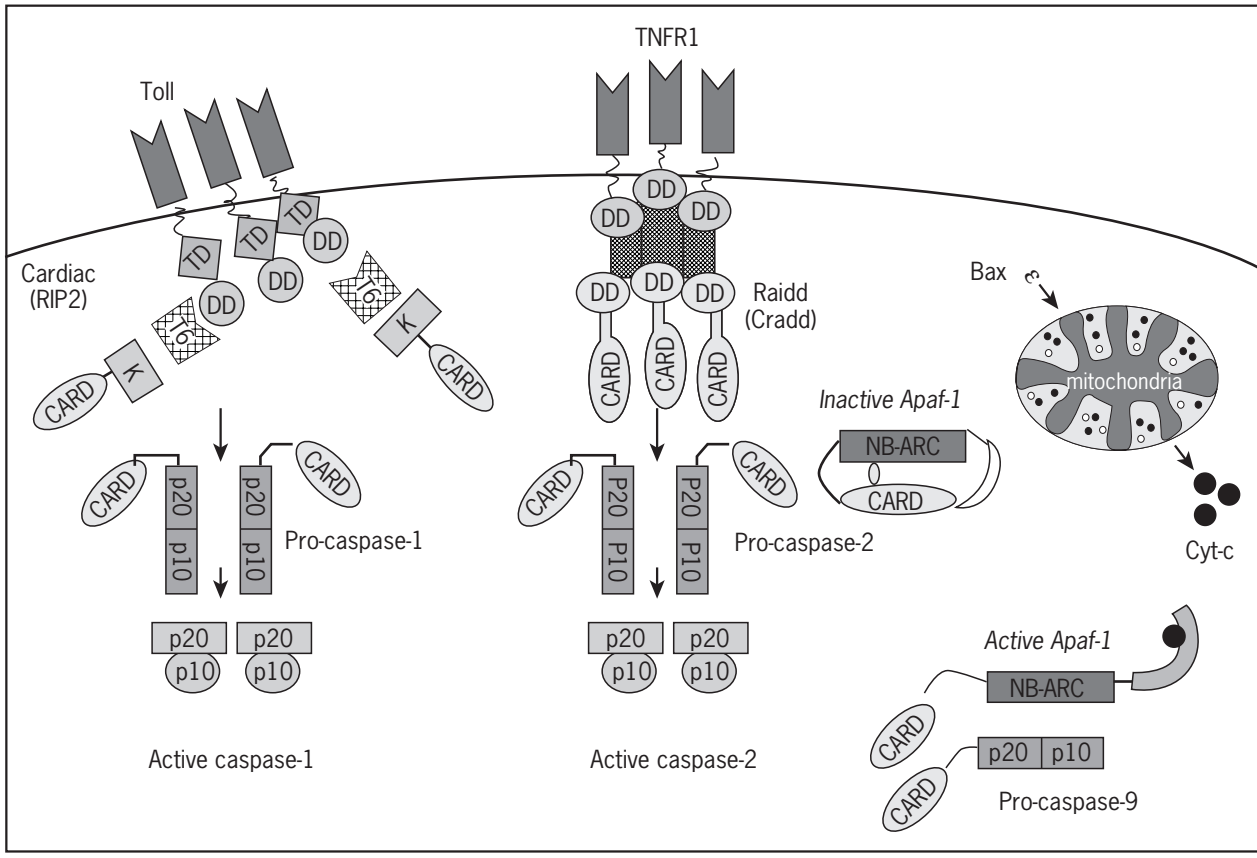


Figure 3 CARD domains and caspase activation. Examples of involvement of CARD-domain proteins in activation of CARD-containing caspases are presented, including examples from the Toll-receptor family (left), tumour necrosis factor receptor-1 (TNFR-1) (middle) and mitochondrial/cytochrome *c* (cyt-c) pathway (right). For each example, the target caspase contains an N-terminal prodomain consisting of a CARD. These CARD prodomains interact with homologous CARDs in various adapter proteins, such as Cardiac (RIP2; Rick) (left) which contains both a CARD and a kinase (K) domain of unknown significance, Raidd (Cradd) (middle) which contains both a CARD domain and a DD, the latter of which binds homologous DD-containing adapter proteins which associate with the TNF-receptor, and Apaf-1 (right), which contains a CARD, followed by a nucleotide-binding oligomerization domain (NB) and then WD repeats which mediates interactions with cyt-c. Release of cyt-c from mitochondria is required for converting Apaf-1 from a latent (inactive) conformation, resulting in activation of Apaf-1 so that it can bind pro-caspase-9 (see **Figure 5** for more details about Apaf-1).

known as BAG4) contains a domain that binds the DDs of TNFR1 and DR3, as well as a conserved Hsp70/Hsc70-binding domain. When overexpressed SODD prevents spontaneous aggregation of DDs, presumably by recruiting Hsp70/Hsc70-family molecular chaperones to these receptors and inducing conformational changes that prevent them from oligomerizing and signalling in the absence of ligand.

Table 4 Mechanisms for interfering with TNF-family death receptors

Mutations in death receptor or ligand genes
Decoy or soluble receptors
SODD (BAG4) suppression of receptor oligomerization
Antiapoptotic DED-family proteins (cFlip; BAR; Bap31)
NF-κB-mediated upregulation of caspase inhibitory proteins

Several postreceptor defects in death receptor signalling have also been uncovered in cancers. For example, FAP-1 is a Fas-binding protein phosphatase that, when overexpressed, can suppress Fas-induced apoptosis through an unidentified mechanism. Some tumours appear to overexpress FAP-1. Moreover, peptidyl antagonists that bind PDZ domains in FAP1 and that block interactions of FAP-1 with Fas have been shown to restore Fas sensitivity to Fas-resistant colon cancer cell lines *in vitro*. Additional antagonists of death receptor signalling include the Flip family of DED-containing proteins. These antiapoptotic proteins contain DEDs that permit them to bind either adapter proteins such as Fadd- or DED-containing procaspases such as pro-caspase-8 and -10. Overexpression of cellular Flip (also known as I-Flice, Cash, Casper, Usurpin, Mrit, Rick) or of its viral homologues can prevent assembly

of death receptor signalling complexes and interfere with caspase activation. Overexpression of c-Flip has been documented in some tumours, and has been associated in animal models with resistance to immune-mediated suppression of cancer. In addition to cytosolic c-Flip and related proteins, membrane-anchored proteins which contain DEDs capable of binding pro-caspase-8 and/or -10 have been identified. Proteins such as BAR and Bap31 localize to internal membranes, primarily endoplasmic reticulum and/or mitochondria, and can modulate apoptosis signalling by TNF-family death receptors. Finally, activation of the transcription factor NF-κB has been associated with resistance to apoptosis induction by several TNF-family death receptors. NF-κB transcriptionally upregulates the expression of several antiapoptotic proteins, including one or more of the IAP-family proteins, resulting in apoptosis resistance. This observation is directly relevant to mechanisms of TNF-induced apoptosis in that several TNF-family receptors recruit proteins that activate caspases as described above, thus promoting apoptosis, but simultaneously bind other proteins which trigger NF-κB induction, thus preventing apoptosis. The net outcome of engaging TNF-family receptors such as TNFR1, DR3 and DR6 is difficult to predict and can be extremely cell-type and cell-context dependent. Taken together, it is clear that malignant cells have many options for developing mechanisms that thwart apoptosis induction by TNF-family death receptors and ligands.

MITOCHONDRIAL PATHWAYS FOR CASPASE ACTIVATION

In addition to receptor-mediated mechanisms for coupling caspase activation to ligation of specific cell-surface receptors, a pathway has been elucidated which links mitochondrial damage to a mechanism for triggering caspase activation. This mitochondrial pathway for caspase activation is engaged in response to growth factor deprivation, genotoxic injury, hypoxia and many other insults, and is commonly referred to as the 'intrinsic' pathway, in contrast to the receptor-mediated caspase activation mechanisms which are sometimes referred to as the 'extrinsic' pathway (**Figure 4**) (Green and Reed, 1998; Kroemer and Reed, 2000).

What allows mitochondria to trigger caspase activation?—release of cytochrome *c* from these organelles into the cytosol. Cytochrome *c* is best known for its role in electron-chain transport, where it transfers electrons from complex III to IV in the respiratory chain. However, following exposure of cells to many apoptotic stimuli, the outer membranes of mitochondria undergo permeability changes that permit cytochrome *c* and other proteins normally sequestered in the space between the inner and outer membranes of these organelles to leak out and enter the cytosol. Once in the cytosol, cytochrome *c* binds a caspase-activating protein, called Apaf-1 (apoptotic protease activating factor-1). Apaf-1 is normally present in the cytosol in an inactive (latent) state in mammals and *Drosophila*. This protein consists of an N-terminal CARD domain, followed by a nucleotide-binding domain, ending with multiple WD repeat domains. The human Apaf-1 protein has been shown to oligomerize, apparently forming octamers, upon binding cytochrome *c*. This oligomerization is mediated by the nucleotide-binding domain and requires adenine triphosphate (ATP) or deoxy-ATP. Cytochrome *c* appears to interact with the WD repeat domains and nucleotide-binding domain, promoting a conformational change that results in oligomerization. Once oligomerized, activated oligomerized Apaf-1 molecules bind via their CARD domains to pro-caspase-9, a member of the caspase family which possesses an N-terminal CARD prodomain that has been shown by X-ray crystallography to possess a complementary interaction surface for association with the CARD of Apaf-1 (the interactions in this case are primarily electrostatic, involving associations of a surface patch of acidic residues on the CARD of Apaf-1 with basic residues on the CARD of pro-caspase-9). Activation of pro-caspase-9 then is achieved by the affinity-approximation model. However, rather than releasing processed caspase-9 to cleave and activate additional downstream effector caspases, instead caspase-9 must remain bound to Apaf-1 for maintaining its optimal protease activity. The 'free' processed caspase-9 molecule has only weak protease activity in the absence of Apaf-1, suggesting that the Apaf-1:caspase-9 complex represents the holoenzyme (**Figure 5**). Thus, the next caspase in the cascade, pro-caspase-3, is recruited to the Apaf-1:caspase-9 holoenzyme, where it becomes activated by cleavage at a conserved aspartyl residues separating the large and small subunits of the protease. Processed caspase-3 is then released from the Apaf-1:caspase-9 complex and autocatalytically cleaves and removes its own prodomain, which consists of a short ($\sim$20 amino-acid) segment, thereby generating the mature enzyme. Active caspase-3 then cleaves and activates additional procaspases, thus amplifying the proteolytic cascade. This and other downstream caspases also cleave a wide variety of specific substrates that ultimately commit the cell to an apoptotic demise, including CIDE-family proteins which control the activity of an endonuclease which appears to be largely responsible for the fragmentation of the nuclear DNA into oligo-nucleosomal length fragments—a hallmark (although non-obligatory event) of apoptosis.

Mechanisms for regulating the cytochrome *c*-dependent activation of Apaf-1 may exist, thus providing a means of fine tuning the coupling of mitochondria to caspases (Reed and Paternostro, 1999). For example, the efficiency of cytochrome *c*-mediated activation of caspases has been shown to be poor under conditions of isotonic salt $\sim$150 mmol L^{-1} (KCl) and neutral pH (7.4). Release of another protein from mitochondria, Smac (second mitochondrial activator

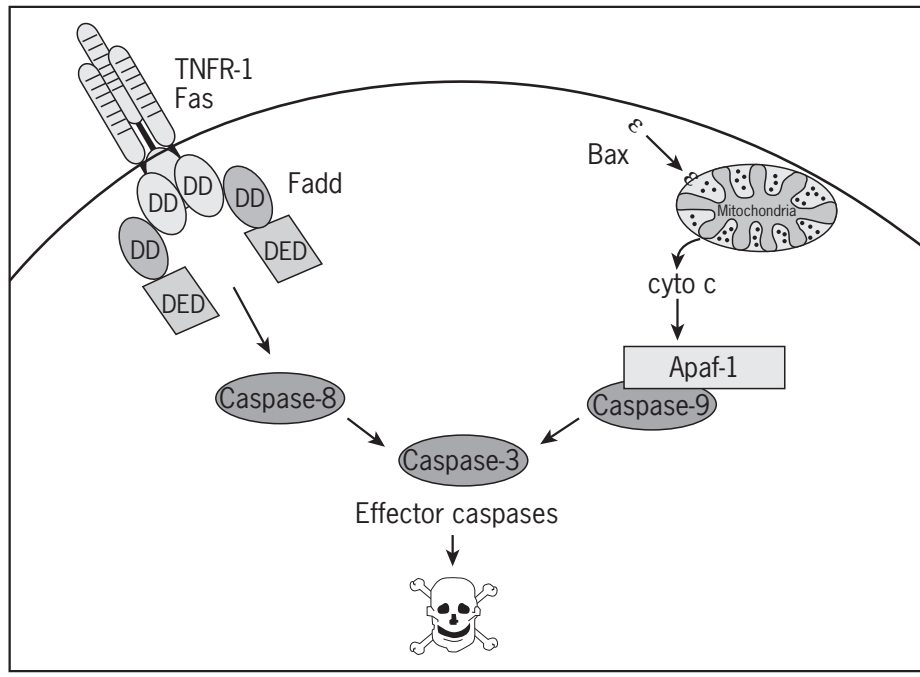


Figure 4 Intrinsic and extrinsic pathways for apoptosis. Two of the major caspase-activation pathways are depicted. The 'extrinsic' pathway is activated by extracellular ligands that bind death receptors such as TNFR-1 and Fas. These death receptors bind, in turn, to adaptor proteins such as Fadd (Mort1) which then bind the initiator pro-caspase, caspase-8. The 'intrinsic' pathway is activated by intracellular signals which induce cyt-c release from mitochondria. Cyt-c binding to the caspase-activator, Apaf-1, results in binding to and activation of pro-caspase-9. Gene ablation studies in mice have shown that pro-caspase-8 and pro-caspase-9 are absolutely required for apoptosis induction by TNF-family death receptors and cyt-c, respectively. Similarly, gene knockout studies in mice indicate that Fadd (Mort1) is critical for caspase activation by most (although perhaps not all) TNF-family death receptors and that Apaf-1 is absolutely required for caspase-activation by cyt-c. Many downstream effector caspases may be cleaved and activated by active caspase-8 and -9, with caspase-3 representing the best documented direct substrate of these upstream initiator caspases.

of caspases), can overcome the dependence on hypotonic salt. Furthermore, some apoptotic stimuli induce mitochondria to extrude protons (H^+), thus acidifying the cytosol and promoting cytochrome-mediated caspase activation. It may be of significance consequently that many growth and survival factors are known to induce cytosol alkalinization as a result of effects on the Na^+/H^+ antiporter, perhaps accounting for some of their antiapoptotic effects. Overproduction of Hsp70 also may interfere with cytochrome *c*-mediated activation of caspases, a finding of possible relevance to cancer given evidence of elevated Hsp70 levels in several kinds of cancer.

Also of relevance to activation of pro-caspase-9, the apical caspase in the mitochondrial pathway, at least two other proteins have been reported to bind pro-caspase-9 and to induce apoptosis through what appears to be cytochrome *c*-independent mechanisms, including an Apaf-1 homologue (CARD4/Nod) and Bcl-10 (hE10; CIPER). Both CARD4 (Nod) and Bcl-10 (hE10; CIPER) possess N-terminal CARD domains that interact with the CARD of pro-caspase-9. Interestingly, chromosomal translocations and somatic point mutations have been described in the *bcl-10* (hE10; CIPER) gene in lymphomas and solid tumours, respectively, which convert it from a pro-apoptotic to an anti-apoptotic protein.

Besides these caspase-9-binding proteins, another way of potentially modulating signalling through caspase-9-dependent pathways involves sequestration of this pro-protease in organellar compartments where it cannot interact with cytosolic proteins such as Apaf-1, CARD4 (Nod) and Bcl-10 (hE10; CIPER). Specifically, pro-caspase-9 has been shown to be stored in the inter-membrane space of mitochondria, along with cytochrome *c*, in some types of terminally differentiated cells. Thus, changes in mitochondrial membrane permeability are required to release pro-caspase-9 into the cytosol where its interacting proteins are located, at least in some cell types. Finally, caspase-9 in humans is a substrate of the kinase Akt, with phosphorylation inhibiting its protease activity through undetermined mechanisms.

Loss of expression of Apaf-1 has been described in melanomas, due to gene hypermethylation. In this regard, in cancers, *in vitro* transformation studies using cells from Apaf-1 and caspase-9 knockout mice suggest that these

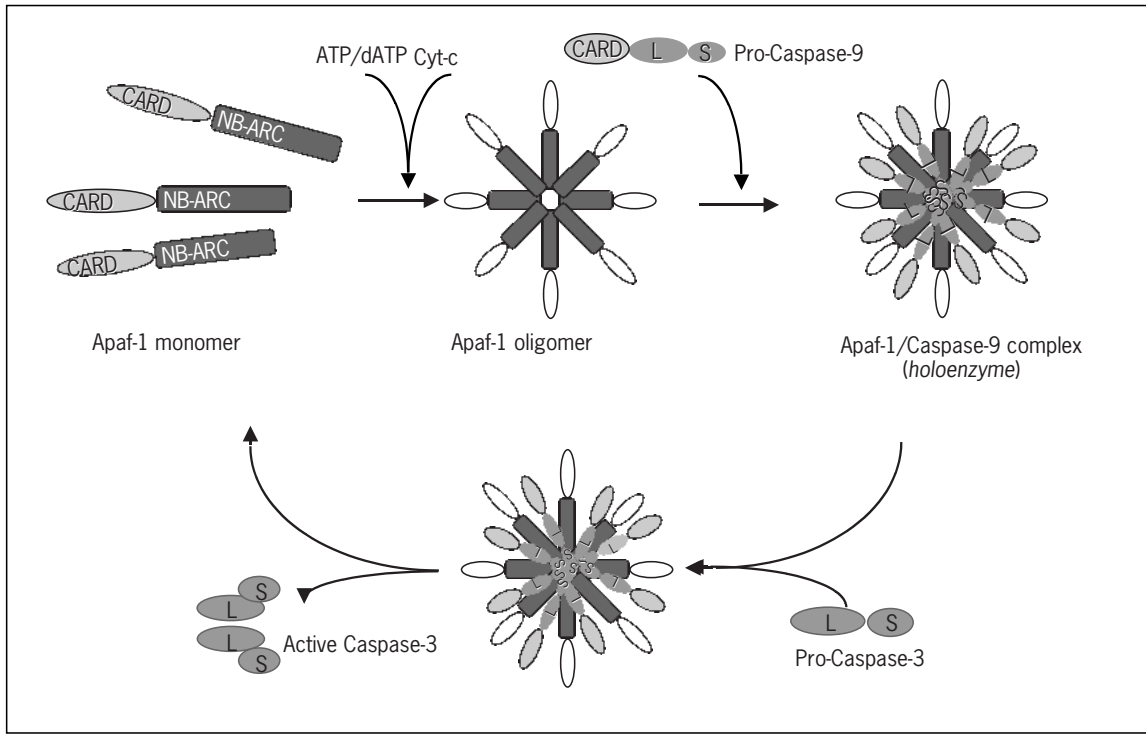


Figure 5 Dynamics of Apaf-1 'Apoptosome.' The sequence of events involved in cyt-c-mediated activation of Apaf-1 and pro-caspase-activation are depicted. Cyt-c plus either ATP or deoxy ATP (dATP) binding to Apaf-1 results in oligomerization mediated by the nucleotide-binding domain (NB-ARC) of this protein. Present estimates suggest that activated Apaf-1 may form an octamer. Oligomerized Apaf-1 binds pro-caspase-9, via a CARD–CARD interaction. Apaf-1-associated pro-caspase-9 molecules are then thought to transprocess each other, cleaving between the large (L) and small (S) subunits. Unlike most other caspases, caspase-9 does not remove its prodomain (CARD) by autoproteolysis. Also, unlike most caspases, after proteolytic processing, caspase-9 must remain associated with its activator Apaf-1 to maintain optimal protease activity. The Apaf-1/caspase-9 complex, therefore, represents the 'holoenzyme' complex. This complex directly recruits pro-caspase-3, which is cleaved by caspase-9, releasing active and processed caspase-3.

pro-apoptotic genes function as tumour suppressors within the p53 pathway. This finding is consistent with other data that suggest that p53 induces apoptosis primarily through a mitochondria-dependent mechanism (intrinsic pathway), although the death receptor (extrinsic pathway) may also be involved in some circumstances, as discussed below.

Bcl-2 FAMILY PROTEINS—REGULATORS OF CYTOCHROME *c* RELEASE AND MORE

A large family of evolutionarily conserved proteins has been identified which regulates the release of cytochrome *c* and other proteins from mitochondria (Adams and Cory, 1998; Reed, 1998b; Gross *et al*., 1999). The founding member of this family is known as Bcl-2, representing an acronym for B-cell lymphoma-2 and reflecting the original discovery of this gene because of its involvement in chromosomal translocations in B-cell lymphomas. In follicular non-Hodgkin lymphomas, the *Bcl-2* gene on the long arm of chromosome 18 at band q21 frequently becomes fused with the immunoglobulin (Ig) heavy-chain gene locus on the long-arm of chromosome 14 at band q32, creating t(14;18) (q32; q21) translocations. Because the Ig gene locus is highly active in B cells, the juxtaposed Bcl-2 gene become transcriptionally deregulated, resulting in continuous production of high levels of *Bcl-2* mRNAs and protein. Bcl-2 is an anti-apoptotic protein that uses a C-terminal membrane-anchoring domain to insert into the membranes of mitochondria as well as some other membranes inside cells. Elevated levels of Bcl-2 prevent cytochrome *c* release from mitochondria following exposure of cells to a wide variety of apoptotic agents and conditions, including growth factor deprivation, oxidants, Ca^{2+} overload, chemotherapeutic drugs and X-irradiation. The mechanism by which Bcl-2 family proteins control cytochrome *c* released from mitochondria remains unknown, although several theories have been advanced and are the subject of several reviews devoted entirely to this important topic (Adams and Cory, 1998; Green and Reed, 1998; Reed, 1998b; Gross *et al*., 1999; Vander Heiden and Thompson, 1999) (**Figure 6**).

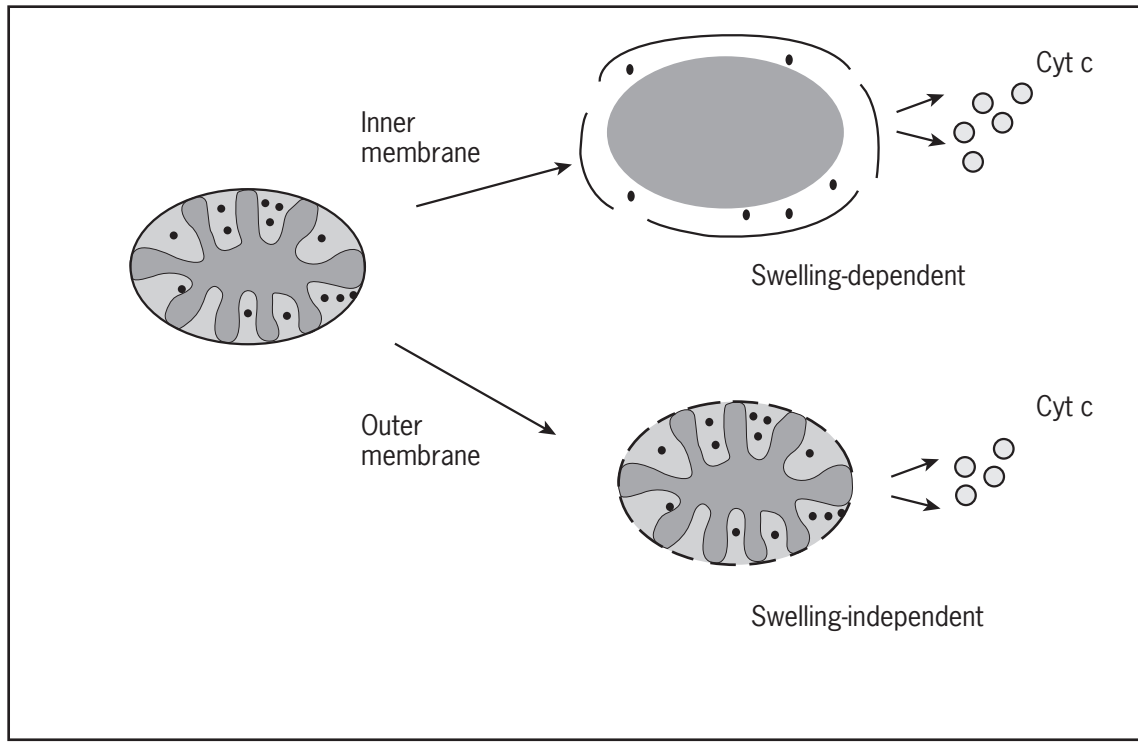


Figure 6 Alternative routes of cytochrome *c* release from mitochondria. Two alternative mechanisms are depicted for inducing release of cyt-c (shown) and other apoptosis-relevant proteins, such as AIF, Smac and intra-mitochondrial caspases (not shown) from the inter-membrane space of mitochondria. The swelling-dependent mechanism (top) involves alterations to the permeability of the inner membrane, causing osmotic disequilibrium and swelling of the matrix space. Because the surface area of the inner membrane, with its folded cristae, is greater than the area of the outer membrane, swelling results in rupture of the outer membrane and release of proteins stored in the inter-membrane space. In the swelling-independent model (bottom), a selective change in outer-membrane permeability occurs, allowing release of cyt-c and other proteins from the inter-membrane space, while inner-membrane integrity and volume homeostasis are preserved.

The Bcl-2 family consists of both anti-apoptotic members, which include in humans Bcl-X_L, Bcl-W, Bfl-1, Bcl-B and Mcl-1, as well as pro-apoptotic members, which include in humans Bax, Bak, Bok, Bad, Bid, Hrk, Bik, Bim, Nip3 and Nix. Homologues of some of these proteins are also found in lower organisms, including *Caenorhabditis elegans* and *Drosophila melanogaster*, where they often play important roles in developmental programmed cell death. Evolutionary conservation of function has been documented by cross-species gene transfer experiments, implying commonality in their mechanisms of action.

Antiapoptotic Bcl-2 family members may possess multiple mechanisms for suppressing cell death. In general, however, these mechanisms can be simplified into two general categories. First, based on determination of their three-dimensional structures or computer modelling predictions of structures, antiapoptotic Bcl-2 family proteins are recognized to be similar to certain types of α-helical ion-channel or pore-forming proteins. Specifically, these Bcl-2 family proteins share structural similarity with the pore-forming domains of certain bacterial toxins which have been implicated in transport of either ions (colicins) or proteins (diphtheria toxin) across membranes. At least *in vitro*, antiapoptotic Bcl-2 family proteins such as Bcl-2 and Bcl-X_L have been documented to form multiconductance ion channels in synthetic membranes, thus providing further experimental evidence in support of a role as channel/pore proteins. Second, antiapoptotic Bcl-2 family proteins also possess at least one hydrophobic pocket on their surface which mediates interactions with other proteins, thus altering the function or intracellular targeting of other proteins which may be relevant in some contexts to cell death. Proteins reported to interact directly or indirectly with antiapoptotic Bcl-2 family members such as Bcl-2 or Bcl-X_L include the caspase activator Apaf-1, the p53-binding protein p53BP2, the Ca^{2+}-dependent phosphatase calcineurin, the protein kinase Raf-1, the Hsp70/Hsc70 molecular chaperone regulators Bag1 and Bag3, the spinal muscular atrophy gene product Smn, the DED-containing proteins BAR and Bap31, and others. It is currently unknown which of these two major functions of antiapoptotic Bcl-2 family proteins is more important for their overall cytoprotective effects in cells. It is clear,

however, that Bcl-2 and Bcl-X_L possess cytoprotective functions which are independent of caspase-family cell death proteases and which can be manifested even in yeast, which lack the various Bcl-2/Bcl-X_L-interacting proteins described above. Hence it seems likely that these proteins possess an intrinsic biochemical function. Additional ancillary functions related to interactions with other protein may have been added over evolutionary time to this core-intrinsic function, as a means of integrating multiple pathways with cell life and death decision making (Reed, 1997).

Proapoptotic Bcl-2 family proteins are more structurally diverse than their antiapoptotic counterparts, probably reflecting differences in the mechanisms by which they promote cell death. A subgroup of proapoptotic Bcl-2 family proteins, which in humans includes Bax, Bak, Bok and Bid, appears to possess a similar pore/channel-like protein fold as the antiapoptotic members. Indeed, where tested, these pro-apoptotic Bcl-2 family members have been documented to form multiconductance, sometimes fairly large, channels in synthetic membranes *in vitro*. They also induce cytochrome *c* release from mitochondria when overexpressed in cells as well as when added (where tested) as recombinant proteins to isolated mitochondria *in vitro*. It remains controversial as to the mechanisms that explain why Bax, Bak, Bok and Bid are proapoptotic and induce cytochrome *c* release from mitochondria, whereas the structurally similar antiapoptotic proteins Bcl-2, Bcl-X_L, Bcl-W, Mcl-1, Bcl-B, and Bfl-1 are cytoprotective and block cytochrome *c* release.

By analogy with the bacterial proteins with which they share structural homology, it has been suggested that Bcl-2 family proteins may exist in two dramatically different conformations—one in which two of more α-helices of these proteins are integrated into membranes and another in which they are not integrated into membranes, although they are often tethered to membranes via their C-terminal membrane-anchoring domains. These two conformational states may then dictate whether these proteins are actively forming channels or perhaps interacting with other channel or pore-like proteins in mitochondrial membranes (inserted state) versus whether they assume conformations necessary for interacting with certain other proteins (noninserted state). Thus, the concept of on/off (active/inactive) conformations has emerged as an element in understanding the regulation of Bcl-2 family proteins—which may not correlate necessarily with the levels of expression of these proteins.

Apoptosis (caspase-dependent death) and necrosis (caspase-independent death) share some elements of overlap and the release of cytochrome *c* from mitochondria represents one of these. When cytochrome *c* exits mitochondria, not only can apoptosis be induced through the Apaf-1-dependent mechanisms described above, but also necrosis can ensue owing to failure of mitochondrial electron chain-transport, resulting in inadequate ATP production, increased generation of oxygen free radicals and other disturbances. Several Bcl-2 family proteins have been shown to regulate both apoptosis and necrosis, probably at least in part owing to their direct caspase-independent effects on mitochondrial membrane permeability. Although many mechanisms remain unclear, it can be deduced that these direct effects of Bcl-2, Bax and certain other Bcl-2 family proteins on mitochondria are also independent of CED-4/Apaf-1-family proteins, given that (1) caspases are required neither for induction of cytochrome *c* release by Bax, Bax, and Bok nor for prevention of cytochrome *c* release by Bcl-2 or Bcl-X_L and (2) the only known function of CED-4/Apaf-1 family proteins is regulation of caspases. Interestingly, a variety of data have provided evidence that clonogenic survival often correlates better with caspase-independent cell death than with caspase-dependent apoptosis, suggesting that a cell death commitment point which is regulated by Bcl-2 family proteins exists upstream of caspase activation in the mitochondrial pathway (Green and Reed, 1998).

Members of another subgroup of pro-apoptotic Bcl-2 family proteins, which in humans includes Bad, Bik, Bim and Hrk, lack similarity to pore-forming domains but do possess a short ($\sim$16–20 amino acid) domain found in nearly all Bcl-2 family proteins known as a Bcl-2 homology-3 (BH3) domain. The BH3 domain participates in protein dimerization among Bcl-2 family members, and therefore represents a functionally important region that permits many pro- and antiapoptotic Bcl-2 family proteins to antagonize each other's functions through physical hand-to-hand combat. The BH3 domain consists of a conserved amphipathic α-helix which inserts into the hydrophobic receptor-like pockets found on (or predicted to exist on) the antiapoptotic Bcl-2 family proteins. Bad, Bik, Bim and Hrk all bind selected antiapoptotic Bcl-2 family members via their BH3 domains, thus antagonizing the cytoprotective functions of Bcl-2, Bcl-X_L or similar proteins, but these killer proteins do not interact with themselves, nor do they bind to pro-apoptotic Bcl-2 family members such as Bax, Bak, Bok and Bid. Specificity exists in the preferences for dimerization partners within the network of Bcl-2 family protein interactions, thus creating opportunities for fine-tuning regulation through tissue-specific and temporally dynamic differences in the repertoires of family members expressed.

Several ways of regulating dimerization among Bcl-2 family proteins have been documented. One obvious way is by controlling the levels of various Bcl-2 family members at the transcriptional or post-transcriptional levels, as has been demonstrated for several members of the family. Another mechanism for controlling dimerization involves protein phosphorylation/dephosphorylation. For example, the proapoptotic protein BAD is a substrate for phosphorylation by several protein kinases, including Akt, Pak1, PKA and Raf-1. When phosphorylated on certain sites, BAD is unable to dimerize with Bcl-2 or Bcl-X_L,

thus abrogating its proapoptotic function. Elevations in Akt activity in cancers, therefore, suppress apoptosis at least in part through effects on BAD, at least in those cases where BAD is expressed. Conversely, sustained elevations in cytosolic Ca^{2+}, resulting in activation of the protein phosphatase calcineurin, have been shown to result in dephosphorylation and reactivation of BAD. In addition to phosphorylation/dephosphorylation, dimerization among some Bcl-2 family proteins can be regulated by proteolysis. For instance, the proapoptotic protein Bid is normally present in the cytosol in a latent state. Removal of an N-terminal 58 amino acid domain by caspase-mediated cleavage activates the Bid protein, exposing its BH3 domain, thus allowing it to bind other Bcl-2 family proteins including Bcl-2, Bcl-X_L, Bax and Bak. The caspase responsible for Bid cleavage is caspase-8, thus providing a mechanism for cross-talk between the death receptor ('extrinsic') and mitochondrial ('intrinsic') cell death pathways. Yet another mechanism of regulating dimerization among Bcl-2 family proteins can be found in the Bim protein. The longer isoforms of Bim are present in a complex with dynein light chain in association with microtubules, thus sequestering Bim proteins in a location where they cannot interact with other Bcl-2 family members. Release of Bim proteins from microtubule-associated protein complexes allows their translocation to the surfaces of mitochondria and other organelles where antiapoptotic Bcl-2 family dimerization partners reside. Release of Bim from microtubules may be relevant to some of the mechanisms by which microtubulin-interacting anticancer drugs induce apoptosis (Huang and Strasser, 2000).

Many examples of altered expression or structure of Bcl-2 family proteins have been documented in human cancers. In addition to its inappropriate overexpression in lymphomas containing the t(14;18) translocations described above, pathological overexpression of Bcl-2 (without attendant structural changes to the gene) has been estimated to occur in roughly half of all human cancers, including most advanced hormone-refractory prostate cancers, two-thirds of breast cancers and over half of all colon and lung cancers. Examples of upregulation of other antiapoptotic proteins such as Bcl-X_L or downregulation of proapoptotic proteins such as Bax and Bak also exist in human cancers. Moreover, mutations that inactivate the *BAX* gene have been detected. One of the more common mutations involves a homopolymeric stretch of eight guanosine residues, which occurs frequently in cancers that manifest the microsatellite instability phenotype due to errors in DNA mismatch repair enzymes or their regulators. Consistent with their important role in controlling apoptosis sensitivity to chemotherapeutic drugs and X-irradiation, the levels of Bcl-2, Bax or other Bcl-2 family proteins are of prognostic significance for some subgroups of patients with certain types of cancer.

INTERACTIONS OF SIGNAL TRANSDUCTION PATHWAYS WITH APOPTOSIS PATHWAYS

Many upstream inputs exist which link cellular responses to various stimuli with the core components of the apoptosis machinery. An exhaustive review of these connections is beyond the scope of this chapter, but a few deserve special mention.

The transcription factor NF-κB has been implicated in apoptosis suppression in many contexts and appears to be elevated in its activity in many types of cancer (Karin and Ben-Neriah, 2000). NF-κB can consist of various dimerizing pairs of Rel-family members, with the best studied representing a heterodimer of p50 and p65. NF-κB directly upregulates the transcription of several antiapoptotic proteins, including (1) the Bcl-2 family members Bcl-X_L and Bfl-1, (2) the IAP-family member cIAP-2, and possibly cIAP-1 and XIAP under some circumstances, (3) the TRAF-interacting protein A20, which displays antiapoptotic activity in some contexts, and (4) IEX-1L, an anti-apoptotic protein of unknown mechanism. Consequently, NF-κB has emerged as an attractive drug-discovery target for cancer therapy, with most efforts aimed at suppression of the activities of upstream kinases such as the IKKα and IKKβ, which are responsible for phosphorylating IκB-family proteins, the endogenous suppressors of NF-κB—resulting in targeting of these proteins for degradation via a ubiquitin/proteasome-dependent pathway.

As described above, the protein kinase Akt serves as an apoptosis-suppressing link between a variety of growth factor receptors and membrane-associated oncoproteins (Datta *et al*., 1999). Several apoptosis relevant substrates of Akt have been identified to date, including (1) the proapoptotic Bcl-2 family protein BAD, (2) the apical protease in the mitochondrial pathway caspase-9, (3) Forkhead-family transcription factors that control transcription of the promoters of apoptosis-inducing genes including Fas-L and (4) the NF-κB-inducing kinase IKKα. In each case, the net effect of phosphorylation of these substrates by Akt is increased protection from apoptosis. Multiple mechanisms for achieving abnormally elevated levels of Akt activity have been documented in tumour cells, perhaps the most common of which is mutations in the *PTEN* gene which result in decreased elimination of the second-messenger polyphosphoinositol phospholipids responsible for initiating Akt activation.

The tumour suppressor p53 is a well-known inducer of apoptosis which becomes inactivated in approximately half of all human tumours as a result of gene mutations and deletions, alterations in p53 kinases (ATM, CHK2), changes in the levels of p53 antagonist proteins and their regulators (MDM2, p19-ARF), ectopic expression of viral oncoproteins (HPV E6; SV40 large T antigen) and other mechanisms (Yan *et al*., 2000). Although p53 has other

functions relevant to cell cycle control, DNA repair responses and genetic instability, analysis of transgenic mice (when crossed with oncogene-bearing transgenic mice) suggest that the apoptosis-inducing activity of p53 is probably its most important attribute for suppression of tumorigenesis *in vivo*. How p53 induces apoptosis remains controversial. Three things however seem clear: (1) apoptosis induction by p53 derives from its function as a transcription factor; (2) p53 possesses multiple potential mechanisms for promoting apoptosis; and (3) the specific mechanisms employed are highly tissue-specific and may vary among clonal neoplasms even of the same lineage. Among the documented apoptosis-regulatory genes that make possible contributions to p53-induced apoptosis are the following: (1) the proapoptotic gene *BAX*, whose promoter in humans contains at least four consensus p53-binding sites and which is directly transcriptionally induced by p53; (2) Fas, which is a direct transcriptional target of p53; (3) DR5, another TNF-family member which may be a direct target of p53 trans-activation; and (4) Fas-L, which reportedly can be induced to translocate by p53 from a sequestered Golgi location to the plasma membrane where its receptor (Fas) primarily resides. Given the importance of p53 in inducing apoptosis of cancer cells (as well as certain types of normal cells) following genotoxic injury, this tumour suppressor has received extensive attention as a possible therapeutic target.

Multiple members of the nuclear receptor (NR) family of ligand-responsive transcription factors have been documented as regulators of the transcription of selected apoptosis genes. For example, retinoids which interact with and activate RAR- and/or RXR-family receptors are known to downregulate the expression of the antiapoptotic *BCL-2* and/or *BCL-X_L* genes in certain malignancies. Attempts to exploit this attribute of retinoids have been made in the clinic and are likely to continue into the future. Conversely, oestrogen is a positive regulator of *BCL-2* gene expression in mammary epithelial cells and in oestrogen receptor (ER)-positive breast cancers, possibly explaining some of the proapoptotic effects of antioestrogens such as tamoxifen on breast cancer cells *in vivo*. PPAR-γ ligands, which include certain prostaglandins produced by cyclooxygenases and other enzymes, as well as synthetic drugs that engage these receptors, have also been shown to either up- or downregulate transcription of *BCL-2*, depending on cellular context. These and other examples illustrate that many opportunities exist to regulate the output of apoptosis-relevant genes via effects on NR-family transcription factors in cancers.

THERAPEUTIC IMPLICATIONS

Knowledge about the core components of the apoptosis machinery and the various upstream inputs into apoptosis pathways has suggested a wide variety of new strategies for devising new therapies for cancer (Reed, 1999; Nicholson, 2000; Reed and Tomaselli, 2000). The full range of therapeutic modalities can be envisioned, including (1) small-molecule drugs that directly bind to and modulate the activities of specific protein targets; (2) antisense, DNAzyme and ribozyme nucleic acid-based therapeutics; (3) gene therapy using proapoptotic proteins; and (4) biologicals such as recombinant protein ligands or monoclonal antibodies, in the case of cell surface targets. Already proof of concept data have been obtained in animal models for many apoptosis-modulating agents, and some of these have advanced into clinical trials. At the time of this writing, for example, phase III trials are underway exploring the efficacy of antisense oligonucleotides directed against *BCL-2* mRNA for patients with a variety of types of refractory tumours. Recombinant Trail (Apo2L) protein is nearing its debut into clinical trials, as an attempt to trigger TNF-family death receptor pathways in tumour cells. Retinoids and PPAR-γ ligands are currently being examined as possible apoptosis sensitizers in leukaemias and solid tumours. Gene therapy trials are underway involving local delivery of p53- or *BAX*-expressing viral vectors, attempting directly to restore p53 or Bax expression in tumour cells. Monoclonal antibodies, recombinant proteins and synthetic peptidyl ligands that induce apoptosis of migrating endothelial cells by interfering with integrin-generated signals for cell survival (probably mediated largely by Akt) are also under investigation in clinical trials, as an approach for inhibiting angiogenesis. Preclinical analysis is also under way of multiple agents, ranging from small molecule organic compounds that restore activity to mutant p53 in malignant cells to synthetic peptides representing BH3 domains of pro-apoptotic Bcl-2 family proteins combined with membrane-penetrating peptides derived from viruses.

Our understanding is also rapidly evolving about the mechanisms by which currently available anticancer agents successfully trigger apoptosis of tumour cells and how resistance develops in all too many instances. A few examples include (1) binding of the triphosphate forms of purine nucleoside analogues to Apaf-1, resulting in improved catalytic efficiency of Apaf-1-mediated activation of caspases relative to endogenous dATP, (2) liberation of the BH3-only proapoptotic Bcl-2 family member Bim from microtubules by agents that disturb normal microtubule polymerization and (3) antioestrogen (tamoxifen)-mediated downregulation of Bcl-2 expression in breast cancers.

Exploiting apoptosis-based therapies for the treatment of cancer must be achieved with an acceptable therapeutic index, resulting in a selective killing of malignant cells. Fortunately, much evidence suggests that the inherent abnormalities found in cancer cells may also render them selectively more vulnerable compared with normal cells when deprived of their roadblocks to apoptosis. For example, several proto-oncogenes which become hyperactive in cancers, such as c-*Myc* and *cyclin-D1 (BCL-1)*,

drive rapid tumour cell division but also promote cell death unless apoptosis is concomitantly blocked (Evan and Littlewood, 1998). Genetically unstable cells also suffer errors in cell cycle checkpoint regulation and DNA/chromosome management which can be triggers for apoptosis when cell death pathways are intact. Finally, metastatic cells are potentially vulnerable because they often depend on defects in apoptosis pathways for avoiding cell death as a result of loss of survival signals from unoccupied cell adhesion receptors and from absence of local growth/survival factors. It seems clear, therefore, that apoptosis-based therapies will eventually find their place in the armentarium of weapons that will be used to wage and eventually to win the war on cancer.

ACKNOWLEDGEMENTS

The author acknowledges research funding from the National Cancer Institute, National Institutes of Health, Department of Defense, California Breast Cancer Research Program and CaP-CURE.

REFERENCES

Adams, J. and Cory, S. (1998). The Bcl-2 protein family: arbiters of cell survival. *Science*, **281**, 1322–1326.

Altieri, D. C., *et al*. (1999). Survivin apoptosis: an interloper between cell death and cell proliferation in cancer. *Laboratory Investigations*, **79**, 1327–1333.

Ashkenazi, A. and Dixit, V. (1998). Death receptors: signaling and modulation. *Science*, **281**, 1305–1308.

Ashkenazi, A. and Dixit, V. M. (1999). Apoptosis control by death and decoy receptors. *Current Opinions in Cell Biology*, **11**, 255–260.

Cryns, V. and Yuan, Y. (1999). Proteases to die for. *Genes and Development*, **12**, 1551–1570.

Datta, S., *et al*. (1999). Cellular survival: a play in three Akts. *Genes and Development*, **13**, 2905–2927.

Deveraux, Q. and Reed, J. (1999) IAP family proteins: suppressors of apoptosis. *Genes and Development*, **13**, 239–252.

Evan, G. and Littlewood, T. (1998) A matter of life and cell death. *Science*, **281**, 1317–1322.

Fesik, S. W. (2000) Insights into programmed cell death through structural biology. *Cell*, **103**, 273–282.

Green, D. R. and Reed, J. C. (1998) Mitochondria and apoptosis. *Science*, **281**, 1309–1312.

Gross, A., *et al*. (1999). BCL-2 family members and the mitochondria in apoptosis. *Genes and Development*, **13**, 1899–1911 (1999).

Huang, D. C. and Strasser, A. (2000). Bh3-only proteins—essential initiators of apoptotic cell death. *Cell*, **103**, 839–842.

Karin, M. and Ben-Neriah, Y. (2000). Phosphorylation meets ubiquitination: the control of NF-κB activity. *Annual Reviews of Immunology*, **18**, 621–663.

Kroemer, G. and Reed, J. C. (2000) Mitochondrial control of cell death. *Nature Medicine*, **6**, 513–519.

LaCasse, E. C., *et al*. (1998). The inhibitors of apoptosis (IAPs) and their emerging role in cancer. *Oncogene*, **17**, 3247–3259.

Lowin, B., *et al*. (1995). Perforin and granzymes: crucial effector molecules in cytolytic T lymphocyte and natural killer cell-mediated cytotoxicity. *Current Topics in Microbiology Immunology*, **198**, 1–24.

Metzstein, M., *et al*. (1998). Genetics of programmed cell death in *C. elegans*: past, present and future. *Trends in Genetics*, **14**, 410–416.

Miller, L. (1999). An exegesis of IAPs: salvation and surprises from BIR motifs. *Trends in Cell Biology*, **9**, 323–328.

Nicholson, D. W. (2000). From bench to clinic with apoptosis-based therapeutic agents. *Nature*, **407**, 810–816.

Reed, J. C. (1997). Double identity for proteins of the Bcl-2 family. *Nature*, **387**, 773–776.

Reed, J. (1998a). Caspases and cytokines: roles in inflammation and autoimmunity. *Advances in Immunology*, **73**, 265–287.

Reed, J. (1998b). Bcl-2 family proteins. *Oncogene*, **17**, 3225–3236.

Reed, J. (1999). Dysregulation of apoptosis in cancer. *Journal of Clinical Oncology*, **17**, 2941.

Reed, J. and Paternostro, G. (1999). Post-mitochondrial regulation of apoptosis during heart failure. *Proceedings of the National Academy of Sciences of the USA*, **96**, 7614–6.

Reed, J. C. (2000). Mechanisms of apoptosis. *American Journal of Pathology*, **157**, 1415–1430.

Reed, J. C. and Tomaselli, K. (2000). Drug discovery opportunities from apoptosis research. *Current Opinions in Biotechnology*, **11**, 586–592.

Salvesen, G. S. and Dixit, V. M. (1997). Caspases: intracellular signaling by proteolysis. *Cell*, **91**, 443–446.

Salvesen, G. S. and Dixit, V. M. (1999). Caspase activation: the induced-proximity model. *Proceedings of the National Academy of Sciences of the USA*, **96**, 10964–10967.

Thornberry, N. and Lazebnik, Y. (1998). Caspases: enemies within. *Science*, **281**, 1312–1316.

Tschopp, J., *et al*. (1998). Inhibition of Fas death signals by Flips. *Current Opinions in Immunology*, **10**, 552–558.

Tschopp, J., *et al*. (1999). Apoptosis: silencing the death receptors. *Current Biology*, **9**, R381–R384.

Vander Heiden, M. G. and Thompson, C. B. (1999). Bcl-2 proteins: regulators of apoptosis or of mitochondrial homeostasis? *Nature Cell Biology*, **1**, E209–E216.

Vaux, D. and Korsmeyer, S. (1999). Cell death in development. *Cell*, **96**, 245–254.

Wallach, D., *et al*. (1999). Tumor necrosis factor receptor and Fas signaling mechanisms. *Annual Reviews of Immunology*, **17**, 331–367.

Wyllie, A. H. (1997). Apoptosis: an overview. *British Medical Bulletin*, **53**, 451–465.

Yan, H., *et al*. (2000). Genetic testing—present and future. *Science*, **289**, 1890–1892.

FURTHER READING

Abrams, J. (1999). An emerging blueprint for apoptosis in drosophila. *Trends in Cell Biology*, **9**, 435–440.

Jacobson, M. D., *et al*. (1997). Programmed cell death in animal development. *Cell*, **88**, 347–354.

Metzstein, M., *et al*. (1998). Genetics of programmed cell death in *C. elegans*: past, present and future. *Trends in Genetics*, **14**, 410–416.

Wyllie, A. H., *et al*. (1980). Cell death, the significance of apoptosis. *International Review of Cytology*, **68**, 251–306.

Chapter 10

Signalling by Steroid Receptors

Torsten A. Hopp and Suzanne A. W. Fuqua
Breast Center at Baylor College of Medicine, Houston, TX, USA

CONTENTS

INTRODUCTION

Steroid hormones (androgens, oestrogens glucocorticoids and progestins) play a vital role in the development and maintenance of normal cellular function, as well as regulatory functions within the reproductive organs. These hormones mediate their activities through binding to specific intracellular receptor proteins, called the androgen receptor (AR), oestrogen receptor (ER), glucocorticoid receptor (GR) and progesterone receptor (PR), respectively. The steroid receptors (SRs) are members of a large superfamily of nuclear receptors, which modulate expression of target genes upon binding of their respective hormones (Evans, 1988). One of the important characteristics of this protein family is the highly conserved organisation of functional domains, implying that the underlying mechanism of transcriptional modulation might also be preserved. A highly simplified and schematized model of SR action can be seen in **Figure 1**: upon binding of the steroid hormone, the SR forms dimers and binds to a specific steroid response element (SRE) in the 5′ flanking region of hormone-responsive genes and stimulates/modulates their transcription.

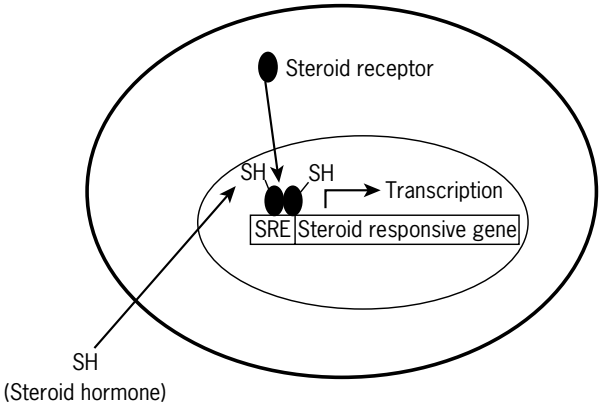


Figure 1 Simplified model of steroid action. Steroids enter a target cell by diffusion through the cell membrane and bind to steroid receptor (filled ovals). The hormone-bound receptor complex stimulates transcription of target genes via interaction with a steroid response element.

Recent research studying the molecular mechanism of transcriptional regulation of target genes by SRs has revealed a very complex network of protein–protein interactions in addition to protein–DNA interactions necessary for proper function. Disruption in this intricately regulated circuitry can perturb SR signalling. Specifically, mutations in the AR and ER, and also altered receptor expression, have been found in prostate and breast cancer, and have been associated with cancer progression and hormone resistance. However, our understanding of how these changes can affect the SR signalling pathway is still very limited in many cases. To comprehend the complex network of SR signalling, we will focus here on the ER, which has been extensively studied owing to its importance in the clinical management of breast cancer (Osborne, 1998).

STRUCTURE AND FUNCTION OF OESTROGEN RECEPTORS α AND β

Like many other members of the nuclear receptor family, the ER has more than one form encoded by separate genes. In the case of ER there are two, called *ER*α and -β. The human *ER*α gene resides on chromosome 6q25.1 and is transcribed in a single mRNA of 6.5 kb that encodes a protein of 595 amino acids with an approximate molecular mass of 66 kDa (Kumar *et al*., 1987). Even though the *ER*α gene is transcribed from at least three different promoters in a cell- and tissue-specific manner, only a single open reading frame appears to exist. However, the three promoters transcribe mRNA isoforms which differ in their 5′ untranslated regions, but no biological differences have yet been reported.

The *ERβ* gene is located on chromosome 14q22–24 and encodes a protein of 530 amino acids with a molecular mass of 60 kDa (Mosselman *et al*., 1996). Unlike *ERα*, several studies indicate that *ERβ* is transcribed as multiple mRNAs and also translated into proteins from at least two reading frames resulting in a second ERβ protein which lacks 53 amino acids of the N-terminus. Overall, ERα and ERβ are highly homologous (49% amino acid identity), but appear to be expressed in different organs and at different developmental stages. Interestingly, ERα is expressed in female organs such as the mammary gland, uterus, ovary, vagina and certain areas of the hypothalamus, whereas ERβ is found in the male in different areas of the hypothalamus when compared with ERα, and in the cerebral cortex. Additionally, *ERβ* mRNA has also been detected in prostate, ovary, uterus, lung, testis and artery.

In Vivo Function of the Oestrogen Receptors

The development of mice lacking the *ERα* (*αERKO*) or *ERβ* (*βERKO*) gene have proved to be valuable tools in evaluating the *in vivo* function of these receptors. The αERKO mice were generated in 1993, and the disruption of ERα expression not only caused infertility in both sexes, but also had profound effects on behaviour (Couse and Korach, 1999). Specifically, pre- and neonatal development of female reproductive organs such as uterus, ovary and mammary gland was almost normal, but maturation of these organs during and after puberty was severely impaired. The αERKO females also failed to display sexual receptivity when treated with the hormonal regime of oestrogen and progesterone that normally induces receptivity in wild-type mice. Surprisingly, adult αERKO males have significantly fewer epididymal sperm than heterozygous or wild-type males, caused by the disruption of spermatogenesis and degeneration of the seminiferous tubules, which becomes apparent 10 weeks after birth. Furthermore, these males develop obesity after sexual maturation, in addition to exhibiting decreased normal male-typical aggressive behaviour, including offensive attacks, and show a reduced number of mount attempts as compared with wild-type animals. Interestingly, both sexes of the αERKO mice also show a 20–25% reduction in bone density, implying that ERα is crucial for proper bone maturation and mineralization. However, the only described case of oestrogen insensitivity in a human male, which was normally masculinized, had incomplete epiphyseal closure with a history of continued growth into adulthood, and also osteoporosis probably induced by increased bone turnover.

More recently, the generation of βERKO mice revealed that ERβ does not affect normal development, and mice lacking ERβ are indistinguishable grossly and histologically as young adults from their littermates (Couse and Korach, 1999). Females are fertile and exhibit normal sexual behaviour, but have fewer and smaller litters owing to reduced ovarian efficiency, and multiple resorbed fetuses. Older males lacking ERβ display signs of prostate and bladder hyperplasia. In contrast to the αERKO animals, the βERKO females exhibit normal breast development and lactate normally, while all components of sexual behaviour in βERKO mice were found to be intact. These observations indicate that unlike ERα, ERβ is essential for normal ovulation efficiency but not for female or male sexual differentiation, fertility or lactation.

Structural and Functional Domains

The ERα and ERβ protein sequences encode all the above-mentioned *in vivo* functions, and can be divided into six functional domains, A–F (**Figure 2**) (Kumar *et al*., 1987). The N-terminal A/B domain, which contains a hormone-independent transactivation function (AF-1), as well as the C-terminal F domain, demonstrate only weak homology (approximately 17%) between the two ERs. In contrast, the central C domain, which contains the DNA-binding domain (DBD) present in all SRs, possesses 97% homology between ERs α and β, suggesting that they recognize and bind to the same or similar oestrogen response elements (EREs) consisting of inverted repeats of the sequence GGTCA separated by three variable nucleotides (Kumar and Chambon, 1988). This DNA-binding domain consists of two functionally distinct zinc-finger motifs, forming a helix–loop–helix motif typical of many transcription factors. Following the C domain is the D domain, which appears to function as a hinge region between the DBD and the ligand-binding domain (LBD), and demonstrates only moderate homology (30%) between the two ERs. However, recent research demonstrates that this domain contains important binding sites for receptor coactivators/corepressors, as will be discussed later, and might even be involved in post-translational regulation of the ER. Finally, the C-terminal E domain contains the LBD and, even though ERα and ERβ possess only moderate homology (59%) in this domain, both receptors bind with nearly the same affinity to oestradiol and to other natural and synthetic ligands. The E domain also contains a ligand-dependent transactivation function (AF-2) and provides a crucial interface for the interaction with many coactivators, as will be discussed later. Recently, a third AF, AF-2a, was identified in the ER hinge and LBD domains (Norris *et al*., 1997). AF-2a has been shown to activate transcription in a ligand-independent manner in the absence of both AF-1 and AF-2.

Structure of the Ligand-binding Domain

The LBD can be viewed as a molecular switch that, upon hormone binding, enables the receptor to activate transcription of target genes by direct interaction of the receptor with DNA in the promoter region of these genes

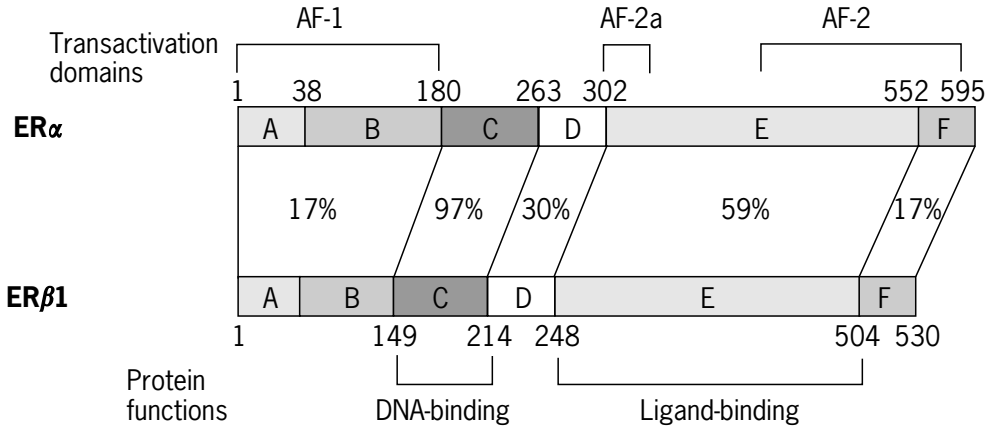


Figure 2 Comparison of ERα and ERβ1 functional domains. The amino acid residues of ERα are shown above domains A–F, and those of ERβ1 are shown below the domains. The degree of homology between them is shown as a percentage. Also shown are the regions for receptor function such as DNA binding, ligand binding and transcriptional activation functions.

and with components of the transcriptional machinery. In addition to the ligand-binding function, the LBD also contains signals necessary for nuclear localization, homo- and heterodimerization, in addition to the above-mentioned AF-2.

Recently, the crystallographic resolution of agonist (oestradiol) and antagonist (raloxifene) bound LBD of ERα revealed a compact structure consisting of 12 α-helices that form the ligand-binding pocket (Brzozowski *et al*., 1997). This structure appears to be a common motif found also in other SR. The binding of oestradiol into the binding pocket induces important structural changes; in particular, helix 12, that prior to ligand binding extends away from the body of the domain, is repositioned over the ligand-binding pocket and seals this pocket like a 'lid.' This process is thought to trap the ligand in a hydrophobic environment, and also forms a coactivator binding surface on the LBD. Helix 12 is fixed in this active position by contact with both the hormone and amino acid residues in helices 3 and 4 on the surface of the LBD. In contrast, binding of the antioestrogen raloxifene into the binding pocket makes helix 12 extend away from the LBD, a conformation similar to one seen in unliganded receptor.

THE ER IS A LIGAND-DEPENDENT TRANSCRIPTION FACTOR

In the absence of hormone, the ER is thought to exist in a complex with chaperone proteins, like heat shock proteins Hsp90 and Hsp70, which may help to maintain the receptor in an appropriate conformation to respond rapidly to hormonal signals. Upon binding of oestrogen, this oligomeric complex dissociates, allowing the ER to homo- or heterodimerize and to interact with target gene promoters through two possible mechanisms. First, the ER can interact directly with DNA by binding to specific EREs, resulting in a bending of the DNA toward the major groove. This bending is thought to facilitate the interaction of ER with proteins of the transcription complex. Second, ERα is also able to interact indirectly with oestrogen-regulated gene promoters through its association with other transcription factors such as AP-1, NF-κB, C/EBPβ, GATA-1 and SP-1. It is currently thought that gene transcription depends on the formation of a preinitiation complex that consists of basal transcription factors. However, it has recently become clear that ER also recruits a host of ancillary factors, which are called coactivators if they enhance and corepressors if they inhibit receptor transcriptional activity.

The Interaction of ER with Basal Transcription Factors

The initiation of transcription by RNA polymerase II requires the assembly of basal transcription factors such as transcription factor IIA (TFIIA), TFIIB, TFIID, TFIIE and TFIIF. The binding of TFIID is the first and the rate-limiting step in this assembly process. TFIID consists of the TATA-box binding protein (TBP) and more than 10 other TBP- associated factors.

There is ample *in vitro* evidence of direct protein–protein interactions between ERα and basal transcription factors. In particular and consistent with the evidence that TBP recruitment is a rate-limiting step in transcriptional initiation, both the N-terminal AF-1 and the C-terminal AF-2 activation function of ERα have been reported to bind to this basal transcription factor. Another component

of TFIID, $TAF_{II}30$, also interacts with the AF-1 of ERα in a ligand-independent manner, and this interaction appears to be required for ERα-mediated transactivation. Following the binding of the TFIID complex to the promoter of target genes, the binding of TFIIB, RNA polymerase II and TFIIF is also required for the assembly of the minimal initiation complex. TFIIB not only contacts DNA sequences both upstream and downstream of the TATA box and other factors of the basal transcription machinery, but also interacts with AF-2 of ERα. However, the significance of all of these interactions is unclear at present since they are not significantly affected by mutations in the ER that are known to disrupt transcriptional activity.

The Interaction of ER with Coactivators

The observation that different classes of SR can interfere with each other's transcriptional activity, termed 'squelching', has indicated that SRs compete for limited amounts of assembly proteins called coactivators or corepressors, in addition to other factors of the basal transcriptional machinery. Over the last 7 years, a large number of potential coactivators have been identified, they have been extensively reviewed by Klinge (Klinge, 2000). Coactivators are generally defined as proteins that can interact in a ligand-dependent manner with DNA-bound SRs and are able to enhance their transcriptional activity. Furthermore, a coactivator should also be able to interact with components of the basal transcription machinery, but not to enhance basal transcriptional activity on their own, although they often contain an autonomous activation function. As examples, we will focus here on the mechanism of action of the intensively studied SRC family of coactivators and also the Creb binding protein (CBP)/p300 in regulating ER's function.

The SRC Family of Coactivators

Several coactivators have been identified as a family of related proteins, called the SRC family, which includes SRC-1 (also termed ERAP-160 or NcoA-1), SRC-2 (also called GRIP1, NcoA-2, or TIF-2) and SRC-3 (also known as ATCR, RAC3, p/CIP, AIB1 or TRAM-1). These coactivators are all able to stimulate oestradiol-mediated gene transcription and promote the interaction between AF-1 and AF-2 to produce full transcriptional activity. The SRC family shares a common domain structure with an overall sequence similarity of 40% between the three members (**Figure 3**). The highest degree of homology is observed in the N-terminal bHLH (for basic helix–loop–helix) and PAS (for Per/Arnt/Sim homology) domains.

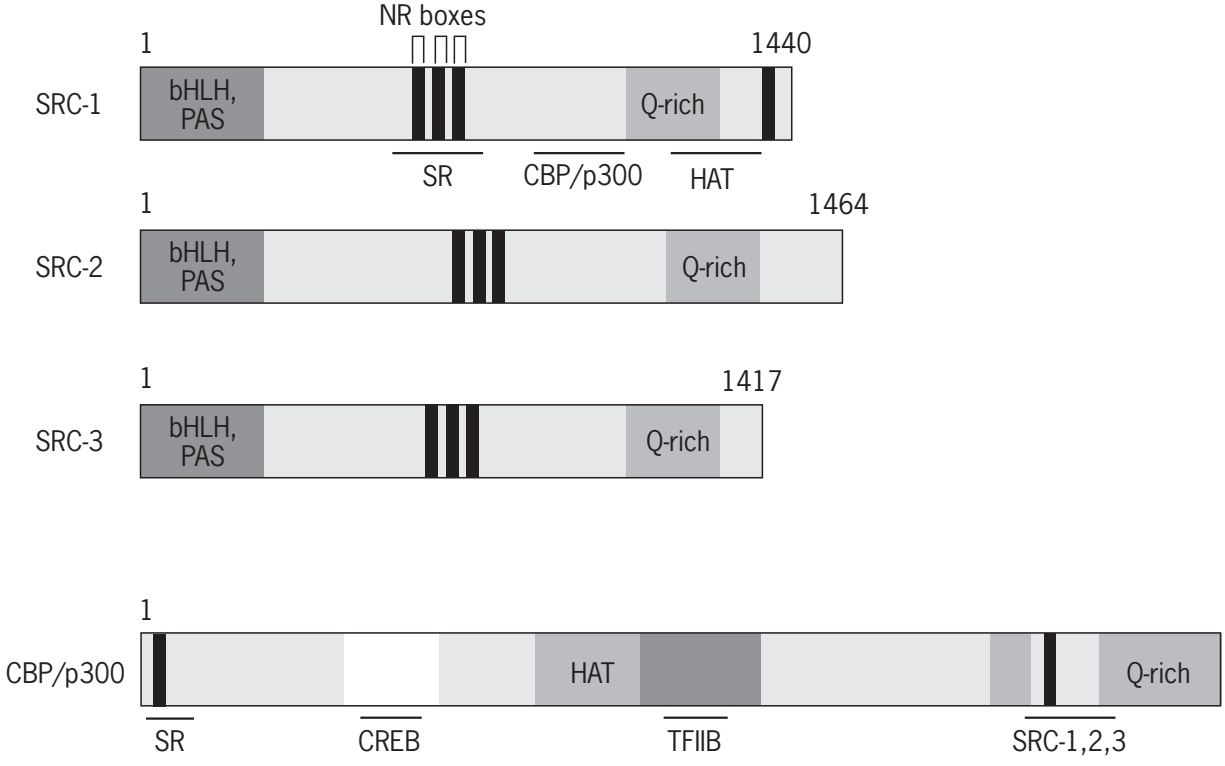


Figure 3 Structural features of the SRC family of coactivators and CBP. Regions to which specific functions of individual coactivators have been assigned are indicated. Below the coactivator representations are also shown the regions of interaction for various nuclear proteins such as steroid receptors (SR), CBP/p300, CREB, TFIIB and SRC1,2,3.

These particular domains mediate homodimeric and heterodimeric interactions between proteins containing these motifs, but their specific function in SRC coactivators remains unknown. It has been speculated that the presence of these motifs might indicate potential cross-talk between the SR pathway and other PAS-containing factors.

Most SR coactivators, including the SRC-family of coactivators, contain one or more copies of a short sequence motif, LXXLL (L stands for leucine and X is any amino acid), that are necessary and sufficient to mediate ligand-dependent direct interaction with the ER. This motif is called the nuclear receptor (NR) box or the receptor-interacting domian (RID). All of the SRC coactivators possess three NR boxes in the central portion of the protein, whereas only SRC-1 has an additional NR box in the extreme C-terminus. Protein structure prediction analyses of these motifs have indicated that they form amphipathic helices with the conserved leucine residues outlining a hydrophobic surface on the face of the helix. Cocrystallization of the LBD of ERα with NR box peptides, as well as systematic mutagenesis of ERα, have revealed that the helix formed by the NR boxes is able to interact with a hydrophobic groove in the AF-2 domain of ER consisting of ERα residues from helices 3, 4, 5 and 12. This hydrophobic groove is the result of conformational changes induced by hormone binding, as discussed earlier.

The SRC coactivators also contain intrinsic activation domains that retain their activity when tethering the coactivator to DNA. Additonally, SRC-1 and SRC-3 contain an intrinsic histone acetyltransferase (HAT) activity which is thought to modulate chromatin structure by histone acetylation, thus facilitating the access of other transcriptional regulators as well as the assembly of the preinitiation complex. Interestingly, these coactivators together with ER are believed to form a ternary complex with CBP/p300, CREB, and other proteins (**Figure 4**), but the function of this complex remains largely unclear. However, a recent study reporting oestradiol-dependent acetylation of SRC-3

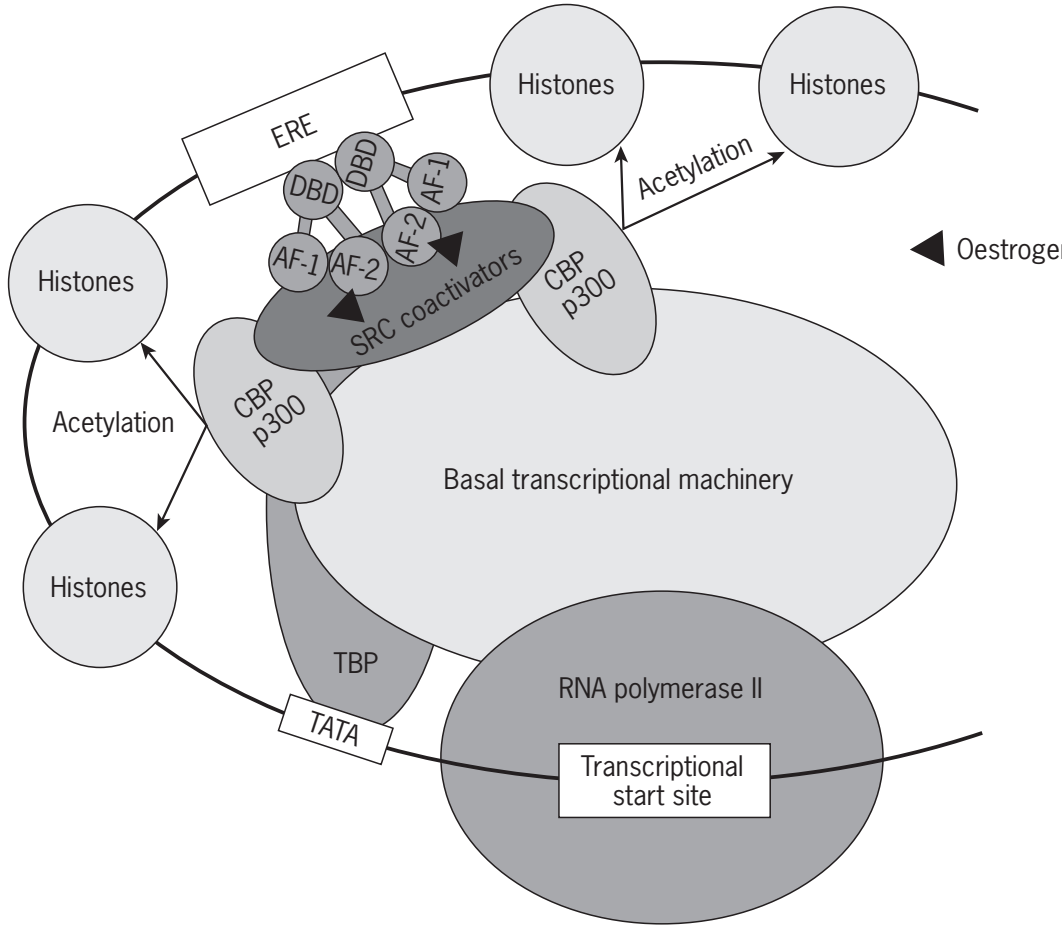


Figure 4 A schematic model of the transcription initiation complex formed at the ERE of an oestrogen-responsive gene. The ER is able to recruit an SRC coactivator upon oestrogen binding, which subsequently results in the recruitment of additional coactivators, such as CBP, and basal transcription factors. However, the precise *in vivo* composition of this complex is still under investigation. Additionally, histone acetylation by CBP/p300 is thought to facilitate the relaxation of chromatin at the target gene promoter, thereby enhancing transcriptional activation.

by CBP/p300 might reveal a new insight into the function of this complex. Surprisingly, acetylation of three lysine residues immediately upstream of the NR box of SRC-3 decreased SRC-3:ERα interactions *in vitro*, and disruption of this interaction appears to be a necessary event for the down-regulation of ER activity. These results indicate that hormone-induced transcription may be dynamically regulated by both histone and SRC-3 acetylation.

The recent development of mice lacking either SRC-1 or SRC-3 provides *in vivo* evidence of a partial functional redundancy between SRC-1 and SRC-2, while the physiological role of SRC-3 in development or disease appears to be different from that associated with SRC-1 expression. In particular, both SRC-1 and SRC-3 knockout mice are viable and fertile. However, in mice lacking SRC-1, oestrogen target organs such as the uterus, prostate, testis and mammary gland display decreased growth and development in response to steroid hormones, as well as increased expression of SRC-2 that is thought to compensate partially for the loss of SRC-1 function. On the other hand, disruption of SRC-3 expression in mice results in dwarfism, delayed puberty, reduced female reproductive function and blunted mammary gland development. This pleiotropic phenotype indicates that SRC-3 plays a critical role in overall growth and sexual maturation.

Interestingly, SRC-1 and SRC-2, but not SRC-3, interact also with ERβ and enhances its hormone-dependent transcriptional activity. In addition, SRC-3 is amplified and overexpressed in many ER-positive breast cancer cell lines, and a large study analysing 1157 clinical breast tumours and 122 ovarian tumours also found amplification of this coactivator gene in a small percentage of breast and ovarian cancers. Expression of SRC-3 seems to correlate with tumour size and with ERα and PR positivity. It has also been reported that SRC-1 and SRC-2 expression is relatively low in breast tumours when compared to normal tissues.

The Cointegrators CBP/p300

The cointegrator CBP and its related functional homologue p300 are thought to be responsible for the integration of numerous environmental stimuli on promoters containing multiple *cis*-acting elements (Goodman and Smolik, 2000). Even though CBP was initially characterized as a coactivator required for efficient activation of cAMP-regulated genes, several studies also implicate this protein as a coactivator for a broad range of transcription factors, including p53, NFκB and the SRs. CBP interacts with ERα in a hormone-dependent manner, and this interaction depends on a crucial NR box in the N-terminal domain of this CBP. Surprisingly, SRC-1 is also able to interact directly with carboxy terminus of CBP and p300, which synergistically enhances ERα-activated gene activity. Despite all the described potential interactions between CBP/p300 and transcription factors, coactivators and the basal transcription machinery, there is little biochemical evidence for the existence of these complexes *in vivo*. There is only one small study analysing the chromatographic elution of SRC-1-containing complexes from T47D breast cancer cells that indicates the existence of distinct coactivator complexes with different properties and activities.

The central portion of both CBP and p300 encodes a relatively large domain that possesses intrinsic HAT activity. Among the major substrates of acetylation are the core histones, particularly the N-terminal tails of histones H3 and H4 (**Figure 4**). The unmodified forms of these histones are thought to maintain DNA packing into highly organized chromatin structures, thus silencing transcription, while a high degree of histone acetylation correlates with increased promoter activity. More recently, it has been shown that nonhistone proteins such as p53, GATA-1, the AR and SRC-3 (that itself possesses HAT activity) are also substrates for CBP/p300 acetylation. However, the precise function of both histone and nonhistone protein modifications in transcriptional regulation is still largely unclear.

Interaction of ER with Corepressors

Unlike coactivators, only very few corepressors of ER action have been reported to date. These include NCoR (nuclear receptor corepressor; also called RIP13), SMRT (silencing mediator for retinoid and thyroid receptor, also termed TRAC2), REA (repressor of oestrogen receptor activity), SHP (short heterodimer partner) and BRCA-1 (breast cancer susceptibility gene). Repression of ER activity can occur in the absence of hormone or when an antagonist is bound to the receptor.

The corepressors NCoR and SMRT were originally identified by biochemical studies of cellular proteins associated with unliganded thyroid hormone receptor (TR) and retinoid acid receptor (RAR). These corepressors are thought to recruit another complex of proteins, including histone deacetylase (HDAC) activity, to DNA-bound TR and RAR (**Figure 5**). This complex is believed to repress gene expression by maintaining chromatin in a more condensed state which impairs the ready access of critical transcription factors. Upon ligand binding, corepressors are released from TR and RAR, and coactivators and basal transcription factors are then recruited to the receptors. In contrast, unliganded or hormone-bound ERα is unable to interact with either NCoR or SMRT, but the antioestrogen tamoxifen (an oestrogen antagonist which is commonly used in the treatment of ER-positive breast cancer) induces the interaction of ER with these corepressors. Interestingly, the antagonist tamoxifen can be switched into in agonist, much like oestrogen, when ERα's ligand-independent AF-1 is activated by the MAPK pathway (this kind of activation will be discussed later in the section

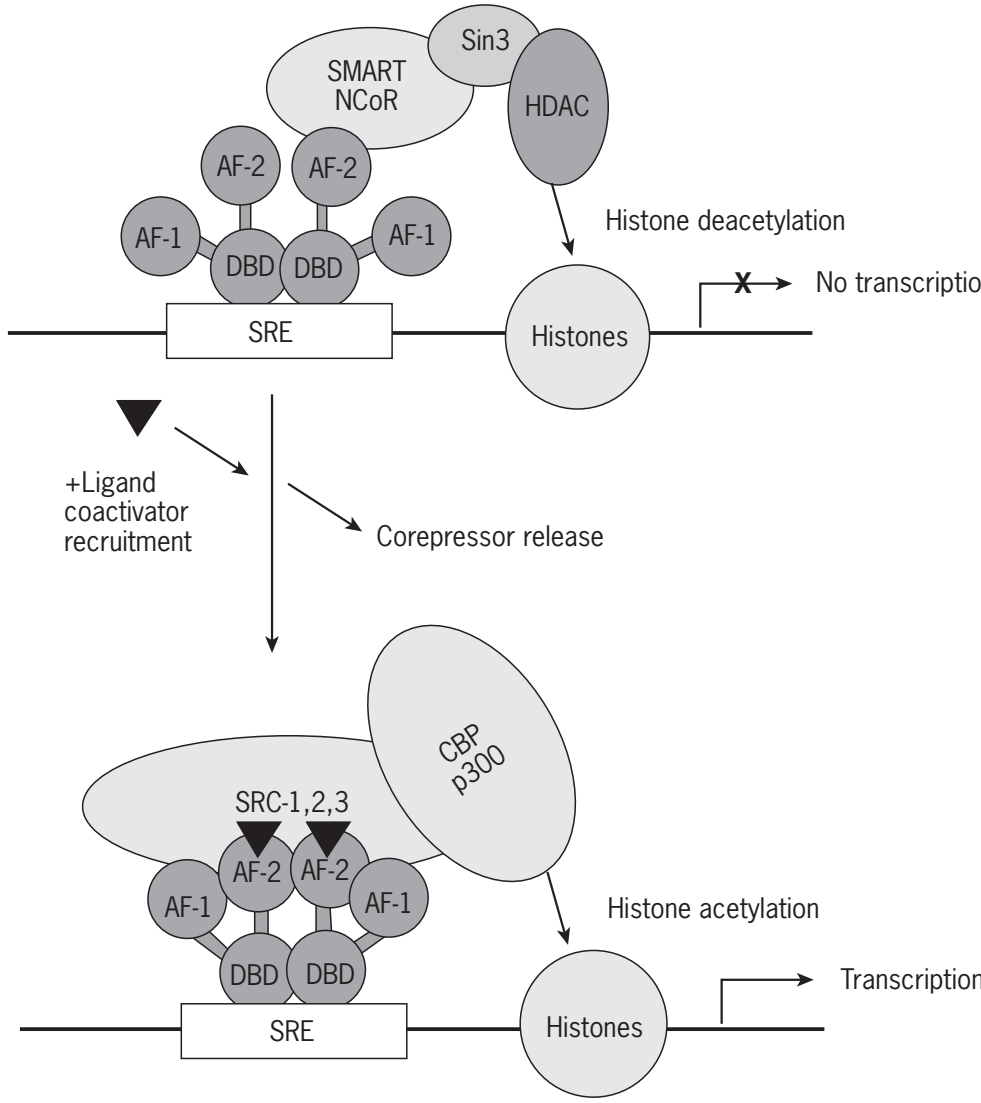


Figure 5 Ligand-dependent switch between an SR associated with either a corepressor or a coactivator complex. The SR is associated with a corepressor (SMRT or NCoR), which in turn recruits a histone deacetylase (HDAC) through its interaction with Sin3. Deacetylation of histones leads to transcriptional repression. Hormone binding disrupts the interaction between SR and repression complex in favour of the association of a coactivation complex. This complex consists in part of a SRC coactivator and CBP/p300, which possess histone acetyltransferase activity that modifies histones resulting in relaxation of chromatin structures and initiation of transcription.

Phosphorylation). This activation results in the release of corepressors from ERα and recruitment of coactivators to this receptor. This phenomenon may well explain why the majority of breast tumours become resistant to tamoxifen.

Regulation of ER Function

The activity of many transcription factors is regulated by post-translational modifications such as phosphorylation, proteasome-mediated degradation and cross-talk with other signal transduction pathways. We will focus upon these three aspects in our discussion of ER function.

Phosphorylation

Stimulation of a number of growth factor receptors and/or protein kinases leads to the phosphorylation of ER resulting in ligand-independent and/or a synergistic increase in transcriptional activation in response to hormone (Kato *et al*., 2000).

The AF-1 region of ERα contains phosphorylation sites for a number of kinases including MAPK, cyclin A/cdk2 and PI3/AKT. In particular, much work has been focused on phosphorylation of serine residue 118 within the A/B domain of ERα. Phosphorylation of this particular residue by growth factor activated MAPK leads to the

enhancement of the N-terminal AF-1 function. MAPK is activated by tyrosine kinase cell membrane receptors, that in turn are stimulated by growth factors such as insulin, IGF-1, EGF and TNF-α. Furthermore, phosphorylation at ER residue 118 also enhances the interaction of this receptor with the p68 RNA helicase, resulting in increased AF-1 but not AF-2 activity.

The phosphorylation of ERβ by MAPK in cells treated with exogenous EGF or via overexpression of activated Ras has been shown to enhance binding between SRC-1 and the ERβ AF-1 domain. This suggests that ligand-independent activation of ERβ also depends on phosphorylation of the N-terminal region, and that this event may be important for the recruitment of coactivators such as SRC-1.

Much less is known about the phosphorylation by the cyclin A/cdk2 or the PI3/Akt kinases and their effects on specific ER function. Specifically, phosphorylation of two serine residues at amino acid 104 and 106 of the ERα by the cyclin A/cdk2 enhances the ERα AF-1 function both in the absence of oestrogen or in the presence of tamoxifen. Furthermore, it was recently shown that PI3 kinase is able to increase both AF-1 and AF-2 activity, whereas Akt, a kinase downstream of PI3 kinase, increases only AF-1 activity. Phosphorylation of the ERα serine residue 167 by Akt also results in protection of breast cancer cells from tamoxifen-induced apoptosis, revealing an important potential mechanism for the onset of resistance to tamoxifen in breast cancer.

Another major ligand-independent phosphorylation site within ERα is the conserved tyrosine 537 residue and the homologous tyrosine 443 residue in ERβ, but the exact consequence of phosphorylation at this site is still controversial and remains under intense investigation. Replacement of this tyrosine residue with other amino acids suggests that phosphorylation at this site is important for hormone binding and transcriptional activation. Specifically, the substitution of tyrosine 537 with alanine, asparagine, or serine results in a mutant ERα that is constitutively active and binds to SRC-1 even in the absence of oestrogen. Another recent study also indicated that phosphorylation of the tyrosine 537 is critically involved in ligand-induced conformational changes of the ERα.

Proteasome-mediated Degradation of the ER

Recently, ubiquitin-dependent, proteasomal degradation of ligand-bound ERα was discovered as an additional mechanism involved in the regulation of hormone receptor-mediated gene transcription. The SRC coactivator family is also a target for degradation via the 26S proteasome. It is well known that the ubiquitin pathway is involved in the degradation of many short-lived proteins (Hershko and Ciechanover, 1998). Through a series of enzymatic reactions, ubiquitin is covalently linked to proteins targeted for degradation, marking them for recognition by the 26S proteasome, a large multisubunit protease. Abnormalities in ubiquitin-mediated degradation have been shown to cause several pathological conditions, including malignant transformation. In particular, it has recently been shown that ERα is ubiquitinated preferentially in the presence of hormone. It is thought that ERα protein degradation, which occurs through the 26S proteasome complex, is required for continued transcriptional activation by this receptor. ERα degradation could well be an important requisite to dissociate the preinitiation complex resulting in the release of the components necessary for another round of transcription. On the other hand, hormone-induced degradation may also serve as a negative feedback to down-regulate the transcription of oestrogen-responsive genes.

THE IMPORTANCE OF OESTROGEN RECEPTOR EXPRESSION IN CANCER INITIATION AND PROGRESSION

Oestrogen receptor gene expression in breast epithelium is an intricately regulated event, and is thought to play a central role in normal breast development, and also breast cancer evolution. ERα expression is significantly increased in both premalignant and malignant breast lesions, and many of these ERα-positive cells proliferate as compared with normal breast. Furthermore, normal breast epithelium, in addition to breast cancer tissue, contains alternatively spliced *ERα* and *ERβ* mRNA variants, but it is still unclear whether changes in the levels of these variants impact upon tumour development or the progression to hormone-independent tumour growth. Single amino acid mutations within the ERα are relatively rare, but may contribute to the progression of breast cancer or metastatic disease. We will next describe the potential role of ERα expression in premalignant disease, as well as the role of specific ER variants and mutations in breast cancer development and progression.

Oestrogen Receptor Expression in Normal Breast and Breast Cancer

In normal nonpregnant, premenopausal human breast, only about 5–10% of the total luminal epithelial cell population expresses ERα, and this expression tends to be highest in the follicular phase of the menstrual cycle. The highest percentage of ERα-expressing cells are found in undifferentiated lobules type 1 (Lob1), with a progressive reduction in the more differentiated Lob2 and Lob3 types. The highest level of cell proliferation is also observed in Lob1, but expression of ERα occurs in cells other than these

proliferating cells, indicating that they represent at least two separate cell populations. These data also suggest that oestrogen might stimulate ERα-positive normal cells to produce a growth factor that in turn stimulates neighbouring ERα-negative normal cells to proliferate. In premalignant and malignant breast lesions, however, ERα expression is significantly increased in the proliferating cell compartment, suggesting that ERα may be involved in the earliest changes to malignancy. Additionally, approximately two-thirds of breast tumours, at least initially, express abundant levels of ERα, and this expression is associated with lower risk of relapse and prolonged overall survival. Unfortunately, we still understand very little about the precise role of ERα expression in tumour progression.

Owing to its recent discovery, only limited data are available on the expression and function of ERβ in normal breast and its potential role in breast carcinogenesis. Studies of ERβ knockout mice suggest that ERβ plays a limited role in normal breast development and function. However, ERβ expression appears to be important for the growth control of urogenital tract epithelium, and may even afford a protective role against hyperproliferation and carcinogenesis in this particular tissue. This interesting hypothesis might also apply to the mammary gland, and is supported by a recent study reporting that ERβ expression in breast tumours is positively associated with ERα and progesterone receptor expression, as well as negative axillary nodes, DNA diploidy and low S phase fraction, all of which imply that ERβ-positive tumours may have a more favourable prognosis. On the other hand, two studies examining a relatively small number of tumours using RT-PCR determined that coexpression of ERβ and ERα is frequently associated with poor prognostic biomarkers, such as positive axillary nodes and higher tumour grade, and also that ER is significantly elevated in tumours resistant to tamoxifen treatment.

Oestrogen Receptor Variant Forms

Both of the *ER* genes undergo alternative splicing in normal and neoplastic oestrogen-responsive tissues. Alternative splicing results in *ER* mRNA variants with single or multiple exons skipped, and are usually coexpressed along with the wild-type receptor (Hopp and Fuqua, 1998). It is still unclear whether any or all of the *ER* splicing variants are indeed stably translated *in vivo*, and to what extent the formation of heterodimers of these splice variants with ERα and ERβ perturb the ER signalling pathway. *ER*α and *ER*β variant forms fall into four major groups: (1) transcripts containing precise single or multiple exon deletions, (2) transcripts containing single nucleotide deletions and others in which several hundred nucleotides have been deleted within known exon sequences, (3) truncated transcripts and (4) transcripts containing insertions.

The most frequently observed *ER*α mRNA splice variants are those lacking exon 4, which has been detected in normal and neoplastic tissue, or exon 7, detected in many breast tumours regardless of their receptor expression status. The exon 4-deleted *ER*α variant is missing the nuclear localization signal and part of the hormone-binding domain, thus potentially encoding a protein whose cellular distribution and oestrogen-binding affinity may be different from those of the wild-type ERα. On the other hand, the exon 7-deleted splicing variant potentially translates into a receptor protein missing the C-terminal part of the hormone-binding domain, which includes the hormone-dependent AF-2 function and the F-domain. However, both *ER*α variants, when transfected into mammalian or yeast cells, can block normal ER signalling in certain cell types, but not in others. In addition, expression of the exon 4-deleted *ER*α variant is associated with two biological markers of good clinical outcome (PR positivity and low histological grade), and thus may prove useful as biological marker of good prognosis in clinical samples.

Another dominant-negative inhibitor of ERα signalling is the exon 3-deleted variant, which encodes a variant receptor lacking portions of the DNA-binding domain. This variant is found in both normal tissue and primary breast cancer, and the ratio of the exon 3-deleted variant to wild-type ERα is reduced about 30-fold in breast cancer cell lines as compared with normal tissue. Stable expression of the exon 3-deleted variant in MCF-7 breast cancer cell lines, which contain high levels of endogenous ERα, results in reduction of both invasiveness and anchorage-independent growth.

One of the best studied *ER*α mRNA splice variants is the exon 5-deleted receptor, which is the only variant so far detected at the protein level in breast cancer cell lines and breast tumours. This variant is a truncated 40-kDa protein missing most of the hormone-binding domain, but it retains AF-1 function and demonstrates variable strengths of hormone-independent transcriptional activity depending on the cell type. Expression of this variant was also significantly increased in cancers from patients relapsing after tamoxifen treatment as compared with the respective primary tumour, suggesting that tumours expressing high levels of the exon 5-deleted variant may acquire resistance to tamoxifen. However, it seems likely that other *ER*α splice variants could also be involved in acquired tamoxifen resistance, since multiple *ER*α variants can occur in the same tumour sample and tamoxifen resistance is thought to be multifactorial.

ER Mutations

In contrast to the abundant expression of *ER*α mRNA splice variants in both normal and neoplastic tissue, mutations of the *ER*α gene are seldom found in primary breast cancer. Changes in the *ER* nucleotide sequence

fall into at least two groups: (1) polymorphisms, which do not change the amino acid sequence, and (2) missense mutations, which do alter the amino acid sequence (Hopp and Fuqua, 1998).

Polymorphisms have been detected in both primary and metastatic breast cancer, but these silent changes do not appear to correlate with clinical parameters, such as tumour type, size, grade or stage. Like polymorphisms, missense mutations in the *ER*α, which potentially affect normal function, have been found in primary and metastatic breast cancers. However, functional characterization of most of these mutations is still missing.

Recently, a specific somatic mutation in the ERα has been found in many typical hyperplasias, a type of premalignant lesion that carries an increased risk of breast cancer development. The mutation substitutes a lysine with an arginine residue at amino acid 303, at the border between hinge domain and the beginning of hormone-binding domain. This mutant ERα shows much higher sensitivity to oestrogen than wild-type ERα, resulting in markedly increased proliferation at subphysiological levels of hormone. Additionally, the mutation enhances the ability of the SRC-2 coactivator to bind at physiological levels of hormone. These data suggest that this mutant receptor may promote or accelerate the development of cancer from premalignant breast lesions.

Another missense mutation, where tyrosine 537 is substituted by asparagine, was isolated from a metastatic lesion from a breast cancer patients. This mutant ERα exhibits a potent, oestradiol-independent transcriptional activity that is only weakly affected by oestrogen, and variably by antioestrogens. As mentioned earlier, phosphorylation of the tyrosine residue at codon 537 is thought to be required for efficient oestrogen binding, and substitution of this amino acid with asparagine may induce conformational changes mimicking hormone binding. Stable expression of this mutant ERα in an ERα-negative breast cancer cell line caused increased production of PTHrP, a known stimulator of osteoclastic bone resorption and a major mediator of the osteolytic process. Furthermore, TGFβ, which is abundant in bone marrow, significantly enhanced the transcriptional activity of this mutant receptor, resulting in further stimulation of PTHrP production. These data indicate a central role for the ERTyr537Asn mutant in the pathogenesis of osteolytic bone metastases from breast carcinoma.

The Role of Oestrogen Receptors in Hormone Resistance and Independence

As mentioned earlier, a large number of breast cancers express high levels of ERα, and the ER status of patients is highly predictive of response to long-term tamoxifen therapy, an antioestrogen frequently used in the treatment of breast cancer. Unfortunately, most breast cancers eventually acquire tamoxifen resistance, resulting in disease recurrence and the frequent emergence of more aggressive disease. Tamoxifen-resistant tumours often continue to express ERα, so that mechanisms for antioestrogen resistance other than the loss of ERα must exist. Interestingly, in some cases antioestrogen resistance is also reversible, e.g. tamoxifen-resistant patients who have been switched to a different type of therapy may later once again respond to tamoxifen.

Potential mechanisms of resistance include alterations in the expression levels of ERα and ERβ as well as *ER* splicing variants, ER mutations, interaction with other growth factor signal transduction pathways (such as erbB-2 and AP-1), abnormal expression or function of coactivators and corepressors and metabolic tolerance as a result of altered systemic antioestrogen metabolism.

CONCLUSION

The purpose of this chapter was to review ER signalling and its perturbation in cancer as an example of SR signalling in general. The ER is an important transcriptional activator for genes involved in many essential processes. Cloning and sequencing of the ERs and the resolution of the crystal structures of hormone- and antioestrogen-bound ERα, and also the development of mice lacking these receptors, have all increased our understanding of the structure–function relationships of these important transcription factors. However, the fairly recent discovery of receptor coactivators and corepressors has provided another level of complexity to models of oestrogen action. The understanding of the molecular mechanisms of action of the ER is beginning to provide an explanation for the function of clinically useful antiestrogens, and is also suggesting potential new therapeutic targets. Further progress in understanding the fine details of transcriptional activation may also provide new insights into mechanisms of hormone resistance.

ER expression is tightly regulated in normal breast epithelium, but increased ERα and expression of mutant ERs may drive abnormal proliferation in premalignant hyperplasias, providing a fertile environment for genetic alterations which, in turn, are associated with tumorigenesis. Unfortunately, very little is known about the precise role of ER expression during the transition of premalignant disease to cancer and the eventual development of hormone resistance. Continued exploration of the basic molecular mechanisms of ER signalling, as an example of steroid receptor action, will certainly enhance our understanding of underlying causes of some of the most prevalent human cancers, and may also provide new treatment approaches, as well as and new mechanisms to prevent these diseases.

REFERENCES

Brzozowski, A. M., *et al*. (1997). Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature*, **389**, 753–758.

Couse, J. F. and Korach, K. S. (1999). Estrogen receptor null mice: what have we learned and where will they lead us? *Endocrinology Reviews*, **20**, 358–417; Erratum, **20**, 459.

Evans, R. M. (1988). The steroid and thyroid hormone receptor superfamily. *Science*, **240**, 889–895.

Goodman, R. H. and Smolik, S. (2000). CBP/p300 in cell growth, transformation, and development. *Genes and Development*, **14**, 1553–1577.

Hershko, A. and Ciechanover, A. (1998). The ubiquitin system. *Annuual Review of Biochemistry*, **67**, 425–479.

Hopp, T. A. and Fuqua, S. A. (1998). Estrogen receptor variants. *Journal of Mammary Gland Biology and Neoplasia*, **3**, 73–83.

Kato, S., *et al*. (2000). Molecular mechanism of a cross-talk between oestrogen and growth factor signalling pathways. *Genes and Cells*, **5**, 593–601.

Klinge, C. M. (2000). Estrogen receptor interaction with co-activators and co-repressors. *Steroids*, **65**, 227–251.

Kumar, V. and Chambon, P. (1988). The estrogen receptor binds tightly to its responsive element as a ligand-induced homodimer. *Cell*, **55**, 145–156.

Kumar, V., *et al*. (1987). Functional domains of the human estrogen receptor. *Cell*, **51**, 941–951.

Mosselman, S., *et al*. (1996). ER beta: identification and characterization of a novel human estrogen receptor. *FEBS Letters*, **392**, 49–53.

Norris, J. D., *et al*. (1997). Identification of a third autonomous activation domain within the human estrogen receptor. *Molecular Endocrinology*, **11**, 747–754.

Osborne, C. K. (1998). Steroid hormone receptors in breast cancer management. *Breast Cancer Research and Treatment*, **51**, 227–238.

FURTHER READING

Elledge, R. M. and Fuqua, S. A. W. (2000). Estrogen and progesterone receptors. In: Harris, J. R., *et al*. (eds), *Diseases of the Breast*. 471–488 (Lippincott Williams & Wilkins, Philadelphia).

Freedman, L. P. (ed.) (1998). *Molecular Biology of Steroid and Nuclear Hormone Receptors* (Birkhauser, Boston).

Chapter 11

Signalling by Cytokines

William L. Farrar, Lihua Wang, Xiaoyi Yang, Weihua Xiao, O. M. Zack Howard
National Cancer Institute, National Institutes of Health, Frederick, MD, USA

Colin Duckett
National Cancer Institute, National Institutes of Health, Bethesda, MD, USA

CONTENTS

STRUCTURE AND FUNCTION OF γ_c-DEPENDENT CYTOKINES AND RECEPTORS

Members of the cytokine haematopoietic superfamily often share a common receptor subunit while retaining their own private receptor subunits. One well-documented example is that five cytokines, IL-2, IL-4, IL-7, IL-9 and IL-15, form one group, which is characterized by utilizing the common γ chain (γ_c) as a receptor subunit. The common γ_c subunit was initially cloned as the γ chain of the IL-2R complex. Soon it was discovered that this γ subunit also participates functionally in the receptors for IL-4, IL-7, IL-9 and IL-15 and, therefore, was designated γ_c, where c represents 'common.' This protein is constitutively expressed on essentially all cells of haematopoietic origin. It functions to enhance the binding of cytokines to their receptor, presumably by direct interaction with the ligand and to induce intracellular signals transduction events such as JAK–STAT signal pathway.

γ_c-Dependent Cytokines

γ_c-Dependent cytokines include IL-2, IL-4, IL-7, IL-9 and IL-15. They are peptides or glycoproteins with molecular masses of 14–20 kDa. Crystal structure analysis reveals that IL-2 is an α-helical protein, lacking β-sheet structure, with a four-fork core stabilized by a single intrachain disulfide bond. IL-4 contains six Cys residues that are all involved in intramolecular disulfide bridges. The secondary structure of IL-4 was shown to consist of a four-helix bundle with a unique up–up–down–down helix topology. IL-7, IL-9 and IL-15 contain a similar α-helical structure. Each cytokine is secreted by particular cell types in response to a variety of stimuli and produces a characteristic constellation of effects on the growth, motility, differentiation or function of its target cells. These cytokines exert multiple biological functions (**Table 1**). It is interesting that IL-2, IL-4 and IL-9, all produced by activated T cells, are important immune regulatory cytokines, whereas IL-7 and IL-15, which are primarily produced by nonlymphoid cells, have also been implicated in the regulation of lymphocyte development.

γ_c-Dependent Cytokine Receptors

With the exception of IL-2R and IL-15R, all γ_c-dependent cytokine receptor subunits are members of the cytokine receptor superfamily that contains an evolutionary-related extracellular region that results in a conserved structural fold for binding to helical cytokines. These receptor subunits are type I membrane glycoproteins with a single hydrophobic transmembrane domain. The extracellular domain contains two major regions of homology. The first is a region having four Cys residues located in the N-terminal half of that extracellular domain. The second region of homology is Trp-Ser-X-Trp-Ser(WSXWS), which is referred to as the 'WS motif.' This motif is close to the transmembrane region. The extracellular region also contains two fibronectin type III-like domains found in a series of cell surface molecules with adhesive properties. The functional significance of these domains remains to be clarified.

Table 1 Major properties of human γ_c-dependent cytokines

Cytokine	Mature protein (kDa)	Cellular source	Functional activities
IL-2	15	Activated T_H2 cells	T cell growth
		Tc cells	Enhance B cell growth and Ig secretion
		NK cells	Augment NK activity
		T_HO cells	Induce LAK
			Programme T cells for apoptosis
			Reverse T cell anergy
IL-4	20	Activated T_H2 cells	T cell growth
		Mast cells	B cell growth
		Basophils	IgG_1 and IgE class switch
		MK1+ CD4+ T cells	Enhance expression of MHC class II and CD23
IL-7	17	Bone marrow stroma	T cell growth
		Thymic stromal cells	Proliferation of pre-B cells
		Intestinal epithelial cells	Viability of TN thymocytes
		Keratinocytes	Promotes development of CTL
IL-9	14	Activated T cells	Promotes the growth of mast cells
			Enhances mast cells secretion of IL-6 and expression of granzyme A and B and FcRε
IL-15	15	Placenta, epithelial cells	T cell growth
		Skeletal muscle	Enhanced NK activity
		Kidney, lung, fibroblasts	Induce LAK
		Activated monocytes	Promote B cell growth and Ig secretion

JAK–STAT SIGNAL PATHWAY

All known γ_c-containing receptors signal through the associated Janus protein tyrosine kinases, JAK1 and JAK3 proteins, although not all γ_c-dependent cytokines activate the same STAT molecules. Phosphorylated tyrosines and flanking amino acid residues in the activated cytokine receptors determine this specificity by providing specific docking sites for the SH2 domains of STATs. Most likely, tyrosine phosphorylation of the receptor proteins is also directly mediated by JAKs. The JAK/STAT signal pathway, therefore, connects activation of the receptor complexes directly to transcription of genes. Upon receptor oligomerization, JAKs are activated, presumably by *trans*-'auto' phosphorylation on tyrosines. Subsequently, JAKs phosphorylate STAT proteins, which form homodimeric or heteromeric complexes via their SH2 domains. These complexes translocate to the nucleus, where they bind to specific targeting sequences and influence gene transcription (Horvath and Darnell, 1997).

Janus Kinases

The Janus kinases (JAKs) are cytoplasmic tyrosine kinases which mediate signalling from a number of cell surface receptors which lack intrinsic tyrosine kinase activity. Four mammalian members of the JAK family are known, JAKs 1–3, and TYK2 (Ihle, 1995). Whereas JAK1, JAK2 and TYK2 are expressed ubiquitously, expression of JAK3 is confined to haematopoietic and lymphoid cells. Characteristic of the structure of JAKs is the presence of two JAK homology (JH) domains, of which the C-terminal (JH1) domain has tyrosine kinase activity. Studies of knockout mice have provided important insights into the function of JAKs *in vivo*. The analysis of JAK3 knockout mice and JAK3-deficient humans has clearly demonstrated the essential, nonredundant role of JAK3 in several cytokine signalling pathways. The similarity with γ_c-deficient mice and humans strongly suggests that the major role of γ_c is the recruitment of JAK3 to each γ_c-receptor. In many cases, other JAKs, such as JAK1, that are found in association with the additional subunits of γ_c-containing cytokine receptors, are not sufficient to initiate signalling. JAK3-deficient mice are viable but exhibit severe defects in the development of lymphoid cells, the residual T cells being functionally deficient. Like JAK3-deficient mice, JAK1 deficiency leads to reduced numbers of T and B lymphocytes. Embryonic fibroblasts from JAK1 knockout mice do not respond to class II cytokine receptor ligands IFNγ and IFNα.

Signal Transducer and Activator of Transcription (STAT)

The STATs (signal transducers and activators of transcription) constitute a family of signal transduction proteins that are activated in the cytoplasm by the binding of extracellular polypeptides to transmembrane receptors and which then regulate the transcription of immediate-response genes. Following their obligatory tyrosine phosphorylation, induced by a cytokine ligand, STATs dimerize, translocate to the nucleus and bind directly to response elements present in the promoters of target genes in order to trigger induction of transcription. Thus far, six mammalian STAT proteins (plus several isoforms) have been identified (Darnell, 1997). Two homologues of STAT5 exist (STAT5A and STAT5B) that are encoded by different genes. Expression of STAT proteins is ubiquitous, except for STAT4, which is expressed in several tissues including spleen, heart, brain, peripheral blood cells and testis. Most STATs are activated by many different ligands. IL-2, IL-7, IL-9 and IL-15 activate STAT3 and STAT5, in contrast to IL-4, which activates STAT6. STAT knockout mice mainly show defects in a single or a few cytokine-dependent processes. Embryonic stem cells deficient for STAT3, or with dominant negative STAT3 proteins, fail to stay in an undifferentiated state in the presence of leucocyte inhibitory factors. STAT5A and STAT5B double knockout mice show loss of function with regard to prolactin and growth hormone receptors, i.e. disturbed ovary and mammary gland development and growth retardation. In addition, these mice lack NK cells, develop splenomegaly and have T cells with an activated phenotype, thus resembling IL-2 receptor β-chain-deficient mice. STAT6 knockout mice lack Th2 function as a consequence of impaired IL-4 and IL-13 signalling.

DISEASES ASSOCIATED WITH PERTURBATIONS IN γ_c RECEPTOR/ JAK/STAT SIGNALLING

Because γ_c-dependent cytokines orchestrate a variety of immune system responses via activating the γ_c receptor/ JAK/STAT signalling pathway, it is not surprising that most circumstances causing an inappropriate inhibition of this signalling pathway have generically immunosuppressive consequences. A number of pathological conditions have been identified with mutations or deregulation in γ_c cytokine receptors or associated signalling molecules.

SCID

Severe combined immunodeficiency (SCID) is a hereditary human disease characterized by functionally inactive T and B cells. The ensuing susceptibility to opportunistic infections is the prevalent cause of premature mortality in young patients suffering from this disease. More than 50% of SCID cases are X-linked. XSCID is commonly associated with mutations, which chromosomally map to Xq13, in the γ-chain of the IL-2, IL-4, IL-7, IL-9 and IL-15 receptors. Strikingly, a form of autosomal SCID exists with clinical symptoms identical with X-SCID, in which the gene encoding JAK3 is affected. A newly identified form of SCID with slightly different clinical features involves mutations in the interleukin 7 receptor chain. γ_c-Deficient mice were shown to have hypoplastic thymuses. Thymic cellularity was reduced by 10–25-fold when compared with normal littermates. CD4 and CD8 staining revealed the presence of all thymocyte subpopulations with a slightly increased proportion of CD4+ 'single positive' cells. In the bone marrow, B cell development was blocked at the pre-B cell stage. Although the cellularity of the spleen was reduced approximately 10-fold, mature T and B cells were detected. Both CD5+ and CD5− B-1 cells were identified in γ_c-deficient mice. The numbers of granulocytes, monocytes/macrophages and erythrocytes were normal or increased.

The phenotype of these γ_c-deficient mice indicated that signalling through γ_c is required for the development of multiple lymphoid lineages but not myeloid and erythroid lineages. When compared with human XSCID, a striking difference in B cell development was observed in mouse models of XSCID. In the mouse, B cell development was substantially inhibited at the pro-B cell stage, whereas in human XSCID, the production of B cells is outwardly normal. IL-7/IL-7R appears to represent the γ_c-dependent cytokine for mouse B cell development. The failure to block B cell development in human XSCID suggests a γ_c-independent pathway for the production of B cells in these patients.

Immunosuppressive Diseases and Suppression of JAK/STAT Signal Transduction

Cytokine receptor signalling substrates, in particular the JAKs and STATs, contribute to tumorigenesis. In *Drosophila*, a dominant mutant Jak kinase causes leukaemia-like abnormalities. In mammals, JAKs and STATs are known to be constitutively activated in haematopoietic cells transformed by diverse oncogenic tyrosine kinases and in a variety of lymphomas and leukaemias. Expression of a constitutively active STAT3 molecule in immortalized fibroblasts causes cellular transformation. Together these data are indicative of a role for constitutive activation of the JAK-STAT pathways in leukaemogenesis.

INTERFERON SIGNALLING PATHWAY

Interferons (IFNs) play a key role in mediating antiviral and antigrowth responses and in modulating the immune response (Seder, 1994; Trinchieri and Scott, 1995; Young and Hardy, 1995). Their signalling pathways provided the first evidence of, and have been used as the model for, the JAK–STAT pathway, which is utilized by many cytokines (Darnell *et al*., 1994). IFNs can be subdivided into two functional classes, and constitute the largest and most divergent subfamily of cytokines. There are more than 20 members in the type I IFN class (e.g. IFNαs, -β, and -τ) and one member in the type II IFN class, i.e. IFNγ. Type I and II IFNs function via related but distinct signal transduction pathways. In both classes of IFNs, signalling is initiated by the binding of the IFNs to their specific membrane receptors, that are expressed in many different cell types.

Interferon-γ Signalling Pathway

Type II, immune or IFNγ, is secreted by thymus-derived (T) cells under certain conditions of activation and by natural killer (NK) cells. The proximal events of IFNγ signalling require the obligatory participation of five distinct proteins, two IFNγ specific receptors, IFNGR1 and IFNGR2, two Janus kinase family members, JAK1 and JAK2, and one STAT family member, STAT1 (Schindler and Darnell, 1995).

IFNγ receptors contain a minimum of two peptide chains that are expressed in nearly all cell types, displaying a strict species specificity in their ability to bind IFNγ. The 90-kDa IFNGR1 consists of three domains, an extracellular (229aa), an intracellular (223aa) and a transmembrane (21aa). The IFNGR1 contains a constitutive and specific JAK1 binding site in the membrane-proximal region of the intracellular domain and a STAT1 binding and phosphorylation site that is activated upon phosphorylation by JAK1. The 62-kDa IFNGR2 differs from the 90-kDa IFNGR1 in that it contains a JAK2 binding site in its intracellular domain. Both IFNGR1 and IFNGR2 are required to activate functionally the IFNγ signalling pathway. IFNGR1 plays more important roles in mediating ligand binding, ligand trafficking through the cell and signal transduction, whereas IFNGR2 plays only a minor role in ligand binding.

Based on the available data and observations, Stark and colleagues have proposed a model, which has been broadly accepted for the IFNγ signalling pathway (**Figure 1**) (Stark *et al*., 1998). When a functional IFNγ homodimer binds two IFNGR1s on their extracellular domains, the IFNGR1 and IFNGR2 are brought into close proximity, forming more stable heterodimers together with their preassociated, inactive JAKs, JAK1 and JAK2. The

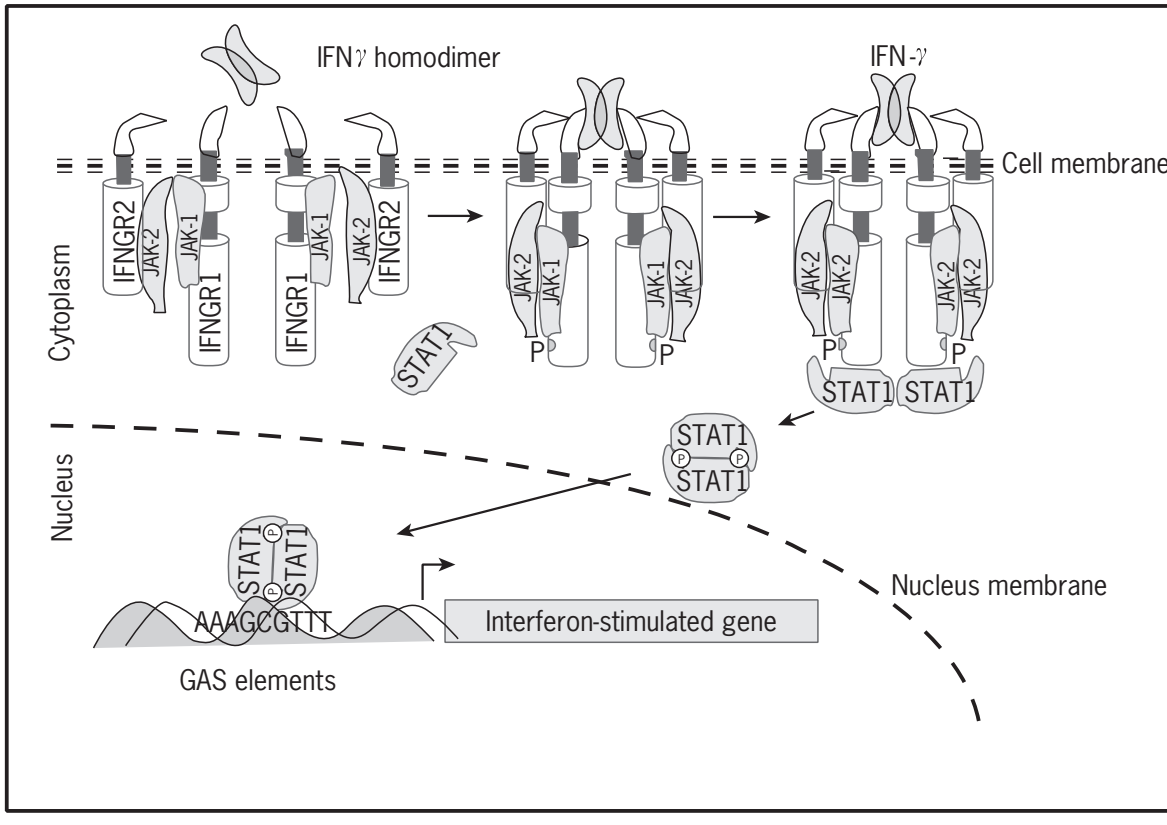


Figure 1 IFNγ signal pathway.

intracellular membrane-proximity domains of IFNGRs are then activated through auto- and transphosphorylation. The activation of the JAKs occurs in sequence, such that JAK2 activates first and is required for JAK1 activation.

Once activated, the JAKs phosphorylate a tyrosine-containing sequence near the C-terminus of IFNGR1, where paired ligand-induced docking sites for STAT1 are formed. Two inactive, monomeric STAT1 proteins then bind to these sites through their SH2 domain and are phosphorylated by the receptor-bound JAK kinases at tyrosine 701, near their C-terminus. After phosphorylation, the STAT1 proteins dissociate from the receptor and form a reciprocal homodimer, which then translocates to the nucleus, inducing the transcription of a set of IFN-stimulated genes (ISGs) via binding to the specific DNA elements residing within the ISGs promoters, and designated as either the interferon stimulated response element (ISRE) or the IFNγ activated site (GAS) (Darnell *et al*., 1994; Schindler and Darnell, 1995; Stark *et al*., 1998) (**Figure 1**).

A similar signalling pathway mechanism in which a ligand-induced, tyrosine-phosphorylated docking site on a receptor, and its association with the transcription factor STAT, has been found to mediate responses to other cytokines. As a result, this JAK–STAT signalling pathway mechanism is now the accepted paradigm that illustrates the important mechanism of how cytokine receptors are coupled to their specific STAT signalling systems.

The negative regulation of the IFNγ signalling pathway has also been defined recently by the discovery of a family of proteins known as SOCS/JAB/SSI, which are induced by IFN-γ (and also several other cytokines) and bind to and inhibit activated JAKs (Yasukawa *et al*., 2000). These discoveries have provided new insights into how JAK–STAT pathways are regulated in response to specific stimuli, and how they function in various tissues and environments.

Interferon-α/β Signalling Pathway

The common pathway for IFNα/β requires seven distinct proteins, which include two IFNα receptors, two JAKs, two STATs and the IRF-family transcription factor p48. The IFNα/β signalling pathways are comparatively illustrated in **Figure 2**, but the fine details of the mechanisms are lacking because information regarding the detailed interactions that play the crucial role for the pathway remain to be elucidated.

The IFNα/β receptors, designated IFNRAs, are composed of a multichain structure on both normal and malignant haematopoietic cells (Novick *et al*., 1994). All type I interferons, including IFNα, IFNβ, and IFNω, bind to the same receptors (Pestka *et al*., 1987). It has been clearly established that there are two distinct components comprising the type I IFNRAs, namely the IFNRA1 and

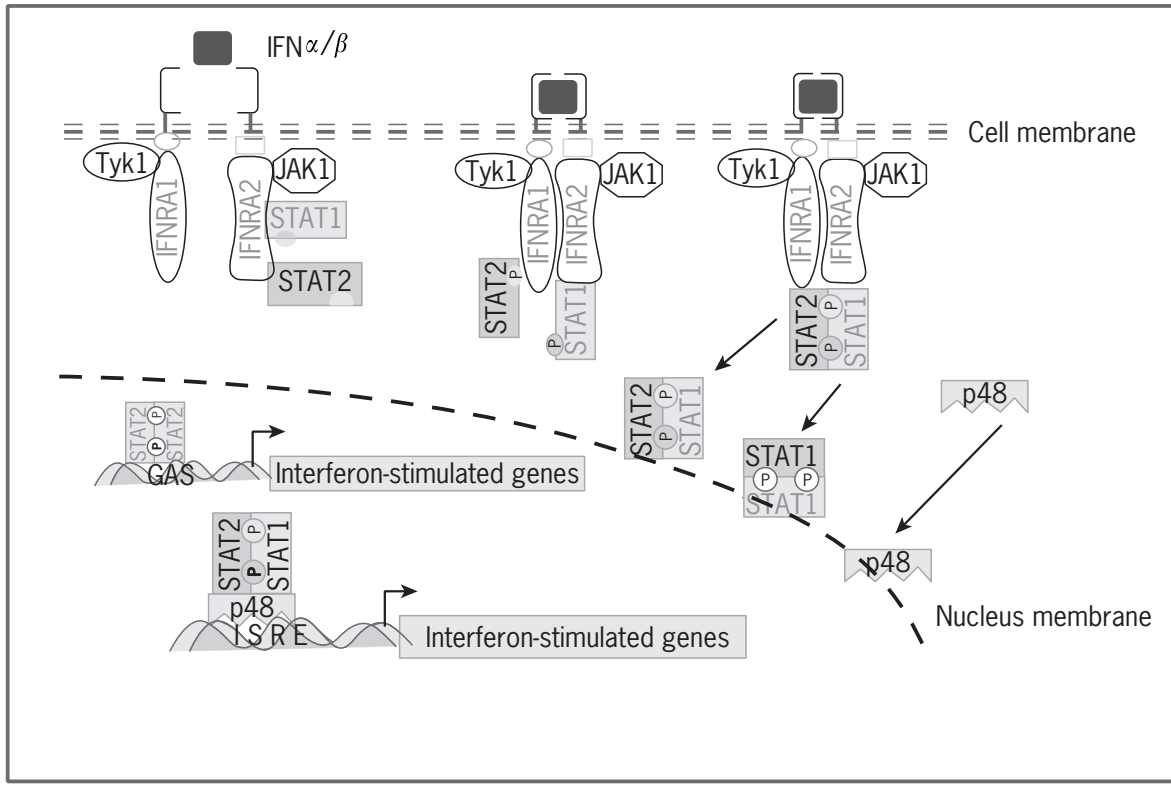


Figure 2 IFNα/β signal pathway.

the IFNRA2. IFNRA1 is a 110-kDa protein. The IFNRA2 subunit occurs in three different forms that are differentially spliced products of the same gene (Colamonici *et al*., 1994a; Novick *et al*., 1994), the soluble form of the extracellular domain, IFNRA2a; the alternatively spliced variant with a short cytoplasmic domain, IFNAR2b, which can have dominant negative activity and the only normal, fully functional form, IFNRA2c protein, with a relative molecular mass of 90–100 kDa. Neither IFNAR1 nor IFNAR2 alone binds to IFNα/β with the high affinity of the two-subunit combination (Cohen *et al*., 1995).

Two members of the Janus family of tyrosine kinases involved in the IFNα/β signalling pathway, Tyk2 and JAK1, are both constitutively associated with the IFNAR1 and IFNAR2c subunits, respectively (Colamonici *et al*., 1994c, 1995; Domanski *et al*., 1995, 1997). Upon the binding of IFNα/β to the IFNARs, the tyrosine kinases are activated and IFNAR1 and IFNAR2c are rapidly phosphorylated on tyrosines. Tyk2 also plays a role in stabilizing the IFNAR1 structure because the amount of IFNAR1 is low in Tyk2-null cells. However, the domains required for this role are different from those required to transduce the cytokine induced signal (Velazquez *et al*., 1995). The activation of tyrosine kinases results in tyrosine phosphorylation of two STAT proteins, STAT1 and STAT2, both preassociated with IFNAR2c in untreated cells (Fu, 1992; Darnell, 1997, 1998). When IFNAR1 is phosphorylated, the SH2 domain of STAT2 binds to it, followed by the phosphorylation of both STATs and subsequent dissociation of the phosphorylated heterodimer from the receptors. In addition to the STAT1–STAT2 heterodimer, much evidence also supports the IFNα/β-mediated induction of STAT1 homodimers (Darnell, 1997; Stark *et al*., 1998).

The phosphorylated STAT proteins form homo- and heterodimers and then translocate to the nucleus, associating with p48 to regulate gene transcription via binding to specific sequences present in the promoters of interferon-stimulated genes (ISGs) (Darnell, 1997; Ihle, 1996). The association of STAT1–STAT2 heterodimers with p48 to form the mature interferon-stimulated gene factor-3 (ISGF3) complex represents the major transcription factor complex formed in response to IFNα/β, and is required to drive the expression of most ISGs by binding to ISREs. While the STAT1–STAT2 heterodimer and STAT1 homodimer form in response to IFNα/β independently of p48, and can each drive the expression of a minority of ISGs, such as the IRF-1 gene, through GAS elements, the relative amount of STAT protein homo- and heterodimer and its complex formation with p48 depends on the level of p48, which can vary greatly among different cell types (Li *et al*., 1998). The functionality of p48 in the IFNα/β pathway seems to be that of an adaptor between STAT proteins and the DNA binding sites to redirect gene regulation and achieve the specificity of biological functions of IFNs.

TUMOUR NECROSIS FACTOR: RECEPTORS AND SIGNAL TRANSDUCTION PATHWAYS

The tumour necrosis factor (TNF) receptor superfamily comprises a group of cell surface receptors whose members generally bind ligands that are structurally related to TNF. The TNF ligand is structurally related to lymphotoxin-α (LTα; sometimes referred to as TNF-β), which is secreted from activated T cells, but binds the same receptors as TNF and has similar biological properties. TNF and LTα, however, are the prototype members of a large family of related proteins which includes CD30, CD40, Fas ligand and TRAIL ligand. TNF is a major physiological mediator of inflammation. It initiates the response to Gram-negative bacteria that produce lipopolysaccharide (LPS). TNF has been shown to induce fever, activate the coagulation system, induce hypoglycaemia, depress cardiac contractility, reduce vascular resistance, induce cachexia and activate the acute phase response in the liver.

Ironically, attempts to use TNF in the clinic actually predate its discovery and characterization. Towards the end of the nineteenth century, it was noticed that a small number of cancer patients experienced disease regression after suffering systemic bacterial infections. Subsequently a mixture of killed *Streptococcus pyogenes* and *Serratia marcescens* ('Coley's toxins') were administered to patients with advanced cancer, albeit with very occasional success. This approach became the treatment of choice for over three decades until superseded by advances in radiotherapy, chemotherapy and surgery. With hindsight, the most likely explanation for the results observed with Coley's toxins was the production of TNF, largely by macrophages in response to bacterial lipopolysaccharide present in the cell wall of Gram-negative bacteria such as *Serratia* sp.

TNF and Its Receptors

TNF and LTα are closely related homotrimeric proteins (32% identity). Human TNF is synthesized as a 233 amino acid glycoprotein, containing a long (76 residue) N-terminal leader sequence which anchors it to the cell membrane as a 25-kDa type II membrane protein. A secreted 17-kDa form of TNF is generated through the enzymatic cleavage of membrane-bound TNF by a metalloproteinase termed TNF-α-converting enzyme (TACE). Both soluble and membrane-bound forms of TNF are biologically active, although they have different affinities for the two TNF receptors, and probably as a consequence exhibit different biological properties (see below).

LTα differs from TNF in that it is synthesized as a secreted glycoprotein. Human LTα is synthesized as a 205 amino acid glycoprotein, which in native form exists as a 25-kDa homotrimer. As mentioned above, LTα can bind

both TNF receptors with affinities comparable to those of TNF, and has similar biological effects. However, a membrane-bound form of LT has been identified which consists of a heterotrimeric complex containing one LTα subunit noncovalently linked to two molecules of an LTα-related type II membrane protein termed LTβ. The LT$\alpha_1\beta_2$ heterotrimer (also known as mLT) is not cleaved by TACE and is thought to exist exclusively as a membrane-bound complex. mLT does not bind either of the two TNF receptors, but rather exerts its effects on another member of the TNF receptor superfamily, the lymphotoxin β receptor (LTβR). TNF and the two LT subunits are encoded by closely linked single copy genes, which are situated in the class III major histocompatibility locus, within a 25-kb region on the short arm of chromosome 6 in humans, at p21.

The two receptors for TNF (and LTα) are type I transmembrane glycoproteins designated TNFR1 (also termed p60 in humans, p55 in mice) and TNFR2 (also known as p80 in humans, p75 in mice). These receptors are characterized by cysteine-rich repeats of about 40 amino acids in their N-terminal extracellular domains. Each extracellular domain consists of three or four cysteine-rich regions containing four to six cysteines involved in intra-chain disulfide bonds. The cytoplasmic domains of these receptors have no obvious similarity to any known kinase and are thought to lack any intrinsic enzymatic activity. Signal transduction is therefore achieved by the recruitment and activation of adaptor proteins which recognize specific sequences in the cytoplasmic domains of these receptors. Recruitment of adaptor molecules activates a number of characteristic signalling pathways that can lead to a remarkably diverse set of cellular responses including differentiation, activation, release of inflammatory mediators and apoptosis. (See also chapter *Apoptosis*.)

Signal Transduction

Death Domains TRADD and FADD

The principal molecule thought to be involved in TNFR1 signal transduction is TNF receptor-associated death domain (TRADD), which is recruited to TNFR1 after activation by TNF. The interaction between TNFR1 and TRADD is mediated by the death domain, a motif found in both adaptor molecules such as TRADD and the cytoplasmic domains of the receptor itself (**Figure 3**). The binding of TRADD to TNFR1 leads to the recruitment and activation of numerous associated signalling molecules. TNF-induced apoptosis is generally thought to be achieved by the interaction of TRADD with FADD (Fas-associated death domain; also known as MORT1), a ~27-kDa protein which oligomerizes with TRADD through the death domains contained in both molecules. Recruitment of FADD is thought to activate a cascade of events which ultimately lead to apoptosis. This is brought about by the coordinate activation of several members of the caspase family. Caspases are cysteine aspartate proteases which are originally synthesized as zymogens, and are typically converted to their activated form by proteolytic cleavage, often by a distinct caspase upstream in the proteolytic cascade. Caspase-8, which is generally considered to be the apical caspase in the TNF and Fas pathways, is recruited to FADD in the activated complex, and is thought to be activated by self-cleavage induced by an increase in its local concentration. Cleaved caspase-8 can subsequently activate downstream caspases, notably caspase-3, and thereby induce apoptosis.

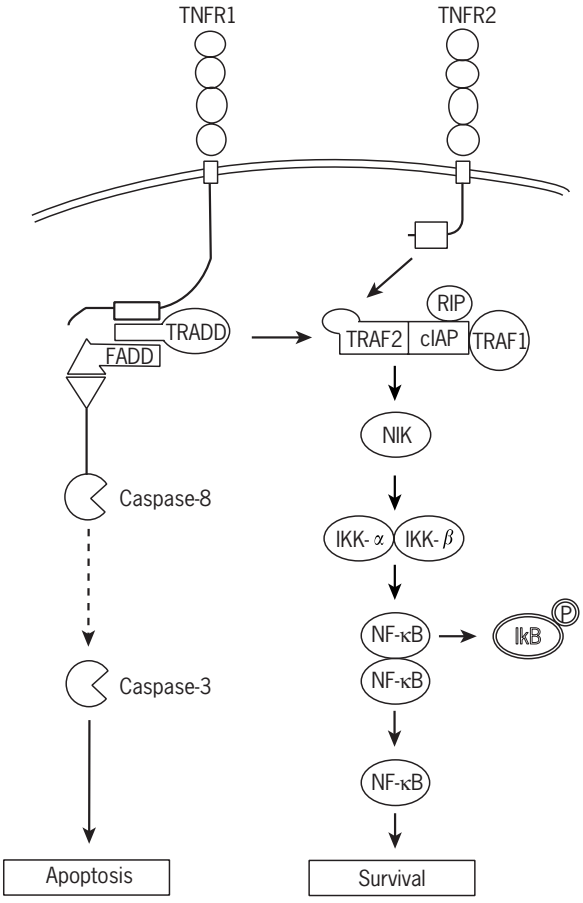


Figure 3 The role of the death domain- and death effector domain-containing molecules in signalling by TNFR1 and TNFR2.

TRAFs

The cytoplasmic domain of TNFR2 does not contain a death domain. In fact, under many circumstances, TNFR2 signalling can induce proliferation. This is thought to be mediated by the direct interaction of the intracellular signalling intermediate, TRAF2, with the cytoplasmic tail of TNFR2, which leads to the activation of NF-κB. In cells

which respond to TNFR2 signals by proliferating, no caspase activation is observed, and presumably no interaction of TRAF2 with TRADD. However, in cells which respond to TNFR2 signalling by undergoing apoptosis, there is mounting evidence that this is effected by a signal crosstalk mechanism with TNFR1, possibly by inducing expression of membrane-bound TNF or by affecting the stability of prosurvival proteins such as TRAF2 (Duckett and Thompson, 1997). As mentioned above, LTα has been shown to bind to both TNFR1 and TNFR2 with comparable affinities and biological outcomes. Membrane LT does not bind to either of these receptors, but binds exclusively to the LTβ receptor. In comparison with TNFR1 and TNFR2, the signal transduction pathways utilized by the LTβ receptor have been less well defined. Signalling by LTβR has been shown to induce apoptosis, although examination of the signalling, cytoplasmic tail has not revealed any obvious homology to the death domain. TRAF3 and TRAF5 have been shown to bind LTβR, but the role of TRAF3 in apoptosis induction is unclear.

TRAF2 is also a central component of the TNFR1 signalling complex, through a direct interaction with TRADD. Although TRAF2 lacks intrinsic enzymatic activity, it has been shown to bind several serine–threonine kinases, including NIK (NF-κB-inducing kinase), RIP (receptor interacting protein) and GCK (germinal centre kinase). Through the recruitment of these kinases, TRAF2 is thought to induce the activation of several transcription factors, particularly NF-κB (nuclear factor-κB), as well as downstream kinases involved in stress responses, notably c-Jun N-terminal kinase (JNK), which are crucial effectors of the TNFR1-mediated proinflammatory reaction. TRADD is also thought to associate with the serine–threonine kinase RIP, and RIP has been shown to interact with another protein, RAIDD (RIP associated Ich-1/CED3 homologous protein with death domain; also known as CRADD), causing the recruitment and activation of caspase-2 and the induction of apoptosis. However, the RAIDD/CRADD pathway is not thought to be the major signalling pathway utilized by TNFR1 to induce apoptosis. Despite the well-defined ability of TNFR1 signalling to induce cell death, the majority of normal cells do not respond to TNF by undergoing apoptosis. It has been proposed that this paradoxical situation can be accounted for by the activation of NF-κB, which has been shown to induce the expression of a number of antiapoptotic proteins. The best described of these are (i) A20, a zinc finger-containing molecule, (ii) A1/BflI, a Bcl-2 homologue, and (iii) c-IAP1/c-IAP2, members of the IAP (inhibitor of apoptosis) family.

Many key signalling intermediates responsible for the induction of NF-κB by the TNF receptors have been identified. NF-κB transcription factors are sequestered in the cytoplasm of cells by a protein called IκB (inhibitor of NF-κB). Phosphorylation of IκB leads to its degradation via ubiquitination by the 26S proteasome. The heterodimeric NF-κB subunits then translocate to the nucleus where they regulate expression of a wide variety of genes involved in inflammatory responses. A complex of two kinases and a regulatory protein are responsible for IκB phosphorylation. The kinases are termed IκB kinases α (IKKα) and β (IKKβ) and are constitutively associated with the regulatory protein IKKγ or NEMO. The serine–threonine kinase NIK, initially identified through its ability to associated with TRAF2, is thought to activate the IKK complex. A naturally occurring mouse mutation termed alymphoplasia (*aly*) is the result of a point mutation of NIK. *Aly/aly*$-/-$ mice lack lymph nodes and Peyer patches, and also exhibit disorganized splenic and thymic structures.

Other Pathways

Alternative models have been proposed to account for the diverse range of outcomes following TNFR1 activation. Notably, SODD (silencer of death domains) was identifed in the basis of its ability to bind to the death domain of TNFR1. SODD is found in the TNFR1 receptor complex before receptor activation, but then dissociates from the receptor after ligand binding. It is thought that the SODD preassociation with TNFR1 may prevent spontaneous signalling by death domain-containing receptors. Neutral sphingomyelinase (N-Smase) activation is thought to mediate some of the inflammatory and proliferative responses to TNF through the activation of ERK and phospholipase A2. Other factors have been identifed which bind distinct sites in the cytoplasmic tail of TNFR1, such as FAN (factor associated with N-Smase activation), which is thought to couple the TNFR to activation of neutral sphingomyelinase activity, which in turn results in the production of ceramide and is thereby thought to lead to the activation of MAP kinases.

CHEMOKINE RECEPTOR SIGNAL TRANSDUCTION

Chemoattractant cytokines are typically <15-kDa proteins which are secreted by many tissue and cell types. Chemokines were recently reviewed (Zlotnik *et al*., 1999). The classical chemokine-induced biological activity is leucocyte migration, but as the field has matured other chemokine-mediated physiological functions have been identified, including regulation of vascularization in embryogenesis, lymphocyte maturation, cellular activation, regulation of angiogenesis and apoptosis. Although these additional functions have been demonstrated, the components of the distinct signalling cascades are poorly characterized, so this section will focus on chemokine-induced chemotaxis, cell activation and apoptosis.

Seven Transmembrane G-protein-coupled Receptors (GPCRs)

Chemokines bind to and activate seven pass transmembrane G-protein-coupled receptors (GPCRs) that are structurally similar to the rhodopsin (type A) subfamily. The N-terminus of the receptor, which is also the first extracellular domain, is essential for high-affinity ligand binding. The closely packed position of the seven transmembrane domains are maintained by disulfide bonds between the extracellular domains. The disulfide bonding is needed for efficient chemokine-induced signalling but is not necessary for HIV-1 coreceptor activity. The chemokine receptor signal is dependent on a ligand-induced dynamic change in the receptor that results in an increased affinity for Gi and Gq heterotrimeric G-proteins. Fine regulation of the GPCR signal occurs at the membrane where increased phosphatidylcholine in the lipid bilayer enhances GTPase activity of the G-proteins and is essential for the activity of G-protein coupled receptor kinase(s) (GRKs).

Heterotrimeric G-proteins

Chemokine-induced chemotaxis is inhibited by pertussis toxin, indicating that chemokine receptors activate trimeric G proteins in the Gi subfamily. However, activation of phospholipase C (PLC), intracellular calcium (Ca^{2+}) mobilization and cellular exocytosis can be mediated through pertussis toxin-insensitive Gq proteins. The type of cell expressing the GPCR and its activation state regulates which G-protein couples to the receptor, such that chemokine receptors have been reported to couple to several classes of G-proteins.

Heterotrimeric G-proteins are composed of α-, β- and γ-subunits. In the resting G-protein, the α-subunit is bound to GDP. Once activated, GDP is exchanged for GTP and the α-subunit separates from the β–γ-subunits. The α-subunits are apparently necessary for regulation of the GPCR function. Regulators of G-protein signal (RGS) are widely expressed GTPase-activating proteins that contain a 130 amino acid domain that binds Gα-GTP subunits accelerating the hydrolysis to Gα-GDP and blocking Gα interaction with PLC. Additionally, some Gα subunits are substrates for protein kinase C (PKC) resulting in autoregulation of this G-protein-mediated signal. Gαi subunits were shown to be nonessential in chemokine-induced chemotaxis. These data indicated that any Gα linked to β–γ-subunits, which are essential, could transmit a chemotactic signal. In addition to transmitting the chemotactic signal, β–γ-subunits participate in signal component receptor docking and cell activation.

Heterotrimeric G-proteins interact with several intracellular domains of GPCRs found in the cytoplasmic tail and second and third intracellular loops. The β–γ-subunits were shown to guide GRK2 to its phosphorylation site on a GPCR, suggesting that both the correct membrane lipid composition and G-protein components are needed for GRK regulation of GPCR signal. The G-protein β–γ-subunits directly bind to and activate Raf-1, phosphoinositide 3-kinase gamma (PI3Kγ) and some small GTPases. Raf-1 is a serine/threonine kinase that links the mitogen-activated protein kinase cascade to tyrosine kinase-dependent growth factor receptors. Mutagenesis studies showed that β–γ-subunits bind to Raf-1 with an affinity similar to that between β–γ-subunits and GRKs, suggesting that there maybe a competition between receptor inactivation by GRKs and the mitogenic signal. Stromal-derived factor-1 (SDF-1, also known as CXCL12) is a CXC chemokine that activates the CXCR4 receptor. Recently it was shown that SDF-1-induced chemotaxis and extracellular signal-related kinase (ERK) activation was inhibited by PI3K inhibitor treatment of T lymphocytes. These data show a link between G-protein β–γ-subunits and PI3K and suggest an order of activation.

Mutagenesis of G-protein β-subunits resulted in inappropriate organisation of cellular cytoskeleton (Peracino *et al*., 1998), and additional studies have shown that small GTPases of the Rho family are essential for chemotaxis and sequestration of β–γ-subunits leads to rearrangement of the actin cytoskeleton, suggesting a link between β-subunits and these small GTPases.

Phospholipase

An approach to demonstrate the role of PLC in the GPCR signal cascade used mice lacking PLC-β2 and -β3. Neutrophils from animals lacking both PLC-β2 and -β3 did not produce inositol triphosphate (IP_3), flux calcium or superoxide in response to chemokines or chemoattractants. Animals lacking only PLC-β2 clearly showed reduced both IP_3 and Ca^{2+} flux in response to interleukin-8 (IL-8 also known as CXCL8) and macrophage inflammatory protein 1 beta (MIP1β, also known as CCL4), but not to the same extent, suggesting that both PLC isoforms participate in signal transduction in neutrophils. In contrast to the calcium flux signal, the chemotaxic response of the neutrophils from animals lacking both PLC-β2 and -β3 was not reduced; rather, there was an enhanced chemotactic response to IL-8. The PLC-β2- and -β3-deficient animals failed to activate PKC in response to chemoattractants, indicating that the PKC pathway is linked to the Gα and not the Gβ–γ signal. Additionally, PLC-deficient animals failed to phosphorylate the mitogen-activated protein kinase (MAPK) c-JUN N-terminal kinase (JNK). However, PLC deficiency had no effect on Rac activation, suggesting that cytoskeletal modification is a separate signal in neutrophils.

In addition to activation of PLC, there is strong evidence that phospholipase D (PLD) is also activated by chemokines. The position of PLD in the chemokine signal

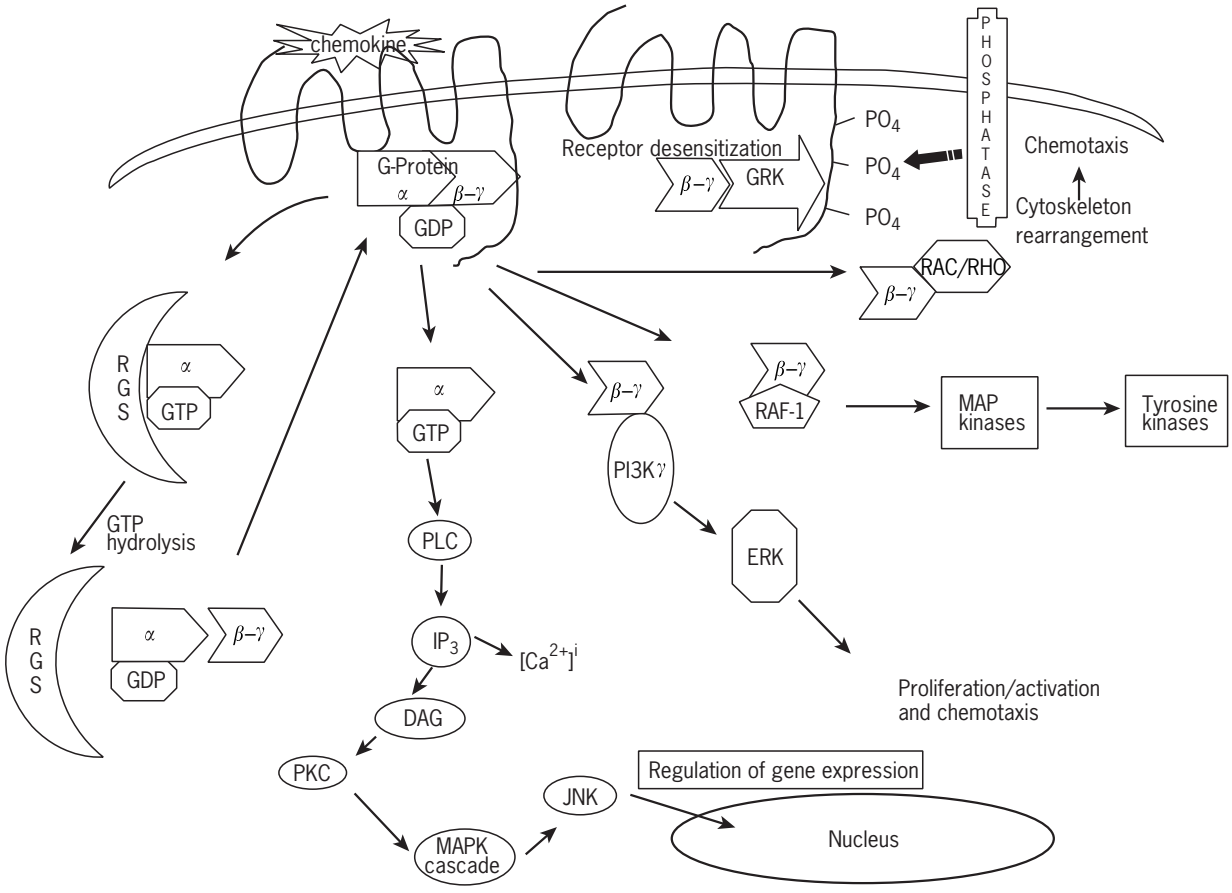


Figure 4 G-protein-coupled chemokine signalling.

cascade is ambiguous because diacylglycerol (DAG) can be interconverted to the lipid hydrolysis product of PLD, phosphatidic acid (PA), suggesting that PLC activity may regulate PLD activity. Additionally, PLD is activated by Ras and Rho family members and PKC. A function for PA in chemokine-induced cell activation has not been identified, but PA is strongly associated with cell vesicle transport, suggesting that PLD may act late in the chemokine-induced cascade by regulating receptor localization to the membrane (**Figure 4**).

Further, PA is in the pathway that leads to respiratory burst (Lennartz, 1999), suggesting that PLD activation by chemokines maybe participate in chemokine-induced NADPH production.

Phosphoinositide 3-Kinase Gamma (PI3Kγ)

PI3Ks have been implicated in many cellular responses, including, proliferation, apoptosis, adhesion and chemotaxis. Recently, a number of groups generated mice deficient in PI3Kγ. All three groups observed a severe reduction ($\leq$85%) in chemokine-induced chemotaxis, but there was some activity left, indicating that Gβ–γ may link to other intracellular components and induce chemotaxis. The least effected was MIP-5-induced chemotaxis (Hirsch *et al*., 2000). MIP-5 (also known as HCC-2 or CCL-15) has a unique six-cysteine structure. The effect this chemokine has on other intracellular signalling components has not been evaluated, but mutational analysis indicated that the third set of disulfide bonds were not needed to induce chemotaxis. PI3Kγ deficiency had no effect on chemoattractant-induced Rac activation, actin polymerization or calcium flux. PI3Kγ deficiency had a profound inhibitory effect on chemoattractant-induced activation of PKB, ERK1 or ERK2. The activation of ERK1 and ERK2 by PI3Kγ directly links the chemokine GPCR signal to both proliferation and activation signals. Taken together, these data indicate that PI3Kγ is a major component in chemokine-induced chemotaxis and cell activation, but not the only component.

Focal Adhesion Kinases and Tyrosine Phosphatases

Several tyrosine kinases are activated during chemotaxis, but whether the effect is mediated through a GPCR pathway or by an adhesion molecule-activated signal pathway

is still unclear. Focal adhesion kinase (FAK) activity is clearly stimulated by RANTES (regulated on activation of normal T cell expression and secreted, also known as CCL5) binding to CC chemokine receptor-5 (CCR5).

SDF-1 and stem cell factor appear to activate related adhesion focal tyrosine kinase (RAFTK), a FAK homology expressed in some leucocytes. The FAK family has been shown to activate or be activated by some MAPK and Rho family members. Thus, although there is no disagreement that focal adhesion kinases are activated in cell migration, their role in chemokine-induced cell migration and activation is unclear.

Studies using tyrosine phosphatase, SHIP-1 and SHP-1-deficient mice showed enhanced chemotaxis and reduced suppression of proliferation in response to chemokines. Activation of CD45, a cell surface glycoprotein with tyrosine phosphatase activity, decreased the cell surface expression of IL-8 receptors and IL-8-induced calcium flux. While it is tempting to theorize a role for these phosphatases in G-protein-mediated chemokine signalling, more research is needed to connect these signals clearly. Recent reviews suggest that the components of adhesion, selectins and integrin receptors, have bidirectional signalling cascades, and that the activation of focal adhesion kinases and tyrosine phosphatases may be an 'outside in' signal.

Apoptosis

Studies of cell death in HIV-1-infected cells showed that CXCR4 is required for this apoptotic signal and the signal is G-protein independent. Pertussis toxin, PI3K inhibitor (wortmannin), ERK pathway inhibitors and MAPK pathway inhibitors had no effect on CXCR4-mediated HIV-1-induced lymphocyte apoptosis. CCR5 does signal through ERK and MAPK when activated by HIV-1. Gp120-independent virus–cell fusion and recombinant gp120 did not induce apoptosis, but, cell surface-expressed gp120 did.

Deletion of the cytoplasmic tail of CXCR4 blocked ligand-induced receptor internalization, but did not block HIV-1-induced apoptosis. Since none of the well-characterized chemokine signal components appear to be necessary for the CXCR4-mediated apoptosis signal, what is? Caspase 3 activity correlates with CXCR4-mediated cell death. These data indicate that HIV-1 induces CXCR4-dependent apoptosis by activating caspase 3.

Musings

The chemokine-induced cellular signal is complex and still poorly characterized. The activation of G-protein subunit-mediated signals is clear, but other non-G-protein-mediated signals still require clarification. Receptor desensitization and resensitization are a great mystery, with few well-characterized components, other than serine–threonine phosphorylation of the GPCR carboxyl tail. Several groups have shown tyrosine kinase activation in response to chemokines; however, because the cell shape change and activation of these kinases parallel each other, an integrin receptor-mediated signal cannot be ruled out. Further, activation of Janus kinases (JAKs) by chemokines has been demonstrated in human embryonic kidney and lymphocyte cell lines. These studies took place in membrane raft microdomains, so that other components of the rafts may have contributed to the activation. Characterization of chemokine-induced activation of signal transducers and activators of transcription (STATs) has begun. SDF-1, RANTES and MIP1α have been shown to stimulate STAT phosphorylation. However, it remains to determine the mediator, be it binding to a domain on the chemokine receptor or JAK activation. Further structural analysis of chemokine receptors is needed to link conclusive tyrosine kinases and STAT activation to chemokine receptors. After 12 years, our understanding of chemokine-mediated signal transduction is almost ready to enter adolescence, but a long way from a mature field.

REFERENCES

Ahmad, M., *et al*. (1997). CRADD, a novel human apoptotic adaptor molecule for caspase-2, and FasL/tumor necrosis factor receptor-interacting protein RIP. *Cancer Research*, **57**, 615–619.

Arai, H. and Charo, I. F. (1996). Differential regulation of G-protein-mediated signaling by chemokine receptors. *Journal of Biological Chemistry*, **271**, 21814–21819.

Bach, E. A., *et al*. (1997). The IFN gamma receptor a paradigm for cytokine receptor signaling. *Annual Review of Immunology*, **15**, 563–591.

Bacon, K. B., *et al*. (1996). RANTES induces tyrosine kinase activity of stably complexed p125FAK and ZAP-70 in human T cells. *Journal of Experimental Medicine*, **184**, 873–882.

Beutler, B. A. (1999). The role of tumor necrosis factor in health and disease. *Journal of Rheumatology*, **26**, Suppl. 57, 16–21.

Blanco, J., *et al*. (1999). The implication of the chemokine receptor CXCR4 in HIV-1 envelope protein-induced apoptosis is independent of the G protein-mediated signalling. *Aids*, **13**, 909–917.

Bluyssen, H. A., *et al*. (1995). Combinatorial association and abundance of components of interferon-stimulated gene factor 3 dictate the selectivity of interferon responses. *Proceedings of the National Academy of Sciences of the USA*, **92**, 5645–5649.

Cohen, B., *et al*. (1995). Ligand-induced association of the type I interferon receptor components. *Molecular and Cellular Biology*, **15**, 4208–4214.

Colamonici, O. R., *et al*. (1994a). Interferon alpha (IFN alpha) signaling in cells expressing the variant form of the type I IFN receptor. *Journal of Biological Chemistry*, **269**, 5660–5665.

Colamonici, O. R., *et al*. (1994b). Ligand-independent anti-oncogenic activity of the alpha subunit of the type I interferon receptor. *Journal of Biological Chemistry*, **269**, 27275–27279.

Colamonici, O. R., *et al*. (1994c). p135tyk2, an interferon-alpha-activated tyrosine kinase, is physically associated with an interferon-alpha receptor. *Journal of Biological Chemistry*, **269**, 3518–3522.

Colamonici, O. R., *et al*. (1995). Transmembrane signaling by the alpha subunit of the type I interferon receptor is essential for activation of the JAK kinases and the transcriptional factor ISGF3. *Journal of Biological Chemistry*, **270**, 8188–8193.

Coppolino, M. G. and Dedhar, S. (2000). Bi-directional signal transduction by integrin receptors. *International Journal of Biochemical Cell Biology*, **32**, 171–188.

Darnell, J. E., Jr, *et al*. (1994). Jak-STAT pathways and transcriptional activation in response to IFNs and other extracellular signaling proteins. *Science*, **264**, 1415–1421.

Darnell, J. E., Jr (1997). STATs and gene regulation. *Science*, **277**, 1630–1635.

Darnell, J. E., Jr (1998). Studies of IFN-induced transcriptional activation uncover the Jak–Stat pathway. *Journal of Interferon and Cytokine Research*, **18**, 549–554.

Domanski, P., *et al*. (1997). A region of the beta subunit of the interferon alpha receptor different from box 1 interacts with Jak1 and is sufficient to activate the Jak–Stat pathway and induce an antiviral state. *Journal of Biological Chemistry*, **272**, 26388–26393.

Domanski, P., *et al*. (1995). Homodimerization and intermolecular tyrosine phosphorylation of the Tyk-2 tyrosine kinase. *FEBS Letters*, **374**, 317–322.

Duckett, C. S. and Thompson, C. B. (1997). CD30-dependent degradation of TRAF2 implications for negative regulation of TRAF signaling and the control of cell survival. *Genes and Development*, **11**, 2810–2821.

Fu, X. Y. (1992). A transcription factor with SH2 and SH3 domains is directly activated by an interferon alpha-induced cytoplasmic protein tyrosine kinase(s). *Cell*, **70**, 323–335.

Gonzalez-Amaro, R. and Sanchez-Madrid, F. (1999). Cell adhesion molecules: selectins and integrins. *Critical Reviews in Immunology*, **19**, 389–429.

Gouilleux-Gruart, V., *et al*. (1997). Activated Stat related transcription factors in acute leukemia. *Leukemia Lymphoma*, **28**, 83–88.

Greenlund, A. C., *et al*. (1995). Stat recruitment by tyrosine-phosphorylated cytokine receptorsan ordered reversible affinity-driven process. *Immunity*, **2**, 677–687.

Haque, S. J. and Williams, B. R. (1994). Identification and characterization of an interferon (IFN)-stimulated response element-IFN-stimulated gene factor 3-independent signaling pathway for IFN-alpha. *Journal of Biological Chemistry*, **269**, 19523–19529.

Harada, H., *et al*. (1996). Regulation of IFN-alpha/beta genes-evidence for a dual function of the transcription factor complex ISGF3 in the production and action of IFN-alpha/beta. *Genes and Cells*, **1**, 995–1005.

Hayakawa, F., *et al*. (1998). Differential constitutive activation between STAT-related proteins and MAP kinase in primary acute myelogenous leukaemia. *British Journal of Haematology*, **101**, 521–528.

Hemmi, S., *et al*. (1994). A novel member of the interferon receptor family complements functionality of the murine interferon gamma receptor in human cells. *Cell*, **76**, 803–810.

Hirsch, E., *et al*. (2000). Central role for G protein-coupled phosphoinositide 3-kinase gamma in inflammation. *Science*, **287**, 1049–1053.

Horvath, C. M., *et al*. (1995). A STAT protein domain that determines DNA sequence recognition suggests a novel DNA-binding domain. *Genes and Development*, **9**, 984–994.

Howard, O. M., *et al*. (1999). Naturally occurring CCR5 extracellular and transmembrane domain variants affect HIV-1 Co-receptor and ligand binding function. *Journal of Biological Chemistry*, **274**, 16228–16234.

Ihle, J. N. (1996). STATs signal transducers and activators of transcription. *Cell*, **84**, 331–334.

Ihle, J. N. (1995). The Janus protein tyrosine kinase family and its role in cytokine signaling. *Advances in Immunology*, **60**, 1–35.

John, J., *et al*. (1991). Isolation and characterization of a new mutant human cell line unresponsive to alpha and beta interferons. *Molecular and Cellular Biology*, **11**, 4189–4195.

Kaul, M. and Lipton, S. A. (1999). Chemokines and activated macrophages in HIV gp120-induced neuronal apoptosis. *Proceedings of the National Academy of Sciences of the USA*, **96**, 8212–8216.

Kotenko, S. V., *et al*. (1995). Interaction between the components of the interferon gamma receptor complex. *Journal of Biological Chemistry*, **270**, 20915–20921.

Lennartz, M. R. (1999). Phospholipases and phagocytosis: the role of phospholipid-derived second messengers in phagocytosis. *International Journal of Biochemical Cell Biology*, **31**, 415–430.

Li, X., *et al*. (1998). Cooperative binding of Stat1–2 heterodimers and ISGF3 to tandem DNA elements. *Biochimie*, **80**, 703–710.

Li, X., *et al*. (1996). Formation of STAT1–STAT2 heterodimers and their role in the activation of IRF-1 gene transcription by interferon-alpha. *Journal of Biological Chemistry*, **271**, 5790–5794.

Lim, L., *et al*. (1996). Regulation of phosphorylation pathways by p21 GTPases. The p21 Ras-related Rho subfamily and its role in phosphorylation signalling pathways. *European Journal of Biochemistry*, **242**, 171–185.

Lopez-Ilasaca, M. (1998). Signaling from G-protein-coupled receptors to mitogen-activated protein (MAP)-kinase cascades. *Biochemical Pharmacology*, **56**, 269–277.

Ma, Q., *et al*. (1998). Impaired B-lymphopoiesis, myelopoiesis, and derailed cerebellar neuron migration in CXCR4- and SDF-1-deficient mice. *Proceedings of the National Academy of Sciences of the USA*, **95**, 9448–9453.

Matsumoto, M., *et al*. (1999). Activation of the transcription factor ISGF3 by interferon-gamma. *Biological Chemistry*, **380**, 699–703.

Monteclaro, F. S. and Charo, I. F. (1997). The amino-terminal domain of CCR2 is both necessary and sufficient for high affinity binding of monocyte chemoattractant protein 1. Receptor activation by a pseudo-tethered ligand. *Journal of Biological Chemistry*, **272**, 23186–23190.

Moriggl, R., *et al*. (1999). Stat5 is required for IL-2-induced cell cycle progression of peripheral T cells. *Immunity*, **10**, 249–259.

Noguchi, M., *et al*. (1993). Interleukin-2 receptor gamma chain mutation results in X-linked severe combined immunodeficiency in humans. *Cell*, **73**, 147–157.

Nosaka, T., *et al*. (1995). Defective lymphoid development in mice lacking Jak3. *Science*, **270**, 800–802.

Novick, D., *et al*. (1994). The human interferon alpha/beta receptor characterization and molecular cloning. *Cell*, **77**, 391–400.

Pease, J. E., *et al*. (1998). The N-terminal extracellular segments of the chemokine receptors CCR1 and CCR3 are determinants for MIP-1alpha and eotaxin binding, respectively, but a second domain is essential for efficient receptor activation. *Journal of Biological Chemistry*, **273**, 19972–19976.

Pellegrini, S., *et al*. (1989). Use of a selectable marker regulated by alpha interferon to obtain mutations in the signaling pathway. *Molecular and Cellular Biology*, **9**, 4605–4612.

Peracino, B., *et al*. (1998). G protein beta subunit-null mutants are impaired in phagocytosis and chemotaxis due to inappropriate regulation of the actin cytoskeleton. *Journal of Cell Biology*, **141**, 1529–1537.

Pestka, S., *et al*. (1987). Interferons and their actions. *Annual Review of Biochemistry*, **56**, 727–777.

Pfeffer, L. M., *et al*. (1997). The short form of the interferon alpha/beta receptor chain 2 acts as a dominant negative for type I interferon action. *Journal of Biological Chemistry*, **272**, 11002–11005.

Platanias, L. C. and Fish, E. N. (1999). Signaling pathways activated by interferons. *Experimental Hematology*, **27**, 1583–1592.

Platanias, L. C., *et al*. (1992). Expression of the IFN alpha receptor in hairy cell leukaemia. *British Journal of Haematology*, **82**, 541–546.

Porter, A. C., *et al*. (1988). Interferon response element of the human gene 6–16. *EMBO Journal*, **7**, 85–92.

Puck, J. M., *et al*. (1993). The interleukin-2 receptor gamma chain maps to Xq13.1 and is mutated in X-linked severe combined immunodeficiency, SCIDX1. *Human Molecular Genetics*, **2**, 1099–1104.

Ram, P. A. and Waxman, D. J. (1997). Interaction of growth hormone-activated STATs with SH2-containing phosphotyrosine phosphatase SHP-1 and nuclear JAK2 tyrosine kinase. *Journal of Biological Chemistry*, **272**, 17694–17702.

Rane, S. G. and Reddy, E. P. (1994). JAK3: a novel JAK kinase associated with terminal differentiation of hematopoietic cells. *Oncogene*, **9**, 2415–2423.

Ray, M., *et al*. (2000). Inhibition of interferon-gamma signaling by *Leishmania donovani*. *Journal of Infectious Diseases*, **181**, 1121–1128.

Richardson, R. M., *et al*. (1998). Multiple signaling pathways of human interleukin-8 receptor A. Independent regulation by phosphorylation. *Journal of Biological Chemistry*, **273**, 10690–10695.

Russell-Harde, D., *et al*. (1995). Reconstitution of a high affinity binding site for type I interferons. *Journal of Biological Chemistry*, **270**, 26033–26036.

Russell-Harde, D., *et al*. (2000). Role of the intracellular domain of the human type I interferon receptor 2 chain (IFNAR2c) in interferon signaling. Expression of IFNAR2c truncation mutants in U5A cells. *Journal of Biological Chemistry*, **275**, 23981–23985.

Schindler, C. and Darnell, J. E., Jr (1995). Transcriptional responses to polypeptide ligands in the JAK–STAT pathway. *Annual Review of Biochemistry*, **64**, 621–651.

Seder, R. A. (1994). Acquisition of lymphokine-producing phenotype by CD4+ T cells. *Journal of Allergy and Clinical Immunology*, **94**, 1195–1202.

Shuai, K., *et al*. (1996). Constitutive activation of STAT5 by the BCR-ABL oncogene in chronic myelogenous leukemia. *Oncogene*, **13**, 247–254.

Singer, W. D., Brown, H. A. and Sternweis, P. C. (1997). Regulation of eukaryotic phosphatidylinositol-specific phospholipase C and phospholipase D. *Annual Review of Biochemistry*, **66**, 475–509.

Soh, J., *et al*. (1994). Identification and sequence of an accessory factor required for activation of the human interferon gamma receptor. *Cell*, **76**, 793–802.

Stark, G. R., *et al*. (1998). How cells respond to interferons. *Annual Review of Biochemistry*, **67**, 227–264.

Takahashi, M., *et al*. (1997). Mechanotransduction in endothelial cells: temporal signaling events in response to shear stress. *Journal of Vascular Research*, **34**, 212–219.

Takeda, K., *et al*. (1997). Targeted disruption of the mouse Stat3 gene leads to early embryonic lethality. *Proceedings of the National Academy of Sciences of the USA*, **94**, 3801–3804.

Takeda, K., *et al*. (1996). Essential role of Stat6 in IL-4 signalling. *Nature*, **380**, 627–630.

Tau, G. and Rothman, P. (1999). Biologic functions of the IFN-gamma receptors. *Allergy*, **54**, 1233–1251.

Tournamille, C., *et al*. (1997). Close association of the first and fourth extracellular domains of the Duffy antigen/receptor for chemokines by a disulfide bond is required for ligand binding. *Journal of Biological Chemistry*, **272**, 16274–16280.

Trinchieri, G. and Scott, P. (1995). Interleukin-12a proinflammatory cytokine with immunoregulatory functions [editorial]. *Research in Immunology*, **146**, 423–431.

Uze, G., *et al*. (1990). Genetic transfer of a functional human interferon alpha receptor into mouse cells cloning and expression of its cDNA. *Cell*, **60**, 225–234.

Uze, G., *et al*. (1995). Alpha and beta interferons and their receptor and their friends and relations. *Journal of Interferon and Cytokine Research*, **15**, 3–26.

Velazquez, L., *et al*. (1995). Distinct domains of the protein tyrosine kinase tyk2 required for binding of interferon-alpha/beta and for signal transduction. *Journal of Biological Chemistry*, **270**, 3327–3334.

Wong, M. and Fish, E. N. (1998). RANTES and MIP-1alpha activate stats in T cells. *Journal of Biological Chemistry*, **273**, 309–314.

Wu, G., *et al*. (1998). Receptor docking sites for G-protein beta-gamma subunits. Implications for signal regulation. *Journal of Biological Chemistry*, **273**, 7197–7200.

Wu, M. Y., *et al*. (1999). The cytoplasmic domain of the lymphotoxin-beta receptor mediates cell death in HeLa cells. *Journal of Biological Chemistry*, **274**, 11868–11873.

Xia, Z., *et al*. (1998). Expression of signal transducers and activators of transcription proteins in acute myeloid leukemia blasts. *Cancer Research*, **58**, 3173–3180.

Yamamoto, K., *et al*. (1997). cDNA cloning, expression and chromosome mapping of the human STAT4 gene: both STAT4 and STAT1 genes are mapped to 2q32.2q32.3. *Cytogenetics and Cell Genetics*, **77**, 207–210.

Yasukawa, H., *et al*. (2000). Negative regulation of cytokine signaling pathways. *Annual Review of Immunology*, **18**, 143–164.

Young, H. A. and Hardy, K. J. (1995). Role of interferon-gamma in immune cell regulation. *Journal of Leukocyte Biology*, **58**, 373–381.

Zheng, B., *et al*. (1999a). Divergence of RGS proteins: evidence for the existence of six mammalian RGS subfamilies. *Trends in Biochemical Sciences*, **24**, 411–414.

Zheng, J., *et al*. (1999b). Intracellular CXCR4 signaling, neuronal apoptosis and neuropathogenic mechanisms of HIV-1-associated dementia. *Journal of Neuroimmunology*, **98**, 185–200.

Zhou, N., *et al*. (2000). Molecular modeling and site-directed mutagenesis of CCR5 reveal residues critical for chemokine binding and signal transduction. *European Journal of Immunology*, **30**, 164–173.

Zlotnik, A., *et al*. (1999). Recent advances in chemokines and chemokine receptors. *Critical Reviews in Immunology*, **19**, 1–47.

Chapter 12

Signalling by Tyrosine Kinases

Kermit L. Carraway and Coralie A. Carothers Carraway
University of Miami School of Medicine, Miami, FL, USA

CONTENTS

INTRODUCTION

Historical Perspective

Few molecules have been more closely linked to cancer than the protein tyrosine kinases. Tyrosine phosphorylation and the tyrosine kinase enzymes which create phosphotyrosine residues in their substrate proteins were discovered during studies of the oncogenic factors in tumour viruses. The identification of v-Src as the transforming factor of the Rous sarcoma virus and the recognition of its activity as a protein tyrosine kinase conceptualized the oncogene theory of tumorigenesis, creating a model which has dominated much of cancer research over the past two decades. The importance of the work on viral oncogenes in the early research on tyrosine phosphorylation is shown by an accounting of landmark events in this area (**Table 1**) (Hunter, 1998). The discovery of a cellular counterpart (c-Src) of the viral Src (v-Src) tyrosine kinase indicated a much broader role for these enzymes and suggested a major role for c-Src and tyrosine phosphorylation in mediating normal cell behaviours. For example, c-Src has been implicated in such diverse cell functions as platelet aggregation, cell cycle control and cell motility. The diversity of cellular functions of tyrosine kinases was further indicated by early observations that both the epidermal growth factor (EGF) receptor, important in epithelial cell growth, and the insulin receptor, a key component in metabolic regulation, are tyrosine kinases. Another tyrosine kinase receptor, called Sevenless for its role in determining cell fate of a particular cell type, was found to be necessary for normal eye development in the fruit fly (*Drosophila*). These and many other discoveries have demonstrated the critical role that the tyrosine kinases and tyrosine phosphorylation play in normal cell regulation and communication and in developmental biology.

Evolutionarily, tyrosine kinases are primarily, if not exclusively, a product of eukaryotic cells. The emergence of protein tyrosine kinases with the appearance of multicellular

Table 1 Important discoveries in history of tyrosine phosphorylation

1979	Tyrosine phosphorylation of polyoma tumour virus middle T antigen
1980	Protein tyrosine kinase activities of v-Src, c-Src, v-Abl and EGF receptor
1981	Insulin-stimulated protein tyrosine kinase activity of insulin receptor
1982	Sequence similarity of v-Src to cAMP-dependent protein kinase catalytic subunit
	c-Abl gene rearrangement in chronic myelogenous leukaemia
1983	Polyoma middle T association with and activation of c-Src
1984	v-ErbB oncogene derived from EGF receptor
1985	Negative regulation of c-Src by tyrosine phosphorylation
1986	Neu oncogene as EGF receptor family member with activating point mutation
1987	*Drosophila* Sevenless is receptor tyrosine kinase
1988	Acetylcholine receptor regulated by tyrosine phosphorylation
1989	Cell cycle regulatory kinase negatively regulated by Tyr phosphorylation
1990	Src homology-2 domains bind phosphotyrosines
1992	Individual receptor Tyr phosphorylation sites bind distinct SH2-containing proteins
	STAT transcription factors activated by Tyr phosphorylation

(Adapted from Hunter, 1998.)

(metazoan) organisms is indicative of their importance in cell communication and organisation. Sequence information from the Human Genome Project indicates that there are 90 tyrosine kinases encoded in the human genome.

The Kinase Superfamily

Tyrosine kinases are members of a much larger family of protein kinases (Hunter, 1998), which can be categorized by two classifications, one based on specificity for the target amino acid and the other on structure and cellular localization. The major specificity classes are the serine or threonine-specific (Ser/Thr) kinases and the tyrosine-specific (Tyr) kinases which catalyse the phosphorylation of serine and threonine or tyrosine residues, respectively. In addition, a few mixed function kinases, which catalyse both Ser/Thr and Tyr phosphorylation, have been described. The discovery of serine/threonine kinases and their importance in regulating metabolic pathways preceded the discovery of tyrosine kinases by about two decades. The regulation of gluconeogenesis by phosphorylation–dephosphorylation established as a paradigm the reversible regulation of enzymes by the covalent addition and removal of phosphate groups. Further, the discovery of cAMP-dependent kinases was a catalyst for the study of cAMP as a second messenger in the transduction of signals through certain types of receptors. Structural analyses of serine/threonine kinases, particularly the cyclic AMP-stimulated kinase (protein kinase A), have provided models for the catalytic domains and enzymic mechanisms of all of the kinases (Taylor *et al*., 1995).

Structurally, there are two major classes of kinases, receptor (**Figure 1a**) and nonreceptor (**Figure 1b**), based on elements of their primary sequences which determine their cellular localization. Almost all receptor kinases characterized to date are tyrosine kinases, with the exception of the TGF-β receptor, a Ser/Thr kinase (see chapter on *Signalling by TGF-*β). Receptor kinases are defined by a hydrophobic transmembrane domain, which passes through the plasma membrane, an extracellular ligand-binding domain and a cytoplasmically oriented kinase domain (**Figure 1a**). In contrast, nonreceptor kinases have no transmembrane or extracellular domains, although they may be associated with the cytoplasmic surfaces of cellular membranes by one of two mechanisms. The first is constitutive membrane localization via a lipid modification which anchors the protein to the phospholipid bilayer. The lipid-linked kinases identified to date have been tyrosine kinases, and Src is the prototype of this kinase type, having an N-terminal consensus site for myristoylation (**Figure 2a**). The second membrane localization mechanism involves binding to a nonenzymic membrane receptor to give a binary tyrosine kinase (**Figure 2b**).

Tyrosine phosphorylation is a relatively rare event in normal, unactivated cells, representing $<0.1\%$ of total protein phosphate groups (serine, 90%; threonine, 10%). However, tyrosine phosphorylation is transiently increased during normal cell activation and often substantially increased, sometimes constitutively, in tumour cells. The functions of tyrosine kinases are significantly different from those of the serine/threonine kinases. Conceptually, the primary role of tyrosine kinases is to provide a mechanism for transmitting information from a factor outside a cell to the interior of the cell without requiring that the factor cross the cell's exterior membrane barrier. That role is exemplified by the ability of a growth factor or polypeptide hormone in the extracellular milieu, which cannot pass through the membrane, to activate specific gene transcription in the nucleus of a target cell (see **Figure 3a**). This example is only one of the many cell functions mediated by tyrosine kinases (**Table 2**), and these functions must be performed in a variety of cell and organismal contexts. Thus, this versatility of cellular tyrosine kinases requires that they be complex proteins and carry a variety of ancillary domains in addition to their tyrosine kinase domain to be able to interact with various cellular proteins in the performance of these various functions. Src is a prototype for these multiple interactions because it has two domains in addition to that encoding its tyrosine kinase (**Figure 1b**), Src homology 2 (SH2) and Src homology 3 (SH3), as noted below, which are frequently found in associations among signalling components.

Mechanistic

The transfer of a signal from an extracellular factor through the membrane via a tyrosine kinase is usually performed by one of two types of mechanisms, exemplified in **Figures 2b** and **3a**. These two mechanisms use receptor and nonreceptor tyrosine kinases. In the first mechanism (**Figure 3a**) the receptor tyrosine kinase is solely responsible for the transfer of signal across the membrane. In the second mechanism (**Figure 2b**) a nonreceptor tyrosine kinase is coupled to a transmembrane receptor cytoplasmic domain in a binary receptor. In either case it is the activation of the tyrosine kinase accessible to the cytoplasm which is the key step in transferring the signal across the membrane. These mechanisms also define the two types of tyrosine kinases which have evolved (**Figure 1**) and the minimal domains necessary for each type of tyrosine kinase. The receptor tyrosine kinase must have an extracellular ligand-binding domain(s), a transmembrane domain, a kinase domain and sites for docking the cytoplasmic molecules to which the signal is transferred. The cytoplasmic tyrosine kinases must have the kinase domain, a binding domain for attaching to the receptor and sites for docking the cytoplasmic molecules to which the signal is transferred.

The transmission of signal from the kinase at the membrane to the nucleus for activation of transcription can

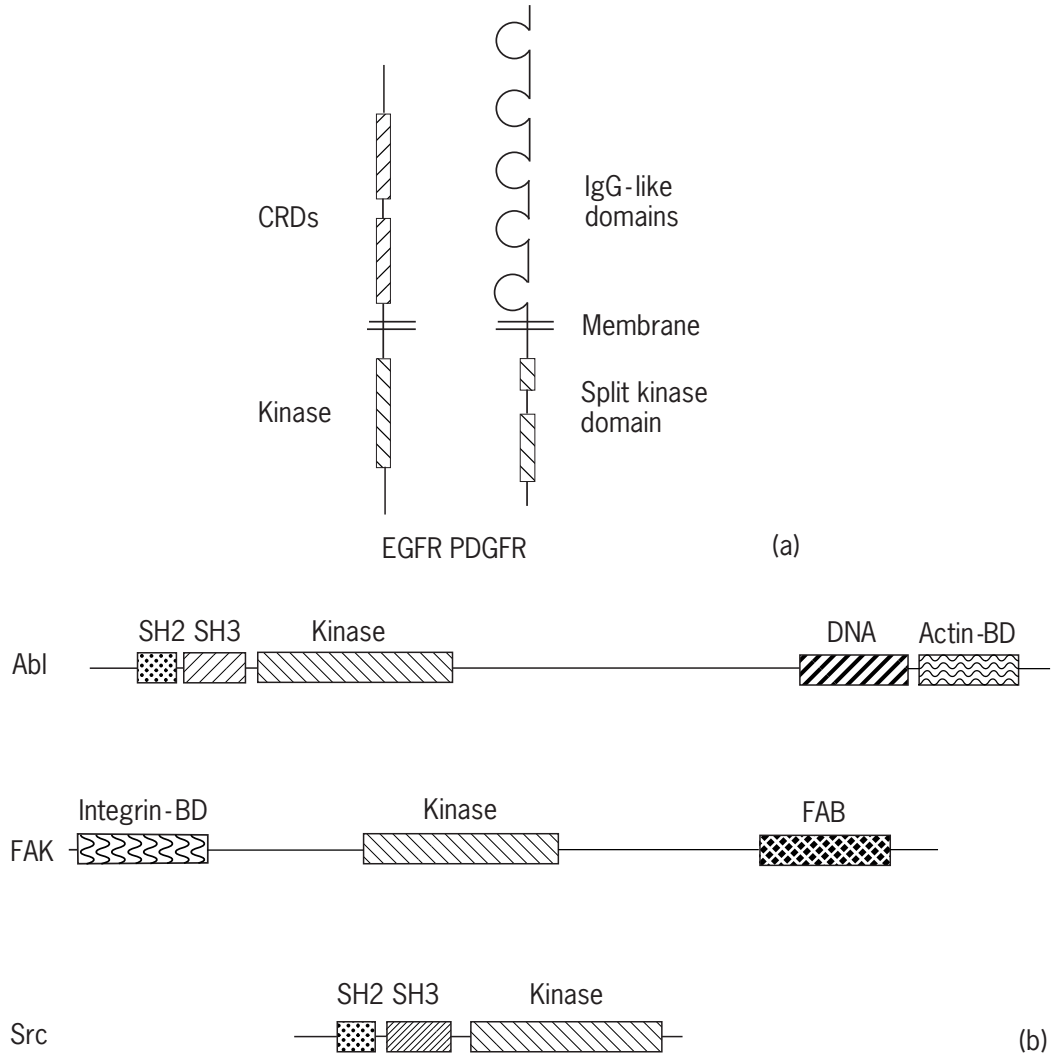


Figure 1 Schematic structures for receptor and nonreceptor protein tyrosine kinases, showing variations in the two classes. (a) Receptor tyrosine kinases ErbB1 (EGF receptor) and PDGF receptor, showing common variations in structure (see also **Table 6**). Ligand-binding sites are contained in the extracellular domains, which are highly variable between different receptor tyrosine kinase families (**Table 6**). CRD, cysteine-rich domain; IgG, immunoglobulin. (b) Nonreceptor (cytoplasmic) tyrosine kinases Abl, FAK and Src. Note the multiple binding domains in the nonreceptor kinases which link them to other functional components in the cell as part of their signalling functions. BD, binding domain; FAB, focal adhesion binding domain. (Adapted from Hunter, 1998.)

also occur by multiple pathways. Two of the most important are illustrated in **Figure 3b**, Ras–MAP kinase and JAK–STAT. The receptor–Ras-MAP kinase pathway consists of four main elements (**Figure 3a** and **b**; **Table 3**): the receptor which receives and transmits the signal through the membrane; an adaptor system (Grb2-SOS) which couples the receptor to a membrane switch (Ras); the Ras switch, which is activated by that coupling mechanism, then activates downstream kinases; and the MAP kinase cascade (Raf–MEK–Erk in **Figure 3**), which ultimately transmits the signal to the nucleus. The central feature of this pathway is the small G protein/GTPase Ras. It is switched on and off in response to its binding of the nucleotide GTP. The binding is controlled by three different types of regulatory proteins: activators which promote the Ras GTPase activity to convert bound GTP to GDP; exchange proteins which replace GDP with GTP; and inhibitors which reduce the GTPase activity. In addition, Ras is not simply an on/off switch. Depending on the cellular context, it can also switch the signal to a different pathway, activating different cell functions. The versatility of this mechanism is illustrated in **Table 4**, which shows some of the multiple downstream components to which Ras can be coupled, again illustrating the multiple cellular functions in which the tyrosine kinases can participate (**Table 2**). Obviously, the complexity of the Ras pathway provides many sites for regulation and for integration of signals by interactions with other pathways.

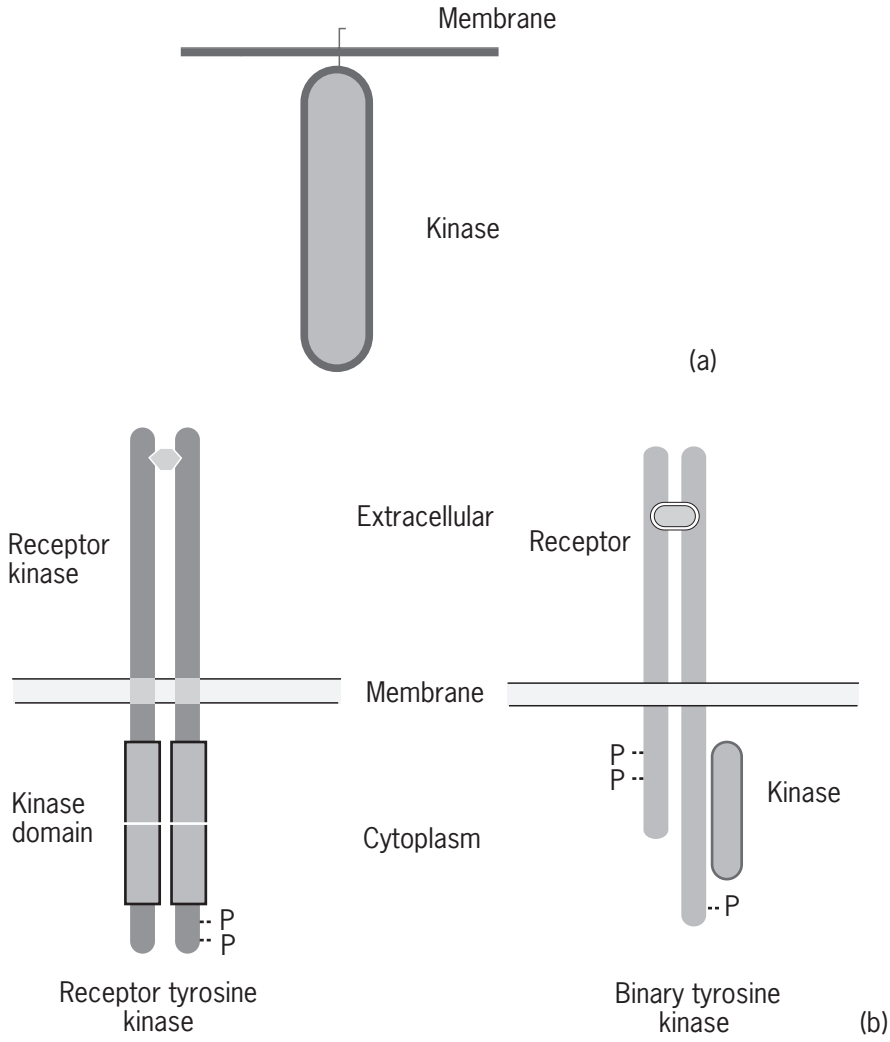


Figure 2 Comparison of mechanisms for associating tyrosine kinases with membrane. (a) Lipid-linked tyrosine kinase; (b) Comparison of structures for transmembrane receptor tyrosine kinases and binary receptor tyrosine kinases. The latter consists of a nonreceptor tyrosine kinase noncovalently linked to a nonenzymic transmembrane receptor.

In the JAK–STAT pathway (**Figure 3b**, also described in the previous chapter, Signalling by Cytokines) the tyrosine kinase JAK phosphorylates a STAT molecule, which translocates to the nucleus. This pathway has the advantage of being much simpler (**Figure 3**), but also has fewer opportunities for regulation.

RECEPTOR TYROSINE KINASES

Mechanism of Signal Transduction

The receptor tyrosine kinase provides a single transmembrane molecule which can transfer information across the membrane from the extracellular milieu to the cytoplasm. This transfer involves two related events: stimulation of the kinase activity and phosphorylation of tyrosine residues accessible to the cytoplasm of the cell. In the most common model for this process, ligand binding to the extracellular domain results in dimerization of the receptor in the membrane (**Figures 2b**, **3a** and **4**). This mechanism has been demonstrated by high-resolution X-ray crystallographic analysis for the growth hormone receptor activated by its dimeric ligand (Wells, 1996). In some cases the receptors may form larger complexes (e.g. tetramers). Regardless, the receptor association leads to potentiation of the kinase activity, probably through structural (conformational) changes in the kinase domain. The second phase of the information transfer involves the phosphorylation of one or more tyrosine residues of the cytoplasmic domain of the receptor (**Figure 4**). Available evidence suggests that this occurs primarily through cross-phosphorylation (*trans*-phosphorylation). In a receptor

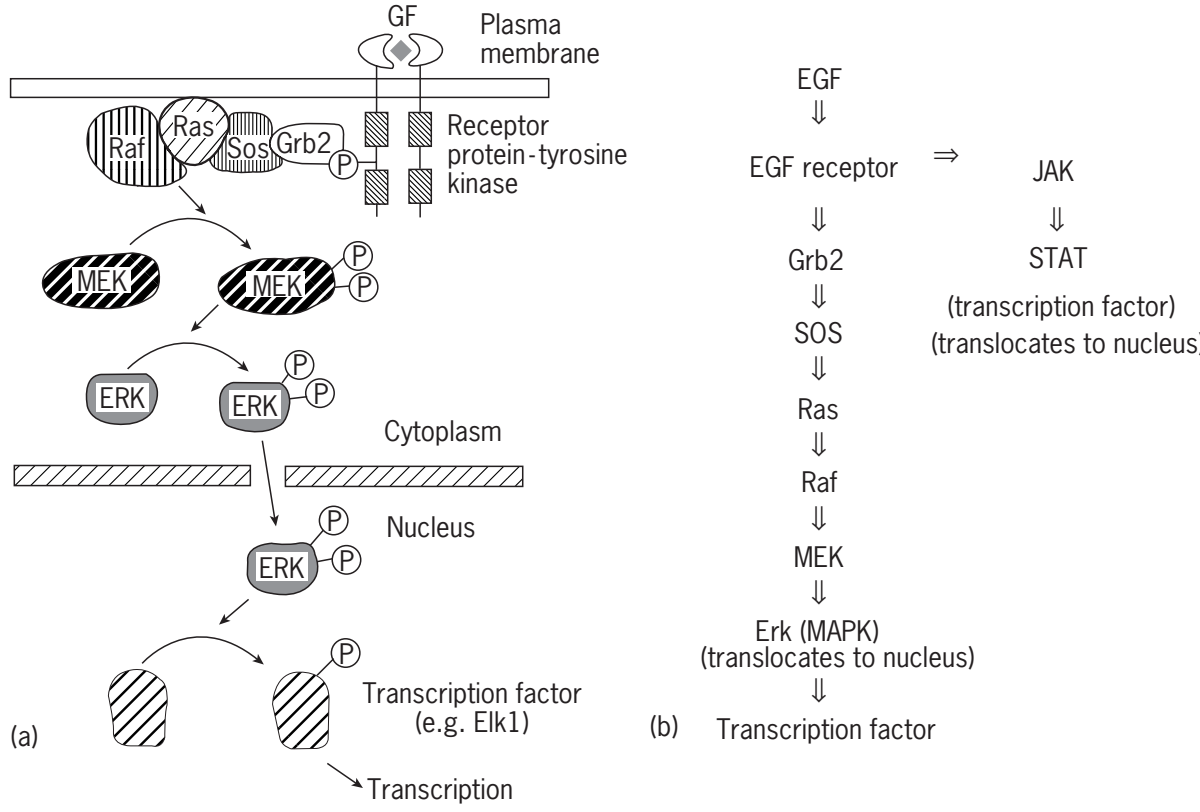


Figure 3 Signal transmission from the extracellular space to the nucleus to activate transcription via the Ras-MAP kinase pathway. (a) Signal translocation; GF, growth factor (b) Comparison of Ras-MAP kinase pathway with JAK–STAT pathway. (Adapted from Hunter, 1998.)

Table 2 Examples of cellular functions regulated by tyrosine kinases

Cell function	Receptor[a]	Downstream effectors[a]
Proliferation	**ErbB**	Ras–Erk (MAPK)
Cell–matrix adhesion	Integrin	**FAK–Src**
Cell–cell adhesion	Cadherin	Catenin-**Src** family
Movement	**PDGFR**	Rac–Rho
Apoptosis control	**IGFR**	PI3K–Akt
Transcription	Cytokine R	**JAK**–STAT
Membrane transport	Channel	**Src** family

[a]Tyrosine kinases are indicated in bold. FAK, focal adhesion kinase; PDGFR, platelet-derived growth factor receptor; IGFR, insulin-like growth factor receptor; PI3K, phosphoinositide 3-kinase; STAT, signal transduction and transcription.

dimer the kinase on one receptor (half-dimer) would phosphorylate tyrosines on the associated receptor (half-dimer) and vice versa. Thus, the ligand-induced association not only activates the enzyme, but also brings the enzyme and its substrate tyrosine residues into proximity. The significance of the tyrosine phosphorylation is that the phosphotyrosine residues created form sites at which cytoplasmic proteins can bind to initiate intracellular signalling (see **Figure 3a**), thus passing a signal across the membrane without the passage of the activating ligand.

The primary function of the phosphorylated tyrosine residues on the activated receptors is to recruit signalling components from the cytoplasm or cytoplasmic surface of the plasma membrane to initiate signalling pathways (**Figures 3a** and **4**) (Panayotou and Waterfield, 1993). The particular pathway initiated is determined by two complementary factors: the site of the tyrosine on the receptor polypeptide chain and the specificity of the phosphotyrosine-binding domains on the cytoplasmic proteins. The sites of tyrosine phosphorylation on a particular receptor are determined by the ligand–receptor and receptor–receptor interactions, which regulate the accessibility of any particular tyrosine to the kinase catalytic site. Two different types of domains have been found in signalling components which bind to phosphotyrosines to initiate their recruitment to the receptor: SH2 (see **Figures 1b** and **5**) and PTB (**Figure 5**). The SH2 domain is a compact globular unit of about 100 amino acids, which was originally discovered in Src and is found in a large number of signalling components (**Table 5**). Its binding specificity is determined primarily by the phosphotyrosine and 1–5 amino acid residues following it (C-terminal direction) in the receptor amino acid sequence. The PTB domain is also globular and has about 150 amino acids; its specificity is determined by the phosphotyrosine and 1–8 amino acid residues preceding it (N-terminal direction) in the receptor

Table 3 Components of pathways regulating transcription

Receptor–Ras–MAPK pathway		Receptor–JAK–STAT pathway	
Component	**Function**	**Component**	**Function**
EGFR	Receptor	EGFR	Receptor
Grb2	Adaptor	JAK	Nonreceptor tyrosine kinase
SOS	GTP exchange protein	STAT	Transcription factor
Ras	G protein switch		
Raf	Ser/Thr kinase		
MEK	Dual-function Tyr/Thr kinase		
Erk (MAPK)	Ser/Thr kinase		
Transcription factor	Transcription factor		

Table 4 Examples of signalling elements and cellular functions coupled to Ras activation

Signalling component	Potential cell function[a]
Rac	Shape, movement
RalGDS	Shape, movement
P120 GAP	Shape, movement
Raf	Transcription, proliferation
PI3K	Proliferation, apoptosis resistance
MEKK	Transcription, stress reaction

[a] Listed functions are illustrative, not inclusive.

sequence. Other types of phosphotyrosine-binding domains have been suggested, but have not been well characterized.

These observations indicate that most components initiating signalling pathways from receptor tyrosine kinases should contain either SH2 or PTB domains. Indeed, identification of proteins with these domains has provided significant insights into signalling. As shown in **Table 5** and **Figure 5**, these proteins include three classes of signalling pathway initiators: enzymes, docking proteins and adaptors. Each of these types of molecules facilitates signalling by a different mechanism. Enzymes actively participate in downstream signalling events. Docking proteins provide additional or surrogate tyrosine phosphorylation sites for further diversification of receptor sites to initiate signalling. Adaptors are involved in assembling complexes of signalling components for initiating and regulating downstream signalling pathways. However, in all three types of mechanisms the key aspect of the pathway is the recruitment and relocalization of the signalling component to the site of the receptor (exemplified by Grb2 in **Figure 3a**) (Panayotou and Waterfield, 1993; Carraway and Carraway, 1995).

Four different rationales can be envisioned for the recruitment of enzymes to initiate signalling pathways. First, recruitment can permit the receptor tyrosine kinase to phosphorylate the recruited molecule, thus changing its activity or binding function. An example of this is phospholipase Cγ, whose recruitment and phosphorylation activates a phospholipid-cleaving enzyme to stimulate a pathway leading to calcium influx into the cytoplasm (**Figure 6**). The calcium plus lipid hydrolysis product diacylglycerol combine to stimulate a potent calcium-dependent serine/threonine protein kinase, protein kinase C, which then has multiple cellular effects. Second, recruitment brings the recruited molecule to the membrane, where it may act upon membrane components as substrates. This mechanism applies to phospholipase Cγ, which cleaves phospholipid molecules found in membranes, and to phosphoinositide 3-kinase, one of the signalling enzymes implicated in oncogenesis, which phosphorylates specific phospholipid molecules in membranes (**Figure 6**). These enzymatic modifications of membrane components may change the organisation of the membranes or produce new signalling molecules, such as the lipid phosphate PIP_2, to initiate and perpetuate downstream signalling pathways. It is important to note that the same signalling molecule may initiate different signalling pathways, depending on the cellular context. Both PI3 kinase and phospholipase Cγ can be involved in multiple cellular functions, as previously indicated for Ras (**Table 4**). Third, binding of a molecule to the receptor phosphotyrosine residues may induce its activation by a conformational change (allosteric effect), a mechanism which contributes to the signalling effects of phosphoinositide 3-kinase (**Tables 2** and **4**; **Figure 6**). Fourth, some of the recruited enzymes are protein tyrosine kinases or protein tyrosine phosphatases. Since the receptor tyrosine kinases are often involved in large, multimeric signalling complexes (Carraway and Carraway, 1995), such as focal adhesion sites (see below), this recruitment brings the additional kinases and phosphatases into proximity of their substrates in these complexes and facilitates regulation of the signalling components involved. An important aspect of such complexes is that they are often located at sites influencing cell behaviour, such as sites for membrane-microfilament interactions involved in cell shape and movement (Carraway *et al*., 1997).

Recruitment of proteins can also amplify signalling potential through the docking protein mechanism. The

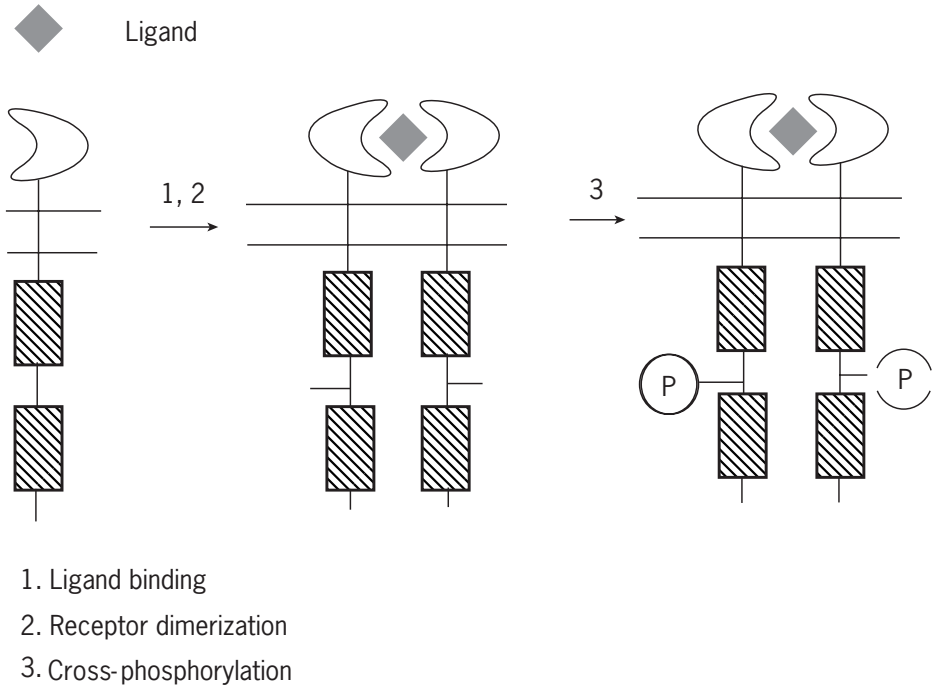


Figure 4 Mechanism for transmission of signal across the membrane by binding of ligand, dimerization of receptor and transphosphorylation of cytoplasmic domain to create sites for recruitment of cytoplasmic signalling components to initiate signalling pathways.

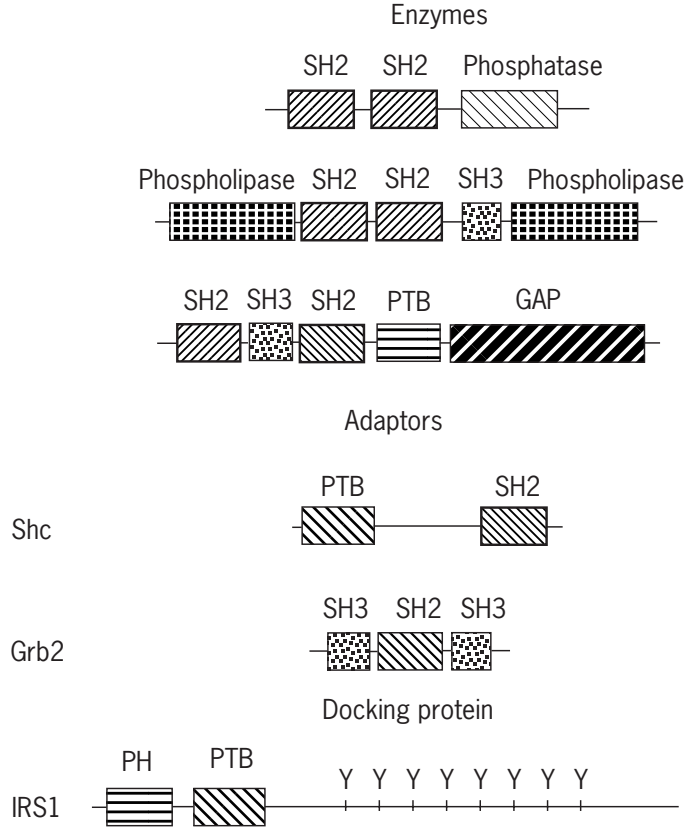


Figure 5 Schematic structures for enzymes, adaptors and docking proteins involved in initiating cellular signalling pathways by binding to tyrosine-phosphorylated receptors or binary receptor complexes. GAP, GTPase activating protein. (Adapted from Hunter, 1998.)

best characterized docking proteins are the insulin receptor substrates (IRSs, **Figure 5**), which regulate a variety of cellular functions, including membrane transport, gene expression, protein synthesis and lipid and carbohydrate metabolism. IRSs contain a PTB domain (**Figure 5**) which allows them to bind the activated insulin receptor. Formation of this insulin–IRS complex increases the number of tyrosines available for phosphorylation and increases the insulin receptor signalling potential. Moreover, IRSs act via multiple receptors, including the insulin-like growth factor (IGF) receptor and various cytokine receptors. Phosphorylated tyrosines on IRSs can interact with a variety of SH2-containing components to activate multiple downstream signalling pathways. Among these are the adaptors Grb2, Crk and Nck (**Figure 5**), the phospholipid kinase PI3 kinase, the tyrosine kinase Fyn and the phosphotyrosine phosphatase SHP2 (**Figure 5**).

Table 5 Phosphotyrosine-binding proteins which could activate or diversify signalling pathways by binding phosphorylated receptors

Protein	Class/function	Domains
PLC-γ	Phospholipase	SH2 (2), SH3
GAP	GTPase activator	SH2, SH3
Src family	PTK	SH2, SH3
ZAP/SYK	PTK	SH2 (2)
Shp1/2	Tyrosine phosphatase	SH2 (2)
PI3K	Lipid kinase	SH2 (2), SH3
Ship	Lipid phosphatase	SH2
Vav	GTP exhange factor	SH2, SH3
Shc	Adaptor	SH2, PTB
Nck	Adaptor	SH2, SH3 (3)
Grb2, Crk	Adaptor	SH2, SH3 (2)
IRS1/2	Docking protein	PTB
Talin	Focal adhesion protein	SH2
STAT	Transcription factor	SH2, SH3

Adaptor molecules participate in the recruitment and organisation of signalling components from the cytoplasm into membrane complexes. The best studied example is the initiation of the Ras–MAP kinase mitogenic pathway by the adaptor Grb2 (**Figure 3a**). Grb2 contains both SH2 and SH3 domains, allowing it to link a tyrosine-phosphorylated receptor to a cytoplasmic SH3-binding protein containing a proline-rich motif. In the cytoplasm of unactivated cells Grb2 is present as a complex through its SH3 domain with a second protein SOS, a GTP exchange protein which activates Ras. Activation of ErbB1 (EGF receptor) by its ligand EGF creates a phosphotyrosine binding site for the Grb2 SH2 domain on the receptor, recruiting the Grb2–SOS complex from the cytoplasm to the plasma membrane. When associated with the membrane, SOS can bind to and activate the proto-oncogene product Ras. Activated Ras stimulates a series of serine/threonine protein kinases, culminating in the activation of a specific MAP kinase (Erk), which can migrate to the nucleus as a signal for inducing gene transcription (**Figure 3a** and **b**). The key step in initiation is the linkage of the Ras switch protein to the ErbB receptor via the adaptor Grb2. As noted previously, Ras also has the potential to activate other signalling pathways (**Table 4**).

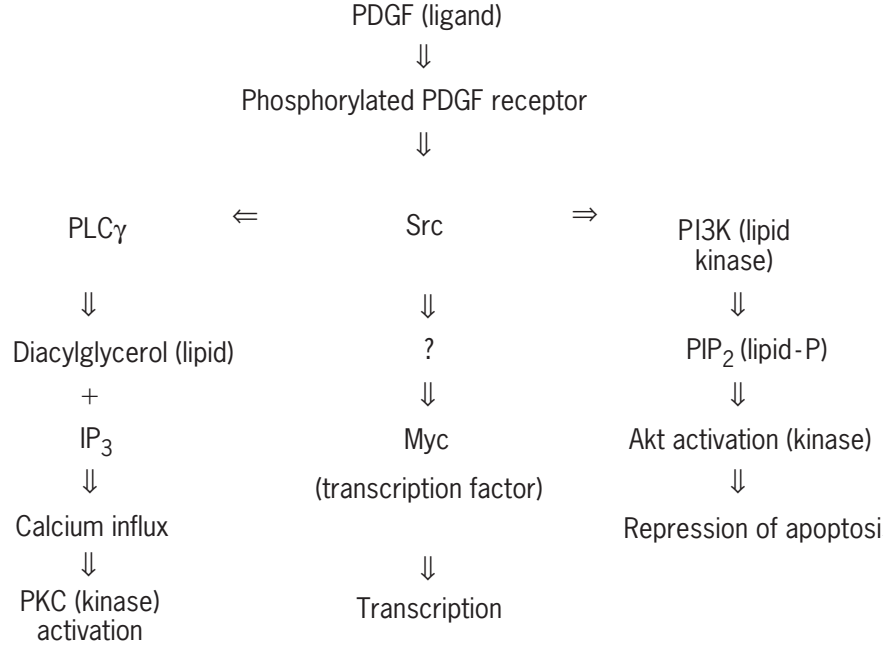


Figure 6 Examples of multiple downstream pathways coupled to receptor phosphorylation of PDGF receptor.

Versatility of Receptor Tyrosine Kinase Signalling

One of the hallmarks of the receptor tyrosine kinase mechanism is its versatility, mediating signal transduction via multiple pathways in different organisms, tissues and cells. This versatility allows these enzymes to regulate many of the vast number of cellular processes and interactions required in the development and function of complex multicellular organisms. The bases for the versatility are several-fold, the first being the multiplicity of receptors. There are a minimum of 20 families of mammalian receptor tyrosine kinases, each with different ligand-binding domains and most with multiple members (**Table 6**). Thus, part of the versatility of the receptor tyrosine kinase mechanism arises from the diversity of ligand–receptor interactions available in different tissues of different organisms at different times. Second, many of these individual receptor kinases can bind more than one ligand, as exemplified by the ligand binding patterns of the class I (ErbB) family (**Table 7**), leading to signal diversification. Third, the distribution of receptors in the organism is tissue and cell dependent. Some receptors are widely distributed; others are specifically localized to a small number of sites. Tumours may have aberrant distributions of receptors. Fourth, the distribution of many receptor tyrosine kinases and their ligands is developmentally regulated, being found only during specific stages of the life history of the organism. Fifth, ligands are produced at different sites from the target cells to exert different levels of control, as shown by examples in **Table 8**. Endocrine ligands are produced in different organs and transmitted through the circulation to the site of action. Paracrine ligands are produced by a different cell type or tissue from the target cells, but near the target. The juxtacrine mechanism is a special case of the paracrine ligand in which the ligand and receptor are expressed on the plasma membranes of adjacent cells. Autocrine responses arise from ligands produced by the same cell (cell type) as the receptor-bearing target cell. The intracrine and intramembrane mechanisms are special cases of the autocrine response, in which ligand and receptor are produced in the same cell.

The receptor structure itself contributes to the diversification of signals. Each receptor contains multiple

Table 6 Multiplicity of receptor tyrosine kinases

Receptor family	Extracellular domain[a]	Number in family
ErbB	Cysteine rich	4
PDGF	Ig	5
FLT	Ig	3
Insulin	Cysteine rich, FN III	2
TRK	Ig, leucine rich	3
FGF	Ig, acid box	4
Eph	FN III, CR, Ig	14
Axl	Ig, FN III	3
TIE	Ig, EGF, FN III	2
RET	Cadherin	1
MET	Sema	2
ROR	Ig, CR, kringle	2
MCK10	Factor VIII	2
MuSK	Ig	1
CCK4	Ig	1
ROS	FNIII	1

[a] Ig, immunoglobulin-like; FNIII, fibronectin III-like; CR, cysteine rich.

Table 7 Ligand–receptor couples for the ErbB family

Receptor	Soluble ligands
ErbB1	EGF, TGFα, amphiregulin, HB-EGF, betacellulin, epiregulin
ErbB2	None known
ErbB3	Neuregulins
ErbB4	Neuregulins, HB-EGF, betacellulin, epiregulin

Table 8 Examples of ligand sources and binding mechanisms

Mechanism	Ligand	Receptor	Definition	Example
Endocrine	Insulin	Insulin R	Ligand reaches receptor through circulation	Pancreatic cell insulin; fat cell receptor
Paracrine	Neuregulin	ErbB3	Ligand produced by adjacent cell or tissue	Mesenchymal NRG; epithelial ErbB3
Autocrine	TGF-α	ErbB1	Ligand secreted by same cell bearing receptor	TGF-α secretion by tumour cells with ErbB1
Intracrine	v-sis	PDGF R	Ligand and receptor interact inside cell	v-sis binding to PDGF R in cells
Juxtacrine	Bride of Sevenless (BOSS)	Sevenless	Ligand and receptor in plasma membranes of adjacent cells	BOSS-Sevenless interaction in control of *Drosophila* eye development
Intramembrane	MUC4	ErbB2	Ligand and receptor in same membrane	MUC4/ErbB2 complexes in tumour cells and epithelia

Table 9 Phosphorylated tyrosine-binding sites for signalling components in PDGF

Sequence position	Binding protein	Class[a]
579	Src, STAT	PTK, TF
581	STAT	TF
716	Grb2	Adaptor
740	P13K	Lipid kinase
751	Nck, PI3K	Adaptor, LK
771	GAP	GTPase activator
775	STAT	TF
778	Grb7	Adaptor
1009	SHP-2	Tyrosine phosphatase
1021	PLC-γ	Phospholipase

[a] PTK, protein tyrosine kinase; TF, transcription factor; LK, lipid kinase.

tyrosine residues in its cytoplasmic domains which can be phosphorylated to form different binding sites for cytoplasmic signalling proteins. For example, the PDGF receptor contains at least 10 tyrosine residues which have been shown to be phosphorylated in response to ligand activation (**Table 9**). Each different phosphorylated tyrosine potentially represents a different signalling pathway which could be initiated (Claesson-Welsh, 1994), some of which are shown in **Figure 6**, though overlaps and redundancies inevitably reduce that number. By comparing **Table 9** with **Tables 2** and **4**, it is clear that the PDGF receptor has the potential for participating in a large number of cellular functions. Diversity and specificity of signalling are achieved in part by phosphorylation of different combinations of the tyrosine residues in response to different extracellular ligands, an example of a combinatorial mechanism for regulation. A second diversification and specificity mechanism arises from the type of receptor association during receptor activation. Formation of a heterodimer by two different molecules of a receptor family obviously can yield more different phosphorylated tyrosine residues than formation of a homodimer by two identical molecules from the same family. Thus, the number of potential signals is increased. In the case of the ErbB family of receptors (**Table 7**), the four receptors can form 10 different combinations of homodimers and heterodimers in response to different ligands, all of which have been observed and have potentially different signalling capabilities (Riese and Stern, 1998). For this family of receptors, heterodimerization appears to be the preferred mechanism of activation in many physiological contexts (Riese and Stern, 1998).

Regulation of Receptor Tyrosine Kinase Signalling

The complex signalling pathway from the receptor to the nucleus (**Figure 3**, ErbB1 through MAP kinase) involves at least seven different components and provides multiple levels of control of the signal. In any phosphorylation-dependent system, one obvious control mechanism is the removal of the phosphate(s). In a complex chain involving multiple phosphorylations, reversal of any phosphorylation event can potentially break the chain and block the signal. Cells contain a large variety of phosphatases to hydrolyse phosphotyrosine and phosphoserine/threonine residues (Streuli,1996; Cohen, 1997). Contrary to early expectations that phosphatases would provide nonspecific 'off' switches for kinase signalling, both serine/threonine and tyrosine phosphatases exhibit considerable specificity. They even participate directly in the activation of protein tyrosine kinases such as Src family members, as described below. As with the tyrosine kinases, both membrane and nonmembrane forms of tyrosine phosphatases have been observed. Both also contain multiple domains which regulate their associations and locations in cells. Although membrane tyrosine phosphatases contain multiple types of extracellular domains (Streuli, 1996), it is not clear whether they can act as true receptors, since no ligand activation mechanisms for their enzyme activities are known.

Ligand-activated receptor tyrosine kinases can also be regulated by controlling the availability of the ligand to the receptor. Ligands are often synthesized from high molecular weight membrane precursors, which are cleaved proteolytically to release the ligand. Thus, ligand release and availability are determined in part by the activity of the protease(s) involved in the cleavage. One of the factors regulating these proteases is calcium-dependent protein kinase C. Ligand degradation may also contribute to determining ligand availability. Ligand binding to some receptors triggers endocytosis of the ligand–receptor complex into the cell (Sorkin and Waters, 1993), which can lead to three possible fates: (1) the complex may dissociate and the ligand transfer to lysosomes for degradation, while the receptor recycles to the cell surface; (2) the complex may transfer to the lysosome for degradation of both components; (3) both components may recycle to the cell surface. Either mechanism for degrading the ligand reduces its availability. Finally, ligand availability may be determined by interactions with extracellular components, both positively and negatively (Schlessinger *et al*., 1995). Many ligands contain positively charged amino acid sequences which interact with glycosaminoglycans (GAGs), such as heparin, at the cell surface or in the extracellular matrix. This interaction may affect ligand availability in two ways. In a positive sense cell-surface GAGs or other cell-surface components may recruit ligands to the cell surface and vicinity of the receptors. Cell-surface components, such as GAGs, may also act as coreceptors to form multimeric complexes with ligands and receptors which facilitate the activation of the receptor kinases. In a negative sense GAGs may sequester ligands away from their receptors. By combining the two mechanisms, ligands in a tissue may be held in a GAG 'reservoir' until they are released by an acute event that

frees the ligand and consequently initiates receptor activation. Soluble ligand-binding proteins are produced by some tissues and sequester extracellular ligands. For example, a whole family of binding proteins regulate the availability of insulin-like growth factors.

Receptor activation can also be determined by receptor availability, often dictated by receptor turnover, as described above for ligand–receptor complex endocytosis and degradation. These mechanisms appear to be rather receptor specific. For example, ligand binding induces turnover of ErbB1 (EGF receptor) more readily than other ErbB family members. A second aspect of receptor availability is localization. This may occur at either the cellular or subcellular level. At the cellular level only certain cell types or cells in specific locations may contain a given receptor in a tissue. At the subcellular level the receptor location may be restricted to a specific region of the cell surface, e.g. to cell–cell contacts. Localization is particularly important to receptors which act by juxtacrine mechanisms (**Table 8**), since they require appropriate cell–cell contacts for their activation. One example is the Eph family of receptors (**Table 6**), which guide movements of cells and neuronal growth cones in establishing neuronal connections. The mechanism for guidance control involves a repulsive reaction to juxtacrine association of an Eph receptor with its complementary Ephrin ligand on an adjacent cell. This behaviour not only emphasizes the importance of location, but also suggests the possibility of reciprocal signalling, in which both the receptor and ligand initiate pathways and responses to the contact. Once again, this unusual mechanism underlines the diversity which has evolved in receptor tyrosine kinase functions to regulate cellular behaviours in multicellular organisms.

Although ligand-dependent activation is the most common means for stimulation of receptor tyrosine kinase pathways, other mechanisms have been proposed. Overexpression of ErbB2 is sufficient for activation of receptor phosphorylation and cell transformation in some types of cells and has been implicated in neoplasia. One explanation for this effect is that the receptors associate when their concentration reaches a certain level in the membrane, triggering kinase activation and cross-phosphorylation, although the participation of other factors cannot be ruled out. One possible example of indirect activation of ErbB2 through multimerization involves the cell surface molecule CD44, a receptor for the extracellular GAG hyaluronic acid. CD44 has been observed to associate with ErbB2. Hyaluronic acid, an extracellular matrix component, can bind and aggregate CD44 in the cell membrane, thus also aggregating ErbB2 and inducing its activation. Other indirect mechanisms for initiating receptor tyrosine kinase signalling pathways are less easily rationalized. These include stimulation of G protein-coupled receptors, activation by cytokines, cellular stress responses, cell adhesion and membrane depolarization. One possible explanation for these other ligand-independent signalling mechanisms is that the receptors are not acting as enzymes in these instances, but merely serving as docking proteins such as the IRSs (see above) which are phosphorylated by cytoplasmic tyrosine kinases, such as Src, activated by these other signalling mechanisms. An alternative possibility is that these additional mechanisms trigger proteolytic activities which release ligands from their precursors or sequestration sites to activate receptor tyrosine kinases.

Not surprisingly, ligand mimics have evolved which can modulate receptor signalling. In the fruit fly the secreted protein Argos contains an EGF-like domain and blocks signalling through the ErbB family tyrosine kinase receptor, probably acting as a competitor for the activating ligands. In contrast, mammalian epithelia contain a mucin Muc4 with an EGF-like domain which binds to ErbB2 and potentiates ligand-activated phosphorylation through heterodimer ErbB2/ErbB3. The sequence of each of these mimics differs slightly from the sequences of activating ligands of the receptors. Other receptor-binding proteins associate with the receptor extracellular domains to influence ligand binding and signalling or with the cytoplasmic domain to modulate downstream signalling, although the exact mechanisms have not been well studied. Receptor 'desensitization' is a common phenomenon for many types of receptors, including tyrosine kinases. A frequent mechanism involves phosphorylation of the receptor at specific sites to repress its activity. For example, ErbB1 (EGF receptor) can be desensitized by phosphorylation of serine residues in its cytoplasmic juxtamembrane domain by protein kinase C.

BINARY RECEPTORS

Activation of the JAK/STAT Pathway

Binary receptors consist of a nonmembrane tyrosine kinase subunit noncovalently associated at the cytoplasmic surface of the plasma membrane with a transmembrane receptor (**Figure 2b**). One advantage of the dual subunit system is the ability to 'mix and match' different gene products for the two subunits to create a broader array of different signalling units from the same amount of genetic information, increasing the diversity of signalling potential. Moreover, binary receptors often contain multiple receptor (nonkinase) subunits to provide further specificity and diversity. However, the mechanism of signal transduction across the membrane is remarkably similar to that of the single subunit receptor tyrosine kinases. Binding of ligand to the receptor induces dimerization or multimerization of the receptor with concomitant activation of the kinase. Cross-phosphorylation can then occur on both the receptor and kinase subunits, creating binding sites for cytoplasmic proteins with SH2 or PTB domains. In the case of the binary cytokine receptors (reviewed in detail in the previous chapter, *Signalling by Cytokines*),

the associated tyrosine kinases are the Janus kinases (JAKs). The JAKs participate with signal transducing and transcription (STAT) factors in a direct mechanism for regulating transcription (**Figure 3b**). STATs are recruited to phosphorylated receptors via their SH2 domains. Tyrosine phosphorylation of the STATs by receptor-associated JAKs creates SH2 binding sites which can interact reciprocally with the STAT SH2 domains to induce STAT dimerization. The STAT dimers translocate to the nucleus, where they are able to regulate gene transcription (see the previous chapter, *Signalling by TGF-β* for details). The JAK/STAT pathway is thus more direct than the Ras–MAP kinase pathway, which also regulates cell transcriptional activity. Moreover, the JAK/STAT pathway can also be initiated by some receptor tyrosine kinases, such as ErbB1. Conversely, adaptors such as Grb2 are able to bind to some cytokine receptor phosphotyrosines to couple them to Ras and the MAP kinase signalling pathway, and IRS docking proteins can link them to multiple signalling elements. Thus, there appears to be substantial redundancy in signalling pathways. A major concern in signalling research is the molecular mechanisms for coordinating different signalling pathways, whether they derive from the same receptor (pleiotropy) or from different receptors (cross-talk).

Lymphocyte T cell Complex

Even more complex binary receptor tyrosine kinase complexes participate in lymphocyte regulation (Peterson and Koretzky, 1998), as exemplified by the T cell receptor complex. T lymphocytes are the immune cells responsible for defence against invading organisms, such as bacteria and viruses. The T cell has evolved a very sophisticated mechanism for recognizing invader cells, involving recognition of specific proteins unique to the invaders. However, rather than recognizing the intact protein, specialized cells of the immune system degrade the foreign proteins to peptides, which can then be 'presented' for recognition at their cell surfaces. This antigen presentation is performed by a cell-surface protein complex called the major histocompatibility complex (MHC). It is the recognition of the foreign peptide on the MHC molecule by a receptor on the T cell (T cell receptor) that triggers T cell activation. The recognition and activation processes are finely tuned to provide a graded response to the foreign material.

The interaction of the peptide on the MHC of the 'antigen-presenting cell' with the T cell receptor on the T cell is a juxtacrine response (**Table 8**). Moreover, the T cell receptor itself is a highly complex moiety, a multicomponent binary receptor. The full receptor contains at least nine polypeptides which can interact with four tyrosine kinases and one or more tyrosine phosphatases. Presumably, this complexity facilitates the regulation of the receptor signalling to provide a graded response instead of a switch (on/off) type of response. The signal development can be best presented as a timed series of events. The T cell receptor initially contains two types of moieties. The peptide recognition moiety is a heterodimer of two transmembrane proteins (α/β or γ/δ, depending on the T cell type) which bind the peptide on the juxtaposed antigen-presenting cell (**Figure 7**). This receptor

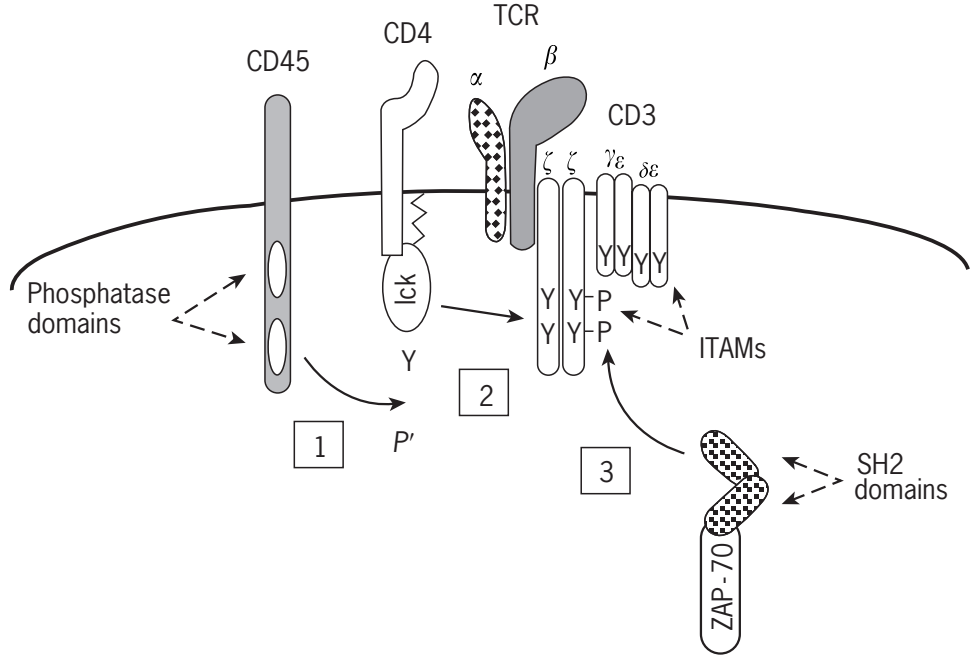


Figure 7 Schematic structure of T cell receptor and some key events in T cell receptor signalling. ITAM, immunoreceptor tyrosine-based activation motif. (From Peterson and Koretsky, 1998, *Clinical and Experimental Rheumatology*, **17**, 107-114.)

heterodimer is associated with two additional transmembrane complexes, a heterodimer and a heterotetramer (CD3). These complexes appear to serve a dual function. First, they act as the tail of a binary receptor to bind specific tyrosine kinases of the Src family for signal initiation. Second, they contain specific sequences called immunoreceptor tyrosine-based activation motifs (ITAMs) which can be phosphorylated to provide binding sites for proteins with SH2 domains, thus serving the function of docking proteins. The kinetic model for T cell activation proceeds when the T cell receptor is engaged by the peptide on the MHC complex of the adjacent cell. This engagement leads to dimerization of the T cell receptor and activation of an associated Src family kinase Fyn. The activated kinase can then phosphorylate the ITAM motifs of the docking proteins to provide recruitment sites for additional molecules. One of the recruited molecules is CD4 (or CD8, depending on the cell type), a binary coreceptor containing a transmembrane component and a second member of the Src family Ick. The CD4 association stabilizes the T cell receptor dimer and provides additional sites for recruitment of SH2-containing components. Among the components recruited to the phosphotyrosine sites created by the Src family members are two additional tyrosine kinases of another family, Zap and Syk.

The result of this complex series of manoeuvers is to recruit a collection of initiators of signalling pathways (**Figure 5** and **Table 5**) to activate the multiple functions of the T cell. The initiators include PLCγ, PI3 kinase, a Ras activator called Vav and the adaptor proteins Shc and Grb2. The activities of the tyrosine kinases are tightly regulated by ancillary molecules. Since Src family kinases can be both positively and negatively regulated by tyrosine phosphorylation (see below), these regulators include the membrane tyrosine phosphatases CD45 and SHP-1 and the nonreceptor tyrosine kinase Csk. CD45 is an activator, while SHP-1 and Csk are both inhibitors of Src family members. The activation mechanisms for Src family members are described in more detail below. Thus, the balance of these activating and inhibiting activities provides mechanisms for generating graded responses. Many of the specifics of the temporal associations and the regulation remain to be discovered, but the T cell receptor provides a clear example of how the binary receptor tyrosine kinase model has evolved to provide a very specific, highly regulated function for higher organisms.

Focal Adhesions

The focal adhesion complex provides another example of the application of the binary receptor principle to a primary cell function. The key tyrosine kinase in this case is the focal adhesion kinase (FAK, see **Figure 1b**). Focal adhesions are sites of cell attachment to the extracellular matrix (see chapter on *the Extracellular Matrix: The Networking Solution*) and of microfilaments to the plasma membrane. Thus, they play a critical role in the determination of cell shape and cell movement (Carraway *et al*., 1997). Furthermore, cell attachment through focal adhesions has been shown to be necessary for normal progression through the cell cycle for cell proliferation. One of the hallmarks of neoplastic transformation is the ability of the transformed cells to escape this adhesion requirement. Focal adhesions are extremely complex structures, containing dozens of components (Carraway *et al*., 1997). Precise assembly and disassembly of these structures is necessary for their roles in cell movements and growth regulation. The objective here is to show how the binary model and tyrosine phosphorylation can contribute to the dynamics of these structures.

The receptors involved in the formation of focal adhesions are integrins, a family containing two types of subunits (α and β), which combine to form a large number of different heterodimeric receptors with different specificities for extracellular matrix components. Ligand binding and aggregation of integrin heterodimers induces their oligomerization in the membrane and recruitment of the nonreceptor tyrosine kinase FAK (**Figure 1B**) from the cytoplasm. This event also stimulates tyrosine phosphorylation of FAK, providing SH2 binding sites for the recruitment of Src. Additional phosphorylation of FAK creates sites for binding docking proteins paxillin and Cas. FAK appears to serve as both kinase and docking protein in these complexes, similar to the cytoplasmic domain of a receptor tyrosine kinase. The combined binding sites in these proteins can initiate further recruitment through SH2 domains and other mechanisms of both structural components of the focal adhesion, such as the actin-binding proteins tensin, talin, vinculin and α-actinin, and signalling pathway components. Included among the latter are PLCγ, PI3 kinase, members of the MAP kinase cascade, members of the Src family and the small G proteins Rac and Rho. These small G proteins are relatives of Ras and act as signalling switches for the processes involved in the organisation of microfilaments in cell movements. One of the enzymes recruited is Csk, which can then serve as a regulator of the Src family members. Thus, the tyrosine kinase FAK, acting through a binary receptor mechanism, is able to initiate the assembly of an extremely complex multimeric structure which is critical to multiple cell functions required for cell movements, such as those involved in tumour invasion.

Src Regulation

Since Src family members play such important roles in many cellular processes involving binary receptors, one of the critical questions in understanding these signalling processes is how the Src activity is regulated. This regulation can most easily be considered as an autorepression mechanism (Hubbard *et al*., 1998) which limits access to the kinase site for catalysis by folding the Src molecule

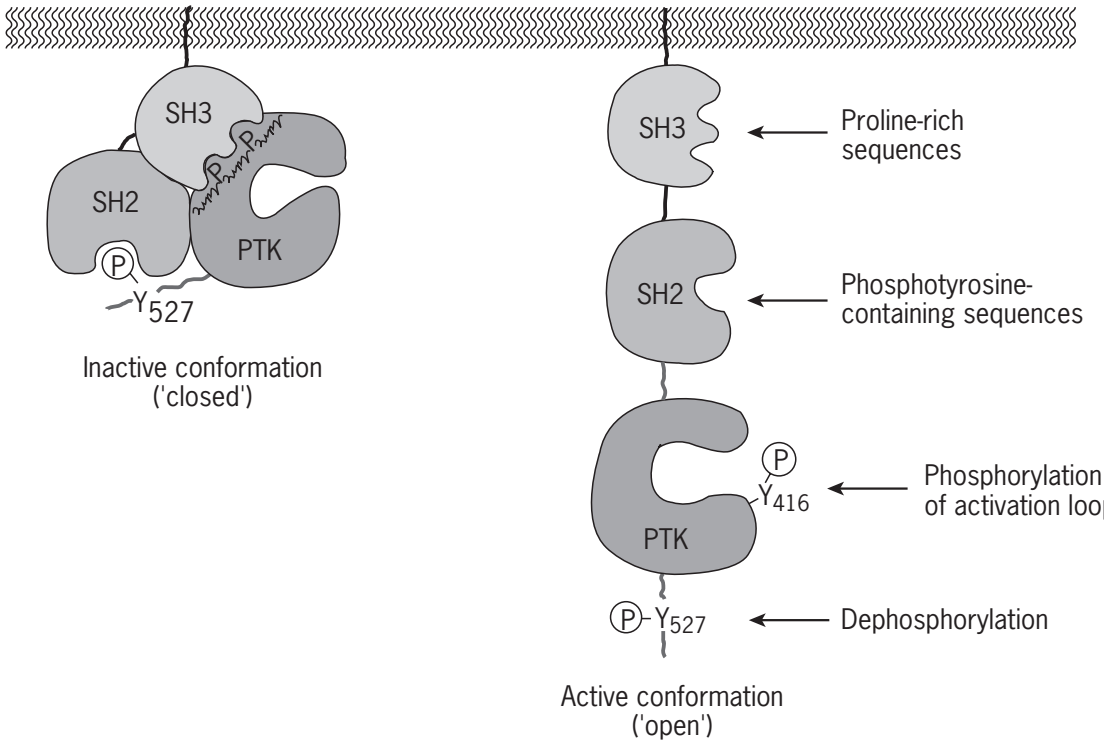


Figure 8 Autorepression/autoactivation model for Src family kinases. (From Hubbard *et al.*, 1998, *Journal of Biological Chemistry*, **273**, 11987-11990.)

into a more compact structure (**Figure 8**). Two interactions are particularly important: (1) the binding of the Src SH3 domain to the peptide linking the SH2 and kinase domains and (2) the binding of the SH2 domain to a phosphorylated tyrosine in the C-terminal tail. Both of these intramolecular interactions must be weak, because they can be displaced by external agents with activation of the enzyme. Thus, Src can be activated by binding appropriate SH3-containing proteins or proteins with SH3-binding motifs. Src can also be activated by protein tyrosine phosphatases, such as CD45 (see **Figure 7**) which cleave the C-terminal phosphotyrosine. Conversely, active Src can be inactivated by phosphorylation of the C-terminal tyrosine by Csk. Activation of Src is also promoted by phosphorylation of a tyrosine near the kinase site to prevent folding of the protein into the compact, inactive form.

NONRECEPTOR TYROSINE KINASE MECHANISMS

Abl

The large number of signalling pathways and cellular functions regulated by the receptor tyrosine kinase and binary receptor mechanisms raises the question whether tyrosine kinases act by any other mode. Unfortunately, it is often difficult to distinguish nonreceptor from receptor mechanisms at the cell level. One highly studied nonreceptor kinase is Abl. Abl is a multidomain molecule whose N-terminal half is similar to the Src family (kinase, SH3 and SH2 domains), but whose C-terminal half has DNA-binding, microfilament-binding and nuclear localization motifs (**Figure 1b**) (Zou and Calame, 1999). Abl is most familiar as the oncogene associated with chronic myelogenous leukaemia. In this cancer Abl is activated as a consequence of a chromosome translocation encoding a fusion protein of Abl and Bcr (breakpoint cluster region). This chimaeric protein appears to be primarily localized to microfilaments in cells. In contrast, normal unactivated Abl is primarily a nuclear protein, although it is able to shuttle between the nucleus and cytoplasmic microfilaments in response to cell adhesion. In the nucleus Abl has been implicated in responses to DNA damage and in cell cycle progression by virtue of its ability to bind p53 and pRb, respectively, both of which are transcription regulators and tumour suppressors. The specific mechanisms for these responses are yet unclear. Abl can also bind RNA polymerase II, the Crk adaptor (**Table 5**) and the protein tyrosine phosphatase SHP1 (**Table 5**). Abl has been implicated in numerous signalling pathways, including those involving PI3K, PKC, Ras, JAK/STAT and Rac. Thus, it has the capacity for participating in multiple cellular functions, arising from both the binding of SH2-containing proteins to its phosphotyrosines and other proteins to other domains or motifs. Other tyrosine kinases which function by nonreceptor mechanisms probably act similarly, although probably not as globally.

Membrane transport and tyrosine kinases

A more direct action of nonreceptor tyrosine kinases is suggested by studies of nonreceptor tyrosine kinase effects on channel activities (Thomas and Brugge, 1997). For example, Src binds K^+ ion channels and induces their phosphorylation, which decreases channel activity. In other cases the relationship between channel function and tyrosine kinase activity is less clear. Aggregation of acetylcholine receptor ion channels can lead to tyrosine phosphorylation of the receptor and association of nonreceptor kinase with the receptor. This and other kinase-channel interaction mechanisms resemble the binary receptor tyrosine kinases, although the consequences for downstream signalling in the specific cases are often unclear. In a number of cases Src family kinases can be activated via ion fluxes into cells through regulated channels. The mechanisms of these activations are varied, involving both receptor and nonreceptor processes. These examples again underline the versatility of the tyrosine kinase functions in many aspects of cell function.

SIGNALLING PATHWAYS AND CELLULAR FUNCTIONS OF TYROSINE KINASES

It should be obvious from the previous sections that tyrosine kinases can initiate multiple signalling pathways. In trying to understand the cellular functions of these kinases, it is useful to describe the downstream effectors and biological effects of these pathways. Such analyses are complicated by the fact that information about some of the pathways is incomplete. Moreover, many of the pathways are branched and have multiple effects. Finally, there are intersections of some pathways which cannot easily be represented by a linear or even a two-dimensional diagram. The major pathways described in this chapter include PLCγ, PI3 kinase and Ras–MAP kinase. However, there are variants of both the PI3 kinases and MAP kinases which involve different cellular functions. For example, at least three different classes of mammalian MAP kinases contribute to cell behaviour by regulating transcription: Erk, p38 and JNK. Each of these responds to different stimuli and results in a different cellular response. Moreover, it is important to remember that these individual pathways are not independent, but form a dynamic network.

The key to understanding the roles of tyrosine kinases is to be able to link these signalling pathways to the cellular functions elaborated by the tyrosine kinases. A preliminary and simplistic effort is shown in **Table 2**. However, all of these complex functions are usually the consequence of multiple pathways. As a result there is not necessarily a linear relationship between any kinase and a function. Moreover, the same pathway may be involved in different, almost contradictory, functions in different cellular contexts. EGF activation of the Ras–Erk pathway in PC12 cells induces proliferation, but NGF stimulation of the same pathway in these cells leads to differentiation. The difference appears to reside in the temporal aspects (kinetics) of the pathway. Another problem in analysing kinase functions is redundancy. This issue is illustrated from 'gene knockout' studies, in which the gene for a particular protein has been eliminated from the mouse genome for analysis of the phenotype of the mutant. For example, the gene for Src can be eliminated from the mouse without severe consequences for the reproduction of the animal. Only when three genes of the Src family are eliminated do the genetic defects become lethal. In contrast, gene deletions are lethal for ErbB2, ErbB3 or ErbB4 and the ErbB ligand heregulin because of a failure of heart development at about embryonic day 10. These studies indicate that the function of a particular kinase in an organism depends on its time and place of expression as well as the consequences of its downstream signalling in the cell and tissue of origin.

TYROSINE KINASES AND CANCER

The functions listed in **Table 2** are important because they include many of the cellular behaviours which are modified in neoplastic transformation of cells (Nicolson, 1976). Thus, the phenotypic changes in tumour cells correspond closely to the functions regulated by tyrosine kinases in cells. However, the relationship of tyrosine kinases and human cancer is not so simple. Formation of diagnosable human tumours appears to require about five genetic changes in a single cell lineage. Contrary to the original oncogene hypothesis, many of these genetic lesions are not in tyrosine kinase-related pathways for cell proliferation. This situation results because tissues have evolved 'tumour suppressors' to act as brakes for cell proliferation and tumour progression. Thus, removal of these suppression mechanisms is as important as enhancement of the progression mechanisms. Surprisingly, reversal of tyrosine phosphorylation by phosphatase action does not appear to be an important tumour-suppressor mechanism. The only tumour-suppressor phosphatase (PTEN) to be identified to date acts more robustly on inositol (lipid) phosphates than on protein phosphates. It may therefore be more important in countering the effects of PI3 kinase than tyrosine kinases.

To understand the role of tyrosine kinases in human cancer, it is instructive to analyse those cancers in which tyrosine kinases have been implicated as contributors (**Table 10**). The list is necessarily limited because providing evidence that a kinase contributes to tumour progression (cause versus effect) is difficult. Furthermore, downstream effectors in tyrosine kinase pathways, such as

Table 10 Tyrosine kinases implicated in neoplasia

Tyrosine kinase	Mechanism of activation	Cancer
ErbB1	Truncation	Glioma
ErbB2	Amplification	Breast
PDGFRβ	Chromosome translocation	Leukaemia
Kit	Mutation	Leukaemia
Met	Mutation, overexpression	Multiple cancers
FGFR	Mutation	Multiple myeloma
Ret	Chromosome translocation	Multiple endocrine
Alk	Chromosome translocation	Lymphoma
Src	Overexpression, activation	Multiple
Yes	Overexpression, activation	Multiple
Abl	Chromosome translocation	CML
JAK2	Chromosome translocation	Leukaemia

R, receptor.

Ras and PI3 kinase, are more proximal to the phenotypic changes in cells and tissues and may thus be more potent oncogenes. However, some tyrosine kinases are very potent oncogenes. A good example is ErbB2, which is an inducer of mammary cancer when its gene is expressed in the mammary gland of mice. Although the reason for this potency is not entirely clear, two observations are likely important. One is that ErbB2 is susceptible to mutations and other events which induce the activation of its tyrosine kinase. The second is that ErbB2 can activate multiple downstream pathways to change the mammary cell phenotype, including both Ras–MAP kinase and PI3 kinase pathways.

The list in **Table 10** suggests that these two factors are important in many instances in which tyrosine kinases contribute to cancer. Particularly noteworthy are activations of tyrosine kinase oncogenes by chromosome translocations. Since tyrosine kinases are frequently autoregulated, displacement of the regulatory regions by truncation (ErbB1) or by chromosome translocation (Abl) can remove the autorepression and activate the kinase. This mechanism is familiar in viral oncogenes. In the instance of v-erbB (ErbB1), the extracellular domain has been truncated, facilitating dimerization and activation of the intracellular kinase. In the case of v-Src, the C-terminal tail is missing. Since this sequence contains the phosphorylated tyrosine which interacts with the SH2 domain (**Figure 8**), phosphorylation of the enzyme at this site is no longer possible as a repression mechanism.

An alternative mechanism for activation of tyrosine kinases in cancer is the autocrine growth loop (Kolibaba and Druker, 1997), in which overexpression or mislocation of a ligand for the receptor kinase aberrantly induces its activation. The cause versus effect relationship of this mechanism in human cancer is also difficult to establish, although it has been well characterized in cell culture studies. Most of the evidence is based on statistical analyses of clinical outcomes. Such studies have implicated TGF-α in oesophageal cancer and insulin-like growth factor in prostate cancer.

OVERVIEW

Tyrosine kinases regulate a large number of cellular functions via an array of signalling pathways. Most, but not all, of these pathways originate at the cell plasma membrane via transmembrane receptor tyrosine kinases or binary receptor–tyrosine kinase couples. Both the tyrosine kinases and their effector pathways are highly regulated and must be integrated both spatially and temporally into the organisation of cellular functions. Mutational activation or overexpression of tyrosine kinases can lead to aberrant cellular behaviours, including neoplastic transformation and tumour progression.

REFERENCES

Carraway, K. L. and Carraway, C. A. C. (1995) Signaling, mitogenesis and the cytoskeleton: where the action is. *BioEssays*, **17**, 171–175.

Carraway, K. L., *et al*. (1997) Signaling and the Cytoskeleton, (Springer, Berlin).

Claesson-Welsh, L. (1994) Platelet-derived growth factor receptor signals. *Journal of Biological Chemistry*, **269**, 32023–32026.

Cohen, P. T. W. (1997) Novel protein serine/threonine phosphatases: variety is the spice of life. *Trends in Biochemical Science*, **22**, 245–251.

Hubbard, S. R., *et al*. (1998) Autoregulatory mechanisms in protein-tyrosine kinases. *Journal of Biological Chemistry*, **273**, 11987–11990.

Hunter, T. (1998) The phosphorylation of proteins on tyrosine: its role in cell growth and disease. *Philosophical Transactions of the Royal Society of London, Series B*, **353**, 583–605.

Kolibaba, K. S. and Druker, B. J. (1997) Protein tyrosine kinases and cancer. *Biochimica Biophysica Acta*, **1333**, F217–F248.

Nicolson, G. L. (1976) Transmembrane control of the receptors on normal and tumor cells. II. Surface changes associated with transformation and malignancy. *Biochimica Biophysica Acta*, **457**, 57–108.

Panayotou, G. and Waterfield, M. D. (1993) The assembly of signalling complexes by receptor tyrosine kinases. *BioEssays*, **15**, 171–177.

Peterson, E. J. and Koretzky, G. A. (1998) Signal transduction in T lymphocytes. *Clinical and Experimental Rheumatology*, **17**, 107–114.

Riese, D. J., II and Stern, D. F. (1998) Specificity within the EGF family/ErbB receptor family signalling network. *BioEssays*, **20**, 41–48.

Schlessinger, J. Lax, I. and Lemmon, M. (1995) Regulation of growth factor activation by proteoglycans: what is the role of the low affinity receptors? *Cell*, **83**, 357–360.

Sorkin, A. and Waters, C. M. (1993) Endocytosis of growth factor receptors. *BioEssays*, **15**, 375–382.

Streuli, M. (1996) Protein tyrosine phosphatases in signalling. *Current Opinions in Cell Biology*, **8**, 182–188.

Taylor, S. S., Radzio-Andzelm, E. and Hunter, T. (1995) How do protein kinases discriminate between serine/threonine and tyrosine? Structural insights from the insulin receptor protein–tyrosine kinase. *FASEB Journal*, **9**, 1255–1266.

Thomas, S. M. and Brugge, J. S. (1997) Cellular functions regulated by Src family kinases. *Annual Reviews of Cellular Development*, **13**, 513–609.

Wells, J. A. (1996) Binding in the growth hormone receptor complex. *Proceedings of the National Academy of Sciences of the USA*, **93**, 1–6.

Zou, X. and Calame, K. (1999) Signaling pathways activated by oncogenic forms of Abl tyrosine kinase. *Journal of Biological Chemistry*, **274**, 18141–18144.

FURTHER READING

Alberts, B., *et al*. (1994). *Molecular Biology of the Cell*, 3rd edn. (Garland Publishing, New York).

Hesketh, R. (1997). *The Oncogene and Tumor Suppressor Gene Factsbook* (Academic Press, San Diego).

Kellie, S. (1994). *Tyrosine Kinases and Neoplastic Transformation* (R. G. Landes, Austin, TX).

Lodish, H., *et al*. (1995). *Molecular Cell Biology*, 3rd edn. (Scientific American Books, New York).

Mendelsohn, J., *et al*. (eds) (1995). *The Molecular Basis of Cancer* (W. B. Saunders, Philadelphia).

Peters, G. and Vousden, K. H. (eds) (1997). *Oncogenes and Tumor Suppressors* (IRL Press, Oxford).

Ruddon, R. W. (1995). *Cancer Biology*, 3rd edn. (Oxford University Press, Oxford).

Websites

http://www.kinase.com/.

http://www.sdsc.edu/kinases/.

http://www.stke.org/.

http://www.mshri.on.ca/pawson/research2.html.

Chapter 13

Signalling by TGF-β

Jing Qing and Rik Derynck
University of California at San Francisco, San Francisco, CA, USA

CONTENTS

INTRODUCTION

While normal, differentiating cells closely coordinate their proliferation and differentiation programmes, this balance is deregulated during malignant transformation into tumour cells. Compared with normal cells, transformed cells acquire a higher proliferative index and decreased differentiation, concomitant with a lower degree of dependence on extracellular cues. These changes in cell behaviour and responsiveness often result in part from genetic alterations, whereby defined mutations activate oncogenes or inactivate tumour-suppressor genes. Additionally, changes in the production of stimulatory or inhibitory growth and differentiation factors and/or in the cellular responsiveness to these factors greatly contribute to the behavioural and phenotypic changes in tumour cells. In fact, tumour cells often display autocrine responsiveness to increased endogenous expression of growth factors, and this autocrine stimulation contributes to tumour formation and cancer progression. Similarly, the invasive and metastatic phenotype of the tumour cells also results from mutations, changes in gene expression and/or altered production of cell surface-associated or extracellular mediators, and altered responsiveness to these factors.

Various growth factors and cytokines have been implicated in the progression and behaviour of tumour cells and cancers. Among these, transforming growth factor-β (TGF-β) and its downstream effectors are key determinants of the tumour cell behaviour of carcinomas. The TGF-β production by tumour cells and the responsiveness of tumour cells to autocrine TGF-β and TGF-β in the tumour microenvironment exert both positive and negative effects on cancer development. Accordingly, TGF-β and the TGF-β signalling pathway have been considered as both a tumour suppressor and a promoter of tumour progression and invasion. This chapter introduces the role of TGF-β in tumour development, specifically of carcinomas, and summarizes our knowledge of the TGF-β signalling mechanisms, i.e. the cell surface receptors and downstream effector proteins.

TGF-β EXPRESSION AND ACTIVATION

TGF-β is a secreted polypeptide, which in its receptor-binding, fully active form is a disulfide-bonded and stable dimer. The TGF-β precursor polypeptide comprises an N-terminal signal peptide followed by a large prosegment, which is about twice as long as the C-terminal, active TGF-β sequence. Protein purification and cDNA cloning have revealed the existence of three TGF-βs, TGF-β1, -β2 and -β3, each encoded by a different gene and made as a precursor with similar structural organisation. TGF-β1 was the first identified and best characterized TGF-β, and is therefore considered as the prototype of these three TGF-βs, which act through the same receptor system and have similar biological activities in cell culture (Derynck and Choy, 1998). TGF-β1 is also the prototype of the TGF-β superfamily, which comprises about 60–70 structurally related proteins, characterized by seven characteristically spaced cysteines in the C-terminal segment of the precursor polypeptide (Derynck and Feng, 1997; Piek *et al*., 1999; Massagué *et al*., 2000). All TGF-β superfamily members are thought to act as dimers that bind to structurally related receptors. TGF-β-related proteins are found in multicellular eukaryotes from *C. elegans* and *Drosophila* to all vertebrates and regulate a variety of developmental and differentiation processes.

TGF-β1 expression is often upregulated in tumour cells and is therefore the focus of most studies on the role of TGF-β in tumorigenesis. With few exceptions, all cells in culture express TGF-β1, but this may be due to the absence of an extracellular matrix substrate. Thus, TGF-β1 expression is repressed when cells are grown on extracellular matrix or once they have deposited their extracellular matrix, a process strongly activated by TGF-β (Derynck and Choy, 1998). TGF-β1 expression in cell culture is therefore not necessarily a reflection of TGF-β expression *in vivo*, but may be a response to injury, consistent with increased TGF-β1 expression at sites of tissue injury. TGF-β itself activates TGF-β1 expression, a positive feedback and amplification response of relevance in tumour development.

Following an intracellular proteolytic cleavage between the prosequence and the active TGF-β sequence, TGF-β is released as a 'latent' complex, which is incapable of binding to the TGF-β receptors and is consequently biologically inactive. This latent complex consists of an active TGF-β dimer in a noncovalent complex with two prosegments, to which one of several 'latent TGF-β binding proteins' is often disulfide linked (Munger *et al*., 1997). This latent complex represents an important safeguard against 'unneeded' or 'inadvertent' activation and these binding proteins may stabilize and target latent TGF-β to the extracellular matrix, where it is stably sequestered (Munger *et al*., 1997). The extracellular matrix thus acts as a reservoir, from which TGF-β can readily be made available to cells without the need to induce TGF-β synthesis.

The secretion of TGF-β as a latent complex necessitates the existence of a regulated activation process. While latent TGF-β is efficiently activated by acidic conditions, its physiological activation occurs most likely through proteases, which degrade the TGF-β prosegments and thereby release the highly stable, active TGF-β dimer. Since plasmin activates latent TGF-β and plasminogen is converted into plasmin at sites of cell migration and invasion, we assume that plasmin-mediated activation of TGF-β occurs at sites of angiogenesis and tumour development, thus exposing endothelial and tumour cells to active TGF-β. Matrix metalloproteases may also play a key role in activation of latent TGF-β. For example, the matrix metalloproteases MMP-9 and MMP-2 have the ability to activate latent TGF-β (Yu and Stamenkovic, 2000). Since metalloproteases are frequently expressed by malignant cells, this mechanism may locally activate TGF-β at sites of tumour cell invasion. Other mechanisms of activation may not depend on proteases. For example, the extracellular matrix protein thrombospondin (Ribeiro *et al*., 1999) and the $\alpha v\beta 6$ integrin, which is expressed at the surface of epithelial cells in response to inflammation (Munger *et al*., 1999), may activate TGF-β through a conformational change in the TGF-β complex. Thus different mechanisms may regulate TGF-β activation in different physiological contexts, and tumour cells are well equipped to activate TGF-β locally. Finally, high-energy irradiation, which is often used in cancer therapy, also activates TGF-β (Barcellos-Hoff *et al*., 1994).

CELLULAR RESPONSES TO TGF-β

TGF-β exerts a large variety of cellular responses, which often depend on the cell type and physiological conditions, thus making the responses highly context dependent. Since the role of TGF-β in tumour development is best studied in carcinoma development, we will briefly describe the cellular responses to TGF-β that play a role in the development of carcinomas from normal epithelial cells.

While TGF-β stimulates proliferation of various mesenchymal cell types, including fibroblasts, it is a very potent inhibitor of proliferation of epithelial cells and cells of haematopoietic origin, including various immune cells (Derynck and Choy, 1998; Massagué *et al*., 2000). The TGF-β signalling process that leads to growth inhibition is considered as a tumour suppressor pathway, which is often inactivated to allow tumour development and cancer progression (de Caestecker *et al*., 2000; Massagué *et al*., 2000). Cell cycle progression and proliferation are driven by a complex and interdependent regulation by cyclins, cyclin-dependent kinases (cdks) and cdk inhibitors, which form complexes with each other (Sherr and Roberts, 1999). Accordingly, TGF-β-induced growth arrest, which occurs in the late G_1 phase of the cell cycle, induces a variety of alterations in the levels and activities of cyclins, cdks and cdk inhibitors, although many of these effects could be considered as indirect. Several direct TGF-β effects may play key roles in TGF-β-mediated growth arrest.

The key event that leads to TGF-β-induced growth arrest is the induction of expression of the cdk inhibitors $p15^{Ink4B}$ and/or $p21^{Cip1}$, depending on the cell type and context (Massagué *et al*., 2000) (**Figure 1**). $p21^{Cip1}$, similarly to the related $p27^{Kip1}$, interacts with complexes of cdk2 with cyclin E or A and thereby inhibits the cdk2 activity, thus preventing cell cycle progression. $p21^{Cip1}$ and $p27^{Kip1}$ also interact with cdk4 and cdk6 and stabilize their interaction with cyclin D, thus playing a role in the activation of these complexes (Sherr and Roberts, 1999). Since the complexes of cdk2 with cyclin E or A act downstream from the complexes of cdk4 or 6 with cyclin D, TGF-β-induced expression of $p21^{Cip1}$ leads to cell cycle arrest in late G_1. In contrast, $p15^{Ink4B}$ interacts with and inactivates cdk4 and 6 and prevents their complex formation with cyclin D. In addition, $p15^{Ink4B}$ binds to the complexes of cdk4 or 6 with cyclin D and thereby not only inactivates the catalytic activity of these cdks, but also displaces $p21^{Cip1}$ and $p27^{Kip1}$ from these complexes, thus allowing them to bind to and inactivate the cdk2 complexes with cyclin A and E. Consequently,

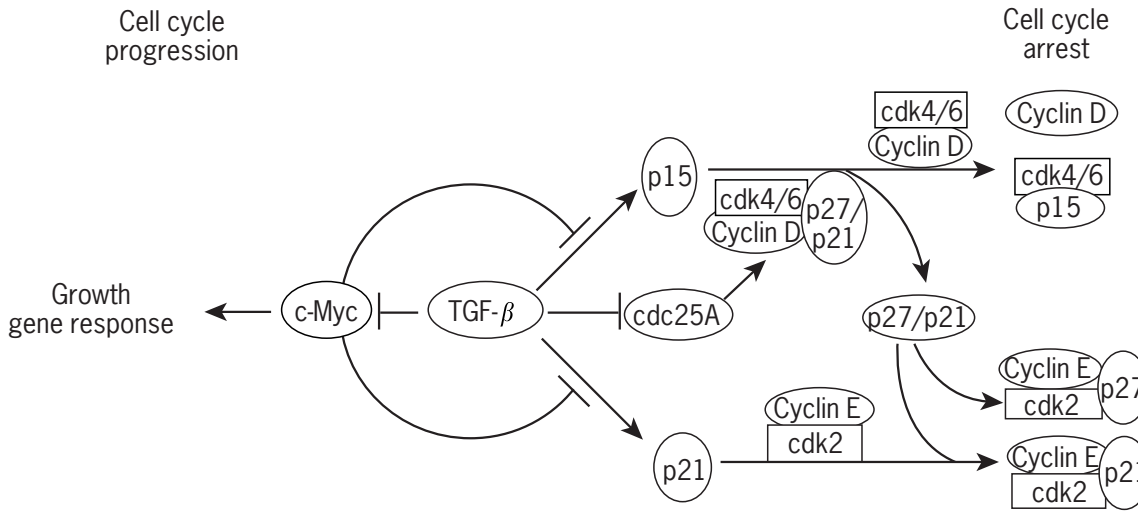


Figure 1 Schematic representation of TGF-β-induced mechanism of growth arrest in late G_1. TGF-β induces the expression of p15Ink4B and p21Cip1 and down-regulates the expression of c-myc and cdc25A. The down-regulation of c-myc expression relieves the inhibition of p15Ink4B and p21Cip1 expression, thus allowing induction of these cdk inhibitors by Smads.

TGF-β-induced expression of $p15^{Ink4B}$ inactivates cell cycle progression at two levels, $p15^{Ink4B}$-mediated inactivation of cdk4 and cdk6 and, as a result of $p21^{Cip1}$ or $p27^{Kip1}$ displacement, inactivation of cdk2 through $p21^{Cip1}$ or $p27^{Kip1}$ binding (Massagué *et al*., 2000).

Additional mechanisms may also contribute to TGF-β-mediated growth arrest, again depending on the cell type. For example, TGF-β inhibits cdk4 expression, presumably at the translational level (Ewen *et al*., 1993). TGF-β also down-regulates the levels of cdc25A (Iavarone and Massagué, 1997), a tyrosine phosphatase, and this decreased activity leads to increased tyrosine phosphorylation of cdk4 and cdk6 and consequent inhibition of their kinase activity. TGF-β also inhibits c-Myc expression in normal epithelial cells, and c-Myc plays a role in growth arrest in response to TGF-β (Massagué *et al*., 2000; Chen *et al*., 2001). High levels of c-Myc repress $p15^{Ink4B}$ and $p21^{Cip1}$ expression, and a decrease in c-Myc levels results in derepression, thus allowing TGF-β-induced transcription of the $p15^{Ink4B}$ and $p21^{Cip1}$ genes. The TGF-β-mediated decrease in c-Myc levels may also play a role in the down-regulation of cdc25A expression in response to TGF-β, since c-Myc may positively regulate cdc25A expression. Finally, the TGF-β-induced interaction of protein phosphatase 2A with S6 kinase and consequently decreased S6 kinase activity may contribute to TGF-β-mediated growth arrest (Petritsch *et al*., 2000).

TGF-β regulates the expression of a large variety of genes through activation or repression of transcription. Among the many genes, TGF-β regulates the expression of transcription factors, secreted cytokines and growth factors, extracellular matrix proteins, proteases and integrins (Derynck and Choy, 1998; Massagué *et al*., 2000). Which genes are regulated by TGF-β and the extent of this regulation are highly cell type and context dependent. The induction of transcription factors by TGF-β results in a variety of indirect responses, thus enhancing the complexity of the TGF-β response. Similarly, the induction of cytokine and growth factor expression, e.g. interleukin 1 and PDGF, by TGF-β results in indirect, yet physiologically very important, cellular responses. The potent ability of TGF-β to induce the expression of extracellular matrix proteins stands in contrast to other growth factors. TGF-β induces the expression of some, but not all, common extracellular matrix proteins, including collagens and fibronectin. Consequently, TGF-β expression and activation are major determinants of extracellular matrix synthesis and deposition. This activity is often complemented by a TGF-β-induced decrease in protease activity and increased expression of protease inhibitors, which together enhance the increased extracellular matrix deposition. TGF-β also enhances the expression of a variety of integrins, depending on the cell type, often resulting in increased cell adhesion to the extracellular matrix and presumably also increased integrin signalling. Finally, TGF-β is a potent chemoattractant of monocytes, macrophages and fibroblasts (Derynck and Choy, 1998). Consequently, TGF-β activation often results in localized inflammation and chemoattraction of fibroblasts, which, together with the mitogenic effect of TGF-β on fibroblasts and stimulation of extracellular matrix deposition, results in fibrosis (Roberts *et al*., 1986). These responses may also explain radiation-induced fibrosis, a consequence of TGF-β activation (Barcellos-Hoff *et al*., 1994).

TGF-β RECEPTORS AND INTERACTING PROTEINS

TGF-β signals through a cell surface receptor complex of two types of transmembrane serine/threonine kinases, named type I and type II receptors (Derynck and Feng, 1997; Massagué, 2000) (**Figure 2**). The type II and type I receptors are structurally related with a high level of sequence conservation in their cytoplasmic kinase domains, besides similarities in their extracellular domains. They autophosphorylate and phosphorylate target proteins on serines and threonines, yet also have tyrosine kinase activity. Both receptor types form cell surface dimers in the absence of TGF-β binding. These type II and type I receptor dimers also have an intrinsic affinity for each other, thus resulting in a heterotetrameric complex, which is stabilized by TGF-β binding. Among the receptors for TGF-β superfamily members, several heteromeric combinations of type II and type I receptors form functional signalling complexes, thus providing a variety of combinatorial type II/type I receptor interactions. In the case of TGF-β, only one type II TGF-β receptor, i.e. TβRII, has been identified. Three different type I receptors have been proposed as type I TGF-β receptors. Most gene expression responses in most cell types are mediated by the TβRI receptor, while the potential roles of the ALK-1/TSR-1 and ALK-2/ActRI/Tsk7L type I receptors in TGF-β signalling remain to be better explored. TGF-β receptors are expressed on most, if not all, cell types (Derynck and Feng, 1997).

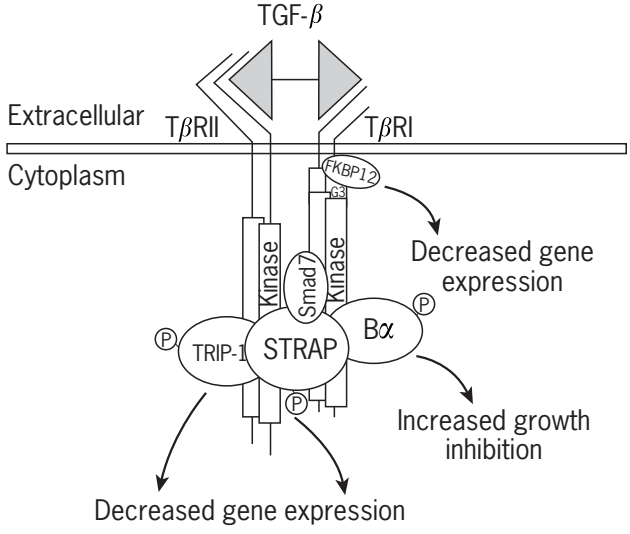


Figure 2 The TGF-β receptor complex with interacting regulatory proteins. The receptor complex consists of two TβRII and two TβRI polypeptides. FKBP12 interacts constitutively with TβRI, while the WD-repeat proteins TRIP-1, STRAP and the Bα subunit of protein phosphatase 2A interact following ligand activation of kinases and are phosphorylated. TRIP-1 interacts with TβRII, while the Bα subunit of protein phosphatase 2A interacts with TβRI. STRAP can interact with either receptor and also interacts and synergizes with Smad7.

All three TGF-βs interact with the same TGF-β receptor complex (Derynck and Feng, 1997). TGF-β1 and -β3 interact primarily with the TβRII receptor, yet also contact the type I receptor in the complex. Thus, TβRII binds TGF-β1 and -β3 without a requirement for a type I receptor. In contrast, TGF-β2 does not bind to either TβRII or TβRI alone, but binds efficiently to the heteromeric receptor complex. TGF-β binding to the receptor complexes is enhanced in the presence of the type III receptors β-glycan and endoglin. β-Glycan is an abundant cell surface proteoglycan with a short cytoplasmic domain and no known signalling function. β-Glycan binds all three TGF-βs with high efficiency and may enhance the efficiency of receptor binding of TGF-β, most notably TGF-β2. Endoglin, a structurally related glycoprotein, is primarily expressed by vascular endothelial cells. Endoglin binds TGF-β1 and -β3 with high affinity, but not TGF-β2, and may enhance TGF-β's presentation to the TGF-β signalling receptor complex, similarly to β-glycan (Derynck and Feng, 1997).

Following TGF-β binding to the receptor complex, signalling is initiated through the activities of the cytoplasmic kinase domains of both receptors. The TβRII kinase, when overexpressed, is constitutively active and phosphorylated. Whether this is also the case at endogenous expression levels or whether TGF-β induces TβRII autophosphorylation remains to be clarified. A key event in receptor activation is the phosphorylation, and consequent activation, of the TβRI kinase by the TβRII kinase. TβRII phosphorylates the cytoplasmic domain of TβRI on serine and threonine residues in the 'GS sequence', a sequence which is conserved among the type I receptors and is located immediately upstream from the kinase domain. This phosphorylation then presumably induces a conformational change in the TβRI cytoplasmic domain that activates the kinase and consequently allows TβRI autophosphorylation and phosphorylation of downstream target proteins (Derynck and Feng, 1997; Massagué, 2000).

Several proteins have been shown to interact with the cell surface TGF-β receptor complexes (**Figure 2**). Among these, FKBP12 interacts with the juxtamembrane domain of type I receptors and may regulate its conformation, and dampens the TGF-β receptor activation. FKBP12 interacts constitutively with, yet is not phosphorylated by, the TGF-β receptor (Chen *et al*., 1997). In contrast, three WD-repeat containing proteins associate with and are phosphorylated by the receptor complex following ligand-induced activation. TRIP-1 interacts with TβRII (Choy and Derynck, 1998), while the Bα subunit of the protein phosphatase 2A interacts with type I receptors (Griswold-Prenner *et al*., 1998). STRAP, on the other

hand, interacts with both TβRII and TβRI and in turn interacts with Smad7 (see below) to decrease TGF-β signalling (Datta and Moses, 2000). Since WD repeats mediate protein–protein interactions, it is likely that these proteins allow interactions of multiprotein complexes with the receptors. These proteins regulate the TGF-β receptor response, but there is currently no solid evidence that they act as effectors of TGF-β responses.

SMAD AND NON-SMAD SIGNALLING

The only characterized signalling effector pathway, initiated by activated TGF-β receptors, is provided by the Smads (Piek *et al*., 1999; Itoh *et al*., 2000; Massagué, 2000; Massagué *et al*., 2000). The Smads, a family of structurally related proteins, are directly activated by the receptors and then translocate into the nucleus to act as ligand-dependent transcriptional regulators of target genes. The Smads are structurally related to each other in two domains, an N- or MH1 domain which corresponds to the N-terminal third of the protein, and a C- or MH2 domain which corresponds to the C-terminal third of these proteins. Based on structural and functional characteristics, the Smads are divided into three subfamilies. Smad1, -5 and -8 and Smad2 and -3 are 'receptor-activated' Smads that are phosphorylated on C-terminal serines by the activated type I receptor. The activated TβRI phosphorylates and thereby activates Smad2 and -3, whereas Smad1, -5 and -8 are activated by BMP receptors. Following their release from the receptors, the activated Smads form a heterotrimeric complex with Smad4, which serves as a common mediator for all receptor-activated Smads. This complex consists of two receptor-activated Smads and one Smad4, raising the possibility of combinatorial interactions among receptor-activated Smads. This Smad complex translocates into the nucleus and cooperates with other transcription factors to activate or repress transcription of defined genes in response to TGF-β (Piek *et al*., 1999; Itoh *et al*., 2000; Massagué, 2000; Massagué *et al*., 2000) (**Figure 3**).

Besides the receptor-activated Smads and Smad4, which act as ligand-induced effectors, two other 'inhibitory' Smads regulate Smad signalling. Smad6 and -7 have much less sequence conservation in their MH1 domain than the other Smads, and interfere with the activation of effector Smads (Piek *et al*., 1999; Itoh *et al*., 2000; Massagué, 2000). Smad6 and -7 are able to interact with type I receptors, thus competitively preventing the 'receptor-activated' Smads from being phosphorylated, whereas Smad6 additionally interferes with the heteromeric complex formation of Smad1 with Smad4. Although Smad6 and -7 seem to interact nonspecifically with type I receptors, Smad7 may primarily inhibit TGF-β signalling, whereas Smad6 primarily inhibits BMP signalling.

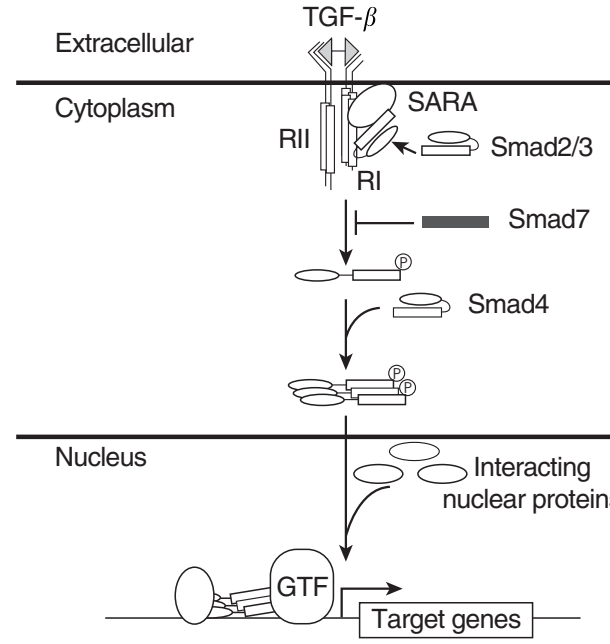


Figure 3 TGF-β-induced signalling through Smads. Following ligand-induced receptor activation, Smad2 and/or -3 interact transiently with the TβRI receptor and this interaction is stabilized by the FYVE-protein SARA. Smad2 and -3 are C-terminally phosphorylated by TβRI and then dissociate from the receptor to form a heterotrimeric complex consisting of two receptor-activated Smads and Smad4. This complex is then translocated into the nucleus where it interacts at the promoter with other transcription factors to regulate gene expression. Smad7 inhibits activation of Smad2 and/or -3 by the receptors.

Accordingly, TGF-β signalling induces Smad7 expression, thus providing a TGF-β induced negative feedback loop, whereas BMP signalling induces Smad6 expression. While the inhibitory functions of Smad6 and -7 can be easily conceptualized as a mechanism of competitive inhibition, the functions of Smad6 and -7 may be more complex (Piek *et al*., 1999; Itoh *et al*., 2000). This is suggested by the observations that Smad6 and -7 are primarily localized in the nucleus and that both Smads can cooperate with TGF-β signalling in inhibiting adipocyte differentiation. Smad7 may also act as an effector of TGF-β-induced cell death, whereas Smad6 can function as a corepressor (Piek *et al*., 1999; Itoh *et al*., 2000). Further studies are needed to define the functions of Smad6 and -7.

Once inside the nucleus, the Smad complexes function as ligand-dependent transcriptional regulators of target genes (Derynck *et al*., 1998; Piek *et al*., 1999; Massagué, 2000; Massagué and Wotton, 2000). TGF-β activates or represses transcription of defined target genes, and many of these responses are direct, i.e. immediate early, responses to TGF-β receptor activation. Smads have

mainly been characterized as TGF-β-induced transcription activators, although they may also mediate TGF-β-induced repression of gene expression. Smads activate transcription from a variety of promoters through physical interaction with the transcriptional coactivator CBP/p300, a variety of DNA-binding transcription factors, and defined sequences in the promoter DNA (Derynck *et al*., 1998; Massagué, 2000; Massagué and Wotton, 2000) (**Figure 4**). Most TGF-β-inducible genes are transcriptionally activated through Smad3, even though TGF-β activates both Smad2 and -3, and some responses, e.g. the induction of $p15^{Ink4B}$ and $p21^{Cip1}$, require the participation and cooperation of both Smad2 and -3 (Feng *et al*., 2000; Pardali *et al*., 2000).

The receptor-activated Smads link to the general transcriptional machinery through a direct association with the transcriptional coactivator CBP/p300. This interaction, which is stabilized by Smad4, occurs through the C-terminal sequence of the Smads, and requires receptor-mediated phosphorylation of the C-terminal serines. CBP/p300 thus serves as transcriptional coactivator for Smads, and this interaction is essential for transcriptional activation (Derynck *et al*., 1998; Massagué, 2000; Massagué and Wotton, 2000). Smads also interact with a variety of DNA binding transcription factors and it is this interaction, together with the DNA binding of Smads, which specifies the promoter binding and transcriptional activation (Derynck *et al*., 1998; Itoh *et al*., 2000; Massagué, 2000; Massagué and Wotton, 2000). The versatility of physical interactions with a variety of transcription factors is best illustrated for Smad3. Among others, Smad3 has been shown to interact with bZIP transcription factors (e.g. c-Jun, ATF-2 or CREB), bHLH transcription factors (e.g. TFE3), runt domain transcription factors (AML-1 and -2 and CBFA1/AML3), nuclear receptors (e.g. the vitamin D_3 and glucocorticoid receptors) and Sp1, either through the MH1- or MH2-domains of the Smad, depending on the interacting transcription factor. While this interaction is usually direct, stable interaction with the transcription factor may depend on the concomitant interaction of the Smad and the transcription factor with CBP/p300. This physical interaction not only localizes the interacting Smad to a defined promoter sequence, but also allows the Smad to enhance the activity of the interacting transcription factor, presumably as a result of the increased interaction with CBP/p300.

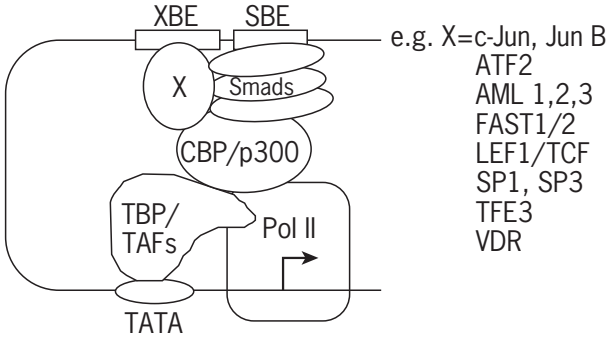


Figure 4 Physical and transcriptional cooperation of TGF-β-activated Smad2 or -3 with other transcription factors at the promoter. The heterotrimeric Smad complex interacts with the CBP/p300 coactivator, which connects to the general transcription machinery. While also interacting with DNA at a Smad binding element (SBE), the Smad complex interacts with one of several possible transcription factors with sequence-specific DNA binding (XBE).

Thus, Smads often serve as coactivators of other transcription factors. This, however, may not always be the case. For example, Smad2 interacts with the DNA binding protein FAST1/2 at an activin-responsive promoter sequence. FAST1/2 does not have transcription activity by itself and therefore serves as a DNA sequence-specific scaffold to allow transcriptional activation by Smad2. Also, the TGF-β-induced transcription of the Smad7 gene is mediated by Smad3, presumably without involvement of other DNA binding, interacting transcription factors. This opportunity may be provided by tandem Smad-binding DNA sequences, which enable permit Smad3 binding to the Smad7 promoter. Finally, while the interacting transcription factor provides high-affinity, sequence-specific binding to the promoter, Smad3 and -4, but not Smad2, are also able to bind DNA. However, Smads have a much lower DNA binding affinity and sequence specificity than most DNA binding transcription factors. Thus, a Smad binding sequence may primarily provide a sequence context conducive to Smad binding in close proximity to the sequence for the high-affinity DNA binding, interacting transcription factor. In this way, the DNA context-dependent binding of a Smad to both the interacting transcription factor and the promoter DNA may explain why TGF-β activates only a select set of the promoters, which show productive DNA binding of the Smad-interacting transcription factor (Derynck *et al*., 1998; Itoh *et al*., 2000; Massagué, 2000; Massagué and Wotton, 2000). This cooperation of Smads with other DNA-binding transcription factors also explains why, prior to the characterization of Smads, no consensus TGF-β response element could be defined in TGF-β-responsive genes and why AP-1 and Sp1 binding sequences were shown to be required for TGF-β responsiveness of various promoters.

The TGF-β-dependent recruitment of Smad complexes to the transcription machinery also allows for interactions with additional coactivators or corepressors, which regulate the amplitude of TGF-β-dependent transcriptional activation through Smads (Itoh *et al*., 2000; Massagué, 2000; Massagué and Wotton, 2000). Besides the interaction of the Smad complex with the CBP/p300 coactivator, Smad4 is able to engage the MSG1 coactivator into the transcription complex to enhance the Smad response. In contrast, recruitment of a corepressor into the complex decreases or inhibits the Smad and TGF-β response.

Several corepressors can interact directly with TGF-β Smads to decrease TGF-β responses. For example, the proto-oncogene product Evi-1 interacts with Smad3 and represses the gene expression and growth-inhibitory responses of TGF-β. Similarly, c-Ski also interacts with receptor-activated Smads and recruits a histone deacetylase to repress Smad-mediated transcription. A similar mechanism may also explain the inhibition by other Smad-interacting corepressors, such as the c-Ski-related SnoN, TGIF, SNIP and SIP1, although, in some cases, their interaction with CBP/p300, or interference with the Smad interaction with CBP/p300, may also play a role (Itoh *et al*., 2000; Massagué, 2000; Massagué and Wotton, 2000). Enhanced expression of corepressors in tumour cells could selectively block the TGF-β response and its growth-inhibitory effect in tumour cells and thus contribute to cancer progression.

The physical interactions and functional cooperativity of Smads with other transcription factors allows for cross-talk with other signalling pathways (Zhang and Derynck, 1999; Massagué, 2000; ten Dijke *et al*., 2000) (**Figure 5**). For example, UV radiation, stress and mitogens activate MAP kinase pathways that activate Jun N-terminal kinase, which phosphorylates and activates c-Jun. Activation of the c-Jun–c-Fos complex by MAP kinase–JNK cascades is therefore likely to regulate TGF-β-induced transcription from promoters with TGF-β-responsive AP-1- and Smad-binding sites. These promoters thus represent targets of convergence for these two types of signalling pathways. TGF-β signalling can also converge with Wnt signalling. This pathway is mediated by β-catenin, which shuttles between the cytoplasm and the nucleus, and serves as coactivator for the nuclear LEF/TCF transcription factors. Smad4 forms a complex with β-catenin and LEF, while Smad3 associates with LEF/TCF, and both Smads can cooperate with these transcription factors at promoters with LEF/TCF and Smad binding sites. Finally, Smad signalling can also converge with STAT signalling. STATs are activated by receptors and translocate into the nucleus to act as DNA binding transcription factors. BMP-activated Smad1 has been shown to form a complex with STAT3, which was activated in response to LIF. This interaction required the presence of the coactivator p300, with which both transcription factors interact. This complex formation is presumably at the basis of the functional cooperativity between Smad and STAT signalling.

The cross-talk of TGF-β/Smad signalling with other pathways can also occur prior to nuclear translocation of

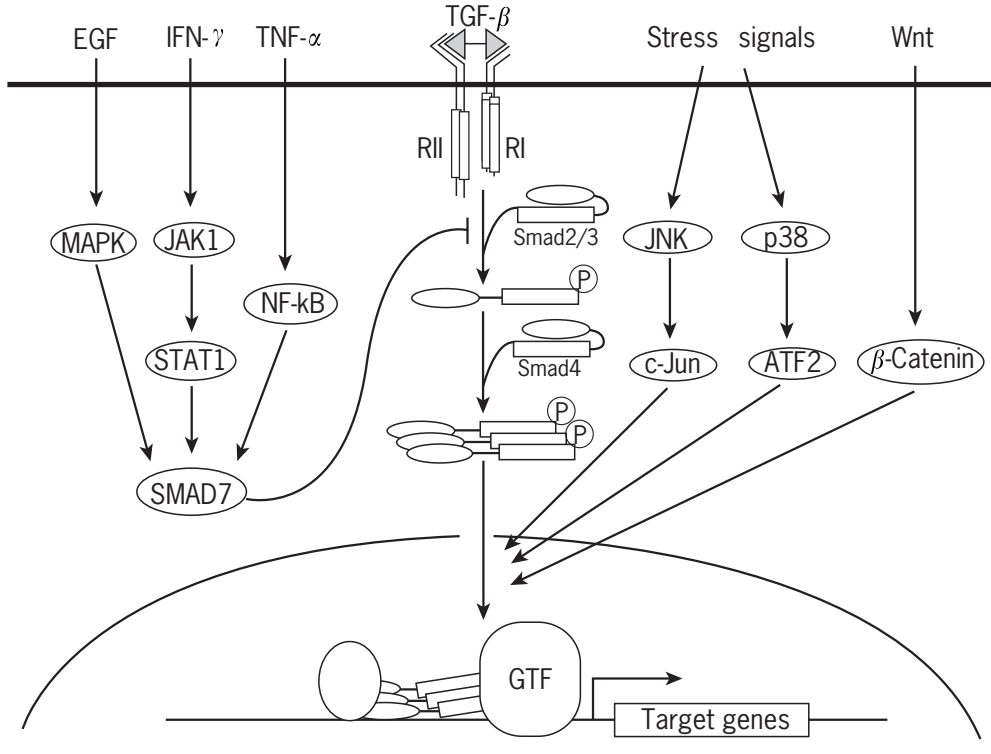


Figure 5 Cross-talk with TGF-β-induced signalling through Smads. A variety of extracellular signals activate JNK and/or p38 MAP kinase signalling, which leads to phosphorylation of c-Jun or ATF-2, two transcription factors, with which Smads can cooperate at defined promoters. Wnt signalling also cross-talks with Smad signalling, through the ability of activated Smads to associate with β-catenin and/or LEF/TCF at some promoters. Activation of MAP kinase signalling by growth factors, such as EGF, and activation of STAT1 by interferon-γ and of NF-κB by TNF-α all induce Smad7 expression, thus leading to inhibition of TGF-β-induced Smad2/3 activation by Smad7.

Smads (Piek *et al*., 1999; Zhang *et al*., 1999; Massagué, 2000; ten Dijke *et al*., 2000). For example, the linker regions of Smad1, -2 and -3 can be phosphorylated by Erk MAP kinase, which is activated in response to receptor tyrosine kinase and Ras signalling, and this phosphorylation then regulates ligand-induced nuclear translocation (Piek *et al*., 1999). Some signalling pathways activate expression of an inhibitory Smad, thus leading to decreased Smad activation. While BMPs and TGF-β induce the expression of the inhibitory Smad6 and -7, thus activating autoregulatory negative feedback loops, EGF can also induce Smad6 and -7 expression. This observation suggests that receptor tyrosine kinases regulate Smad signalling through induction of an inhibitory Smad, although other observations suggest Smad activation in response to receptor tyrosine kinase activation. Finally, interferon-γ-induced signalling through STATs and TNF-α-induced activation of NF-κB also activate Smad7 expression, which in turn inhibits TGF-β/Smad signalling. Thus, upregulation of Smad7 may represent a convenient mechanism used by several signalling pathways to inhibit TGF-β responsiveness (Piek *et al*., 1999; Zhang and Derynck, 1999; Itoh *et al*., 2000; Massagué, 2000; ten Dijke *et al*., 2000).

The mechanisms of Smad signalling and regulation, outlined above, now also explain the versatility and context dependence of the TGF-β responses. Indeed, the levels of interacting transcription factors and their activation state, as regulated by other signalling pathways, are important determinants of the TGF-β response. For example, TGF-β-induced transcription resulting from interactions of Smads with AP-1 transcription factors are likely to be regulated by MAP kinase signalling (Piek *et al*., 1999; Zhang and Derynck, 1999; Itoh *et al*., 2000; Massagué, 2000; ten Dijke *et al*., 2000). In contrast, the interaction of the heteromeric Smad2–Smad3–Smad4 complex with Sp1 induces expression of the cdk-inhibitors $p15^{Ink4B}$ or $p21^{Cip1}$ in a manner that is unlikely to require MAP kinase signalling (Feng *et al*., 2000; Pardali *et al*., 2000). Finally, the presence and identity of corepressors may determine the amplitude of the Smad response and even whether TGF-β receptor activation results in transcriptional activation or repression.

Finally, recent evidence strongly suggests that TGF-β signals through other pathways, distinct from the Smad-mediated regulation of gene expression (Piek *et al*., 1999; Itoh *et al*., 2000; Massagué, 2000) (**Figure 6**). TGF-β is able to activate MAP kinase signalling, although the extent and kinetics of activation differ substantially among different cell lines and types. These activation events have been shown to result in activation of Erk MAP kinase, p38 MAP kinase and Jun N-terminal kinase (JNK). p38 MAP kinase and JNK activation occurs through phosphorylation by MKK3, MKK4 and/or MKK6. Activation of p38 MAP kinase and JNK by TGF-β may enhance Smad signalling through Smad phosphorylation or phosphorylation of c-Jun and ATF-2, transcription factors that cooperate with TGF-β Smads, thus resulting in cross-talk with Smad-mediated transcription. TAK1, a MAP kinase, which is rapidly activated by TGF-β, yet is also involved in other unrelated signalling pathways, may serve as a TGF-β-dependent initiator of these signalling cascades. TGF-β may also activate or stabilize the small GTPases RhoA and RhoB and these may in turn play roles in several TGF-β responses, e.g. a requirement of RhoB for JNK activation (Itoh *et al*., 2000; Massagué, 2000). Finally, TGF-β also induces an interaction of protein phosphatase 2A with p70/S6 kinase, a kinase known to regulate protein translation and growth control, thus decreasing its activity (Petritsch *et al*., 2000). While the mechanisms of activation by TGF-β and roles of these non-Smad signalling cascades remain to be characterized, these observations strongly suggest that inactivation of Smad pathways may not leave the cell unresponsive to TGF-β.

INACTIVATION OF TGF-β's TUMOUR-SUPPRESSOR FUNCTIONS IN CARCINOMAS

The growth-inhibitory response of epithelial cells to TGF-β strongly suggests that TGF-β signalling may exert tumour suppression. On the other hand, TGF-β1 expression is often upregulated in carcinomas, which then would suggest that tumour cells benefit from TGF-β expression. Accordingly, there is substantial evidence for both tumour promoting and tumour-suppressor roles of TGF-β in carcinoma development. The tumour-suppressor role of TGF-β signalling is best supported by the presence of inactivating mutations in TGF-β receptors and Smads in human carcinomas and by tumour development studies in mouse models.

Somatic mutations in TβRII are common in tumours from patients with hereditary nonpolyposis colorectal cancer (HNPCC), who have germ-line defects in their capacity for DNA mismatch repair (de Caestecker *et al*., 2000; Massagué *et al*., 2000) (**Figure 7**). Nucleotide additions to or deletions from the stretch of adenines within the *TβRII* coding sequence give rise to a truncated TβRII, which is incapable of signalling. Consequently, the cells acquire a selective growth advantage, allowing them to progress for tumour development. While TβRII mutations occur frequently in colon cancers, gastric cancers and gliomas with microsatellite instability, they are less common in tumours from the endometrium, pancreas, liver and breast with microsatellite instability. Missense and inactivating mutations in the kinase domain of TβRII have also been reported in colon cancers, which do not display microsatellite instability. Together, inactivating TβRII mutations may be present in 20–25% of all colon cancers.

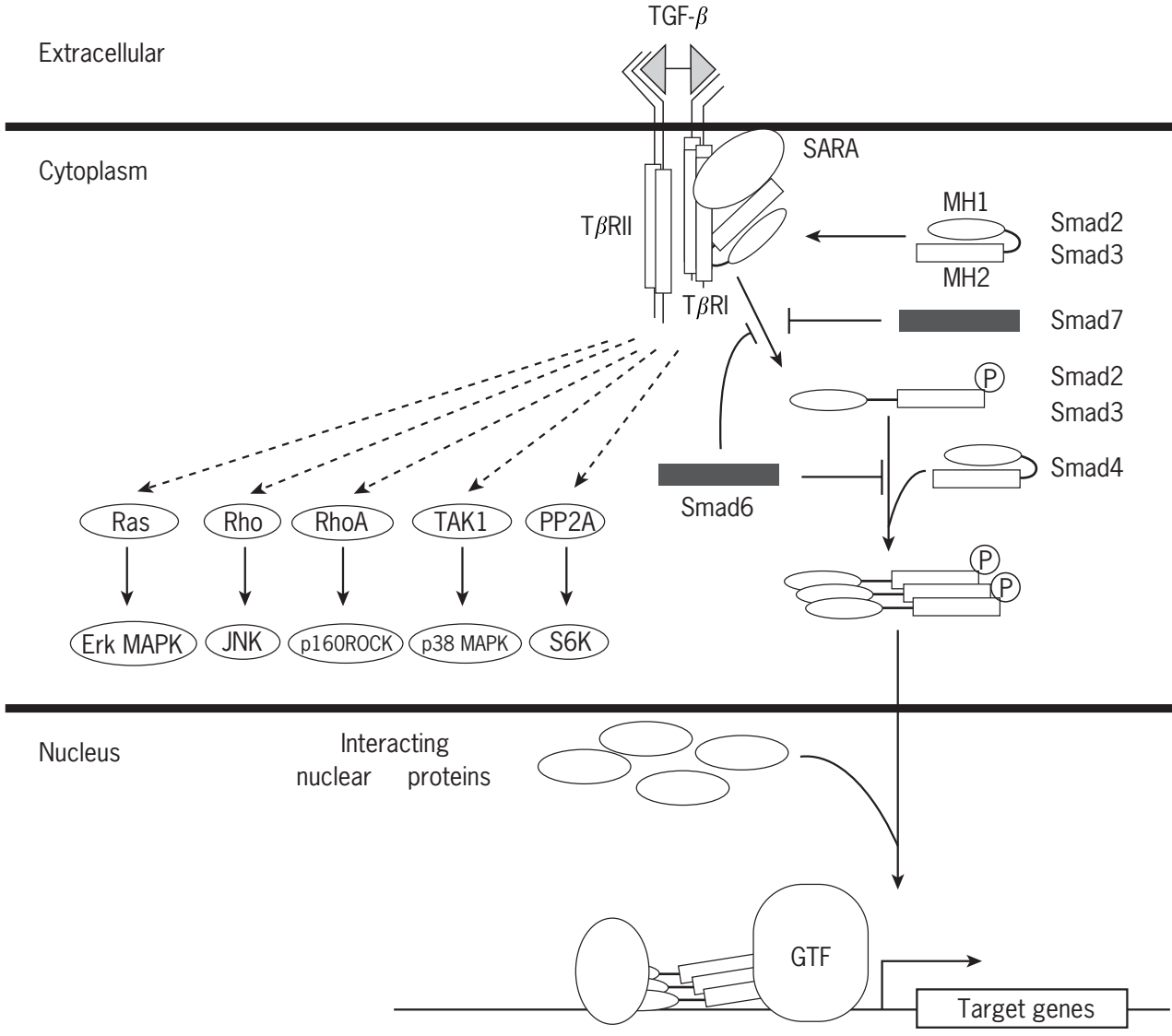


Figure 6 TGF-β-induced signalling through Smads is complemented by several non-Smad signalling mechanisms activated by the TGF-β receptors. Although the exact mechanism of activation remains to be better characterized, TGF-β induces activation of Ras, RhoB and RhoA, as well as TAK1 and the protein phosphatase 2A, thus leading to activation of several MAP kinase pathways, and down-regulation of S6 kinase activity.

Although less common, inactivating TβRI mutations also occur in ovarian cancers, metastatic breast cancers, pancreatic carcinomas and T cell lymphomas. Together, these mutations suggest a function of TβRII and TβRI as tumour suppressors in carcinoma development (de Caestecker *et al*., 2000; Massagué *et al*., 2000).

TGF-β receptor expression is often decreased in carcinomas and this down-regulation may allow cells to escape growth inhibition by TGF-β. Several Ets transcription factors regulate the expression of TβRII and reduced expression of these factors correlates with reduced receptor expression in gastric cancers. Transcriptional silencing may also result from hypermethylation of CpG islands in the *TβRI* or *TβRII* promoters, or mutations in the *TβRII* promoter which interfere with transcription factor binding. Decreased TβRII function has been shown to confer resistance against the growth-inhibitory effect of TGF-β, whereas other TGF-β responses, e.g. extracellular matrix protein expression, may not be similarly affected, and increased TβRII expression correlates with sensitivity to the growth inhibitory response of TGF-β. Thus, different signalling threshold requirements may be the basis for the observation that a decrease in TβRII function may primarily affect the growth responsiveness (de Caestecker *et al*., 2000; Massagué *et al*., 2000).

Expression of wild-type TβRII in colon or breast carcinoma cell lines, which lack a functional *TβRII* allele, provided evidence that the TβRII acts as a tumour

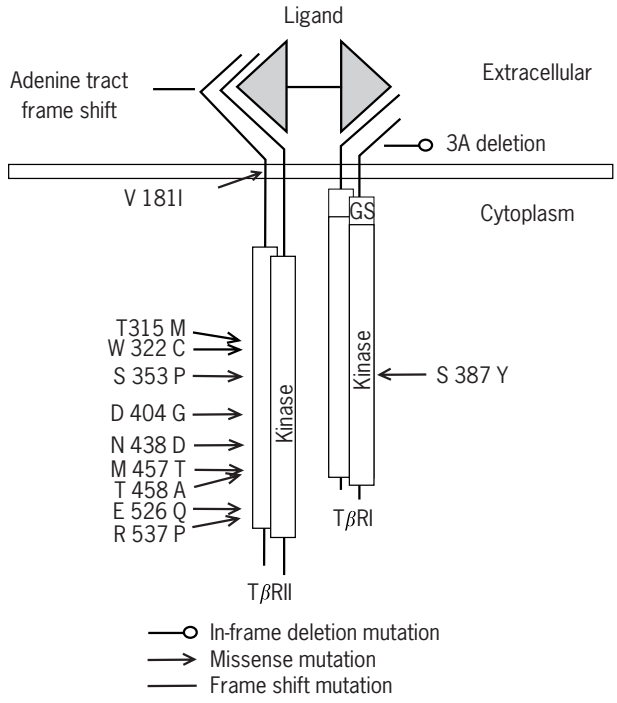


Figure 7 Summary of mutations in TβRII and TβRI found in various cancers.

suppressor (de Caestecker *et al*., 2000; Massagué *et al*., 2000). The TβRII-expressing cells were growth inhibited and had suppressed anchorage independence and strongly reduced tumour formation in nude mice, when compared with the parental cells. Transgenic expression of dominant negative forms of TβRII in the skin or mammary gland increased tumour formation, further supporting a tumour suppressor role of TβRII. In addition, mice with an inactivated *TGF-β1* gene also show increased carcinoma development, either spontaneously (Engle *et al*., 1999) or after carcinogen treatment (Tang *et al*., 1998). In the latter case, haploinsufficiency of the *TGF-β1* gene increased tumour susceptibility, since the tumours retained one wild-type allele. Consistent with these results, decreased TβRII expression correlated with high tumour grade of human breast cancers. While these observations suggest that loss of TGF-β responsiveness provides a distinct advantage for tumour development, most tumours do not have inactivated TGF-β receptors, and HNPCC patients, who frequently have TβRII mutations in their tumours, have a better prognosis than patients with sporadic colon cancer, who do not have TβRII mutations. Therefore, abrogation of TGF-β signalling, while leading to loss of the growth inhibition by TGF-β and early tumour onset, paradoxically protects against tumour progression, since the tumours do not adopt an invasive phenotype in response to autocrine or paracrine TGF-β (see below). This possibility is supported by mouse studies, in which a functional TβRII was expressed in colon cancer cells that lack TGF-β receptor expression. Although the transfected cells showed reduced growth rate, they had a strongly increased invasive and metastatic capacity (Oft *et al*., 1998). Together these findings illustrate the tumour-suppressor role of TGF-β receptors in carcinoma development, and the distinct advantages, provided by TGF-β responsiveness, for cancer progression at later stages.

Mutations in Smads have been found in a variety of carcinomas (**Figure 8**), and, even though generally uncommon, they suggest that some Smads act as tumour suppressors (de Caestecker *et al*., 2000; Massagué *et al*., 2000). While no mutations in Smad3 or the inhibitory Smad 6 or -7 have been reported, inactivation of Smad2 or -4 genes occurs by loss of chromosome segments, small deletions, frameshift, nonsense or missense mutations. Smad4 mutations occur primarily in pancreatic carcinomas, in which Smad4 was originally identified as *DPC4* (deleted in pancreatic carcinoma), and colon carcinomas, and less frequently in other carcinomas. While biallelic inactivation of Smad4 is most commonly observed, haploinsufficiency of the Smad4 locus may also contribute to cancer progression, although this interpretation may be confounded by the presence of the Smad2 and DCC loci in close proximity to the Smad4 gene. In contrast to Smad4, Smad2 mutations are rare and occur primarily in colorectal and lung carcinomas. Finally, enhanced Smad7 levels are observed in pancreatic carcinomas and may also decrease Smad responsiveness.

Tumour-associated mutations in Smad4 and Smad2 occur most frequently in the MH2 domain, which mediates heteromeric complex formation and transcriptional activation. C-terminal deletions or mutations often inactivate the Smad, and provide dominant negative interference with wild-type Smad function and TGF-β-induced nuclear translocation. Many mutations also map at the interfaces of Smad heteromerization, suggesting interference with heteromerization. Other mutations decrease the stability, e.g. through increased ubiquitin-mediated degradation of the Smad proteins. Finally, mutations in the MH1 domain of Smad4 interfere with its DNA binding. While most if not all mutations impair Smad functions, some mutations may alter TGF-β signalling to the tumour's advantage. Indeed, the Smad2 mutations *D450E* and *P445H*, found in colorectal carcinomas, enhance the invasive behaviour of the tumour cells, when overexpressed (de Caestecker *et al*., 2000; Massagué *et al*., 2000).

While homologous inactivation of the Smad2 or -4 genes results in embryonic lethality and therefore prevents an assessment in tumour progression, several mouse studies suggest a role of Smad4 as tumour suppressors. Mice heterozygous for an inactivated Smad4 gene develop intestinal polyps, which can progress into carcinomas. When combined with an inactivated allele of the adenomatous polyposis coli (*APC*) gene, this double heterozygosity allows the development of highly invasive colon carcinomas with both the *APC* and Smad4 genes

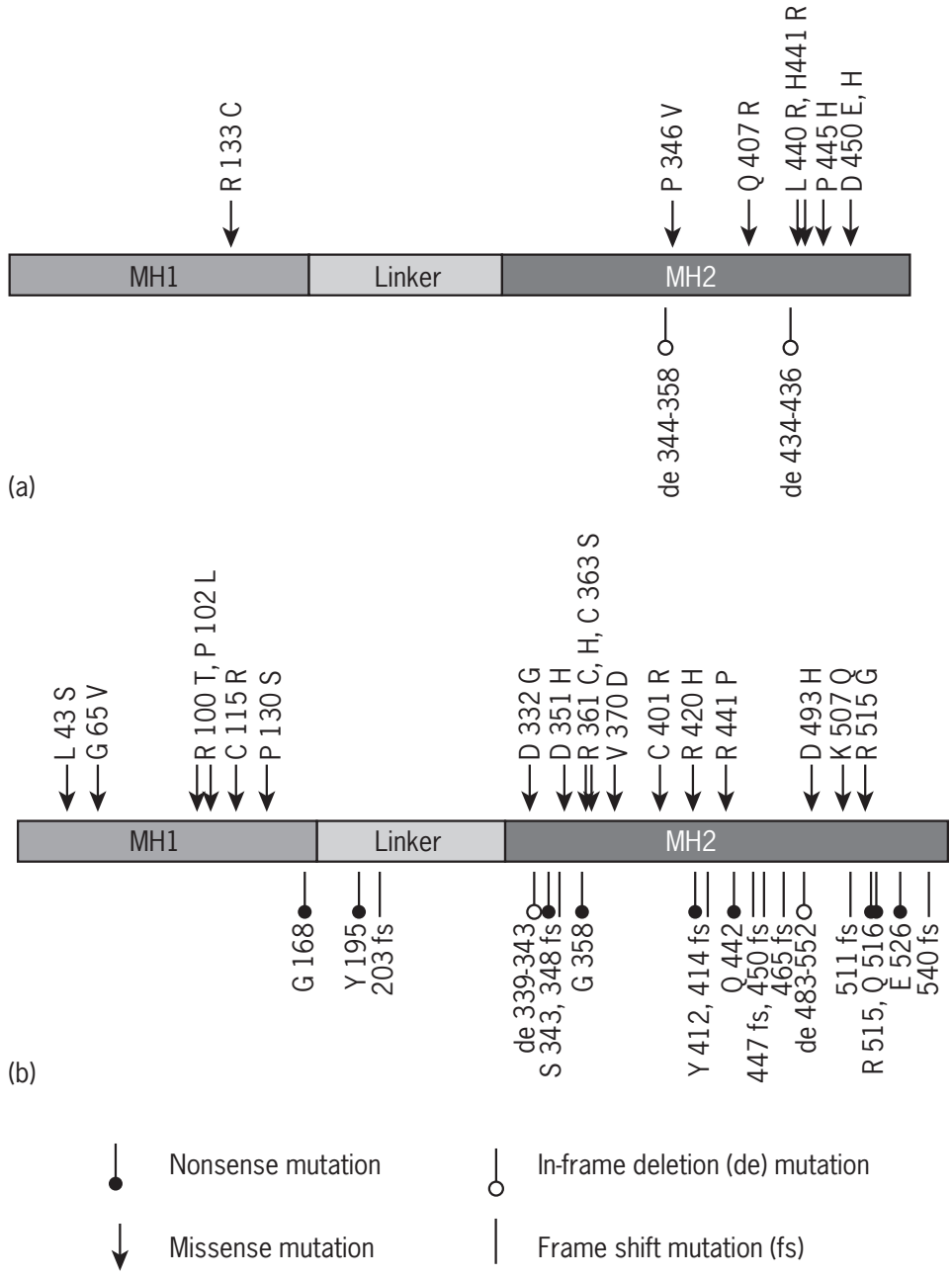


Figure 8 Summary of mutations in (a) Smad2 and (b) Smad4 found in various carcinomas.

inactivated (Takaku *et al*., 1998). Mice with homozygous inactivation of the Smad3 gene have also been shown to develop colon carcinomas (Zhu *et al*., 1998), although this phenotype was not seen in two other similar studies. While a role of Smad3 as tumour-suppressor gene is conceivable, no inactivating Smad3 mutations have been observed in human tumours. This could be rationalized by the observation that Smad3 activates most gene expression responses to TGF-β, whereas Smad2 cooperates with Smad3 for TGF-β-induced expression of the $p21^{Cip1}$ or $p15^{Ink4B}$ cdk inhibitors (Feng *et al*., 2000; Pardali *et al*., 2000). Smad2 inactivation may therefore inactivate the growth-arrest response, without affecting Smad3 responses, e.g. induction of extracellular matrix proteins, which provide an advantage for tumour development. Importantly, while Smad4 is required for many if not most TGF-β gene expression responses, Smad4-deficient cells may have retained a variety of responses (Sirard *et al*., 2000). Thus, inactivation of Smad4 or Smad2 function may not abolish, but rather perturb, the complex TGF-β response, and thereby primarily targets the growth inhibitory response to TGF-β. Moreover, different TGF-β responses may have differential sensitivity to Smad signalling. Accordingly, Smad7 overexpression in a colon

carcinoma cell line suppressed TGF-β-induced growth arrest, without affecting TGF-β-induced expression of plasminogen activator inhibitor 1, and enhanced anchorage-independent growth and tumorigenicity. Finally, inactivating Smad4 mutations have been found in conjunction with mutations in TβRII or TβRI, strongly suggesting that Smad4 has tumour-suppressing activities which are unrelated to TGF-β signalling.

Oncogene expression or increased proto-oncogene expression may also down-regulate the TGF-β responsiveness and allow for escape of TGF-β's tumour-suppressor effects. For example, *Evi-1*, an oncogene involved in leukaemic transformation of haematopoietic cells, inhibits Smad3 function and thus decreases or abolishes TGF-β responsiveness. The oncogenic effect of *Evi-1* may therefore result in part from interference with TGF-β signalling. Conceptually similarly, c-Ski represses Smad-mediated transcription, thus raising the possibility that its interference with TGF-β responsiveness may contribute to its oncogenic activity. Accordingly, c-Ski expression is often elevated in melanomas and carcinomas, and this has been correlated with decreased TGF-β responsiveness. Finally, c-Myc also inhibits the antiproliferative response to TGF-β. TGF-β downregulates c-Myc expression in epithelial cells, but this repression is lost in various cancer cell lines, concomitant with the loss of the growth inhibitory response to TGF-β (Massagué *et al*., 2000; Chen *et al*., 2001). Thus, downregulation of c-Myc by TGF-β may be required for the growth-inhibitory response to TGF-β, and the absence of this downregulation may confer resistance to the tumour-suppressor activity of TGF-β. Clearly, carcinoma cells have developed several strategies to escape the growth-inhibitory response of TGF-β.

CELL-AUTONOMOUS, STIMULATORY EFFECTS OF TGF-β ON TUMOUR DEVELOPMENT

Even though TGF-β signalling exerts tumour-suppressor activity, TGF-β1 expression is often increased in tumour cells and stimulates tumour development and cancer progression. Indeed, autocrine TGF-β signalling can induce morphological changes and invasive behaviour of the tumour cells, while increased TGF-β1 production and activation make the microenvironment more permissive for tumour progression (Akhurst and Balmain, 1999; Dumont and Arteaga, 2000).

Tumour metastasis depends on various factors, including the ability of tumour cells to migrate and invade the stroma and to migrate in and out of the blood vessels. Epithelial to mesenchymal differentiation of tumour cells plays an important role in this invasive phenotype. Fibroblastoid or 'spindle cell' tumours of epithelial origin have been characterized as highly malignant and invasive (Cui *et al*., 1996). TGF-β and TGF-β-related proteins have been implicated in epithelial to mesenchymal transdifferentiation both in normal development and in tumour progression. This phenotypic transition is characterized by extensive changes in expression of cell adhesion molecules and a switch from a cytoskeleton of mainly cytokeratin intermediate filaments to one predominantly composed of vimentin. Epithelial to mesenchymal differentiation in culture is thought to correlate with these cell changes that facilitate invasion and metastasis *in vivo*.

TGF-β has been shown to induce epithelial to mesenchymal transition in culture of normal and transformed breast epithelial cells, squamous carcinoma, ovarian adenosarcoma and melanoma cells. The phenotypic changes have been best characterized in NMuMG cells, in which TGF-β induces cell shape changes, down-regulates expression of E-cadherin, ZO-1, vinculin and keratin and induces expression of vimentin and N-cadherin, which has been shown to increase cell motility and scattering (Miettinen *et al*., 1994; Bhowmick *et al*., 2001). Although the mechanism of TGF-β-induced epithelial to mesenchymal differentiation is presumably complex, PI-3-kinase signalling and TGF-β-induced Smad and RhoA activation appear to play a role (Bhowmick *et al*., 2001). In addition, the morphological changes occur rapidly in response to TGF-β, and are reversible. *In vivo*, the effect of TGF-β on the spindle cell phenotype and invasive behaviour of carcinomas has been well documented for skin carcinomas in mice (Cui *et al*., 1996). Consistent with a tumour-suppressor role, increased expression of activated TGF-β1 reduced the hyperplastic response to the tumour promoter PMA and the number of papillomas after treatment with DMBA. On the other hand, increased TGF-β1 expression enhanced the malignant conversion of skin carcinomas in a carcinogenesis mouse model, and increased the incidence of spindle cell carcinomas (Cui *et al*., 1996). In another study, expression of a dominant negative TβRII prevented squamous carcinoma cells from undergoing mesenchymal differentiation in response to TGF-β both *in vitro* and *in vivo*. Consequently, the tumours had a differentiated epithelial phenotype and were less malignant and invasive than the parental cells, which developed fibroblastoid spindle cell carcinomas (Portella *et al*., 1998). A similar role of TGF-β signalling was also apparent in Ras-transformed mammary epithelial cells and a fibroblastoid colon carcinoma cell line (Oft *et al*., 1998). Expression of a dominant negative TβRII prevented TGF-β-induced epithelial to mesenchymal changes and reverted the colon carcinoma cells to an epithelial phenotype. In addition, blocking TβRII function suppressed invasion *in vitro* and the metastatic phenotype of this colon carcinoma cell line. Finally, restoration of TβRII signalling in HNPCC cells with a mutated TβRII rendered the cells invasive, in contrast to their normally, noninvasive phenotype.

TUMOUR PROGRESSION STIMULATED BY EFFECTS OF TGF-β ON THE TUMOUR ENVIRONMENT

The increased expression and activation of TGF-β1 by tumour cells also makes the microenvironment more conducive to tumour development. Increased TGF-β1 expression by tumour cells, presumably a result of activated Ras/MAP kinase signalling and signal amplification in response to TGF-β1 itself, enhances the TGF-β levels in the tumour microenvironment. The increased protease expression and plasmin generation by tumour cells and the TGF-β1-induced expression of collagenases and other proteases, such as the matrix metalloproteases MMP-2 and MMP-9, result most likely in TGF-β activation and degradation of the extracellular matrix with consequent release of stored TGF-β. Increased TGF-β1 production stimulates synthesis of extracellular matrix proteins and chemoattraction of fibroblasts. All these changes together result in a microenvironment that is conducive for tumour growth and invasion, and for angiogenesis (Akhurst and Balmain, 1999; Dumont and Arteaga, 2000).

Tumour angiogenesis is critical for tumour growth and invasion, since blood vessels are required to deliver nutrients and oxygen to the tumour cells and allow tumour cells to intravasate the blood system, leading to metastatic spread. TGF-β1 acts as a potent inducer of angiogenesis in several assays (Roberts *et al*., 1986), while mouse models defective in TGF-β signalling illustrate the important role of TGF-β1 in normal vascular development. For example, targetted inactivation of the *TGF-β1* or *TβRII* genes results in embryonic lethality due to defective vasculogenesis and angiogenesis (Dickson *et al*., 1995; Oshima *et al*., 1996), while angiogenesis-defective phenotypes are also apparent in mice with null mutations of the genes for *Alk-1* (Oh *et al*., 2000), a TGF-β type I receptor that is expressed in endothelial cells, or endoglin, a TGF-β type III receptor expressed by endothelial cells (Arthur *et al*., 2000). (See chapter on *Angiogenesis*.)

Several tumour models illustrate the importance of tumour cell-secreted TGF-β1 in tumour angiogenesis. Increased TGF-β1 secretion by transfected pancreas carcinoma (Stearns *et al*., 1999) or CHO (Ueki *et al*., 1992) cells enhanced tumour angiogenesis in immunodeficient mice, whereas local administration of neutralizing antibodies to TGF-β1 strongly reduced tumour angiogenesis (Ueki *et al*., 1992). TGF-β antibodies also reduced angiogenesis and tumorigenicity of a renal carcinoma cell line in T, NK and B cell-deficient mice. These cells did not have cell-autonomous responses to TGF-β since they lacked TβRII (Ananth *et al*., 1999). In humans, histological studies of breast tumours correlate high levels of TGF-β1 mRNA with high microvessel density, and each of these factors correlated with poor patient prognosis. Diagnostic studies on other carcinoma types correlate high tumour burden and circulating plasma levels of TGF-β1, and enhanced tumour angiogenesis and poor patient prognosis. In one study, TGF-β1 levels were also correlated with expression levels of the angiogenic growth factor VEGF (Saito *et al*., 1999).

The mechanisms whereby TGF-β1 stimulates angiogenesis remain to be further characterized, but presumably combine direct and indirect effects. TGF-β induces expression of VEGF, a potent angiogenic growth factor, which directly stimulates endothelial cell proliferation and migration. TGF-β also induces capillary formation of endothelial cells, cultured on collagen matrix (Choi and Ballermann, 1995). Indirect stimulation of angiogenesis may also occur, since TGF-β is a potent chemoattractant for monocytes, which release angiogenic cytokines, and the TGF-β1-induced changes in the microenvironment stimulate endothelial cell migration and capillary formation. Moreover, the TGF-β-induced expression of MMP-2 and MMP-9, and down-regulation of the protease inhibitor TIMP in both tumour cells and endothelial cells, are expected to enhance the migratory and invasive properties of endothelial cells required for angiogenesis. Thus, both direct effects of TGF-β and effects on the microenvironment stimulate tumour angiogenesis.

Local immunosuppression in response to increased TGF-β1 levels allows tumour cells to escape from immunosurveillance and thus stimulates tumour development and progression (Dumont and Arteaga, 2000). TGF-β1 is the most potent immunosuppressive cytokine known to date, and inhibits proliferation and functional differentiation of T lymphocytes, lymphokine-activated killer cells, natural killer cells, neutrophils, macrophages and B cells (Derynck and Choy, 1998; Letterio and Roberts, 1998; de Visser and Kast, 1999). Several findings illustrate the role of TGF-β-induced, local immunosuppression in tumorigenicity. For example, increased TGF-β1 expression in a tumour cell line prevented cytotoxic T lymphocyte activation and enhanced tumorigenicity, in contrast to parental cells. Such repression can be elaborated by the ability of TGF-β to inhibit the expression and function of interleukin 2 and interleukin 2 receptors (Letterio and Roberts, 1998; de Visser and Kast, 1999). Accordingly, mammary tumour cells, which produce high TGF-β1 levels, inhibited the cytotoxic T lymphocyte response, but this inhibition was overcome by expressing interleukin 2 in these cells. These and other results suggest that TGF-β-mediated suppression of the cytotoxic T cell response promotes tumour development.

TGF-β1 also inhibits other immune functions of relevance to tumour development. Increased TGF-β1 expression by tumour cells decreased natural killer cell activity and promoted tumour formation in nude mice that lack T cells. In addition, anti-TGF-β antibodies suppressed tumour formation and metastasis of a breast carcinoma cell line in nude mice, while enhancing natural killer cell function. This suppression was not seen in beige mice,

which lack natural killer cells, thus implicating TGF-β1-induced suppression of natural killer cells in cancer progression (Arteaga *et al*., 1993). TGF-β-mediated suppression of neutrophil function may also be involved in tumour progression. Indeed, Fas-ligand expressing carcinoma cells underwent neutrophil-mediated rejection, but this rejection did not occur at a site with high TGF-β levels or when TGF-β1 was injected at the tumour site (Chen *et al*., 1998). Finally, TGF-β down-regulates the expression of the major histocompatibility complex (MHC) class II antigens (Letterio *et al*., 1996), suggesting that TGF-β1 expression renders the tumour cells less immunogenic. This regulation contributes to the local immunosuppression and to the escape from immune surveillance.

In summary, the mechanisms through which TGF-β signals and exerts its multiple responses are rapidly being characterized. Smads exert multiple gene expression responses, but cross-talk with other signalling pathways and non-Smad TGF-β signalling provide further complexity to the TGF-β response. TGF-β signalling plays an important role in tumour cell behaviour and cancer progression. The growth-inhibitory activity of TGF-β acts as a tumour suppressor in carcinoma development, whose function is often eliminated through mutations in receptors or Smads. In contrast, the increased TGF-β1 expression by tumour cells stimulates tumour development, both in a cell-autonomous manner and through effects on its environment, e.g. stimulation of angiogenesis and localized immunosuppression.

REFERENCES

Akhurst, R. J. and Balmain, A. (1999). Genetic events and the role of TGF-β in epithelial tumour progression. *Journal of Pathology*, **187**, 82–90.

Ananth, S., *et al*. (1999). Transforming growth factor β1 is a target for the von Hippel–Lindau tumour suppressor and a critical growth factor for clear cell renal carcinoma. *Cancer Research*, **59**, 2210–2216.

Arteaga, C. L., *et al*. (1993). Anti-transforming growth factor (TGF)-β antibodies inhibit breast cancer cell tumorigenicity and increase mouse spleen natural killer cell activity. Implications for a possible role of tumor cell/host TGF-β interactions in human breast cancer progression. *Journal of Clinical Investigation*, **92**, 2569–2576.

Arthur, H. M., *et al*. (2000). Endoglin, an ancillary TGFβ receptor, is required for extraembryonic angiogenesis and plays a key role in heart development. *Developmental Biology*, **217**, 42–53.

Barcellos-Hoff, M.-H., *et al*. (1994). Transforming growth factor-β activation in irradiated murine mammary gland. *Journal of Clinical Investigation*, **93**, 892–899.

Bhowmick, N. A., *et al*. (2001). Transforming growth factor-β1 mediates epithelial to mesenchymal transdifferentiation through a RhoA-dependent mechanism. *Molecular Biology of the Cell*, **12**, 27–36.

Chen, Y. G., *et al*. (1997). Mechanism of TGF-β receptor inhibition by FKBP12. *EMBO Journal*, **16**, 3866–3876.

Chen, J. J., *et al*. (1998). Regulation of the proinflammatory effects of Fas ligand (CD95L). *Science*, **282**, 1714–1717.

Chen, C. R., *et al*. (2001). Defective repression of c-myc in breast cancer cells: A loss at the core of the transforming growth factor β growth arrest program. *Proceedings of the National Academy of Sciences of the USA*, **98**, 992–999.

Choi, M. E. and Ballermann, B. J. (1995). Inhibition of capillary morphogenesis and associated apoptosis by dominant negative mutant transforming growth factor-β receptors. *Journal of Biological Chemistry*, **270**, 21144–21150.

Choy, L. and Derynck, R. (1998). The type II transforming growth factor-β (TGF-β) receptor-associated protein TRIP-1 acts as a modulator of the TGF-β response. *Journal of Biological Chemistry*, **273**, 31455–31462.

Cui, W., *et al*. (1996). TGFβ1 inhibits the formation of benign skin tumors, but enhances progression to invasive spindle carcinomas in transgenic mice. *Cell*, **86**, 531–542.

Datta, P. K. and Moses, H. L. (2000). STRAP and Smad7 synergize in the inhibition of transforming growth factor β signaling. *Molecular and Cell Biology*, **20**, 3157–3167.

de Caestecker, M. P., *et al*. (2000). Role of transforming growth factor-β signaling in cancer. *Journal of the National Cancer Institute*, **92**, 1388–1402.

Derynck, R. and Choy, L. (1998). Transforming growth factor-β and its receptors. In Thompson, A. (ed.), *The Cytokine Handbook*, 3rd edn 593–636 (Academic Press, Boston).

Derynck, R. and Feng, X.-H. (1997) TGF-β receptor signaling. *BBA Reviews in Cancer*, **1333**, F105–F150.

Derynck, R., *et al*. (1998). Smads: transcriptional activators of TGF-β responses. *Cell*, **95**, 737–740.

de Visser, K. E. and Kast, W. M. (1999). Effects of TGF-β on the immune system: implications for cancer immunotherapy. *Leukemia*, **13**, 1188–1199.

Dickson, M. C., *et al*. (1995). Defective haematopoiesis and vasculogenesis in transforming growth factor-β1 knock out. *Development*, **121**, 1845–1854.

Dumont, N. and Arteaga, C. L. (2000). Transforming growth factor-β and breast cancer: tumor promoting effects of transforming growth factor-β. *Breast Cancer Research*, **2**, 125–132.

Engle, S. J., *et al*. (1999). Transforming growth factor β1 suppresses nonmetastatic colon cancer at an early stage of tumorigenesis. *Cancer Research*, **59**, 3379–3386.

Ewen, M. E., *et al*. (1993). TGF β inhibition of Cdk4 synthesis is linked to cell cycle arrest. *Cell*, **74**, 1009–1020.

Feng, X.-H., *et al*. (2000). Smad2, Smad3 and Smad4 cooperate with Sp1 to induce p15Ink4B expression in response to TGF-β. *EMBO Journal*, **19**, 5178–5193.

Griswold-Prenner, I., *et al*. (1998). Physical and functional interactions between type I transforming growth factor-β receptors and Bα, a WD-40 repeat subunit of phosphatase 2A. *Molecular and Cell Biology*, **18**, 6595–6604.

Iavarone, A. and Massagué, J. (1997). Repression of the CDK activator Cdc25A and cell-cycle arrest by cytokine TGF-β in cells lacking the CDK inhibitor p15. *Nature*, **387**, 417–422.

Itoh, S., *et al*. (2000). Signaling of transforming growth factor-β family members through Smad proteins. *European Journal of Biochemistry*, **267**, 6954–6967.

Letterio, J. J. and Roberts, A. B. (1998). Regulation of immune responses by TGF-β. *Annual Review of Biochemistry*, **16**, 137–161.

Letterio, J. J., *et al*. (1996). Autoimmunity associated with TGF-β1-deficiency in mice is dependent on MHC class II antigen expression. *Journal of Clinical Investigation*, **98**, 2109–2119.

Massagué, J. (2000). How cells read TGF-β signals. *Nature Reviews of Molecular and Cellular Biology*, **1**, 169–178.

Massagué, J. and Wotton, D. (2000). Transcriptional control by the TGF-β/Smad signaling system. *EMBO Journal*, **19**, 1745–1754.

Massagué, J., *et al*. (2000). TGF-β signaling in growth control, cancer, and heritable disorders. *Cell*, **103**, 295–309.

Miettinen, P. J., *et al*. (1994). TGF-β induced transdifferentiation of mammary epithelial cells to mesenchymal cells: involvement of type I receptors. *Journal of Cell Biology*, **127**, 2021–2036.

Munger, J. S., *et al*. (1997). Latent transforming growth factor-β: structural features and mechanisms of activation. *Kidney International*, **51**, 1376–1382.

Munger, J. S., *et al*. (1999). The integrin αv β6 binds and activates latent TGF β1: a mechanism for regulating pulmonary inflammation and fibrosis. *Cell*, **96**, 319–328.

Oft, M., *et al*. (1998). TGF-β signaling is necessary for carcinoma cell invasiveness and metastasis. *Current Biology*, **19**, 1243–1252.

Oh, S. P., *et al*. (2000). Activin receptor-like kinase 1 modulates transforming growth factor-β1 signaling in the regulation of angiogenesis. *Proceedings of the National Academy of Sciences of the USA*, **97**, 2626–2631.

Oshima, M., *et al*. (1996). TGF-β receptor type II deficiency results in defects of yolk sac hematopoiesis and vasculogenesis. *Developmental Biology*, **179**, 297–302.

Pardali, K., *et al*. (2000). Role of Smad proteins and transcription factor Sp1 in p21(Waf1/Cip1) regulation by transforming growth factor-β. *Journal of Biological Chemistry*, **275**, 29244–29256.

Petritsch, C., *et al*. (2000). TGF-β inhibits p70 S6 kinase via protein phosphatase 2A to induce G1 arrest. *Genes and Development*, **14**, 3093–3101.

Piek, E., *et al*. (1999). Specificity, diversity, and regulation in TGF-β superfamily signaling. *FASEB Journal*, **13**, 2105–2124.

Portella, G., *et al*. (1998). Transforming growth factor β is essential for spindle cell conversion of mouse skin carcinoma *in vivo*: implications for tumor invasion. *Cell Growth and Differentiation*, **9**, 393–404.

Ribeiro, S. M., *et al*. (1999). The activation sequence of thrombospondin-1 interacts with the latency-associated peptide to regulate activation of latent transforming growth factor-β. *Journal of Biological Chemistry*, **274**, 13586–13593.

Roberts, A. B., *et al*. (1986). Transforming growth factor type β: rapid induction of fibrosis and angiogenesis *in vivo* and stimulation of collagen formation *in vitro*. *Proceedings of the National Academy of Sciences of the USA*, **83**, 4167–4171.

Saito, H., *et al*. (1999). The expression of transforming growth factor-β1 is significantly correlated with the expression of vascular endothelial growth factor and poor prognosis of patients with advanced gastric carcinoma. *Cancer*, **86**, 1455–1462.

Sherr, C. J. and Roberts, A. B. (1999). CDK inhibitors: positive and negative regulators of G1-phase progression. *Genes and Development*, **13**, 1501–1512.

Sirard, C., *et al*. (2000). Targeted disruption in murine cells reveals variable requirement for Smad4 in transforming growth factor β-related signaling. *Journal of Biological Chemistry*, **275**, 2063–2070.

Stearns, M. E., *et al*. (1999). Role of interleukin 10 and transforming growth factor β1 in the angiogenesis and metastasis of human prostate primary tumor lines from orthotopic implants in severe combined immunodeficiency mice. *Clinical Cancer Research*, **5**, 711–720.

Takaku, K., *et al*. (1998). Intestinal tumorigenesis in compound mutant mice of both Dpc4 (Smad4) and Apc genes. *Cell*, **92**, 645–656.

Tang, B., *et al*. (1998). Transforming growth factor-β1 is a new form of tumor suppressor with true haploid insufficiency. *Nature Medicine*, **4**, 802–807.

ten Dijke, P., *et al*. (2000). Signaling inputs converge on nuclear effectors in TGF-β signaling. *Trends in Biochemical Science*, **25**, 64–70.

Ueki, N., *et al*. (1992). Excessive production of transforming growth-factor β1 can play an important role in the development of tumorigenesis by its action for angiogenesis: validity of neutralizing antibodies to block tumor growth. *Biochimica Biophysica Acta*, **1137**, 189–196.

Yu, Q. and Stamenkovic, I. (2000). Cell surface-localized matrix metalloproteinase-9 proteolytically activates TGF-β and promotes tumor invasion and angiogenesis. *Genes and Development*, **14**, 163–176.

Zhang, Y. and Derynck, R. (1999). Regulation of Smad signaling by protein associations and signaling crosstalk. *Trends in Cell Biology*, **9**, 274–279.

Zhu, Y., *et al*. (1998). Smad3 mutant mice develop metastatic colorectal cancer. *Cell*, **94**, 703–714.

FURTHER READING

Chang, H., *et al*. (1999). Smad5 knockout mice die at mid-gestation due to multiple embryonic and extraembryonic defects. *Development*, **126**, 1631–1642.

Daopin, S., *et al*. (1992). Crystal structure of transforming growth factor-β2: an unusual fold for the superfamily. *Science*, **257**, 369–373.

Derynck, R., *et al*. (1988). A new type of transforming growth factor-β, TGF-β3. *EMBO Journal*, **7**, 3737–3743.

Hata, A., *et al*. (1998). TGF-β signaling and cancer: structural and functional consequences of mutations in Smads. *Molecular Medicine Today*, **4**, 257–262.

Kim, S. J., *et al*. (2000). Molecular mechanisms of inactivation of TGF-β receptors during carcinogenesis. *Cytokine Growth Factor Review*, **11**, 159–168.

Markowitz, S. D. and Roberts, A. B. (1996). Tumor suppressor activity of the TGF-β pathway in human cancers. *Cytokine Growth Factor Review*, **7**, 93–102.

Massagué, J. (1998) TGF-β signal transduction. *Annual Review of Biochemistry*, **67**, 753–791.

McAdam, A. J., *et al*. (1994). Transfection of transforming growth factor-β producing tumor EMT6 with interleukin-2 elicits tumor rejection and tumor reactive cytotoxic T-lymphocytes. *Journal of Immunotherapy with Emphasis on Tumor Immunology*, **15**, 155–164.

Miyazono, K., *et al*. (2000). TGF-β signaling by Smad proteins. *Advances in Immunology*, **75**, 115–157.

Torre-Amione, G., *et al*. (1990). A highly immunogenic tumor transfected with a murine transforming growth factor type β1cDNA escapes immune surveillance. *Proceedings of the National Academy of Sciences of the USA*, **87**, 1486–1490.

Wotton, D. and Massagué, J. (2001). Smad transcriptional corepressors in TGF β family signaling. *Current Topics in Microbiology and Immunology*, **254**, 145–164.

Chapter 14

Wnt Signal Transduction

Jeffrey R. Miller
University of Minnesota, Minneapolis, MN, USA

CONTENTS

INTRODUCTION

Whether it is a meeting with a colleague, a conversation over coffee with a friend or a telephone call to a loved one, interpersonal communication is an essential part of our daily lives. In our bodies, cells also communicate with one another to coordinate their behaviour and determine their specialized role in the body. Cells utilize elaborate systems of proteins to transmit, receive and respond to signals from neighbouring cells. These systems, or signal transduction pathways, utilize secreted signalling proteins, cell surface and intracellular receptor proteins, protein kinases, transcription factors and other intracellular proteins. Intercellular signals control a variety of processes in the body and ensure, for example, that each cell divides only when its neighbours dictate that it should do so. The importance of such signals becomes apparent when this communication breaks down and cells begin to divide uncontrollably, resulting in cancer. For example, colon cells normally divide at a rate that balances the loss of colon cells due to attrition. Colon cells receive signals that tell them to divide when more cells are needed and are told to stop dividing when the appropriate number of cells is reached. In colon cancer, like many other cancers, defects in this regulatory network cause colon cells to divide continuously leading to tumour formation.

The past 25 years of cancer research have revealed that cancer is a complex disease involving dynamic changes in the genome. This was realized through the discovery that mutations in specific genes, called oncogenes and tumour-suppressor genes, played critical roles in tumour formation. Mutations in these genes result in defects in regulatory pathways that control normal cell proliferation and homeostasis. Dominant gain-of-function mutations produce oncogenes, a gene that is locked in the 'ON' position and leads to hyperactivation of a regulatory pathway. Conversely, recessive mutations inactivate tumour-suppressor genes that normally keep a regulatory pathway 'OFF'. Coming back to the example of colon cancer, both types of mutations have been uncovered with the end result being the uncontrolled proliferation of colon cells. Thus, understanding the function of oncogenes and tumour-suppressors is vital to gain insights into the molecular causes of cancer and for designing therapeutic agents to treat cancer.

Through modern cloning techniques, a number of oncogenes and tumour suppressors have been identified. Characterization of these genes led to the realization that many oncogenes and tumour suppressors encode components of evolutionarily conserved signal transduction pathways important for controlling embryonic development. The Wnt signal transduction pathway provides one of the most striking examples of this connection. Inappropriate activation of the Wnt signal transduction pathway is implicated in a variety of human cancers, most notably colon cancer. Many cases of colon cancers are associated with either oncogenic mutations in β-catenin, a positive regulator of Wnt signal transduction, or inactivating mutations in APC (the tumour-suppressor protein encoded by the adenomatous polyposis coli gene), a negative regulator of Wnt signalling. In both cases, these mutations lead to the aberrant activation of the Wnt pathway and tumorigenesis. During development, Wnt signalling plays critical roles in controlling a variety of processes including cell fate determination and cell proliferation. For example, Wnt signalling is required for the patterning of the central nervous system and the establishment of the dorsal–ventral axis in frogs. The involvement of Wnt signalling in both embryonic development and cancer has fuelled an extraordinary explosion of interest in understanding the underlying molecular mechanism of Wnt signal transduction. This chapter summarizes the current model for Wnt signal transduction and then discusses how inappropriate activation of Wnt signalling causes cancer. Owing to constraints on space and an effort to present a simplified and coherent picture

of Wnt signalling, certain aspects of Wnt signalling will not be covered in this review. For more information, readers should refer to the list of further reading located at the end of this chapter. Additional information on Wnt signalling can also be found on the Wnt gene homepage (http://www.stanford.edu/~rnusse/wntwindow.html) and at the connections map at *Science's* STKE Web site (http://www.stke.org).

THE Wnt SIGNAL TRANSDUCTION PATHWAY

The Wnt signal transduction pathway is one of the major developmentally important signalling pathways with well-characterized roles in a variety of organisms including mice, frogs, fish, flies and worms. Like other intercellular signal transduction pathways, the Wnt pathway utilizes a secreted signalling protein a transmembrane receptor protein and complex intracellular machinery to relay signals from the cell surface to the nucleus. The major components of the Wnt signalling pathway include the Wnt family of secreted proteins, the Frizzled family of transmembrane receptor proteins and the intracellular proteins Casein Kinase Iε (CKIε), Dishevelled, GBP/Frat, Glycogen Synthase Kinase 3 (GSK3), APC, Axin, β-catenin (Armadillo) and the TCF/LEF family of transcriptional regulators. **Figure 1** presents a schematic representation of each of these proteins showing their important structural and functional domains.

The first insights into the mechanism of Wnt signal transduction came from pioneering studies in the fruit fly, *Drosophila melanogaster*. Researchers used the awesome power of genetics to characterize several fly genes with mutant phenotypes similar to that seen in embryos lacking *wingless*, the fly ortholog of vertebrate *Wnt-1*. These genes were then ordered into a genetic pathway, which has served as a paradigm for understanding Wnt signal transduction in other model systems. In recent years, additional components of the pathway have been identified through a variety of methods, including protein–protein interaction screens. Now, researchers are working to put the pieces of the Wnt signalling puzzle together, a task that is proving to be challenging. However, recent studies examining the complex biochemical relationships between components of the Wnt pathway have provided exciting new insights into the mechanism of Wnt signalling and provides the working model for Wnt signal transduction shown in **Figure 2** and described below.

Wnt signalling is dependent on the presence or absence of the intracellular protein β-catenin. In the absence of Wnt signal, the destruction complex, a multiprotein machine composed of Axin, APC and GSK3, tags β-catenin for destruction by the addition of phosphate groups to serine and threonine residues near the N-terminus of β-catenin. These phosphoamino acids then act as a binding site for a second complex of proteins, the ubiquitin ligase complex, which covalently adds a small protein called ubiquitin to β-catenin. Proteins tagged with ubiquitin are targetted to the proteasome, the cell's protein incinerator, where they are rapidly destroyed. This process serves to keep the levels of β-catenin in the cell low and the Wnt pathway is OFF. When cells perceive a Wnt signal, a group of proteins including Dishevelled, CKIε and GBP/Frat are activated and together they inactivate the destruction complex. As a result, β-catenin is no longer ubiquitinated and is protected from degradation by the proteasome. These events lead to the accumulation of β-catenin in the cell. As the level of β-catenin rises, it enters the nucleus and interacts with a DNA-binding protein of the TCF/LEF family. Together, β-catenin and TCF/LEF activate expression of specific cassettes of target genes. During development these targets include genes that direct cells to adopt specific cell fates, whereas in human colon cancer cells these targets include genes that control cell growth and proliferation. With this brief introduction in hand, the role that each of these proteins plays in regulating Wnt signalling will now be examined in greater detail.

The Messengers – *Wnt* Genes

W*nt* genes were first identified independently by researchers in two different fields. *Wnt-1* (first called *int-1*) was identified as a preferred integration site for mouse mammary tumour virus (Nusse and Varmus, 1982). Insertion of the mouse mammary tumour virus in regions surrounding the *Wnt-1* gene led to its inappropriate activation and breast cancer. *Wingless (wg)*, the fly counterpart of *Wnt-1*, was identified in the Nobel Prize-winning screen of Nüsslein-Volhard and Weischaus (1980) as a mutation that resulted in segment polarity defects. Cloning of *Wnt-1* and *wg* showed that these genes shared a high degree of sequence identity. This finding brought together researchers in the fields of cancer biology and developmental biology and greatly accelerated our understanding of Wnt signalling. Many additional members of the *Wnt* gene family have since been cloned in many organisms from nematode worms to humans, with each organism possessing multiple related *Wnt* genes. For example, the fruit fly possesses seven and the mouse at least 18 *Wnt* genes. A list of known *Wnt* genes and sequence comparisons can be found on the *Wnt* gene homepage (http://www.stanford.edu/~rnusse/wntwindow.html).

Wnt genes encode secreted glycoproteins typically 350–400 amino acids in length and are characterized by a conserved pattern of 23–24 cysteine residues (**Figure 1**). Wnt proteins appear to act as short-range messengers, acting within several cell diameters of the producing cell. Although Wnt proteins are secreted and can be found associated with the cell surface and extracellular matrix,

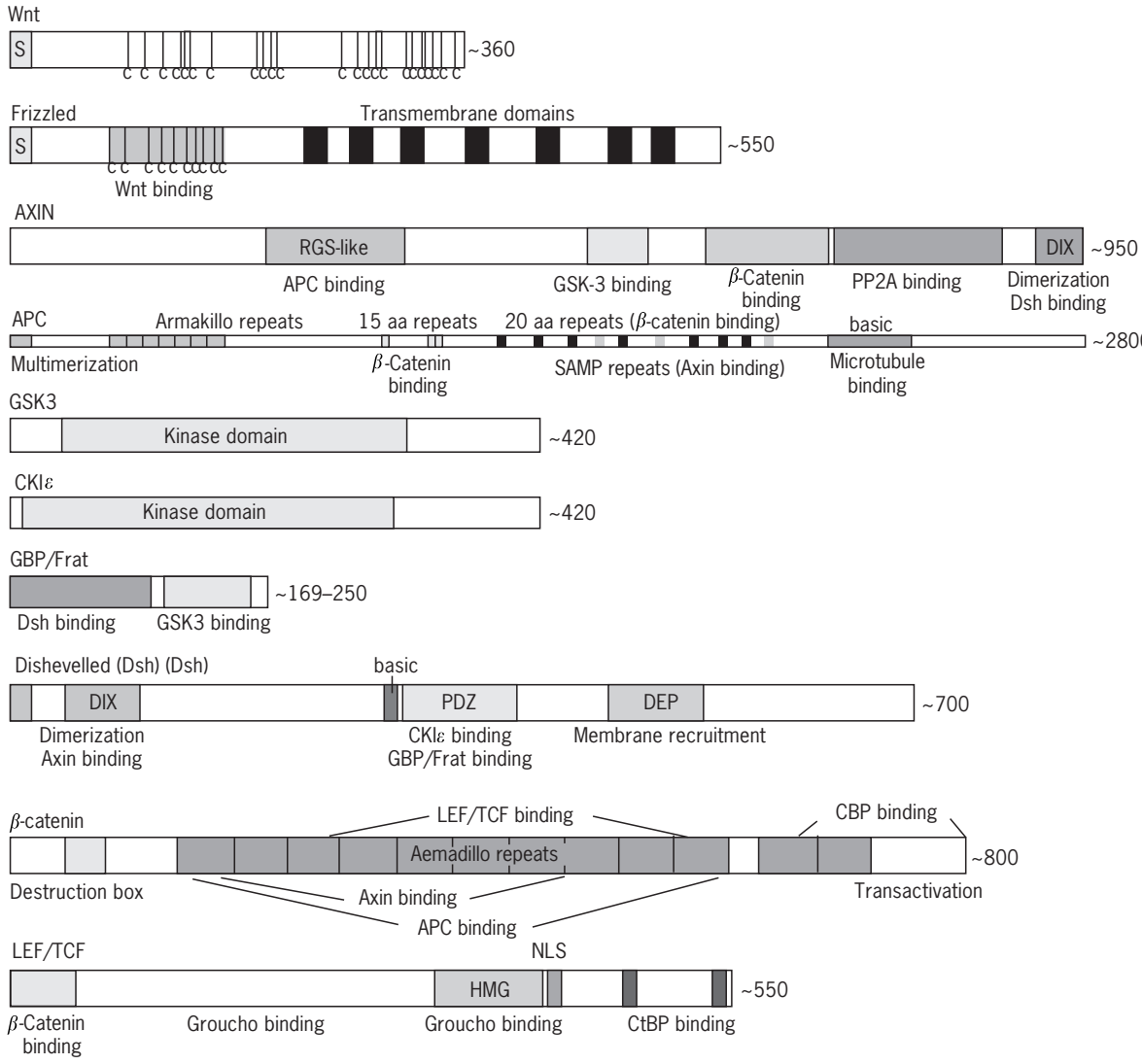


Figure 1 Schematic representation of the major components of the Wnt pathway showing their important structural and functional characteristics. For Wnt and Frizzled: (s) indicates signal sequence and (c) indicates conserved cysteine residues. Axin possesses an RGS domain, found in a family of proteins that regulate signalling via heterotrimeric G-proteins, and a DIX domain, found in both Axin and Dishevelled. Regions implicated in the binding of APC, GSK3, β-catenin, Protein Phosphatase 2A (PP2A) and Dishevelled (Dsh) are shown. APC is very large protein comprised of the following domains: a multimerization domain; seven Armadillo repeats important for localization; a series of β-catenin binding sites (15 amino acid repeats and 20 amino acid repeats), a series of SAMP repeats important for Axin binding; and a basic microtubule binding domain. Both GSK3 and CKIε are serine–threonine protein kinases and the kinase domain is indicated. GBP/Frat is a small protein with an N-terminal Dishevelled binding domain and a C-terminal GSK3 binding domain. Dishevelled possesses the following conserved domains: an N-terminal DIX domain important for dimerization and Axin binding, a centrally located PDZ domain that binds GBP/Frat and CKIε and a C-terminal DEP domain important for membrane recruitment and planar cell polarity signalling. β-Catenin possesses the N-terminal destruction box containing four GSK3 phosphorylation sites and a series of Armadillo repeats. Binding sites for Axin, APC and LEF/TCF have been mapped to the Armadillo repeat region. The N-terminal domain is important for transactivation. TCF/LEF contains a nuclear localization sequence (NLS) and an HMG box that mediates DNA binding. Binding sites for β-catenin, Groucho, CtBP and CBP/p300 are indicated.

they are notoriously insoluble and troublesome to work with biochemically. The difficulty to acquire soluble forms of Wnt proteins has hindered progress in understanding how cells send and receive Wnt signals. However, several forms of Wnt protein have been recovered from the medium of cultured cells and, using *in vitro* assays for activity, these soluble forms have been shown to be biologically active. This work will, it is hoped, provide the tools for

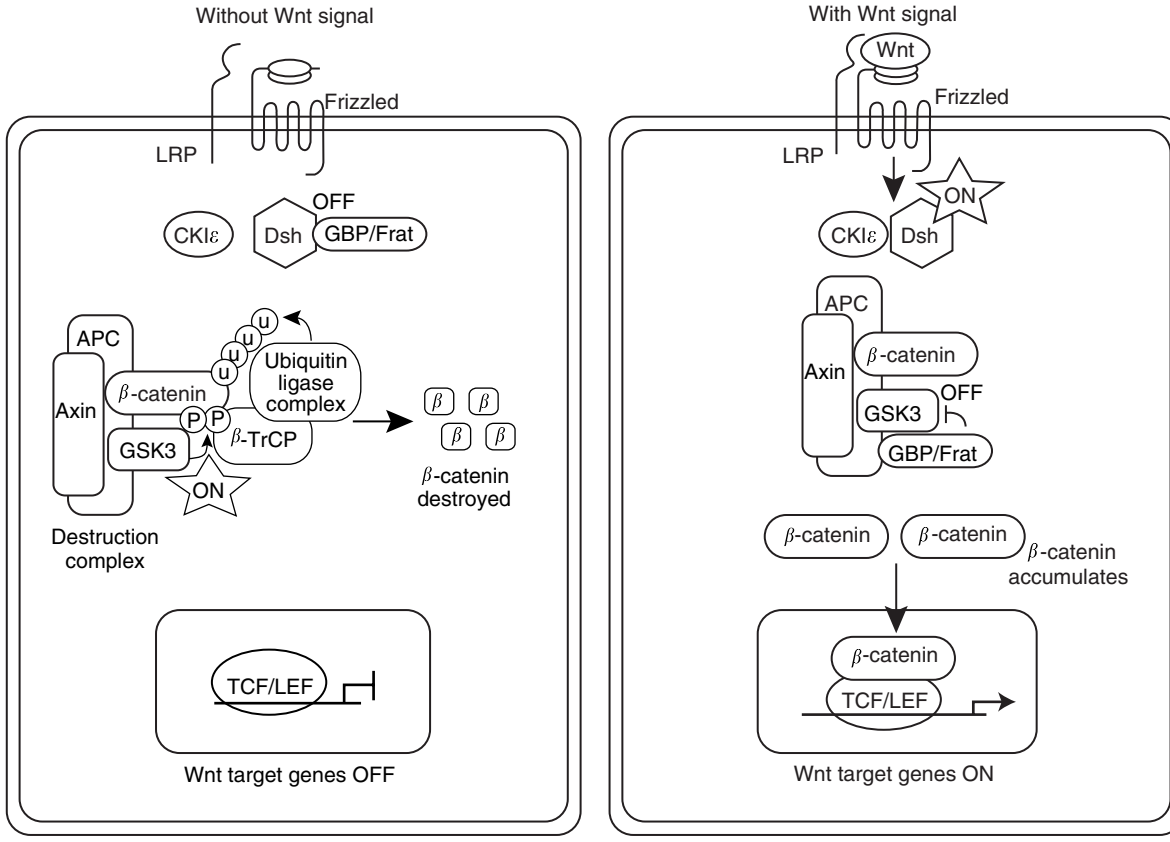


Figure 2 The Wnt signalling pathway. Left: in the absence of Wnt signal, Axin and APC facilitate the addition of phosphate groups to β-catenin by GSK3. Phosphorylated β-catenin binds to β-TrCP and is modified by the addition of a polyubiquitin tag. Proteins tagged with ubiquitin are degraded by the proteasome. The pathway is OFF because β-catenin is rapidly destroyed and its levels in the cell are low. Right: binding of Wnt to cell surface Frizzled and LRP receptors 'activates' Dishevelled (Dsh), CKIε and GBP/Frat by an unknown mechanism. Activation leads to inhibition of the destruction complex and a decrease in the phosphorylation of β-catenin. Thus, β-catenin evades the proteasome and accumulates in the nucleus where it interacts with a DNA-binding protein of the LEF/TCF family. Together, they activate expression of Wnt target genes and the pathway is ON.

understanding how Wnt proteins interact with components of the extracellular matrix and cell surface receptors.

At the Cell Surface – The Reception of Wnt Signals

For many years the identity of Wnt receptors remained elusive, leaving a large gap in our understanding of Wnt signalling. This hole has now been filled with the finding that members of the *Frizzled* gene family can function as Wnt receptors (**Figures 1** and **3**; Bhanot *et al*., 1996; Yang-Snyder *et al*., 1996). *Frizzled* genes encode seven-transmembrane proteins with an N-terminal cysteine-rich domain that binds Wnts with high affinity. Like the *Wnt* gene family, members of the *Frizzled* gene family have been identified in a number of organisms from worms to humans with each organism possessing multiple *Frizzled* genes. A list of known *Frizzled* genes and sequence alignments can be found on the *Wnt* gene homepage (http://www.leland.stanford.edu/~rnusse/wntwindow.html). Work in vertebrate systems has demonstrated that different Wnt proteins preferentially interact with and signal through specific subsets of Frizzled receptors. Given the large number of Wnt and Frizzled genes and, as a consequence, the large number of possible Wnt–Frizzled combinations, deciphering the importance of specific Wnt–Frizzled interactions *in vivo* will be challenging.

Very little is known about how Frizzled proteins function. Structurally, Frizzled receptors resemble other seven-transmembrane receptor proteins that signal through heterotrimeric GTP-binding proteins (G-proteins). Recent evidence suggests that in vertebrates a subset of Frizzled receptors may signal through G-proteins to affect levels of intracellular Ca^{2+}. However, there is no biochemical evidence that Frizzled proteins directly bind G-proteins or that G-proteins are directly involved in promoting the stabilization of β-catenin. Thus, the mechanism of Frizzled

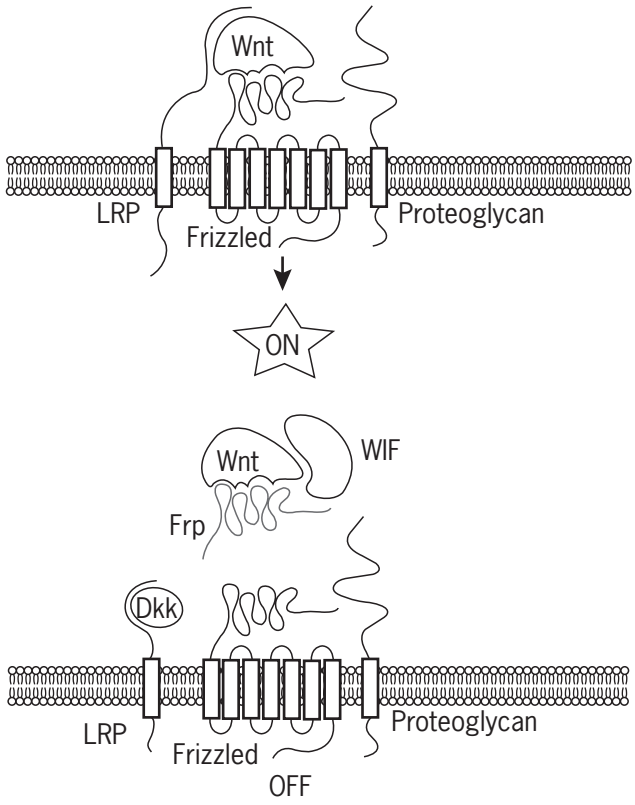


Figure 3 Wnt signalling at the cell surface. Top: Wnt proteins are secreted and bind Frizzled receptors with high affinity at a conserved domain in the N-terminal region. The mechanism by which Frizzled proteins transduce the Wnt signal across the membrane is unclear but may involve signalling through heterotrimeric G-proteins. Wnt proteins also interact with LRP co-receptors. Proteoglycans are also involved in the reception of Wnt signals through an unknown mechanism. Bottom: Wnt signalling is modulated extracellularly by a variety of secreted Wnt inhibitors including members of the secreted Frizzled-related proteins (FRPs), WIF-1 and Dickkopf (Dkk). Binding of these proteins to Wnt is thought to prevent the interaction of Wnt with Frizzled, thereby preventing activation of the pathway.

action and the potential importance of G-proteins in Wnt signalling await further experimentation.

Although compelling evidence exists that Frizzled proteins act as Wnt receptors, it is clear that other cell surface and extracellular molecules also play roles in the reception of Wnt signals (**Figure 3**). For example, members of the LDL-receptor-related family of transmembrane proteins (LRP 5/6) have recently been shown to act as Wnt co-receptors. Proteoglycans, extracellular or cell surface proteins that consist of a protein core and at least one glycosaminoglycan sugar side chain, have been shown to be important regulators of Wnt signalling. In addition, reception of Wnt signals is also modulated extracellularly by a diverse group of secreted Wnt inhibitors. At present, these Wnt inhibitors include members of the secreted Frizzled-Related Protein family (called FrzB or FRP), Wnt-inhibitory factor-1 (WIF-1), Cerberus and Dickkopf (Dkk). FRPs, WIF-1 and Cerberus can directly bind Wnt proteins and are thought to antagonize Wnt function by preventing their interaction with Frizzled receptors. FRP can also interact with Frizzled, suggesting that FRPs may also antagonize Wnt signalling through the formation of a nonfunctional complex with Frizzled receptors. Dkk functions by binding to LRP and preventing the interaction between LRP, Wnt, and Frizzled. Together, these data demonstrate that the reception of Wnt signals is complex, involving the interplay between Wnt protein, Frizzled and LRP cell surface receptors, proteoglycans and secreted Wnt inhibitors. Understanding the nature and importance of these interactions will require further biochemical analyses of the Wnt reception complex and how this complex transduces the Wnt signal from the outside to the inside of the cell.

From the Membrane to the Nucleus – Transduction of Wnt Signals inside the Cell

Wnt signalling is dependent on the levels of β-catenin in the cell. The pathway is OFF when levels of β-catenin are low and the pathway is ON when levels of β-catenin are high. This begs the question of how the levels of β-catenin are regulated in the cell. The answer to this question requires understanding the function of two competing groups of proteins. On one side, the destruction complex, GSK3, Axin and APC work to destroy β-catenin keeping levels low and the pathway OFF. On the other side, CKIε, Dishevelled and GBP/Frat are activated in response to Wnt and work to antagonize the destruction complex, increasing the levels of β-catenin and turning the pathway ON. The following sections examine the role of each of these groups of proteins in Wnt signalling in more detail.

The Destruction Complex: GSK3, Axin and APC

GSK3, a serine–threonine protein kinase, is the central player in the destruction complex (**Figures 1** and **4**). In the absence of Wnt signal, GSK3 is active and adds phosphate groups to four N-terminal sites of β-catenin (S33, S37, T41 and S45 in human β-catenin). These phosphoamino acids act as a tag on β-catenin, marking it for destruction by the proteasome. The amino acid sequence in β-catenin that is phosphorylated by GSK3 is called the 'destruction box' to denote its involvement in regulating the stability of β-catenin. Mutation of the GSK3 phosphorylation sites within the destruction box significantly diminishes the phosphorylation of β-catenin and results in highly stable forms of β-catenin with increased activity (Yost *et al*., 1996; Pai *et al*., 1997). The importance of the destruction box sequence is highlighted by

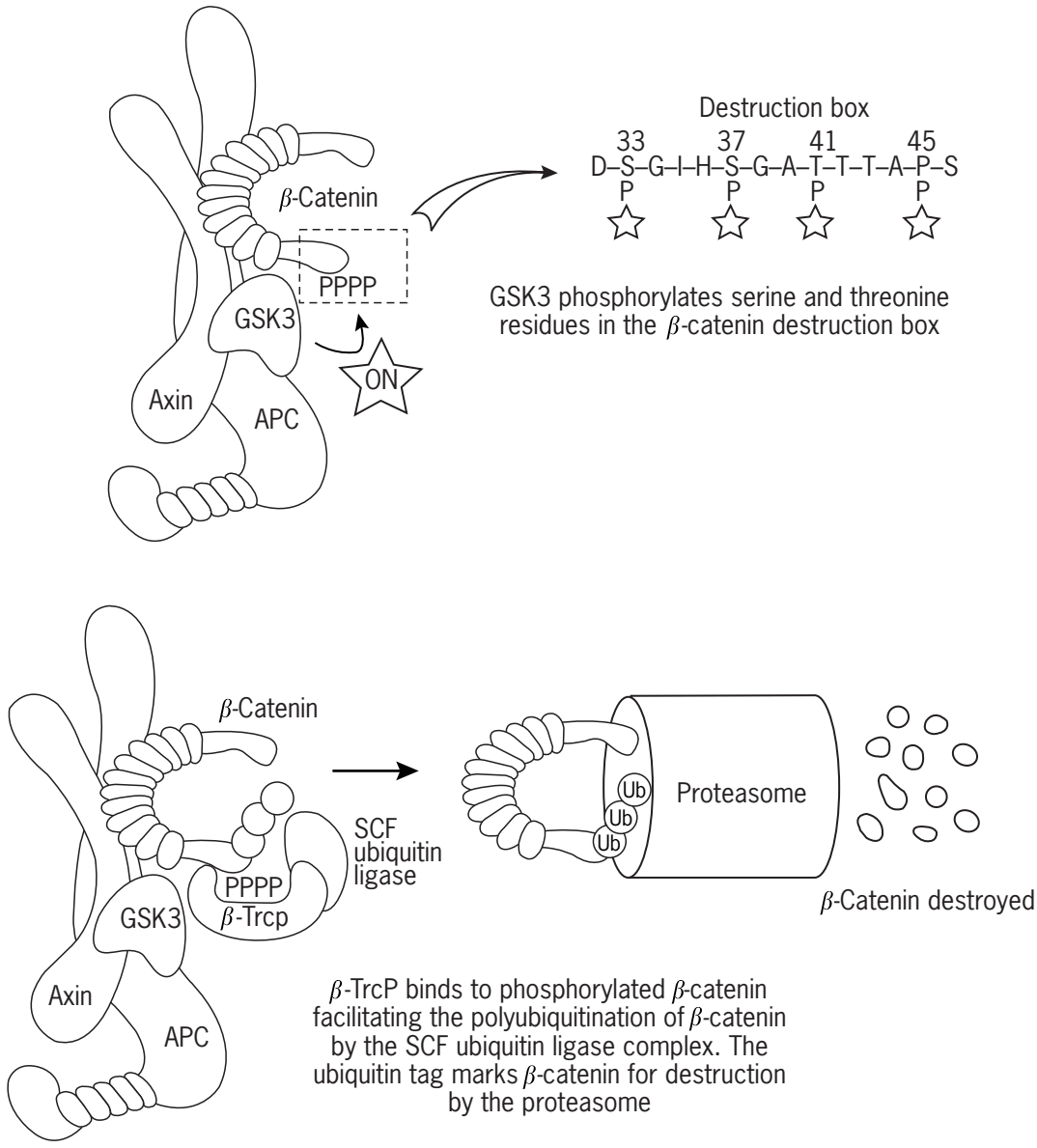


Figure 4 The destruction complex. In the absence of Wnt signal a large multiprotein machine, called the destruction complex, facilitates the rapid destruction of β-catenin. Top: two proteins, Axin and APC, act as scaffold proteins forming the underlying structure of the complex. APC and Axin bind themselves, each other and β-catenin. Additionally, Axin binds GSK3, Dishevelled and PP2A. The binding of GSK3 and β-catenin to Axin is critical for bringing GSK3 in close proximity to β-catenin and stimulating the GSK3-mediated phosphorylation of β-catenin at serine and threonine residues in the destruction box. Bottom: phosphorylated β-catenin is then bound by β-TrCP, which promotes the addition of a polyubiquitin tag to β-catenin. This ubiquitin tag marks β-catenin for rapid destruction by the proteosome, the cell's protein incinerator.

the recent finding that residues within the destruction box are often mutated in human cancers (**Figures 3** and **7**, discussed in detail below).

Axin is a second key component of the destruction complex (**Figures 1** and **4**). Mice lacking functional Axin develop with defects in the patterning of the dorsal–ventral axis, a phenotype similar to that seen following ectopic activation of the Wnt pathway in frogs (Zeng *et al*., 1997). In addition, over-expression of Axin can inhibit Wnt signalling and promote the degradation of β-catenin. Axin possesses multiple protein–protein interaction domains and appears to act as scaffold proteins, i.e. it serves as a building block for the construction of multiprotein complexes. Axin binds several components of the Wnt pathway including Dishevelled, APC, GSK3 and β-catenin. The binding of both GSK3 and β-catenin appears to be critical for the function of Axin as this interaction greatly enhances the phosphorylation of β-catenin by GSK3. Thus, Axin

appears to promote β-catenin degradation by bringing GSK3 and β-catenin into close proximity, thereby facilitating the phosphorylation of β-catenin by GSK3. Recently, mutations in Axin have been found in hepatocellular carcinomas and colon cancer, underscoring the importance of Axin in regulating the activity of the Wnt signalling pathway (Satoh *et al*., 2000; Webster *et al*., 2000).

APC is the third critical component of the destruction complex (**Figures 1** and **4**). APC was originally identified as a tumour-suppressor protein and mutations in APC are found in >80% of all colorectal tumours. The *APC* gene encodes a large, multidomain protein that, like Axin, appears to function as a scaffold protein. APC binds to several components of the Wnt pathway, including Axin, GSK3 and β-catenin. The first clue into the function of APC came from the finding that colorectal adenocarcinoma cell lines harbouring mutations in the *APC* gene possess high levels of β-catenin. Expression of wild-type APC in these cells resulted in a dramatic reduction in β-catenin levels, suggesting that APC is negative regulator of β-catenin stability (Munemitsu *et al*., 1995). How does APC promote the degradation of β-catenin? Current models predict APC functions in a similar fashion to Axin, stimulating GSK3-mediated phosphorylation of β-catenin. An alternative idea is that APC may function to localize the destruction complex to a specific location in the cell. In support of this hypothesis, mutations in APC that perturb its normal cortical location in the cell also perturb its ability to promote β-catenin degradation (McCartney *et al*., 1999). These two models are not mutually exclusive and further analysis of APC in a variety of systems should clarify the role APC plays in regulating β-catenin stability.

How does phosphorylation of β-catenin stimulate ubiquitination? The answer to this question came recently with the finding that β-TrCP/Slimb, a component of the SCF ubiquitin ligase complex, plays a critical role in regulating β-catenin degradation (**Figure 4**; Jiang and Struhl, 1998; Kitagawa *et al*., 1999). β-TrCP/Slimb specifically binds the phosphorylated destruction box of β-catenin, resulting in ubiquitination of β-catenin and subsequent proteolysis by the proteasome. What remains unclear, however, is how β-catenin is delivered to its final destination, the proteosome. Together, these data shed light on how mutations in APC, Axin and β-catenin lead to hyperactivation of the Wnt pathway and cancer. Recessive mutations in the tumour suppressors APC and Axin would lead to the inability of the destruction complex to target β-catenin for degradation in the absence of Wnt signals. Oncogenic mutations in the destruction box of β-catenin would prevent phosphorylation by GSK3 and/or the interaction of β-catenin with β-TrCP/Slimb. In each of these cases, β-catenin would evade proteasomal degradation and accumulate in the cell, leading to inappropriate activation of the pathway. This idea is supported by a number of studies showing that primary human tumour cells harbouring mutations in APC, Axin or β-catenin display elevated levels of β-catenin. This knowledge provides potential targets for clinical intervention and will be invaluable for the designing and testing new therapeutic agents for the treatment of cancer in humans.

Antagonizing the Destruction Complex: CKIε, Dishevelled and GBP/Frat

How does activation of the Wnt pathway promote the stabilization of β-catenin in the cell? The answer to this question hinges on understanding how Wnt signals, once transduced across the plasma membrane, act to antagonize the destruction complex. Until recently, our knowledge of the molecular events that occur upon Wnt stimulation was limited, but major gaps have been filled with the characterization of the biochemical relationships between intracellular components of the Wnt pathway.

Dishevelled is the most upstream intracellular component of the Wnt pathway and is 'activated' in response to Wnt signals (**Figures 1** and **5**). Dishevelled appears to function as a scaffold protein, acting through its association with other signalling proteins. Dishevelled possesses three highly conserved domains important for its function. The first is an N-terminal DIX domain that shares sequence identity to a C-terminal domain of Axin. The DIX domain has been shown to be important for homodimerization of Dishevelled and the binding of Axin. Second, is a centrally located PDZ domain. PDZ domains are present in a variety of proteins and serve as sites for protein–protein interactions. A number of proteins have been shown to bind Dishevelled through the PDZ domain including CKIε, GBP/Frat and Casein Kinase II. Both the DIX and PDZ domains are required for Dishevelled function in Wnt signalling. Third is the DEP domain found in the C-terminal third of Dishevelled. The DEP domain shares sequence similarity to the *Caenorhabditis elegans* gene *egl-10* and pleckstrin. Although the DEP domain is not required for the ability of Dishevelled to stabilize β-catenin, it is required for the function of Dishevelled in regulating cell polarity in flies and cell movements during gastrulation in vertebrates. It remains unresolved how Dishevelled transduces Wnt signals, but recent studies showing that Dishevelled interacts with Axin suggest that it may play a direct role in antagonizing the destruction complex.

Dishevelled function may also be dependent on its localization within the cell. In *Xenopus*, activation of the Wnt pathway is required for the establishment of dorsal cell fates. Examination of the localization of Dishevelled in early embryos revealed that it associates with small vesicle-like organelles (0.5–1.0 μm in diameter) that are enriched on the dorsal side of the embryo (Miller *et al*., 1999). This localization appears to be important since treatments that prevent dorsal development also prevent the dorsal enrichment of Dishevelled. Interestingly, time-lapse confocal microscopy analysis of Green Fluorescent Protein tagged Dishevelled (Dishevelled-GFP)

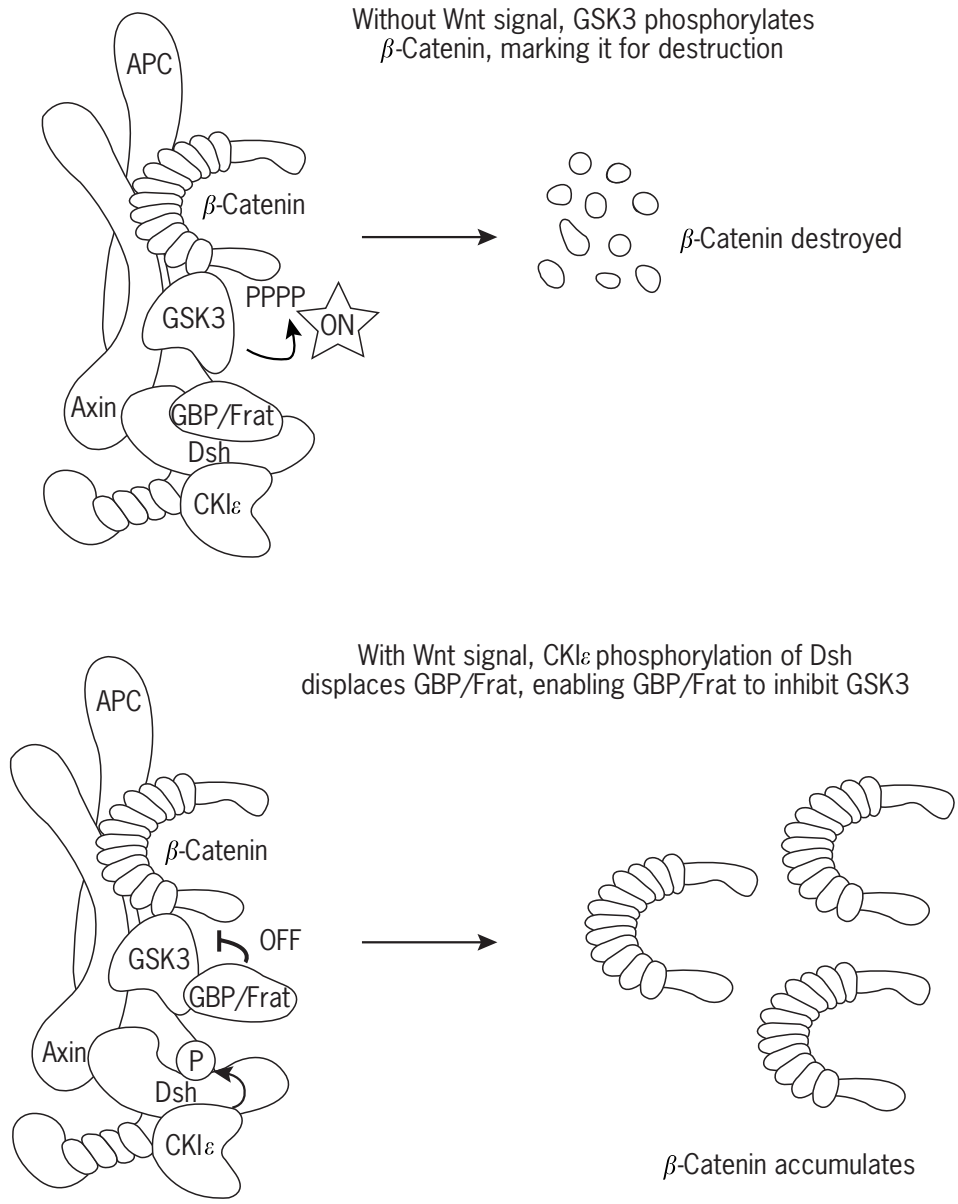


Figure 5 A current model of Wnt signalling and stabilization of β-catenin. Top: without Wnt signal, it is thought that Dishevelled might bind Axin and GBP/Frat, sequestering GBP/Frat from GSK3. Bottom: Wnt signal leads to antagonism of the destruction complex. One current model predicts that Dishevelled, through its interaction with Axin, might bring GBP/Frat to the destruction complex. Upon Wnt stimulation, CKIε might phosphorylate Dishevelled, displacing GBP/Frat, allowing it to bind and inhibit GSK3. Through this, or a similar mechanism, Wnt signal may also promote the dissolution of the destruction complex. Inhibition of GSK3 protects β-catenin from degradation and promotes the accumulation of β-catenin in the cell. This model is supported by recent studies but there are other possible mechanisms. Further characterization of the biochemical relationships between components of the destruction complex will help to resolve how Wnt signals stabilize β-catenin.

localization in early embryos demonstrated that Dishevelled-GFP associates with and is transported along the microtubule cytoskeleton towards the prospective dorsal side of the embryo. To view movies of Dishevelled transport in frog embryos, visit the *Journal of Cell Biology* web page at http://www.jcb.org/cgi/content/full/146/2/427/DC1. In addition, Dishevelled has also been shown to be associated with the actin cytoskeleton in embryonic kidney cells (Torres and Nelson, 2000). Together these data suggest that Dishevelled localization, perhaps through its association with the cytoskeleton, plays an important role in modulating the activity of the Wnt pathway.

GBP (GSK3 Binding Protein) and its mammalian orthologue Frat function as positive regulators of the Wnt signalling pathway (**Figures 1** and **5**; Yost *et al*., 1998). GBP/Frat can inhibit GSK3 kinase activity, suggesting that it promotes β-catenin stabilization through direct inhibition of GSK3-mediated phosphorylation of β-catenin. However, it is unclear whether this affect is due to a change in GSK3 activity or through steric blockade of GSK3–substrate interactions.

More recently, CKIε, a serine–threonine protein kinase, was identified as a positive regulator of the Wnt pathway (**Figures 1** and **5**; Peters *et al*., 1999; Sakanaka *et al*., 1999). Expression of CKIε stabilizes β-catenin and expression of dominant negative forms of CKIε antagonizes Wnt signalling. Overexpression studies have placed CKIε downstream of Dishevelled and upstream of GSK3. In addition, CKIε can bind to and phosphorylate Dishevelled, suggesting that CKIε could directly affect the activity of Dishevelled.

How do Dishevelled, GBP/Frat and CKIε antagonize the destruction complex? In the absence of Wnt signals, Dishevelled might associate with Axin and bind GBP/Frat, sequestering it from GSK3. Upon Wnt stimulation, CKIε might phosphorylate Dishevelled, causing the displacement of GBP/Frat. Released GBP/Frat would then bind GSK3 that is associated with Axin, thereby inhibiting GSK3 activity. This model is consistent with current data, but so are many others and many questions are unresolved. How do Frizzled receptors transduce Wnt signals? How are Dishevelled, GBP/Frat and CKIε activated in response to Wnt signals? Where in the cell is the destruction complex located and is this location important for signal transduction? These and other questions await further experimental analysis.

In the Nucleus – Regulation of Gene Expression by β-Catenin and TCF/LEF

As we have seen, cells have evolved a very elaborate and complex mechanism for regulating intracellular levels of β-catenin. Wnt signalling frees β-catenin from the destruction complex and it now accumulates in the cell. What happens next? It is now dogma that upon Wnt stimulation β-catenin accumulates and enters the nucleus, where it regulates gene expression. However, when first proposed the idea that the nucleus was the primary location of β-catenin function in Wnt signalling seemed hard to believe. At the time, β-catenin was known as a protein that localized to cell–cell junctions and played a crucial role in regulating cell–cell adhesion. This raised the perplexing question of how a cell adhesion protein could also be a signalling protein in Wnt pathway that affected gene expression in the nucleus. The answer came from studies showing that β-catenin's roles in cell adhesion and signalling were separable and involved the interaction of β-catenin with distinct sets of protein partners. Several groups also showed that β-catenin, in addition to its membrane localization, also localizes to the nucleus and that Wnt signalling caused an enrichment of β-catenin in the nucleus. Now this idea is so prevalent that nuclear β-catenin localization is used as a diagnostic tool for Wnt pathway activation in development and oncogenic activation of the Wnt pathway in cancer.

In the nucleus, β-catenin binds to a number of different protein partners to regulate gene expression (**Figure 6**). These partners include members of the LEF/TCF family of transcription factors (**Figures 1** and **6**). The LEF/TCF proteins are sequence-specific DNA binding proteins and serve to localize β-catenin to the promoters of Wnt target genes. A number of these target genes have been identified in the past several years and include developmental regulatory genes such as *siamois*, *twin* and *Xnr-3* in *Xenopus* and *ultrabiothorax* in *Drosophila*. Additional targets include regulators of cell growth and proliferation, c-*myc*

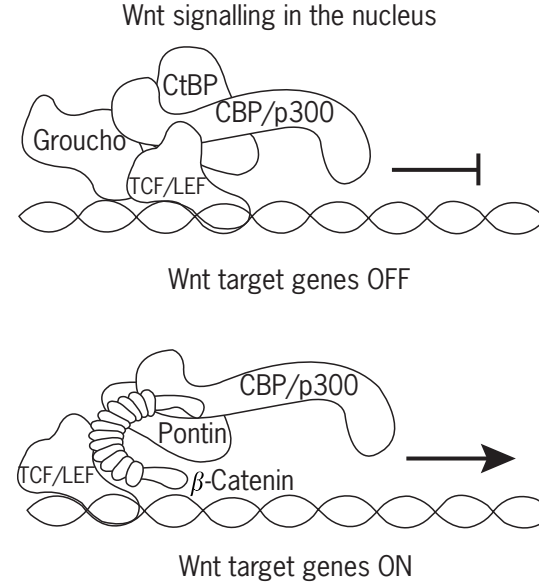


Figure 6 Regulation of gene expression by β-catenin and TCF/LEF. In the absence of Wnt signal, TCF/LEF transcription factors specifically bind to sequences in the promoters of Wnt target genes and act as repressors keeping these genes OFF. This repression is mediated through the interaction of TCF/LEF with a number of transcriptional repressor proteins including members of the Groucho family, CtBP and CBP/p300. Upon Wnt stimulation, β-catenin accumulates in the nucleus and interacts with TCF/LEF. This complex specifically binds to sites in the promoters of Wnt target genes and through interactions with additional transcription factors (e.g. CBP/p300 and Pontin 52) activate transcription. The mechanism by which the β-catenin–TCF/LEF complex activates transcription is unclear, but may involve the displacement of repressors bound to TCF/LEF by β-catenin.

and *cyclin D1*. For a complete list of known targets of the Wnt pathway, see the Wnt gene homepage (http://www.stanford.edu/~rnusse/wntwindow.html).

The mechanism by which β-catenin activates transcription remains unclear. Mutational analyses have identified two regions of β-catenin, one near the N-terminus and the other at the C-terminus, that are important for transcriptional activation. It appears that these sites may serve as protein–protein interaction domains enabling β-catenin to regulate transcription through binding of additional transcriptional regulators (**Figure 6**). Several binding partners have been identified including CBP/p300 (Creb Binding Protein) and Pontin 52. In addition, Lef-1 has been shown to form complexes with several members of the SMAD family of signalling proteins that play a role in Transforming Growth Factor β signalling (see below). Lef-1 and SMADs appear to function cooperatively to stimulate expression of specific developmental target genes such as *twin*.

Another possible mechanism by which β-catenin may activate transcription is by displacing co-repressors from TCF/LEF (**Figure 6**). In the absence of Wnt signals, TCF/LEF proteins can act as transcriptional repressors, preventing transcription of Wnt/β-catenin target genes. TCF/LEF proteins do not appear to act alone but instead require interactions with one of several identified co-repressors. These repressors include members of the Groucho family, CtBP (C-terminal Binding Protein), and CBP/p300. It is thought that the ability of CBP/p300 to act as an activator and a repressor may be due to differences in cellular context. In other words, CBP/p300 can function as either an activator or repressor depending on the situation.

ADDITIONAL REGULATORS OF Wnt SIGNAL TRANSDUCTION

Although the mechanism of Wnt signal transduction described thus far may seem complicated enough, additional players continue to be identified. For example, protein phosphatase 2A appears to play a role in regulating β-catenin stability although it is unclear whether it promotes β-catenin degradation or stabilization. The *Drosophila naked cuticle* gene has recently been shown to be a novel cytoplasmic antagonist that may limit the potency, duration or distribution of Wnt signals. Recent evidence also implicates components of the mitogen-activated protein kinase (MAPK) pathway, transforming growth factor-β-activated kinase (TAK-1) and NEMO-like kinase (NLK) as regulators of Wnt signalling. However, it remains to be determined whether these genes are true components of the Wnt pathway or whether they act in parallel to the canonical Wnt pathway.

In addition to these players, it has become clear that cross-talk between the Wnt pathway and the Transforming Growth Factor β (TGF-β) signalling pathway plays an important role in regulating Wnt/β-catenin signalling both during development and in human disease. For example, the secreted Wnt antagonist Cerberus can also interact with members of the BMP and Nodal families of TGF-β-like signalling molecules, suggesting that Cerberus can function as a multivalent modulator of both Wnt and TGF-β/BMP signalling. In addition, Wnt and TGF-β signalling pathways also cross-talk inside the cell. Several recent papers have shown that TCF/LEF transcription factors interact with members of the SMAD family of TGF-β/BMP signal mediators. Specifically, Lef-1 can form complexes with three different SMAD proteins: SMAD-2 and -3, effectors for TGF-β and Activin signals; and SMAD-4, a ubiquitous effector for all TGF-β/BMP signalling pathways. Through these interactions, SMAD proteins were found to stimulate synergistically transcription of specific Wnt target genes. These data are intriguing since mutations in components of both the Wnt (APC and β-catenin) and TGF-β signalling pathways (TGF-β receptor type II, SMAD-2, -3 and -4) are associated with colorectal cancers. Furthermore, mice double heterozygous for both *APC* and *SMAD-4* display intestinal polyps that develop into more malignant tumours than those in mice heterozygous for *APC* alone (Takaku *et al*., 1998). Together these data argue that the Wnt and TGF-β pathways cross-talk to regulate cooperatively gene expression and that this synergistic interaction may be important both during development and in cancer.

β-Catenin levels can also be regulated by Wnt-independent mechanisms. For example, expression of integrin-linked kinase in mammalian cells promotes the stabilization and nuclear accumulation of β-catenin (Novak *et al*., 1998). Presenilin proteins have also been implicated as regulators of β-catenin stability. Mutations in presenilin associated with the rapid onset of Alzheimer disease decrease the stability of β-catenin in neurons. This effect on β-catenin was also correlated with an increase in the susceptibility of neurons to apoptosis resulting from the accumulation of β-amyloid protein. Given the ability of these signalling pathways to modulate β-catenin stability, it seems likely that β-catenin may regulate many cellular processes independent of its role in Wnt signalling.

ONCOGENIC ACTIVATION OF THE Wnt PATHWAY

Tumour formation results from the loss of control over cell proliferation. This occurs through mutations that produce oncogenes with a dominant gain of function or inactivate tumour suppressor genes through recessive loss of function mutations. Both types of mutations lead to defects in regulatory pathways that normally control cell proliferation. Recently, it has become clear that components of the

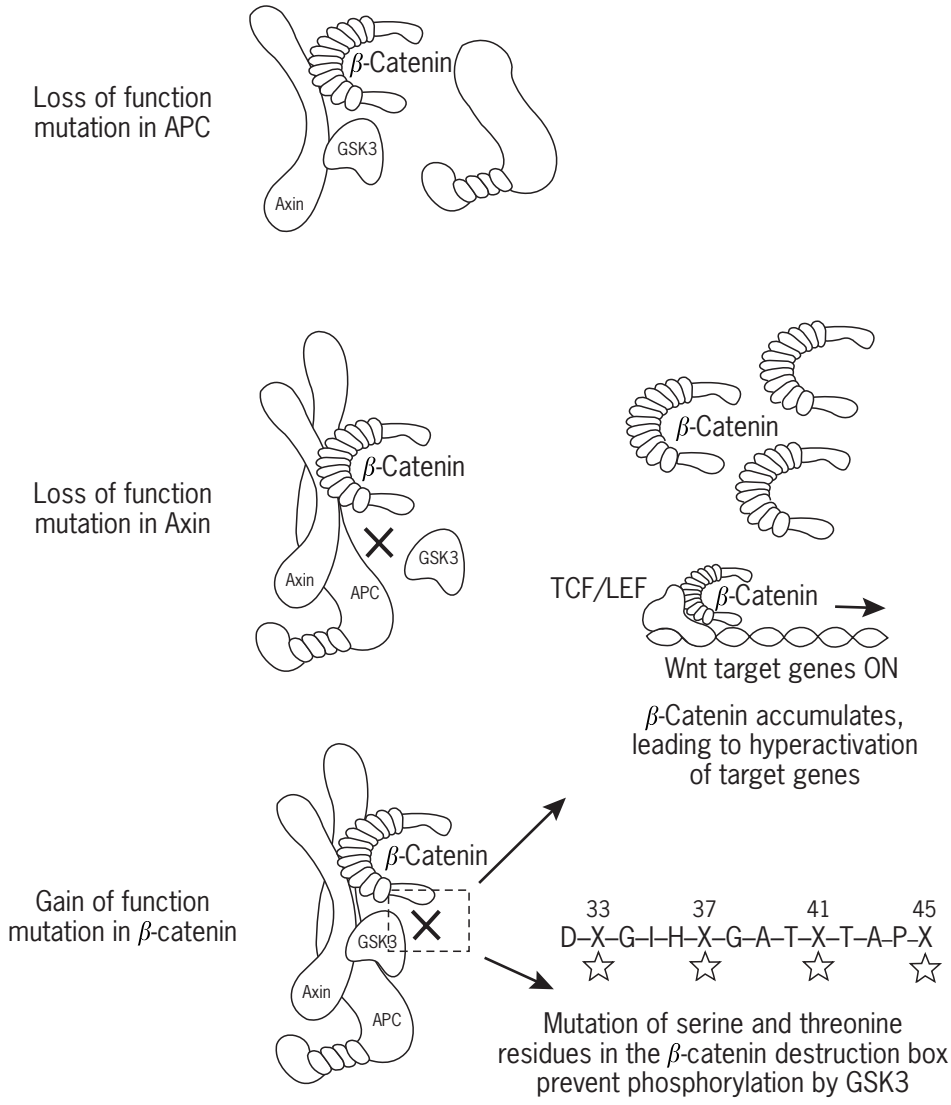


Figure 7 Oncogenic activation of the Wnt signalling pathway. Recessive loss of function mutations in the tumour suppressors APC and Axin disable the destruction complex and lead to the inappropriate accumulation of β-catenin in the cell. Most mutations isolated in *APC* are nonsense mutations that lead to the premature truncation of the protein and loss of critical binding sites for Axin and β-catenin. One mutation found in Axin appears to decrease the binding affinity between Axin and GSK3. Gain of function mutations in β-catenin that allow it to escape regulation by the destruction complex also lead to constitutive activation of the Wnt pathway. The majority of mutations in β-catenin found thus far are missense mutations that alter one of the four potential GSK3 phosphorylation sites in the destruction box. The inappropriate accumulation of β-catenin in the cell then leads to hyperactivation of target genes such as c-*myc* and *cyclin D1*, and uncontrolled cell division.

Wnt signalling pathway are mutated in a variety of human cancers. Thus far, mutations in APC, Axin and β-catenin have been identified in various cancers and these findings have fuelled a great explosion of interest in the relationship between Wnt signalling and cancer. These mutations result in the inability of the cell to regulate appropriately levels of β-catenin (**Figure 7**). Recessive mutations in the tumour suppressor genes *APC* and *Axin* result in defects in the destruction complex allowing β-catenin to escape degradation. Conversely, mutations in β-catenin produce a dominant gain of function protein that evades regulation by the destruction complex. The end result of these mutations is the constitutive activation of Wnt target genes and uncontrolled cell proliferation.

Germline mutations of the *APC* gene lead to familial adenomatous polyposis characterized by the development

Table 1 Current list of human cancers associated with mutations in β-catenin, APC and Axin

β-Catenin	APC	Axin
Colorectal adenoma and carcinoma	Colorectal adenoma and carcinoma	Colorectal
Endometrial carcinoma	Breast cancer	Hepatocellular carcinoma
Hepatoblastoma	Medulloblastoma	
Hepatocellular carcinoma		
Malignant fibrous histiocytoma		
Medulloblastoma		
Ovarian carcinoma		
Pilomatricoma		
Prostate		
Synovial sarcoma		
Uterine		
Wilms' tumours		

of colorectal polyps in the second to third decade of life. In addition, somatic mutations of the *APC* gene are associated with >80% of sporadic colorectal adenomas and carcinomas. More than 95% of germ-line and somatic mutations of the *APC* gene are nonsense mutations that result in the synthesis of a truncated protein that lacks the region of APC important for its function in the destruction complex. Significantly, these truncations in APC remove binding sites for β-catenin and Axin and also putative phosphorylation sites for GSK3 (**Figures 1** and **7**). Colon cancer cells expressing such a truncated form of APC possess very high levels of β-catenin, suggesting that the inability of APC to promote β-catenin degradation is causally linked to tumour formation. In addition to the well-documented link between APC function and colorectal cancer, mutation of *APC* is also implicated in other human cancers including aggressive fibromatosis and breast cancer (**Table 1**). Together these findings highlight the importance of APC in Wnt signalling and oncogenesis.

Recent evidence demonstrates that *Axin*, like *APC*, functions as a tumour-suppressor gene (**Figure 7**, **Table 1**). The *Axin1* gene is mutated in hepatocellular carcinomas and missense mutations in *Axin* have also been isolated in colon cancer cells (Satoh *et al*., 2000; Webster *et al*., 2000). One of these mutations was found to diminish the interaction between Axin and GSK3, providing a potential mechanism for oncogenic activation of the Wnt pathway.

The striking connection between Axin- and APC-mediated regulation of β-catenin and oncogenesis suggested that mutations in β-catenin itself might play a role in tumour formation. Consistent with this idea, targeted mutation or deletion of the destruction box sequence results in highly stable forms of β-catenin that are hyperactive (Yost *et al*., 1996; Pai *et al*., 1997). Cell lines harbouring activating mutations of β-catenin often display high levels of both cytoplasmic and nuclear β-catenin and constitutive activation of TCF/LEF reporter genes. In addition, expression of an N-terminal truncated form of β-catenin lacking the destruction box in the epidermis of transgenic mice resulted in the formation of hair follicle-related tumours. These data demonstrate that increasing levels of β-catenin are sufficient to promote tumour formation and implicate the N-terminal destruction box sequence as potential sites for oncogenic activation of β-catenin.

Over the past several years, many studies have shown that mutations in β-catenin exist in a variety of human cancers (**Table 1**). Strikingly, the identified mutations are missense or deletion mutations in the destruction box and most of these alter one of the GSK3 phosphorylation sites (**Figure 7**). In addition to the GSK3 sites, missense mutations at aspartate-32 and glycine-34 have also been reported. These residues, along with serine-37, have been characterized as a ubiquitination target motif based on its similarity with Iκ-B, another protein targeted for degradation by the $SCF^{\beta\text{-TrCP}}$ ubiquitin ligase complex. Together, these data strongly argue that mutations in β-catenin that enable it to evade regulation by the destruction complex play an important role in tumorigenic transformation of many cell types.

What is the consequence of constitutive β-catenin activity? β-Catenin functions as a transcriptional activator in Wnt signalling, suggesting that its role in tumour formation may be through inappropriate activation of genes important for regulating cell division and growth. This idea has gained support with the finding that the c-*myc* and *cyclin D1* genes, both of which are known to promote cell proliferation, are direct targets of β-catenin. c-*myc* is a potent oncogene that regulates cell cycle progression, promoting the G_1/S phase transition. Similarly, *cyclin D1* also plays an important role regulating movement through the cell cycle. Thus, mutations in Axin, APC or β-catenin that result in the stabilization and accumulation of β-catenin may lead to inappropriate expression of target genes, such as c-*myc* and *cyclin D1*. Expression of c-*myc* and *cyclin D1* would then expedite the G_1 to S transition, leading to uncontrolled cell proliferation. Additionally, evidence also suggests that β-catenin may act as a survival factor protecting cells from cell

death. Since the transition from G_1 to S phase requires the presence of survival factors, β-catenin may also stimulate this transition directly by preventing apoptosis and permitting cell cycle progression.

CONCLUSION

With the recent completion of the human genome sequence, we have entered a new era in biology and medicine. This achievement has been compared to putting a man on the moon and will undoubtedly revolutionize basic biological and medical sciences. This revolution, however, is only in its infancy. What remains is put the pieces of the puzzle together by characterizing the function of each of the estimated 80 000 human genes. How do these genes instruct cells to divide, to migrate or to die? How do these genes control embryogenesis? How do mutations in specific genes contribute to human disease?

Although many aspects of Wnt signalling remain unclear, we are beginning to put the pieces of the Wnt puzzle together to understand how the genes involved in Wnt signalling communicate signals between cells. In particular, great strides have been made towards understanding the molecular and biochemical mechanics of Wnt signal transduction. These findings underscore the important predictive value of analysing the fundamental mechanisms by which cells signal to one another: by establishing how genes function in specific signalling pathways (e.g. whether they act as repressors or activators), one can make educated guesses how these genes might contribute to human disease. A repressor could be a tumour-suppressor gene and an activator could be an oncogene. Thus, a clear understanding the basics of how cells communicate will lead us to an understanding of how inappropriate activation of signalling pathways leads to cancer. With this information, researchers will hopefully be able to design new therapeutic reagents for treating and preventing cancer.

REFERENCES

Bhanot, P., *et al*. (1996). A new member of the *frizzled* family from *Drosophila* functions as a Wingless receptor. *Nature*, **382**, 225–230.

Jiang, J. and Struhl, G. (1998). Regulation of the Hedgehog and Wingless signalling pathways by the F-box/WD40-repeat protein Slimb. *Nature*, **391**, 493–496.

Kitagawa, M., *et al*. (1999). An F-box protein, FWD1, mediates ubiquitin-dependent proteolysis of beta-catenin. *Embo Journal*, **18**, 2401–2410.

McCartney, B. M., *et al*. (1999). *Drosophila* APC2 is a cytoskeletally-associated protein that regulates *wingless* signaling in the embryonic epidermis. *Journal of Cell Biology*, **146**, 1303–1318.

Miller, J. R., *et al*. (1999). Establishment of the dorsal–ventral axis in *Xenopus* embryos coincides with the dorsal enrichment of dishevelled that is dependent on cortical rotation. *Journal of Cell Biology*, **146**, 427–437.

Munemitsu, S., *et al*. (1995). Regulation of intracellular beta-catenin levels by the adenomatous polyposis coli (APC) tumor-suppressor protein. *Proceedings of the National Academy of Sciences of the USA*, **92**, 3046–3050.

Novak, A., *et al*. (1998). Cell adhesion and the integrin-linked kinase regulate the LEF-1 and beta-catenin signaling pathways. *Proceedings of the National Academy of Sciences of the USA*, **95**, 4374–4379.

Nusse, R. and Varmus, H. E. (1982). Many tumors induced by the mouse mammary tumor virus contain a provirus integrated in the same region of the host genome. *Cell*, **31**, 99–109.

Nüsslein-Volhard, C. and Wieschaus, E. (1980). Mutations affecting segment number and polarity in *Drosophila*. *Nature*, **287**, 795–801.

Pai, L.-M., *et al*. (1997). Negative regulation of Armadillo, a Wingless effector in *Drosophila*. *Development*, **124**, 2255–2266.

Peters, J. M., *et al*. (1999). Casein kinase I transduces Wnt signals. *Nature*, **401**, 345–350.

Sakanaka, C., *et al*. (1999). Casein kinase I-epsilon in the Wnt pathway: regulation of beta-catenin function. *Proceedings of the National Academy of Sciences of the USA*, **96**, 12548–12552.

Satoh, S., *et al*. (2000). *AXIN1* mutations in hepatocellular carcinomas, and growth suppression in cancer cells by virus-mediated transfer of *AXIN1*. *Nature Genetics*, **24**, 245–250.

Takaku, K., *et al*. (1998). Intestinal tumorigenesis in compound mutant mice of both *Dpc4* (*Smad4*) and *Apc* genes. *Cell*, **92**, 645–656.

Torres, M. A. and Nelson, W. J. (2000). Colocalization and redistribution of Dishevelled and Actin during Wnt-induced mesenchymal morphogenesis. *Journal of Cell Biology*, **149**, 1433–1442.

Webster, M. T., *et al*. (2000). Sequence variants of the *axin* gene in breast, colon, and other cancers: an analysis of mutations that interfere with GSK3 binding. *Genes, Chromosomes and Cancer*, **28**, 443–453.

Yang-Snyder, J., *et al*. (1996). A *frizzled* homolog functions in a vertebrate Wnt signaling pathway. *Current Biology*, **6**, 1302–1306.

Yost, C., *et al*. (1996). The axis-inducing activity, stability, and subcellular distribution of beta-catenin is regulated in *Xenopus* embryos by glycogen synthase kinase 3. *Genes and Development*, **10**, 1443–1454.

Yost, C., *et al*. (1998). GBP, an inhibitor of GSK-3, is implicated in *Xenopus* development and oncogenesis. *Cell*, **93**, 1031–1041.

Zeng, L., *et al*. (1997). The mouse Fused locus encodes Axin, an inhibitor of the Wnt signaling pathway that regulates embryonic axis formation. *Cell*, **90**, 181–192.

FURTHER READING

Behrens, J. (2000). Cross-regulation of the Wnt signalling pathway: a role of MAP kinases. *Journal of Cell Science*, **113**, 911–919.

Bienz, M. (1999). APC: the plot thickens. *Current Opinions in Genetics and Development*, **9**, 595–603.

Bienz, M. and Clevers, H. (2000). Linking colorectal cancer to Wnt signalling. *Cell*, **103**, 311–320.

Boutros, M. and Mlodzik, M. (1999). Dishevelled: at the crossroads of divergent intracellular signaling pathways. *Mechanics of Development*, **83**, 27–37.

Cadigan, K. M. and Nusse, R. (1997). Wnt signaling: a common theme in animal development. *Genes and Development*, **11**, 3286–3305.

Czech, C., *et al*. (2000). Presenilins and Alzheimer's disease: biological functions and pathogenic mechanisms. *Progress in Neurobiology*, **60**, 363–384.

Dierick, H. and Bejsovec, A. (1999). Cellular mechanisms of wingless/Wnt signal transduction. *Current Topics in Development in Biology*, **43**, 153–190.

Kikuchi, A. (1999). Roles of Axin in the Wnt signalling pathway. *Cell Signalling*, **11**, 777–788.

Kuhl, M., *et al*. (2000). The Wnt/Ca^{2+} pathway: a new vertebrate wnt signaling pathway takes shape. *Trends in Genetics*, **16**, 279–283.

Miller, J. R., *et al*. (1999). Mechanism and function of signal transduction by the Wnt/beta-catenin and Wnt/Ca^{2+} pathways. *Oncogene*, **18**, 7860–7872.

Peifer, M. and Polakis, P. (2000). Wnt signaling in oncogenesis and embryogenesis – a look outside the nucleus. *Science*, **287**, 1606–1609.

Perrimon, N. and Bernfield, M. (2000). Specificities of heparan sulphate proteoglycans in developmental processes. *Nature*, **404**, 725–728.

Polakis, P. (1999). The oncogenic activation of beta-catenin. *Current Opinions in Genetics and Development*, **9**, 15–21.

Polakis, P. (2000). Wnt signaling and cancer. *Genes and Development*, **14**, 1837–1851.

Roose, J. and Clevers, H. (1999). TCF transcription factors: molecular switches in carcinogenesis. *Biochimica Biophysica Acta*, **1424**, M23–37.

Waltzer, L. and Bienz, M. (1999). The control of beta-catenin and TCF during embryonic development and cancer. *Cancer and Metastasis Reviews*, **18**, 231–246.

Wodarz, A. and Nusse, R. (1998). Mechanisms of Wnt signaling in development. *Annual Reviews in Cell and Development Biology*, **14**, 59–88.

Chapter 15

Extracellular Matrix: The Networking Solution

Nancy Boudreau
University of California San Francisco, San Francisco, CA, USA

Mina J. Bissell
Ernest Orlando Lawrence Berkeley National Laboratory, Berkeley, CA, USA

CONTENTS

INTRODUCTION

One of the fundamental properties of functional differentiated tissues is their unique three-dimensional (3D) organisation which allows individual cells to act in a coordinated manner to carry out complex tissue functions. Although individual cells may be capable of eliciting certain responses to external stimuli such as proliferation, many differentiated functions can be achieved only via the coordinated actions of cells within a tissue. For example, endothelial cells which line the walls of functional blood vessels may individually produce necessary anti-clotting factors but alone cannot function as a conduit for blood. Similarly, although rounded single mammary epithelial cells are capable of expressing the tissue-specific milk protein β-casein, they cannot secrete it and expression is significantly enhanced when cells form multi-cellular complexes (Streuli *et al*., 1991). Three-dimensional organisation was also found to be essential for the generation of significant populations of functional T cells from CD34 + progenitors. Only by embedding the progenitor cells into a 3D carbon matrix seeded with thymus extract could mature functional T cells be obtained. (Poznansky *et al*., 2000) Thus 3D organisation is essential not only for developing and establishing functional differentiated phenotypes but also for maintaining tissue specific gene expression and function. Moreover, 3D tissue architecture acts to override the genetic information contained within normal and malignant cells (Weaver *et al*., 1997).

In order for multi-cellular organisms to acquire and maintain a 3D organisation of their tissues, they have elaborated a complex network of extracellular proteins referred to as the extracellular matrix (ECM). This network of extracellular proteins allows cells to adhere, migrate, proliferate and undergo morphogenesis or alternatively instructs them to undergo programmed cell death (Boudreau and Bissell, 1996).

The significance of the ECM in influencing cellular behaviour was initially overlooked as the ECM was considered merely to provide scaffolding to cells and tissues. However, it has become clear that cellular interactions with the ECM also provide essential information to the cell by initiating intracellular signalling cascades which culminate in changes in gene expression. Furthermore, the ECM also provides structural cues to adherent cells. In turn, these ECM-induced changes in cell morphology and gene expression allow cells to subsequently alter their interactions with the extracellular environment. This 'dynamic reciprocity' is the basis for integrated tissue function and allows cells monitor to constantly their extracellular environment and adjust their responses to maintain differentiated tissue phenotypes. In contrast, tumorigenesis results from the loss of this dynamic interaction between cells and their ECM, and the subsequent aberrant dialogue between cells and their microenvironment prevents cells from achieving or maintaining their functional differentiated state. How cells acquire and maintain their 3D organisation is a fundamental question which largely remains to be understood. In order to understand how the various physical and biochemical properties of ECM components contribute to cellular responses, it is necessary to review the components contributing to ECM-mediated responses in cells.

ECM PROTEINS AND INTEGRIN RECEPTORS

ECM Proteins

The ECM proteins comprise a large family of glycoproteins. These large proteins are often comprised of several distinct subunits. For example, laminin, which is composed of α-, β- and γ-chains, and has a molecular mass of over 10^6 kDa, has up to three variants of each of these chains. The final composition of laminin is tissue specific. For example, laminin 5 is found in basement membranes associated with epithelia and endothelium, while laminin-2 is preferentially found in muscle (for a review, see Ekblom, 1996).

Another major component of the extracellular matrices are the collagens. This family is comprised of over 18 members, which are often expressed in tissue-specific patterns (for a review, see Prockop and Kivirikko, 1995). In addition to laminins and collagens, other major ECM components include fibronectin found in association with cells as well as in the serum, vitronectin also found in serum, and tenascin. These matrix proteins are often found in connective tissue matrices which are rich in collagens and to a lesser extent in association with basement membranes.

Another major class of ECM proteins are the proteoglycans, which, with the exception of hyaluronic acid, consist of membrane-associated globular proteins with specialized sulfated N- and O-linked carbohydrate chains called glycosaminoglycans (GAGs). The most common GAG chains found are the heparan and chondroitin chains which are found in almost all mammalian tissues (for a review, see Bernfield *et al*., 1999).

The structural and biochemical composition of many ECM proteins has been known for some time. These large proteins are usually multidomain structures capable of interacting with other matrix proteins as well as acting as a potential reservoir for soluble mitogens and morphogens (Woodley *et al*., 1983; Saskela and Rifkin, 1990; Vu *et al*., 1998). Indeed, specific matrix proteins rarely exist in isolation but instead are found as members of larger complex matrices comprised of various ratios of individual components. It must be emphasized that the net effect of these complex matrices on cell behaviour is equally complex and not simply the sum of the individual components. One fairly well-characterized type of complex ECM is the basement membrane (BM), which consists primarily of laminin-1 and -5, type IV collagen and other minor components including nidogen, fibronectin and proteoglycans (for a review, see Schwarzbauer, 1999). As the name implies, BMs are found at the basal lateral surface of most epithelial tissues including mammary and intestinal epithelium, hepatocytes, keratinocytes and endothelial cells which line the blood-vessel walls. Although laminin can comprise up to 80% of BM, cells behave very differently when in contact with laminin alone as compared with a complete BM. For example, when endothelial cells are plated on a complete reconstituted BM, they rapidly form an anastomosing network of hollow tube-like structures reminiscent of capillaries (Kubota *et al*., 1998). In contrast, when plated on laminin alone, they form a continuous 'cobblestone' monolayer of cells (Madri and Williams, 1983). Similarly, mammary epithelial cells will form 3D alveolar structures and express milk proteins when cultured on a complete BM but not when plated on laminin alone (Roskelley *et al*., 1994). Together these findings emphasize that complex ECMs are necessary to direct complex tissue-type organisation and gene expression. A recent comparative analysis of the *Drosophila* and *C. elegans* genomes indicated that these BM proteins have remained highly conserved throughout evolution from *C. elegans* to vertebrates, emphasizing the essential nature of these proteins in multi-cellular organisms (Hynes and Zhao, 2000).

More often than not, contact with a BM-type ECM leads to growth arrest and differentiation and promotes tissue-specific gene expression. In contrast to BMs, stromal matrices or provisional matrices formed at sites of injury often consist of various ratios of the fibrillar collagens, tenascin, fibronectin as well as other matrix components derived from the circulation, including vitronectin and fibrinogen. These matrices often comprise the bulk of connective tissues surrounding cells other than epithelium such as fibroblasts and chondrocytes. Adhesion to these connective-tissue type matrices often promotes cell migration and proliferation. Interestingly, increased amounts of stromal ECM are often found adjacent to many epithelial tumor cells (Weaver *et al*., 1996).

In attempting to understand how different ECM components give rise to these markedly different phenotypes, much attention has been focused on different cell-surface receptors for these various ECM components.

Integrins

Cells recognize and respond to different ECM matrices by interacting with cell-surface receptors called integrins. Integrins are a large, specialized family of transmembrane heterodimeric proteins which consist of an α subunit and an often larger β subunit with a cytoplasmic domain. These cytoplasmic domains may directly interact with cytoskeletal proteins and thereby serve to 'integrate' the extracellular and intracellular environments.

The α and β subunits can form up to 19 different combinations which have some selective but also overlapping affinities for various ECM components (see **Tables 1** and **2**). For example, the $\alpha 2\beta 1$ heterodimer has been shown to bind both collagen I and laminin, whereas the $\alpha 6\beta 1$ integrin will bind laminin only and not collagen.

Table 1 Integrin heterodimers

Beta subunits	Alpha partners
β1	α1, α2, α3, α4, α5, α6, α7, α8, α9, αv
β4	α6
β3, β5, β6, β8	αv
β2	αL, αM, αX
β7	α4

Table 2 Integrins for common ECM ligands

ECM ligand	Integrin
Laminin	$\alpha1\beta1$, $\alpha2\beta1$, $\alpha3\beta1$, $\alpha6\beta1$, $\alpha7\beta1$, $\alpha6\beta4$
Collagen I	$\alpha1\beta1$, $\alpha2\beta1$, $\alpha3\beta1$,
Fibronectin	$\alpha4\beta1$, $\alpha5\beta1$, $\alpha v\beta1$, $\alpha v\beta3$, $\alpha v\beta6$, $\alpha v\beta8$, $\alpha4\beta7$
Tenascin	$\alpha v\beta3$, $\alpha v\beta6$, $\alpha v\beta8$
Vitronectin	$\alpha v\beta1$, $\alpha v\beta3$, $\alpha v\beta6$, $\alpha v\beta8$

The ligand recognition sequences of the integrins are determined by the combination of α and β subunits of the heterodimeric molecules, rather than by the α or β chains alone, which may account for some of the substrate overlap. The integrin heterodimers can recognize distinct amino acid sequences present in different ECM ligands. For example, the arginine–glycine–aspartic acid (RGD) recognition domain present in many integrin heterodimers including the $\alpha v\beta3$ and $\alpha5\beta1$ integrins can bind to any ECM component which contains an exposed RGD sequence such as fibronectin and tenascin (Pierschbacher and Ruoslahti, 1984; Joshi *et al*., 1993). Furthermore, because most ECM proteins contain multiple domains, they may also contain multiple integrin recognition sites. For example, in addition to the RGD domain, tenascin also contains fibronectin-type repeats which can bind other integrins such as $\alpha9\beta1$ in an RGD-independent manner (Yokosaki *et al*., 1994).

Like the tissue-specific distribution of some ECM components, many of the integrin subunits are often distributed in cell- and tissue-specific patterns. For example, the $\beta2$ integrin is almost exclusively expressed by leukocytes, whereas others such as the $\beta1$ subunit are widely expressed. Although it is tempting to suggest that tissue-specific responses might be determined by the composition of the adjacent ECM as well the particular species of integrin available to interact with it, this relationship is not straightforward. Many integrins exist in inactive conformations, whereby the ligand recognition domains are masked or unavailable to interact with the ECM. Integrin activation can be induced by many different stimuli and provides an additional level of control by which cells can regulate their interaction with the microenvironment (for a review, see Hughes and Pfaff, 1998). This activation phenomenon, referred to as 'inside-out signalling,' is often induced by binding of the ECM to a particular integrin, which in turn induces signalling cascades to allow other surface integrins to become activated. On the other hand, it has also been shown that activating intracellular signalling pathways can also lead to inactivation of integrins (Hughes *et al*., 1997). 'Inside-out signalling' is yet another example of the dynamic interplay between cells and their matrices and emphasizes the reciprocal nature of these interactions.

In addition, proteolytic fragments of ECM molecules appear capable of binding many of the same integrins as their full-length counterparts, yet often give rise to dramatically different cellular responses. For example, both the full-length laminin-5 and the $\alpha3$ fragment of laminin-5 which is generated via proteolytic cleavage can bind the $\alpha3\beta1$ integrin. However, in contrast to full-length laminin, the proteolysed fragments increase cellular migration (Gianelli *et al*., 1997). Thus the net balance of proteolytic activity and structural integrity of the ECM as well as the expression pattern or activation state of integrins all contribute to the coordinated tissue responses to a particular microenvironment.

Signalling by Integrins

Binding of an ECM molecule to its integrin receptor initiates a series of intracellular signalling events which led to changes in cellular behaviour. Signalling via integrins often involves the recruitment of adaptor proteins and activation of a series of cytoplasmic protein kinases. (A detailed description of tyrosine kinase signalling is provided in the chapter on Signalling by Tyrosine Kinases.) The classical integrin-mediated signalling pathways originally described involved the cytoplasmic domains of the β integrin subunit undergoing a conformational change upon ligand binding. This conformational change in turn allows focal adhesion kinase (FAK) to phosphorylate itself. Phosphorylated FAK could then bind to promote the activation of a number of different downstream mediators including members of the Src family of protein kinases, the PI3 kinase pathway, as well as recruitment of adaptor proteins including $p130^{cas}$, Crk and the Grb2–Sos complex which recognize phosphorylated intermediates via their SH2 domains. In turn, assembly of these focal adhesion complexes and phosphorylated intermediates leads to the activation of Ras and Raf. Activated Raf then phosphorylates extracellular-related kinase (ERK). Phosphorylated ERK can then translocate to the nucleus and activate cell proliferation via its ability to modulate the activity of transcription factors such as TCF necessary for expression of cell cycle mediators including cyclin D1 and c-fos (**Figure 1**).

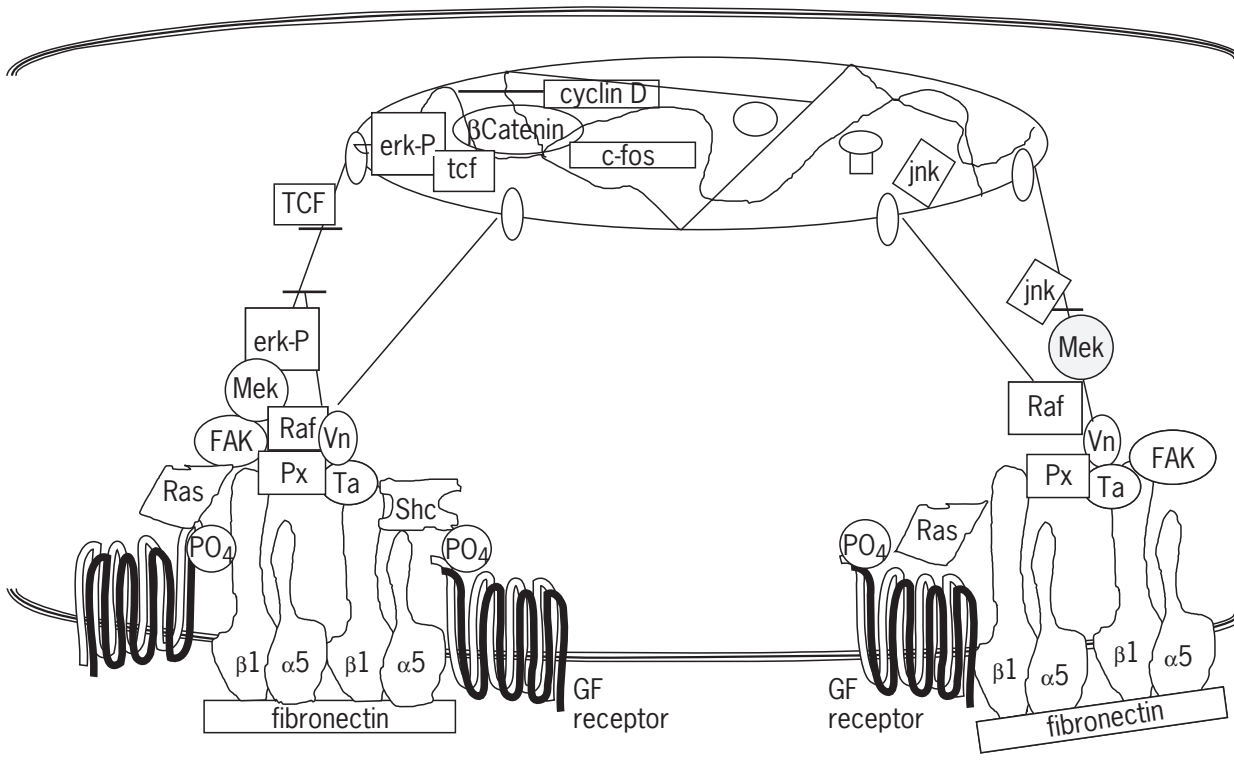


Figure 1 Intracellular signalling pathways which may participate in cell proliferation induced by adhesion to a fibronectin extracellular matrix. Integrins and growth factor receptors are clustered and form focal adhesion complexes which contain vinculin (Vn), paxillin (Px) and talin (Ta), which in turn bind to the cytoskeleton. The growth factor receptors are also phosphorylated (PO_4) and signalling initiated by the focal adhesion complexes allows ERK or JNK to become activated and phosphorylated and translocate to the nucleus where they act as transcription factors to allow expression of genes including *c-fos* or cyclin D1. β-Catenin can also move to the nucleus and form transcriptional complexes with TCF. Note that on this substratum that cells generally become flattened and elongated.

More recently, it has been shown that α integrin subunits could also specifically initiate signalling cascades. Rather than initiating signalling through phosphorylation of FAK, the α chains require the membrane-bound caveolin-1 complex to recruit the Shc adaptor molecule. Shc in turn can interact with the Grb2–Sos complex which then activates Ras and Raf and the MAPK signalling pathways to promote proliferation (Wary *et al*., 1996, 1998). Recent studies with *Drosophila* have indicated that Shc is required only for a subset of receptor tyrosine kinase activities, further suggesting that recruitment of Shc may impart some specificity to signalling cascades (Luschnig *et al*., 2000). What is not clear is how different cell types might differentially recruit signalling mediators as many of the α integrin subunits which recruit Shc are ubquitously expressed.

Another issue which has been raised in studying ECM–integrin-mediated proliferation is the apparent lack of specificity of the signalling intermediates. For example, the signalling mediators employed by integrins to induce cell proliferation are essentially the same factors required for mitogen or cytokine-induced proliferation (Hill and Treisman, 1995). This has led to some debate as to whether integrins were simply permissive for growth factor-induced signalling rather than being capable of independently initiating these signalling cascades. More recent evidence suggests that ECM–integrin binding can independently initiate signalling responses but these signals are relatively transient (Chen *et al*., 1996). What is clear, however, is that the propagation or amplification of MAPK pathways arising from binding of soluble mitogens to their receptors is absolutely dependent upon cell adhesion to the ECM. For example, in isolated non-adherent fibroblasts, addition of the mitogen PDGF could only promote transient MAPK activation and cell cycle progression to the G1/S boundary. Further progression through the cell cycle including DNA synthesis and expression of cyclin E required that cells be attached to an ECM to sustain a threshold of MAPK activity compatible with cell cycle progression. Moreover, they noted that the shape the cells acquired upon adhesion to fibronectin, rather than adhesion *per se*, was critical for proliferative responses (Zhu and Assoian, 1995; Zhu *et al*., 1996).

Complex Matrices and Growth Arrest

Although the signalling pathways induced by the interactions of individual ECM ligands with a particular integrin receptor in homogeneous cell populations have been relatively well defined, it is not clear how signalling cascades are impacted by complex multiligand matrices interacting with multiple cell surface integrins. Whether complex interactions like these which normally occur *in vivo* give rise to qualitative, quantitative or spatially modified signalling processes has not been resolved.

Furthermore, many cells, in particular epithelial and endothelial cells, respond to complex BM ECMs by withdrawing from the cell cycle, acquiring a differentiated phenotype and initiating tissue-specific gene expression. Although MAPK activity is significantly attenuated in cells induced to growth arrest on these matrices, it is not yet clear how this is achieved as initial adhesion to these matrices via integrins also activates MAPK, albeit transiently. It may be possible that BM activates phosphatases which are capable of uncoupling the kinase-dependent activity of the MAPK pathway. For example, PTEN is a phosphatase which attenuates integrin-induced phosphorylation of FAK as well as the PI3-kinase pathway (for a review, see Tamura *et al*., 1999).

Another hypothesis that has been put forward to explain inhibition of cell proliferation by BM binding integrins including $\alpha 3\beta 1$, $\alpha 6\beta 1$ and $\alpha 2\beta 1$ is that they do not efficiently recruit Shc and are not capable of sustaining MAPK activity at a level which would support proliferation (Wary *et al*., 1996). However, other studies have shown that these same α integrins could not only sustain MAPK but also promote proliferation in both endothelial and epithelial cells (Aplin *et al*., 1999; Gonzales *et al*., 1999). Thus other factors which may or may not influence Shc recruitment must be involved in deciding whether cells proliferate or growth arrest using BM integrins.

Interestingly, when cells did proliferate using these same integrins, not only was it independent of Shc but, more importantly, it also appeared that the shape that the cells acquired was a dominant determinant of whether they could sustain MAPK activity and undergo proliferation (Aplin *et al*., 1999). Indeed, recent evidence suggests that the structure imposed on cells through their interaction with the ECM is responsible for modulating these signalling pathways and ultimately determines whether cells proliferate or undergo growth arrest and differentiate (Roskelley *et al*., 1994; Chen *et al*., 1997; Wang *et al*., 1998). Furthermore, evidence is emerging that ECM-induced changes in cell morphology and architecture may in fact influence the activity of phosphatases such as PTEN (Wu *et al*., 2000). The role of ECM in modulating the morphology of cells and its influence on mediating intracellular signalling is discussed in the following sections.

INTEGRATION OF CELL MORPHOLOGY AND SIGNALLING BY ECM

Integrin Clustering and Cell Shape

One of the most critical events required for propagation and maintenance of signals generated following binding of integrins by ECM is clustering of the integrin receptors, often at sites known as focal adhesions or focal contacts. Without the appropriate clustering, ligand occupation of the receptor is not sufficient to induce a full biological response (Miyamoto *et al*., 1996). The clustering is believed to facilitate interactions between the integrin cytoplasmic tails and adaptor proteins to allow focal adhesion complexes to assemble. Not surprisingly, immunoprecipitation of these FA complexes showed that the growth factor (GF) receptors are also found within these integrin-containing complexes. For example, both PDGF-BB and insulin receptors were immunoprecipitated in complexes isolated using antibodies against $\alpha v\beta 3$ integrin (Schneller *et al*., 1997). The colocalization of GF receptors and integrins is believed to facilitate 'crosstalk' between ECM and GF receptors and coordinate or amplify the signals which may be independently generated by the ECM or soluble mitogens.

Clustering of integrins is not only required for integrin-induced migration or proliferation but is also essential for differentiation and tissue-specific gene expression in mammary epithelial cells on BM (Roskelley *et al*., 1994). This absolute dependence on clustering of integrins for either proliferation or differentiation underscores the importance of having structure imposed upon cytoplasmic signalling mediators.

In addition to structural changes at the level of the focal adhesion, adhesion to different ECMs can induce cells either to spread or to become rounded and polarized. It has become increasingly evident that cell-shape changes are a necessary and integral component of how cell–ECM interactions can generate tissue-specific architectures and gene expression, i.e. in order for ECM to evoke the appropriate response, cells must adopt an appropriate morphology (Roskelley *et al*., 1994). In general, cell spreading appears to be required for cells to proliferate while cell rounding is a prerequisite for growth arrest. For example, work by Ingber and colleagues has elegantly demonstrated that when endothelial cells are cultured on FN matrices they often adopt a spread morphology and proliferate. However, using micropatterned substrates which forced cells to become rounded while maintaining a similar degree of integrin mediated contact with the matrix, cells were unable to grow in the presence of mitogens (Chen *et al*., 1997). In contrast, cell rounding, which can be induced by BM-type ECM, is required for other functions such as the expression of the β-casein gene by mammary epithelial cells (Roskelley *et al*., 1994).

Binding of cells to the ECM not only can ligate and cluster integrins to initiate signalling cascades, but also provides the cells with a morphology to sustain the appropriate response.

What has not been directly established in these studies is how the cell shape impacts on intracellular signalling cascades. It is entirely possible that a cell's shape can determine whether integrins recruit signalling intermediates which interact with the growth-promoting MAPK pathways. For example, perhaps cell rounding, which generally suppresses growth, might preclude recruitment of membrane-associated mediators such as caveolin–Shc complexes and thereby attenuate proliferative signals, whereas cell spreading may support this effect. In order to understand how cellular geometry influences intracellular signalling it is necessary to understand the dynamics of integrin–cytoskeletal interactions which underlie these morphological changes. These are discussed briefly below.

Integrin–Cytoskeletal Connections

The observations that β integrin subunits extend into the cytoplasm has long generated speculation that the integrin cytoplasmic domains directly interact with cytoskeletal elements to bring about the changes in clustering and cell morphology. Functional linkages are supported by an extensive literature which shows that disrupting ECM–integrin interactions with function blocking antibodies leads to significant alterations in cell shape and cytoskeletal organisation. Similarly, directly disrupting the cytoskeleton with a variety of agents will impair integrin-mediated functions (for a review, see Schoenwaelder and Burridge, 1999).

It has now been established that integrins, in particular the β1 subunit, can directly interact with either microfilaments (actin) or intermediate filaments of the cell cytoskeleton. This is mediated by binding of proteins including α-actinin, talin, vinculin, filamen and paxillin. Many of these proteins directly interact with integrin cytoplasmic domains as well as actin filaments (for reviews, see Yamada and Geiger, 1997; Critchely, 2000). Furthermore, the ability of integrins to interact with these cytoskeletal proteins also requires that the integrins be clustered in focal adhesions (Calderwood *et al*., 1999). In addition to clustering, ligand occupancy is also required as the ability of integrins to bind these proteins is masked or suppressed in unoccupied integrins (Miyamoto *et al*., 1996).

Although integrins may directly interact with many of these cytoskeletal elements, many biological processes such as cell migration require that these interactions be dynamic. One means by which the ECM and integrins can dynamically reorganize the cytoskeleton is through activation of members of the Rac/Rho GTPase family which modulate actin cytoskeleton dynamics and interactions with integrins (for a review, see Bishop and Hall, 2000). For example, activated Rho can induce phosphatidylinositol-4,5-biphosphate (PtdInsP2), which in turn unmasks talin and actin binding sites in vinculin (Gilmore and Burridge, 1996). In addition to mediating dynamic changes in cell shape and motility, integrin-mediated changes in gene expression are also dependent upon Rho activation as induction of collagenase gene expression in fibroblasts by $\alpha5\beta1$ integrin ligation could not proceed when Rho was mutated (Kheradmand *et al*., 1998).

It is believed that these cytoskeletal rearrangements induced by ECM and integrins act to organize or compartmentalize signalling intermediates in such a way as to facilitate or enhance their interactions. Indeed, the attenuation of MAPK activity observed in suspended cells as compared with adherent cells supports this notion (Zhu and Assoian, 1995; Aplin *et al*., 1999). It has also been suggested that efficient signal transduction also requires that additional factors be recruited to stabilize these interacting protein complexes.

Nonintegrin–Cytoskeletal Linkages and Scaffolding Proteins

Analogous to the adaptor proteins mentioned earlier, scaffolding proteins can act to stabilize the relatively weak interactions between signalling mediators by binding multiple components which interact with each other, as well as with cytoskeleton components. For example, in *Drosophila*, the Ina D scaffold protein involved in photoreceptor signal transduction acts to bind and stabilize up to five partner proteins at a specific subcellular location resulting in the formation of a 'transducisome' (for a review, see Burack and Shaw, 2000). Many of the proteins which carry out these scaffolding functions contain one or more domains capable of facilitating protein–protein interactions. One particularly common domain found in many of these proteins are the PDZ domains (post-synaptic/discs large/Zo-1 domains) (for a review, see Dimitratos *et al*., 1999). In addition to facilitating interactions between proteins, many of these PDZ domain proteins have been shown to undergo extensive interactions with both the plasma membrane and the cytoskeleton, which probably contributes to their ability to localize these multiprotein complexes or receptors at specific cellular locations (Fanning *et al*., 1998; Hildebrand and Soriano, 1999).

Although a role for scaffolding proteins in mitogen-induced signalling is well established, there is as yet no direct evidence that integrin signalling pathways employ these proteins. Interestingly, there is evidence that non-integrin ECM receptors interact with several closely related proteins. Dystroglycan is a component of the dystrophin-associated protein complex found in muscle, neurons and epithelial cells. Dystroglycan not only binds to laminin in the extracellular space but also binds

dystrophin, which in turn binds actin in the cytoplasm (Ervasti and Campbell, 1993; Kachinsky *et al*., 1999). Syndecan, a heparan sulfate proteoglycan, is another nonintegrin ECM receptor capable of binding laminin, while syndecan's cytoplasmic face can bind to CASK. CASK is a specialized type of PDZ protein or MAGUK (membrane associated guanylate kinase), which can also interact with protein 4.1 of the actin cytoskeleton to provide an additional link between the ECM and the cytoskeleton (Cohen *et al*., 1998). Furthermore, when CASK's interaction with syndecan is disrupted, possibly by proteolytic cleavage, CASK can then shuttle directly to the nucleus and interact with T-brain, a T box transcription factor which induces expression of reelin, another ECM protein which is required for neuronal adhesion, migration and pathfinding (D'Arcangelo *et al*., 1995; Hseuh *et al*., 2000). Thus an intact ECM is required to prevent translocation of these factors.

Related members of this PDZ/MAGUK family of proteins have also been implicated in maintenance of cell polarity and cell–cell junctions, a process which is significantly enhanced by adhesion of cells to BM. The relationship between ECM, the establishment of cell junctions and recruitment of these and other proteins in growth arrest and differentiation is discussed in the following section.

Establishment of Polarity and Cell–Cell Junctions

When cells assume their tissue-specific 3D organisation, they also establish extensive cell–cell junctions and exhibit tissue polarity. Not surprisingly, the cytoskeletal and morphological changes induced by cells interacting with the ECM are necessary for the formation of appropriate cell junctions and polarity. Hemidesmosomes, adherens junctions and tight or occludens junctions found in epithelial tissues can be directly influenced by the ECM.

Adherens junction are multiprotein complexes present in most epithelial cells *in vivo*. These complexes are located between adjacent epithelial cells and contain, most notably, E-cadherin in an insoluble complex with α- and β-catenins (for a review, see Fuchs *et al*., 1997). As mentioned, ECM and integrins play a critical role in assembling and establishing adherens junctions in epithelial tissues, as culturing epithelial cells on BM can promote the formation of polarized 3D spheroids with functional adherens junctions containing E-cadherin and α- and β-catenin (for reviews, see Gumbiner 1996; Weaver *et al*., 1997). On the other hand, loss of ECM adhesion or disruption of appropriate cell–ECM contacts and 3D morphology interferes with the assembly of junctional complexes. Like most cell–ECM interactions, this relationship between integrins and adherens junctions is dynamic and reciprocal, as establishment of adherens junctions and recruitment of cadherin–catenin complexes can also feedback to down-regulate expression of the β1 integrin associated with growth in keratinocytes (Hodivala and Watt, 1994).

Although the exact mechanisms which link the ECM and adhesion junctions is not clear, both focal adhesions and adherens junctions contain many of the same molecules which interact with the cytoskeleton. For example, vinculin, which localizes to FA via its interactions with talin and β integrins, is also present in adherens junctional complexes. Vasp, a vinculin-binding protein capable of binding G actin and nucleating actin fibril assembly, is not only found in focal adhesions but also recruited to epithelial junctions along with vinculin (Vasioukhin *et al*., 2000). The use of common cytoskeletal proteins which bind to integrin or cadherin complexes and in turn interact with the actin cytoskeleton suggests an integrated system of cytoskeletal fibres being pulled by potentially competing molecules to generate appropriate cell tension and shape.

What is clear is the requirement for ECM and functional adhesion junctions in maintaining a quiescent differentiated state in epithelial tissues. Interfering with cell-ECM interactions disrupts junctional complex assembly, and β-catenin is no longer retained in insoluble junctional complexes (Weaver *et al*., 1997; Novak *et al*., 1998). Free β-catenin can then migrate to the nucleus and form complexes with the transcription factor TCF/LEF, which in turn promotes expression of a number of genes which are incompatible with a quiescent differentiated phenotype including matrilysin, fibronectin and cyclin D1 and the oncogenic c-*myc* gene (He *et al*., 1998; Crawford *et al*., 1999; Shtutman *et al*., 1999).

In addition to β-catenin, many of the MAGUK/PDZ-domain protein family members also figure prominently in the formation and maintenance of junctional complexes in polarized epithelial tissues. Elegant work on *Drosophila* has identified genes critical for epithelial polarization including the PDZ domain protein *scribble*, and a related MAGUK protein called discs large lethal (*dll*). When these genes were mutated, epithelial cells exhibited a loss of polarity and became round and multilayered, further emphasizing that tissue structure is essential to prevent tumorigenesis (Bilder *et al*., 2000). Although the mammalian homologue of scribble is not yet known, the mammalian homologue of dll, DLG, has been identified as a protein which binds to APC, a junction-associated tumour-suppresser protein which also binds β-catenin to prevent its translocation to the nucleus (Matsumine *et al*., 1996). Recent evidence has shown that the activity of another tumour suppressor, PTEN, a phosphatase which attenuates integrin-mediated FAK and PI3 kinase activity, could be enhanced by binding to MAGI-2, a MAGUK scaffolding protein related to DLG. Like dll and scribble, MAGI-2 is normally recruited to and anchored in membranes at epithelial tight junctions where it binds PTEN via its PDZ domains (Wu *et al*.,

2000). Although a direct link between the ECM and these particular junction-associated proteins has not been demonstrated, the ability of non-integrin ECM receptors to bind closely related proteins and the critical role of the ECM in organizing the cytoskeleton and cell junction assembly suggest that such interactions are likely.

ECM and integrins have also been directly linked to the formation of hemidesmosomes found on the basolateral surface of epithelial cells which contact the BM. The $\alpha6\beta4$ laminin-binding integrin is a major component of hemidesmosomes (Sonnenberg *et al*., 1991). In contrast to other β integrins, $\beta4$ has an unusually long cytoplasmic tail over 1000 amino acids in length. The $\beta4$ tail can directly bind to plectin, a large cytoskeletal protein capable of interacting with both the actin microfilament and intermediate filament cytoskeleton (for a review, see Steinbock and Wiche, 1999). In contrast to other integrins which require activation and clustering to interact with proteins linked to the cytoskeleton, the $\beta4$ integrin cytoplasmic tail appears capable of interacting with hemidesmosome components even in the absence of ligand occupancy (Nievers *et al*., 2000). A critical role for the $\beta4$ integrin in the formation and function of hemidesmosomes is evident in $\beta4$ integrin-deficient mutants, where blistering of the skin and epidermal detachment were observed due to the lack of hemidesmosome formation (Dowling *et al*., 1996; Van der Neut *et al*., 1996) Furthermore, interfering with basolateral localization of $\alpha6\beta4$ in breast epithelial cells not only disrupts hemidesmosome formation but also induces these cells to undergo apoptosis (V. M. Weaver and M. J. Bissell, unpublished work).

Together a picture emerges that beyond simple ligation of integrins and induction of signalling cascades, the ECM also directs the changes in cell and tissue architecture via cytoskeletal linkages from both integrin and nonintegrin ECM receptors. The cytoskeletal reorganisation helps to organize hemidesmosomes and adherens junctions and immobilize multidomain scaffolding proteins. This sequestration can either prevent untoward transcriptional activity or enhance the function of growth-suppressing genes such as *PTEN* or *APC*.

3D ORGANISATION AND INTEGRATED TISSUE RESPONSES

Mammary Gland

One model which has been invaluable for investigating and understanding how interactions between cells and complex ECMs can direct and maintain the functionally differentiated state of tissues is the adult mammary gland. At the onset of pregnancy, the adult mammary gland undergoes a series of morphological and functional changes which culminate in the establishment of organized acinar structures consisting of polarized epithelial cells which produce and secrete milk proteins. This organogenesis is intimately dependent upon the production and deposition of an intact basement membrane.

Many of the cellular and molecular events which contribute to this functional differentiation have been elucidated through the use of a 3D tissue culture model which mimics the normal postnatal mammary gland morphogenesis which occurs *in vivo* (Barcellos-Hoff *et al*., 1989). Either primary or immortalized mammary epithelial cells can be induced to undergo this morphological and functional differentiation by simply plating them on a reconstituted intact BM. The mammary epithelial cells immediately adhere to the BM and within a few days form 3D spheroids which resemble the acinar structure of the gland *in vivo* both in organisation and in size. Contact with the BM also induces cells to exit the cell cycle, become polarized and form adherens junctions. In contrast, when the same cells are plated on tissue culture plastic without the BM matrix, they adhere primarily to non-BM proteins which are present in the serum including fibronectin and vitronectin using the appropriate integrins. These cells will spread to form monolayers and are not able to form spheroids. These cells also cannot polarize, form proper adherens junctions or exit the cell cycle. Not surprisingly, these cells cannot express the tissue-specific milk protein genes (Barcellos-Hoff *et al*., 1989; Schmidhauser *et al*., 1990).

The BM-induced functional differentiation requires the contribution of several different laminin-binding receptors at different stages of this process. Initially the epithelial cells adhere to the BM via laminin-binding $\beta1$ integrins, most likely the $\alpha3$ or $\alpha6$ $\beta1$ heterodimers. Following the initial adhesion, cells begin to round in a manner which is not perturbed when either $\alpha6\beta4$ or other $\beta1$ integrins are blocked. Instead, the E3 fragment of laminin, which probably binds to the nonintegrin ECM receptors dystroglycan or syndecan, appears necessary for this process (J. Muschler and M. J. Bissell, unpublished work). Furthermore, although the E3-dependent rounding is a prerequisite for the cells to begin expressing milk proteins (Streuli *et al*., 1995), milk gene expression in rounded cells could also be blocked by antibodies against both the $\alpha6\beta4$ integrin and $\beta1$ integrins but not $\alpha1$, $\alpha5$ or αv integrins (Muschler *et al*., 1999). This 'division of labour' between both integrin and nonintegrin receptors and the dependence on cell shape emphasize the integrated nature of the response of cells to BM.

In addition to the role for ECM-binding proteins, the organisation of the BM and its presentation to the cells is also critical in eliciting these responses. As mentioned previously, when the BM is cross-linked or fixed so that it is no longer malleable, cells are unable to become rounded and acquire a polarized morphology and cannot express milk proteins despite the cells' ability to adhere strongly to this matrix (Streuli and Bissell, 1990). Furthermore, although the minor BM component nidogen alone is not capable of inducing milk gene expression in these cells,

addition of nidogen can enhance the laminin-dependent induction of milk gene expression (Pujuguet *et al*., 2000). Nidogen normally acts as a bridge between BM laminin and type IV collagen, and thus probably acts to organize the presentation of laminin to the cells (see **Figure 2**).

Furthermore, once these cells have undergone the necessary changes in cell shape, they must also undergo further changes in cell–cell organisation to form 3D alveolar structures, complete with functional adherens junctions and hemidesmosomes. The generation of these 3D structures is particularly critical for maintaining a functional differentiated state, as interfering with this organisation can induce the cells to undergo apoptosis (Boudreau *et al*., 1996). When this organisation is disrupted, the cells are not able to remain quiescent and begin to progress through the cell cycle, a process which often triggers apoptosis in normally quiescent cell (Evan *et al*., 1992). Furthermore, the 3D organisation is also essential for meaningful cross-talk between integrins and growth factor receptors (Wang *et al*., 1998).

Consequences of Disrupted Tissue Organisation

Based on the above, it is clear that 3D tissue-type morphology and the related intracellular organisation imposed by the ECM determine whether cells will differentiate, proliferate or undergo apoptosis. With respect to differentiation, it is clear that in addition to cellular rounding, establishing polarity and forming junctional complexes are equally critical for maintaining this state. It might be predicted

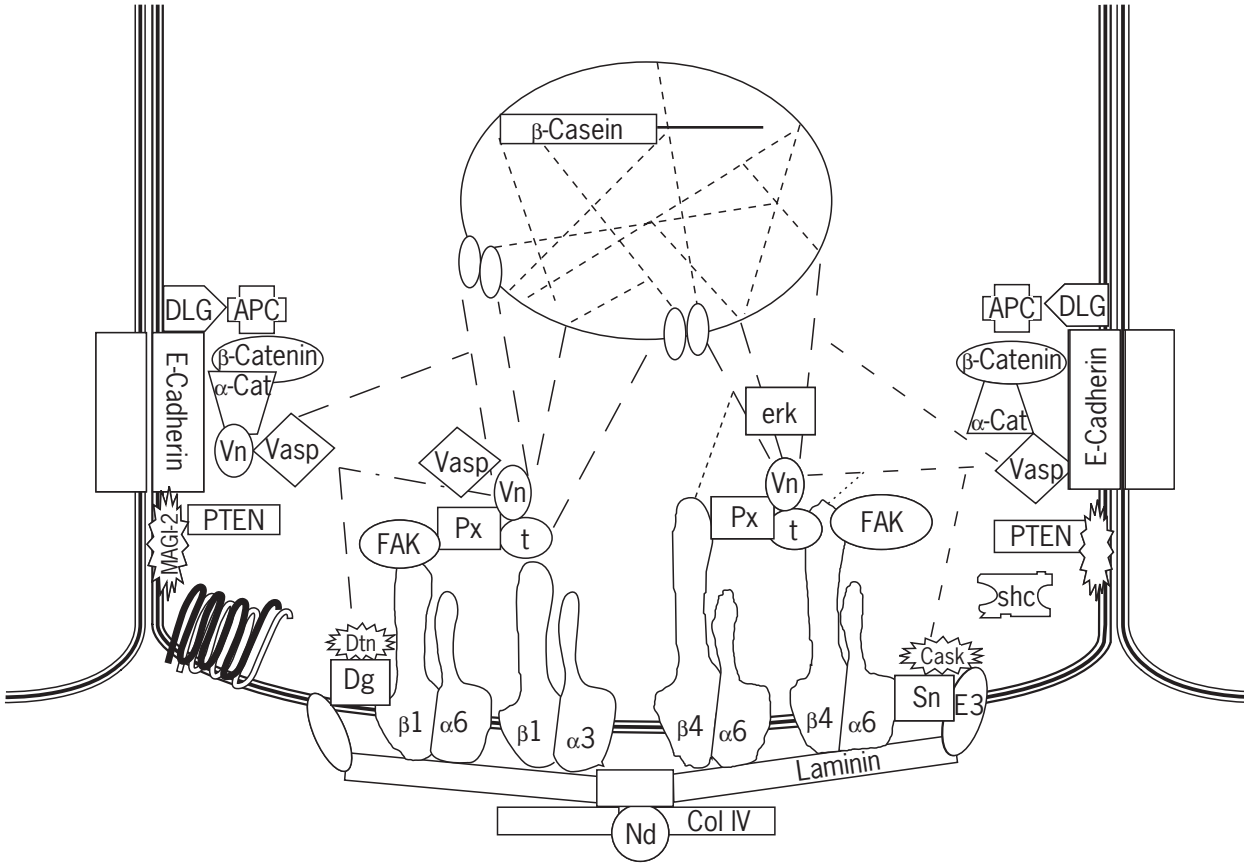


Figure 2 Intracellular organisation of polarized, quiescent epithelial cells in response to basement membrane extracellular matrix. The basement membrane matrix on the basolateral surface of the cells is comprised of several components including laminin, type IV collagen (Col IV) and nidogen (Nd). Laminin binds to $\alpha 6\beta 4$, $\alpha 6\beta 1$ and $\alpha 3\beta 1$ integrins and generates signals which allow transcription of the tissue-specific gene β-casein. Cells also employ non-integrin receptors including dystroglycan (Dg) and syndecan (Sn) to bind to the E3 domain of laminin which help cells to become rounded. Cask and dystrophin (Dtn), which associate with Sn and Dg receptively, can also directly bind the actin cytoskeleton (dashed lines). Additional cytoskeletal links are provided by vinculin (vn) and vasp which associate both with integrin complexes and with adherens junctions which form in the polarizing cells. The adherens junctions contain E-cadherin complexed with β-catenin and α-catenin. Other junction-associated proteins include MAGI-2 and DLG (discs large lethal), which in turn bind the tumour-suppressor proteins PTEN and APC to enhance their activity.

then that interfering with any component of this integrated 3D system not only compromises tissue function but also contributes to the process of deregulated growth control and ultimately tumorigenesis. Perhaps the most informative examples of how critical the ECM and an organized 3D structure are for maintaining differentiated tissue function and preventing tumour development and progression come from studies with tumour cells themselves which lack this organisation.

One system which very clearly demonstrates the strong correlation between loss of tissue organisation and tumour progression is a series of human breast epithelial cells which become increasingly tumorigenic as tissue architecture is progressively lost. The HMT-3522 cell series originated from a purified breast epithelial cell population isolated from a female with fibrocystic breast disease. Although these cells, referred to as S1, can remain relatively stable for over 500 passages, cells could alternatively be induced to undergo tumorigenic progression by selecting cells which would survive in the absence of essential growth factors. These factor-independent cells, designated T4-2, were found ultimately to form tumours when injected into nude mice, while the original S1 population which was maintained at a similar passage number but in defined media containing growth factors were stable and non-tumorigenic (Briand *et al*., 1987). More importantly, it was observed that whereas S1 cells maintained the ability to form organized, polarized alveolar structures which could undergo growth arrest in response to BM, the tumorigenic derivatives formed progressively disorganized and nonpolarized groups of cells which failed to growth arrest despite the presence of an intact BM. The disorganized T4-2 cells were found to have poorly organized nuclei, adherens junctions, f-actin and aberrant integrin localization and expression (Weaver *et al*., 1997) (see **Figure 3**). In fact, the extent to which epithelial cells respond to BM and elaborate organized structures is a relatively accurate means of predicting their degree of tumorigenicity with the disorganized cells being more tumorigenic (Petersen *et al*., 1992).

Aberrations in ECM–Integrins in Cancer

The disruption of normal architecture and morphology in the progressively tumorigenic cells described above could be directly related to the improper expression of high

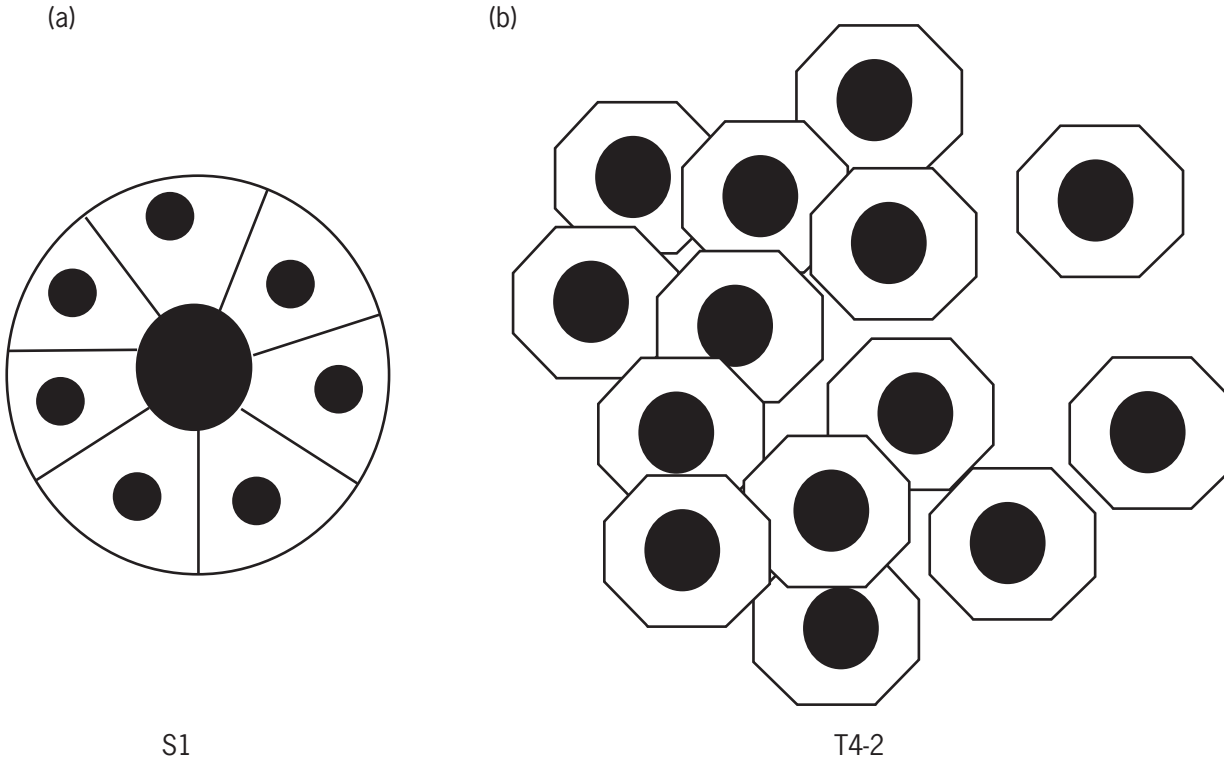


Figure 3 Cell morphology and tumorigenicity. (a) Schematic representation of normal, growth-arrested, breast epithelial cells (S1) organized into a acinar structures when cultured in a 3D basement membrane extracellular matrix. Note the basolateral localization of the nucleus in these polarized cells. (b) Tumorigenic derivatives of the normal breast epithelial cells (T4-2) form disorganized clusters when cultured in basement membrane extracellular matrix. The cells do not become polarized or form proper adherens junctions and continue to proliferate. (Adapted from Bissell *et al*., 1999.)

levels of a number of surface receptors including $\beta 1$ integrin. In addition, these cells display an aberration in surface expression of $\alpha 6\beta 4$ integrin which would otherwise normally encourage growth arrest. Together these defects might be expected not only to disrupt the formation of hemidesmsomes, but also to impair the establishment of adherens junctions and recruitment of β-catenin (Weaver *et al*., 1997).

An extensive literature documents a wide range of abnormal cell–ECM interactions in tumour cells. Aberrant integrin expression has been described in a number of breast tumour cells (see review by Zutter *et al*., 1998). In addition to overexpressing integrins, many tumour cells also lack expression of integrin receptors which might suppress growth, and studies aimed at reintroducing the missing integrins can often result in restoring growth arrest. For example, in tumorigenic intestinal epithelial cells, which lack expression of $\beta 4$ integrin, transfection of the $\beta 4$ integrin induced expression of the Cdk inhibitor $p21^{WAF1}$ and led to growth arrest (Clarke *et al*., 1995). Similarly, introduction of the $\alpha 2\beta 1$ collagen/laminin receptor missing from a poorly differentiated breast tumour line also induced growth arrest (Zutter *et al*., 1995).

Although this approach may compensate for a lack of a functional integrin, often integrins are expressed but not properly localized in the tumour cells. In the T4-2 cell series described above, $\beta 4$ integrin was expressed but not at the cell surface (Weaver *et al*., 1997) (**Figure 4**). It might be expected that integrins localized to different cellular locations may not be able to access their ligands, particularly if the ECM ligands have a restricted distribution, such as BM laminin. In addition, changing the localization of an integrin often results in the receptor adopting a different conformation, which might also prevent binding even if the ligand is accessible (Bishop *et al*., 1998).

In addition to deficiencies in ECM receptors, tumour cells often have altered ECMs. For example, many tumour cells are not able properly to synthesize or assemble BMs, which in turn are required for growth arrest (Petersen *et al*., 1992; Howlett *et al*., 1994). One of the more common reasons for a loss of organized ECM surrounding tumour cells is the finding that many tumour cells express high

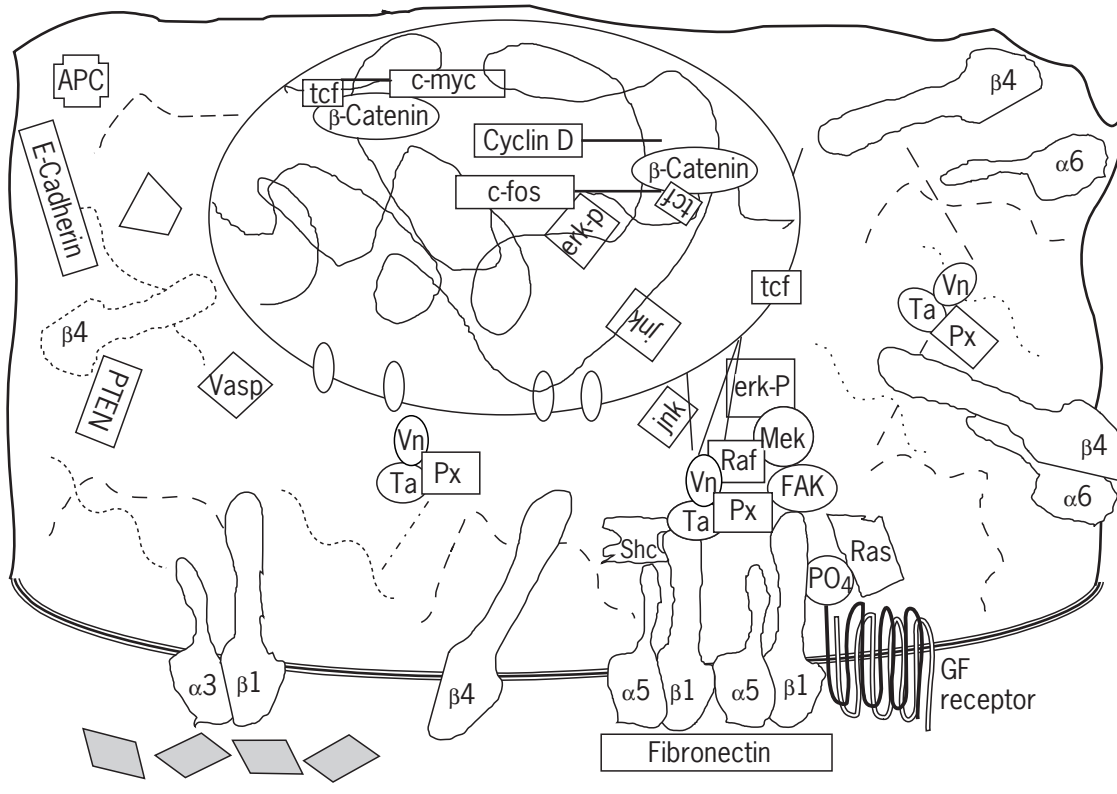


Figure 4 Structural disorganisation of tumour cells. Expression and subcellular localization of integrin subunits including $\alpha 6$ and $\beta 4$ is disrupted in tumorigenic epithelial cells. In addition the basement membrane may be degraded (grey diamonds) by high levels of matrix-degrading protease activity. Structural connections with the cytoskeleton (dashed lines) are also diminished. Polarity is absent and adherens junctions are not properly formed, allowing free β-catenin to translocate to the nucleus and activate transcription of growth-promoting genes including c-*myc*. Growth-promoting $\beta 1$ integrin heterodimers ($\alpha 5\beta 1$ is shown here) are present and continue to generate proliferative signals including phosphorylated erk (erk-P), which may translocate to the nucleus and activate additional growth-promoting genes. Growth factor receptors are also active and phosphorylated.

levels of various matrix-degrading proteinases (Barsky *et al*., 1983; Liotta and Stetler-Stevenson, 1990).

More recent evidence has shown that even in normal cells excessive and sustained proteolytic activity and BM degradation can directly contribute to tumour development and progression. Loss of BM by excessive proteolysis resulted in the transformation of epithelial cells to a mesenchymal phenotype (Lochter *et al*., 1997). The loss of the syndecan which also binds BM laminin also resulted in normal epithelial cells undergoing a similar transition to an anchorage-independent mesenchymal phenotype (Kato *et al*., 1995). These results together emphasize how BM not only promotes but also maintains and stabilizes the epithelial cell phenotype. Furthermore, in transgenic mice engineered to overexpress MMP-3 in differentiated mammary epithelial cells, an unusually high incidence of adenocarcinoma was observed in the BM-compromised mammary epithelial cells (Sternlicht *et al*., 1999). The chronic degradation of the BM in turn was permissive for a number of other genetic alterations which contributed to the tumorigenic phenotype. These findings again emphasize the requirement for BM in maintaining functionally differentiated tissues and for stabilizing the genetic information in the cells.

Apoptosis

It should be emphasized that loss of normal cell–ECM interaction does not immediately result in oncogenic transformation of most cells. The tumours which arise in the MMP-overproducing mice described above require chronic or sustained activity of the protease and loss of basement membrane before overt tumorigenesis is observed. As mentioned earlier, normal mammary epithelial cells initially respond to the loss or degradation of their ECM by undergoing programmed cell death or apoptosis within 48 h (Boudreau *et al*., 1995). This is viewed as a protective mechanism to help eliminate potentially dangerous cells which have lost the growth suppressive signals normally supplied by the BM.

Although complete loss of adhesive contact with any type of ECM can induce apoptosis in normally adhesive cells (Meredith *et al*., 1993; Frisch and Francis, 1994), the ability of a cell to respond to more subtle changes in their microenvironment provides a more effective means to eliminate potentially harmful cells. In fact, normal adherent epithelial cells can be induced to undergo apoptosis by simply disrupting their three-dimensional organisation (Boudreau *et al*., 1996). Even in polarized MDCK epithelial cells, a collagen gel overlay was found to induce apoptosis by simply disorientating cells (Tang *et al*., 1998). Unfortunately, many tumour cells that have compromised interactions with the ECM fail to recognize alterations in cell morphology and orientation, and are often resistant to apoptosis, even in the complete absence of adhesion, a phenomenon referred to as anchorage independence. Attempts to restore their sensitivity to apoptotic signals is a major anti-tumorigenic strategy (see the chapter on *Apoptosis*).

It is interesting that not all normal cells immediately respond to loss or degradation of their BM by undergoing apoptosis, nor do they proliferate uncontrollably. In cells which retain a high regenerative potential, such as vascular endothelium and hepatocytes, the loss of BM allows cells to re-enter the cell cycle and divide. Loss of BM will induce endothelial cells to undergo angiogenesis, while hepatocytes can be induced to proliferate and regenerate portions of the liver following a partial hepatectomy (Fausto, 2000). In these cases, however, the amount or degree of cell proliferation which occurs is limited because proliferating endothelial cells or hepatocytes immediately begin to resynthesize and deposit a new BM, which in turn induces the cells to re-establish a 3D tissue organisation and subsequently withdraw from the cell cycle (Kubota *et al*., 1988; Boudreau *et al*., 1997). Therefore, cells which do not undergo apoptosis when they re-enter the cell cycle can be directed to growth arrest and differentiate by re-establishing and responding to a functional BM.

Restoring a Cell's Balance of Surface Receptors and Interaction with the BM Can Reverse Tumorigenesis

Although attempting to restore a normal dialogue with the ECM might at first seem somewhat naive to apply to genetically destabilized tumour cells, in fact re-establishing a functional BM and 3D organisation, or restoring a tumour cell's ability to recognize and respond to BM, can in fact induce growth arrest in tumorigenic cells. Transfection of the tumorigenic breast epithelial cell line MDA-435 with the NM-23 tumour suppressor resulted in the ability of the cells to resynthesize and deposit a 3D BM and to form organized, acinar structures leading to growth arrest (Howlett *et al*., 1994). In the case of the progressively tumorigenic human breast cells mentioned earlier, blocking the excessive β1 integrin signalling allowed cells to revert completely to a nontumorigenic phenotype when they were cultured in the presence of a complete basement membrane. This reversion was accompanied by reorganisation into acinar-type structures, establishment of normal adherens junctions containing E-cadherin and associated α- and β-catenins, reorganisation of the actin cytoskeleton and growth arrest. Furthermore, blocking β1 integrin also attenuated the high levels of expression and activity of the EGF receptor normally found in these tumorigenic cells. Interestingly, simply blocking β1 integrin was not sufficient to downregulate EGF receptor expression or activity if the cells were also prevented from acquiring the corresponding 3D organisation, underscoring the contribution of the basement membrane to this reversion process (Wang

et al., 1998). Perhaps the most striking observation made using this system was that when β1 integrin was blocked and cells were allowed to resume their normal interaction with the basement membrane, the resulting 3D structure was sufficient to override a host of genetic alterations that had accumulated as cells progressed toward increasing tumorigenicity (Weaver *et al*., 1997).

CONCLUSION

Simply stated, cancer is a problem of tissue organisation. Although it has long been recognized that one of the hallmarks in diagnosing tumour cells is their aberrant morphology, both at the tissue, cell and nuclear level, the molecular consequences of this disorganisation are now beginning to be appreciated. By maintaining a proper 3D organisation through dynamic reciprocal interactions with their microenvironment, tissue structure can act in a dominant manner to override a host of genetic aberrations that may otherwise compromise normal tissue function.

A recent large-scale study emphasized an environmental rather than genetic basis of cancer based on findings that the incidence of identical twins developing similar cancers was almost nondetectable, whereas genetically unrelated individuals exposed to similar environments were more likely to do so (Lichtenstein *et al*., 2000). Although these observations apply to the relationship of an entire organism to its environment, it would appear that the same principles may also apply to tumour development at the cell and molecular level. Thus although cells may harbour genetic alterations from birth or acquire distinct genetic changes over a life span, it is how the cells interact with their immediate extracellular environment which appears to govern cell form, function and fate.

ACKNOWLEDGEMENTS

This work was supported by the Director, Office of Biological and Environmental Research (contract DE-AC03-76SF00098) of the U.S. Department of Energy and by NIH grants CA-57621 and CA-64786 to MJB, and by NIH grant CA-85249 to NB.

REFERENCES

Aplin, A. E., *et al*. (1999). Anchorage-dependent regulation of the Mitogen-activated protein kinase cascade by growth factors is supported by a variety of integrin alpha chains. *Journal of Biological Chemistry*, **274**, 31223–31228.

Barcellos-Hoff, M. H., *et al*. (1989). Functional differentiation and alveolar morphogenesis of primary mammary cultures on reconstituted basement membrane. *Development*, **105**, 223–235.

Barsky, S. H., *et al*. (1983). Loss of basement membrane components by invasive tumors but not by their benign counterparts. *Laboratory Investigation*, **49**, 140–147.

Bernfield, M., *et al*. (1999). Functions of cell surface heparan sulfate proteoglycans. *Annual Review of Biochemistry*, **68**, 729–777.

Bilder, D., *et al*. (2000). Cooperative regulation of cell polarity and growth by drosophila tumor suppressors. *Science*, **289**, 113–116.

Bishop, A. L. and Hall, A. (2000). Rho GTPases and their effector proteins. *Biochemical Journal*, **348**, 241–255.

Bishop, L. A., *et al*. (1998). Lack of intrinsic polarity in the ligand-binding ability of keratinocyte beta1 integrins. *Experimental Dermatology*, **7**, 350–361.

Bissell, M. J., *et al*. (1999). Tissue structure, nuclear organization and gene expression in normal and malignant breast. *Cancer Research (Supplement)*, **59**, 1757–1764.

Boudreau, N. and Bissell, M. J. (1996). Regulation of gene expression by the extracellular matrix. In: Comper, W. E. (ed.), Extracellular Matrix. Volume 2. Molecular Components and Interactions, 246–261 (Harwood Academic Publishers: Amsterdam).

Boudreau, N., *et al*. (1995). Suppression of ICE and apoptosis in mammary epithelial cells. *Science*, **267**, 891–893.

Boudreau, N., *et al*. (1996). Suppression of apoptosis by basement membrane requires three-dimensional tissue organization and withdrawal from the cell cycle. *Proceedings of the National Academy of Sciences of the USA*, **93**, 3509–3513.

Boudreau, N., *et al*. (1997). Hox D3 induces an angiogenic phenotype in endothelial cells. *Journal of Cell Biology*, **139**, 257–264.

Briand, P., *et al*. (1987). A new diploid non-tumorigenic human breast epithelial cell line isolated and propagated in chemically defined medium. *In Vitro Cell Developmental Biology*, **23**, 181–188.

Burack, W. R. and Shaw, A. S. (2000). Signal transduction; hanging on a scaffold. *Current Opinion in Cell Biology*, **12**, 211–216.

Calderwood, D. A., *et al*. (1999). The talin head domain binds to integrin beta subunit cytoplasmic tails and regulates integrin activation. *Journal of Biological Chemistry*, **274**, 28071–28074.

Chen, Q., *et al*. (1996). Integrin-mediated activation of MEK and mitogen-activated protein kinase is independent of Ras. *Journal of Biological Chemistry*, **271**, 18122–18127.

Chen, C. S., *et al*. (1997). Geometric control of cell life and death. *Science*, **276**, 1425–1428.

Clarke, A. S., *et al*. (1995). Activation of the p21 pathway of growth arrest and apoptosis by the β4 integrin cytoplasmic domain. *Journal of Biological Chemistry*, **270**, 22673–22676.

Cohen, A. R., *et al*. (1998). Human CASK/LIN-2 binds syndecan-2 and protein 4.1 and localizes to the basolateral membrane of epithelial cells. *Journal of Cell Biology*, **142**, 129–138.

Crawford, H. C., *et al*. (1999). The metalloproteinase matrilysin is a target of beta-catenin transactivation in intestinal tumors. *Oncogene*, **18**, 2883–2891.

Critchley, D. R. (2000). Focal adhesions – The cytoskeletal connection. *Current Opinion in Cell Biology*, **12**, 133–139.

D'Arcangelo, G., *et al*. (1995). A protein related to extracellular matrix proteins deleted in the mouse mutant reeler. *Nature*, **376**, 719–723.

Dimitratos, S. D., *et al*. (1999). Signaling pathways are focused at specialized regions of the plasma membrane by scafolding proteins of the MAGUK family. *Bioessays*, **21**, 912–921.

Dowling, J., *et al*. (1996). Beta 4 integrin is required for hemidesmosome formation, cell adhesion and cell survival. *Journal of Cell Biology*, **134**, 559–572.

Ekblom, P. (1996). Receptors for laminins during epithelial morphogenesis. *Current Opinion in Cell Biology*, **8**, 700–706.

Ervasti, J. M., and Campbell, K. P. (1993). A role for the dystrophin–glycoprotein complex as a transmembrane linker between laminin and actin. *Journal of Cell Biology*, **122**, 809–823.

Evan, G. I., *et al*. (1992). Induction of apoptosis in fibroblasts by c-myc protein. *Cell*, **69**, 119–128.

Fanning, A. S., *et al*. (1998). The tight junction protein ZO-1 establishes a link between the transmembrane protein occludin and the actin cytoskeleton. *Journal of Biological Chemistry*, **273**, 29745–29753.

Fausto, N. (2000). Liver regeneration. *Journal of Hepatology,* **32**, 19–31.

Frisch, S. M. and Francis, H. (1994). Disruption of epithelial cell–matrix interactions induces apoptosis. *Journal of Cell Biology*, **124**, 619–626.

Fuchs, E., *et al*. (1997). Integrators of epidermal growth and differentiation; distinct functions for β1 and β4 integrins. *Current Opinion in Genetics and Development*, **7**, 672–682.

Gianelli, G., *et al*. (1997). Induction of cell migration by matrix metalloproteinase-2 cleavage of laminin-5. *Science*, **277**, 225–228.

Gilmore, A. P. and Burridge, K. (1996). Regulation of vinculin binding to talin and actin by phosphatidyl-inositol-4,5-bisphosphate. *Nature*, **381**, 531–535.

Gonzales, M., *et al*. (1999). A cell signal pathway involving laminin-5, alpha 3 beta 1 integrin and mitogen activated protein kinase can regulate epithelial cell proliferation. *Molecular Biology of the Cell*, **10**, 259–270.

Gumbiner, B. M. (1996). Cell adhesion: the molecular basis of tissue architecture and morphogenesis. *Cell*, **84**, 345–357.

He, T. C., *et al*. (1998). Identification of c-MYC as a target of the APC pathway. *Science*, **281**, 1509–1512.

Hildebrand, J. D. and Soriano, P. (1999). Shroom, a PDZ domain-containing actin-binding protein, is required for neural tube morphogenesis in mice. *Cell*, **99**, 485–497.

Hill, C. S. and Treisman, R. (1995). Transcriptional regulation by extracellular signals: mechanisms and specificity. *Cell*, **80**, 199–211.

Hodivala, K. J. and Watt, F. M. (1994). Evidence that cadherins play a role in the downregulation of integrin expression that occurs during keratinocyte terminal differentiation. *Journal of Cell Biology*, **124**, 589–600.

Howlett, A. R., *et al*. (1994). A novel function for the nm-23H1 gene; overexpression in human breast carcinoma cells leads to the formation of basement membranes and growth arrest. *Journal of National Cancer Institute*, **86**, 1838–1844.

Hseuh, Y.-P., *et al*. (2000). Nuclear translocation and transcription regulation by the membrane associated guanylate cycles kinase CASK/LIN-2. *Nature*, **404**, 298–302.

Hughes, P. E. and Pfaff, M. (1998). Integrin affinity modulation. *Trends in Cell Biology*, **8**, 359–364.

Hughes, P. E., *et al*. (1997). Suppression of integrin activation; a novel function of a Ras/Raf initiated MAP-kinase pathway. *Cell*, **88**, 521–530.

Hynes, R. O. and Zhao, Q. (2000). The evolution of cell adhesion. *Journal of Cell Biology*, **150**, F89–F86.

Joshi, P., *et al*. (1993). Endothelial cells adhere to the RGD domain and the fibrinogen-like terminal knob of tenascin. *Journal of Cell Science*, **106**, 389–400.

Kachinsky, A. M., *et al*. (1999). A PDZ-containing scaffold related to the dystophin complex at the basolateral membrane of epithelial cells. *Journal of Cell Biology*, **145**, 391–402.

Kato, M., *et al*. (1995). Loss of cell surface syndecan-1 causes epithelia to transform into anchorage-independent mesenchyme-like cells. *Molecular Biology of the Cell*, **6**, 559–576.

Kheradmand, F., *et al*. (1998). Role of Rac1 and oxygen radicals in collagenase-1 expression induced by cell shape change. *Science*, **280**, 898–902.

Kubota, Y., *et al*. (1988). Role of laminin and basement membrane in the morphological differentiation of human endothelial cells into capillary-like structures. *Journal of Cell Biology*, **107**, 1589–1598.

Lichtenstein, P., *et al*. (2000). Environmental and heritable factors in the causation of cancer – Analyses of cohorts of twins from Sweden, Denmark, and Finland. *New England Journal of Medicine*, **343**, 78–85.

Liotta, L. A. and Stetler-Stevenson, W. G. (1990). Metalloproteinases and cancer invasion. *Seminars in Cancer Biology*, **1**, 99–106.

Lochter, A. *et al*. (1997). Matrix metalloproteinase stromelysin-1 triggers a cascade of molecular alterations that leads to stable epithelial-to-mesenchymal conversion and a premalignant phenotype in mammary epithelial cells. *Journal of Cell Biology*, **139**, 1861–1872.

Luschnig, S., *et al*. (2000). The *Drosophila* SHC adaptor protein is required for signalling by a subset of receptor tyrosine kinases. *Molecular Cell*, **5**, 231–241.

Madri, J. A. and Williams, S. K. (1983). Capillary endothelial cell culture: phenotypic modulation by matrix components. *Journal of Cell Biology*, **97**, 153–165.

Matsumine, A., *et al*. (1996). Binding of APC to the human homolog of the *Drosophila* discs large tumor suppressor protein. *Science*, **272**, 1020–1023.

Meredith, J. E., Jr, *et al*. (1993). The extracellular matrix as a cell survival factor. *Molecular Biology of the Cell*, **4**, 953–961.

Miyamoto, S., *et al*. (1996). Integrins can collaborate with growth factors for phosphorylation of receptor tyrosine kinases and MAP kinase activation: roles of integrin aggregation and occu-pancy of receptors. *Journal of Cell Biology*, **135**, 1633–1642.

Muschler, J., *et al*. (1999). Division of labor among the alpha6-beta4 integrin, beta1 integrins, and an E3 laminin receptor to signal morphogenesis and beta-casein expression in mammary epithelial cells. *Molecular Biology of the Cell*, **10**, 2817–2828.

Nievers, M. G., *et al*. (2000). Formation of hemidesmosome-like structures in the absence of ligand binding by the (alpha)6 (beta)4 integrin requires binding of HD1/plectin to the cytoplasmic domain of the (beta)4 integrin subunit. *Journal of Cell Science*, **113**, 963–973.

Novak, A., *et al*. (1998). Cell adhesion and the integrin-linked kinase regulate the LEF-1 and beta-catenin signalling pathways. *Proceedings of the National Academy of Sciences of the USA*, **95**, 4374–4379.

Petersen, O. W., *et al*. (1992). Interaction with the basement membrane serves to rapidly distinguish growth and differentiation pattern of normal and malignant human breast epithelial cells. *Proceedings of the National Academy of Sciences of the USA*, **89**, 9064–9068.

Pierschbacher, M. D. and Ruoslahti, E. (1984). Cell attachment activity of fibronectin can be duplicated by small synthetic fragments of the molecule. *Nature*, **309**, 30–33.

Poznansky, M. C., *et al*. (2000). Efficient generation of human T cells from a tissue-engineered thymic organoid. *Nature Biotechnology*, **18**, 729–734.

Prockop, D. J. and Kivirikko, K. I. (1995). Collagens: molecular biology, diseases, and potentials for therapy. *Annual Review of Biochemistry*, **64**, 403–434.

Pujuguet, P., *et al*. (2000). Nidogen-1 regulates laminin-1-dependent mammary-specific gene expression. *Journal of Cell Science*, **113**, 849–858.

Roskelley, C. D., *et al*. (1994). Extracellular matrix-dependent tissue-specific gene expression in mammary epithelial cells requires both physical and biochemical signal transduction. *Proceedings of the National Academy of Sciences of the USA*, **91**, 12378–12382.

Saksela, O. and Rifkin, D. B. (1990). Release of basic fibroblast growth factor–heparan sulfate complexes from endothelial cells by plasminogen activator-mediated proteolytic activity. *Journal of Cell Biology*, **110**, 767–775.

Schmidhauser, C., *et al*. (1990). Extracellular matrix and hormones transcriptionally regulate bovine 5′ sequences in stably transfected mouse mammary cells. *Proceedings of the National Academy of Sciences of the USA*, **87**, 9118–9122.

Schneller, M., *et al*. (1997). Alphav beta3 integrin associates with activated insulin and PDGF beta receptors and potentiates the biological activity of PDGF. *EMBO Journal*, **16**, 5600–5607.

Schoenwaelder, S. M. and Burridge, K. (1999). Bidirectional signalling between the cytoskeleton and integrins. *Current Opinions in Cell Biology*, **11**, 274–286.

Schwarzbauer, J. (1999). Basement membranes: putting up the barriers. *Current Biology*, **9**, R242–R244.

Shtutman, M., *et al*. (1999). The cyclin D1 gene is a target of the beta-catenin/LEF-1 pathway. *Proceedings of the National Academy of Sciences of the USA*, **96**, 5522–5527.

Sonnenberg, A., *et al*. (1991). Integrin alpha 6 beta 4 complex is located in hemidesmosomes suggesting a major role in epidermal cell-basement membrane adhesion. *Journal of Cell Biology*, **113**, 907–917.

Steinbock, F. A. and Wiche, G. (1999). Plectin; a cytolinker by design. *Biological Chemistry*, **380**, 151–158.

Sternlicht, M. D., *et al*. (1999). The stromal proteinase MMP3/stromelysin-1 promotes mammary carcinogenesis. *Cell*, **98**, 137–146.

Streuli, C. H. and Bissell, M. J. (1990). Expression of extracellular matrix components is regulated by substratum. *Journal of Cell Biology*, **110**, 1405–1415.

Streuli, C. H., *et al*. (1991). Control of mammary epithelial differentiation: basement membrane induces tissue-specific gene expression in the absence of cell–cell interaction and morphological polarity. *Journal of Cell Biology*, **115**, 1383–1395.

Streuli, C. H., *et al*. (1995). Laminin mediates tissue-specific gene expression in mammary epithelia. *Journal of Cell Biology*, **129**, 591–603.

Tamura, M., *et al*. (1999). PTEN gene and integrin signalling in cancer. *Journal of the National Cancer Institute*, **91**, 1820–1828.

Tang, M. J., *et al*. (1998). Collagen gel overlay induces apoptosis of polarized cells in cultures; disoriented cell death. *American Journal of Physiology*, **275**, C921–C931.

Van der Neut, R., *et al*. (1996). Epithelial detachment due to absence of hemidesmosomes in integrin beta 4 null mice. *Nature Genetics*, **13**, 366–369.

Vasioukhin, V., *et al*. (2000). Directed actin polymerization is the driving force for epithelial cell–cell adhesion. *Cell*, **100**, 200–219.

Vu, T. H., *et al*. (1998). MMP-9/gelatinase B is a key regulator of growth plate angiogenesis and apoptosis of hypertrophic chondrocytes. *Cell*, **93**, 411–422.

Wang, F., *et al*. (1998). Reciprocal interactions between beta 1 integrin and epidermal growth factor receptor in three-dimensional basement membrane breast cultures; a different perspective in epithelial biology. *Proceedings of the National Academy of Sciences of the USA*, **95**, 14821–14826.

Wary, K. K., *et al*. (1996). The adaptor protein Shc couples a class of integrins to the control of cell cycle progression. *Cell*, **87**, 733–743.

Wary, K. K., *et al*. (1998). A requirement for caveolin-1 and associated kinase Fyn in integrin signalling and anchorage-dependent cell growth. *Cell*, **94**, 625–634.

Weaver, V. M., *et al*. (1996). The importance of the microenvironment in breast cancer progression: recapitulation of

mammary tumorigenesis using a unique human mammary epithelial cell model and a three-dimensional culture assay. *Biochemistry and Cell Biology*, **74**, 833–851.

Weaver, V. M., *et al*. (1997). Reversion of the malignant phenotype of human breast cells in three-dimensional culture and in vivo by integrin blocking antibodies. *Journal of Cell Biology*, **137**, 231–245.

Woodley, D. T., *et al*. (1983). Interactions of basement membrane components. *Biochimica Biophysica Acta*, **761**, 278–283.

Wu, X., *et al*. (2000). Evidence for regulation of the PTEN tumor suppressor by a membrane-localized multi-PDZ domain containing scaffold protein MAGI-2. *Proceedings of the National Academy of Sciences of the USA*, **97**, 4233–4238.

Yamada, K. M. and Geiger, B. (1997). Molecular interactions in cell adhesion complexes. *Current Opinion in Cell Biology*, **9**, 76–85.

Yokosaki, Y., *et al*. (1994). The integrin alpha 9 beta 1 mediates cell attachment to a non-RGD site in the third fibronectin type III repeat of tenascin. *Journal of Biological Chemistry*, **269**, 26691–26696.

Zhu, X. and Assoian, R. K. (1995). Integrin-dependent activation of MAP kinase: a link to shape-dependent cell proliferation. *Molecular Biology of the Cell*, **6**, 273–282.

Zhu, X., *et al*. (1996). Adhesion-dependent cell cycle progression linked to the expression of cyclin D1, activation of cyclinE–cdk2, and phosphorylation of the retinoblastoma protein. *Journal of Cell Biology*, **133**, 391–403.

Zutter, M. M., *et al*. (1995). Re-expression of the alpha 2 beta 1 integrin abrogates the malignant phenotype of breast carcinoma cells. *Proceedings of the National Academy of Sciences of the USA*, **92**, 7411–7415.

Zutter, M. M., *et al*. (1998). Altered integrin expression and the malignant phenotype: the contribution of multiple integrated integrin receptors. *Journal of Mammary Gland Biology and Neoplasia*, **3**, 191–200.

FURTHER READING

Bissell, M. J., *et al*. (1982). Dynamic reciprocity – How does the extracellular matrix direct gene expression? *Journal of Therapeutic Biology*, **99**, 31–68.

Bissell, M. J., *et al*. (1999). Tissue structure, nuclear organization and gene expression in normal and malignant breast. *Cancer Research (Supplement)*, **59**, 1757–1764.

Boudreau, N. J. and Jones, P. L. (1999). Extracellular matrix and integrin signalling: the shape of things to come. *Biochemical Journal,* **339**, 481–488.

Erickson, A. C. and Couchman, J. R. (2000). Still more complexity in mammalian basement membranes. *Journal of Histochemistry and Cytochemistry*, **48**, 1291–1306.

Giancotti, F. and Ruoslahti, E. (1999). Integrin signaling. *Science*, **285**, 28–32.

Werb, Z. (1997). ECM and cell surface proteolysis: regulating cellular ecology. *Cell*, **91**, 439–442.

Chapter 16

Invasion and Metastasis

Lance A. Liotta and Cloud P. Paweletz
National Cancer Institute, National Institutes of Health, Bethesda, MD, USA

CONTENTS

INTRODUCTION

We usually diagnose and treat cancer when it is too late for local therapeutic strategies and most patients already harbour occult or overt metastasis. In fact, 30% of patients are diagnosed with overt metastases, while an additional 30–40% appear metastasis free during initial diagnosis, but harbour occult metastasis instead. Unfortunately for the cancer patient, the existence of metastasis greatly reduces the success of current surgical, chemotherapy and radiotherapy strategies (Astrow, 1994).

During the course of the disease most patients suffer from metastases at multiple sites, not all of which may be occurring at the same time. Furthermore, metastases have the potential to metastasize further: the presence of large identifiable metastases in a given organ is frequently accompanied by a greater number of micrometastases. And lastly, formation of metastatic colonies is a continuous process that commences early in the growth of the tumour and increases with time.

Cancer metastasis is a highly complex process that involves the deregulation of interacting proteins and genes that are responsible for invasion, angiogenesis, circulation of tumour cells in blood vessels, colonization at secondary organ sites, and finally evasion of host defence systems (**Figure 1**). Metastatic dissemination via the lymphatics and the vascular systems is the culmination, and end stage, of a disease process that evolves over 5–20 years. During most of that time period, microscopic lesions are progressing through a series of hyperproliferative and premalignant states through to carcinoma *in situ* presenting a phenomenally long time period for initial screening and treatment of cancer (**Figure 2**). We know that for most types of human solid neoplasia, microinvasive carcinoma emerges from a carcinoma *in situ* precursor lesion (Gallager and Martin, 1969; Zhuang *et al*., 1995). Preventing the transition from premalignant to invasive/metatastic carcinoma is a major goal for cancer chemoprevention. In order to reach this goal, it will be important to identify molecular targets that are causally associated with the acquisition of cancer invasion.

The malignant tumour is a state that emerges from a host microecology that actually participates in the selection and expansion of the most aggressive neoplastic cells (Aboseif *et al*., 1999; Hanahan and Weinberg, 2000) (**Figure 3**). Instead of being autonomous, malignant cells communicate extensively with other cells and the extracellular matrix. Sustained proliferation is not unique to cancer cells. In fact, growth pressure alone will not cause a neoplasm to be malignant. Malignant tumour cells migrate across tissue boundaries and have the capacity to survive and grow among 'foreign' cell populations. The true life-threatening behaviour of malignant cancer cells is their propensity to infiltrate and usurp the 'sovereignty' of host tissue societies.

Normal cellular physiology is a tightly regulated process with positive and negative feedback loops that decides whether a cell should differentiate, divide, adapt or commit apoptosis. Genetic changes, such as activation of oncogenes, increased production of growth factors, loss of growth inhibitory cytokines or loss of function of tumour-suppressor genes may result in an imbalance of growth regulation, leading to uncontrolled proliferation. However, unrestrained growth by itself does not cause metastasis, and additional genetic mutations over and above those that cause uncontrolled proliferation are needed. Genetic defects in the cancer cell translate into proteomic derangements in signal transduction pathways. The result of such derangements is a persistent pathological communication state between the tumour cell and the host. Tumour cells that successfully invade and metastasize are selected out because somatic genetic progression has resulted in an altered communication circuit that continues to call up and

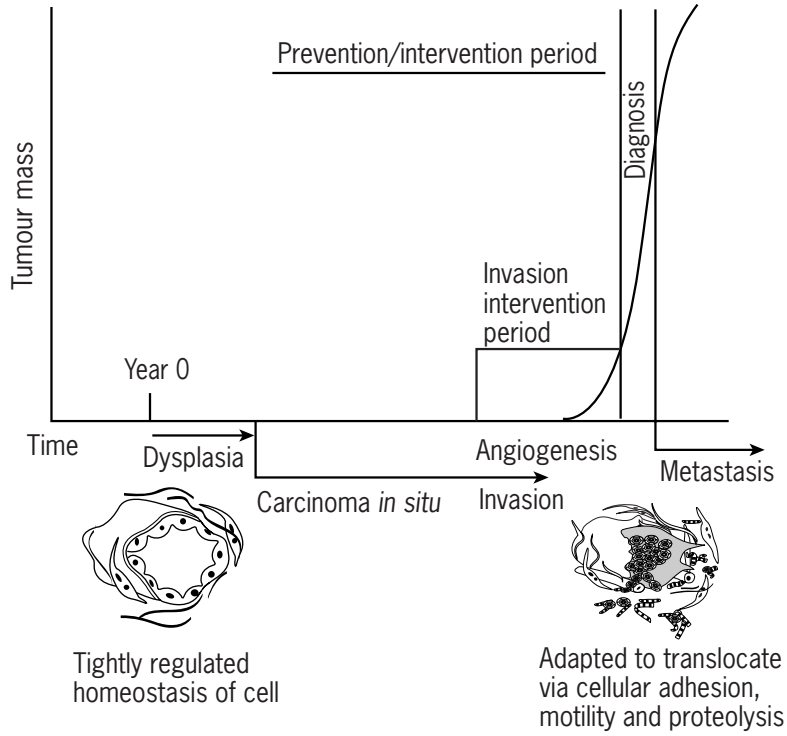


Figure 1 Intervention period for carcinogenesis. Theoretical improvement for intervention strategies as a function of time. Diagnosis of cancer usually is performed late into the disease, when most patients already present with occult metastasis. However, progression of cancer from dysplasia to metastatic dissemination may extend as far back as 10 years, providing a much larger window for intervention strategies before metastasis occurs.

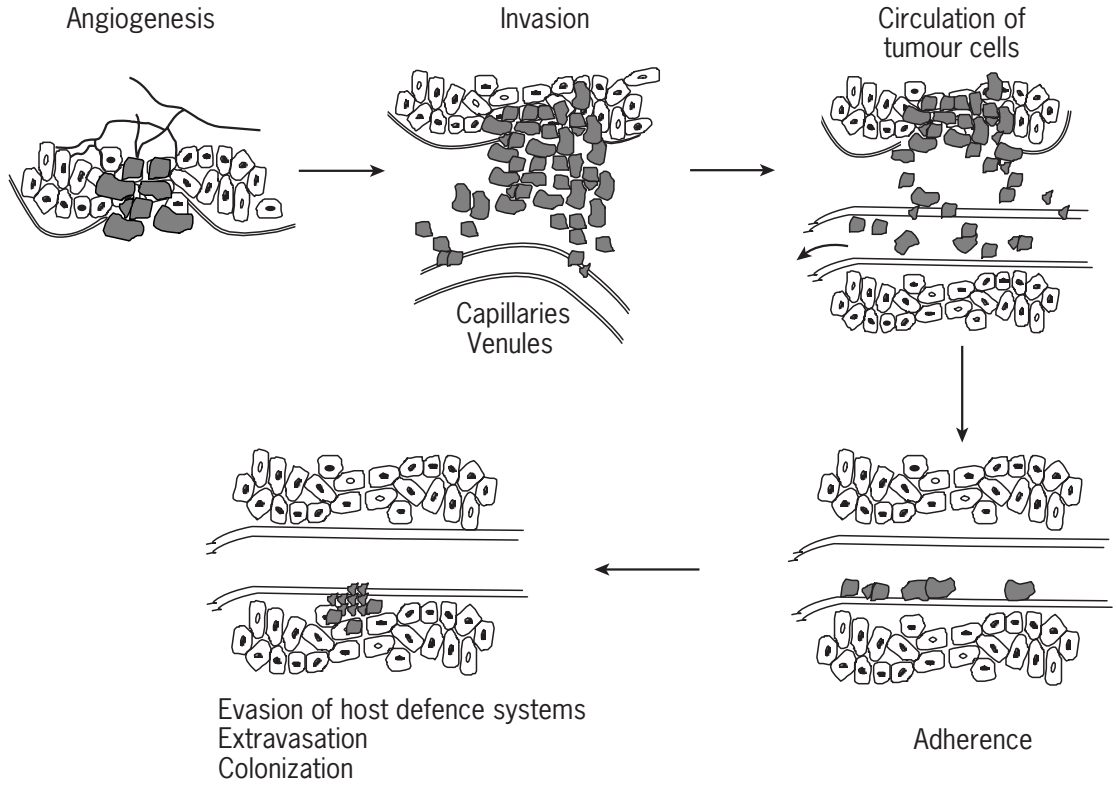


Figure 2 Pathogenesis of metastasis. Cancer metastasis is a highly complex process that involves the deregulation of interacting proteins and genes that are responsible for angiogenesis, invasion, circulation of tumour cells in blood vessels, colonization at secondary organ sites and finally evasion of host defence systems.

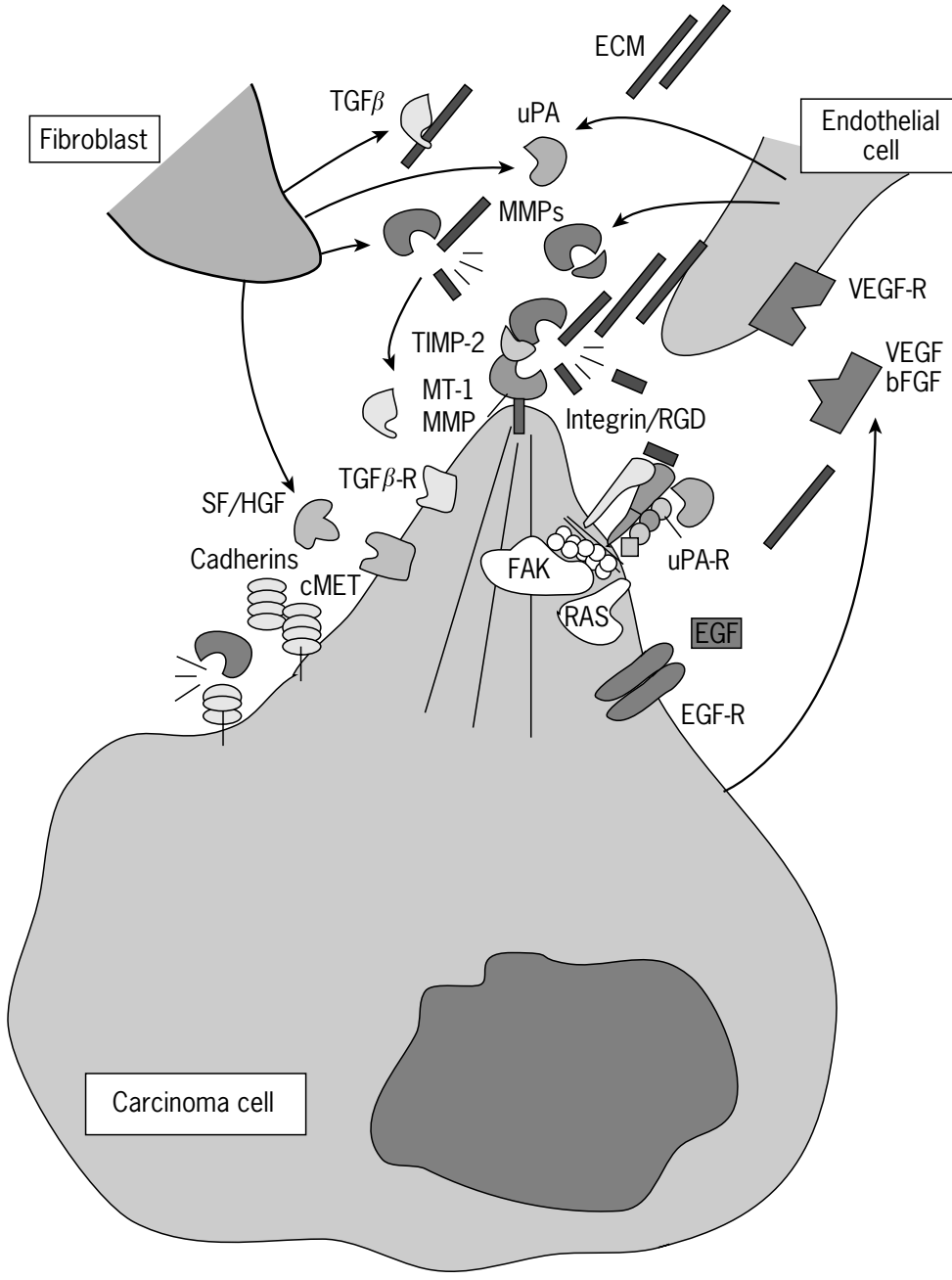


Figure 3 Microecology at the invasion front. Example mediators are shown. Motility and invasion is a bidirectional process. Fibroblasts produce chemoattractants including scatter factor/hepatocyte growth factor (SF/HGF) which stimulates motility of tumour cells by binding to the Met receptor (c-Met). Tumour cells produce angiogenesis factors including vascular endothelial growth factor (VEGF) and basic fibroblast growth factor (bFGF), which bind to receptors on stromal vascular cells causing increased vascular permeability, endothelial proliferation, migration, and invasion. Fibroblasts and endothelial stromal cells elaborate latent enzymes including matrix metalloproteinases (MMPs) and urokinase plasminogen activator (uPA) which dock on the surface of the carcinoma invadopodia and become activated, thereby degrading the ECM, and clearing a pathway. ECM degradation releases bound growth factors such as transforming growth factor beta (TGF-β) and epidermal growth factor (EGF), which bind to cognate receptors (TGFβ-R and EGF-R) on the carcinoma cell. ECM proteolysis also exposes cryptic RGD sites which are recognized by integrins (integrin/RGD). Cross-talk between signal pathways within the carcinoma cells links motility, proliferation and prosurvival.

support invasion and survival. At the biochemical level, the mechanism for initial invasion may parallel, or be similar to, that used by nonmalignant cells that traverse tissue boundaries. Using specialized cell models and new array technology, investigators are uncovering the interplay of specific signal transduction molecules that mediate the malignant state (Clark *et al*., 2000; Paweletz *et al*., 2001). Using protein microarray technology, the authors demonstrated that activation of PI3 kinase substrates and suppression of apoptosis are early events in the microenvironment of prostate cancer evolution. This analysis provided direct quantitative evidence that suppression of apoptosis in human PIN and invasive prostate cancer may be associated with phosphorylation of Akt and its substrate GSK3-β. Moreover, the authors verified that downstream components of the apoptotic cascade (cleaved and noncleaved caspase-7, and also cleaved and noncleaved PARP) are shifted toward prosurvival messages at the cancer invasion front. High-grade PIN exhibited a lower level of phospho-ERK compared with normal-appearing epithelium. Invading carcinoma cells contained phospho-ERK levels that were even more reduced compared with PIN. These data are in keeping with known prosurvival pathways, which emanate from Akt through its substrates. Augmentation of the ratio of phosphorylated Akt to total Akt will suppress downstream apoptosis pathways through intermediate substrates such as GSK3β. Reduction in apoptosis will shift the balance of cell birth and death rates favouring the observed accumulation of cells within the epithelial gland. Prosurvival messages are required for migrating cells to resist the proapoptotic signals that take place during the disruption of integrin-mediated adhesion to extracellular matrix molecules. In parallel, transient ERK activation and augmentation of prosurvival pathways may be associated with cellular migration. Activation of Akt, a substrate of PI3K, can therefore promote cell motility and survival as the invading cancer cells leave the gland, invade the stroma and metastasize.

As the tumour cell invader enters foreign soil, it appropriates the local growth signals and ignores its instructions to undergo apoptosis. Tumour cells escaping the primary tumour mass respond to host signals that call up the capacity for motility (Jo *et al*., 2000), survival (Frisch and Francis, 1994), and proliferation (Brown and Giavazzi, 1995; Kohn and Liotta, 1995). Host cells contribute enzymes and cytokines that aid the tumour cell. For example, tumour cells can penetrate host cellular and extracellular barriers with the help of degradative enzymes produced by the host cells, but locally activated by the tumour (Chambers and Matrisian, 1997; Werb, 1997). The presence of the new malignant cells within the invaded host tissue is associated with a local reorganisation of the stroma, blood vessels, lymphatics and epithelial morphology. We may incorrectly assume that the host response to the tumour is designed to repel an invader. Instead, the host reaction to the tumour cells may be simply an indifferent accommodation process (Wernert, 1997). Regardless, the multifocal disruption and damage associated with the modified microenvironment are ultimately lethal to the host.

Distributions of metastases vary widely with histological type and anatomical location of the primary tumour (**Table 1** and **Table 2**). For some tumour types a frequent

Table 1 Organ preferences of metastasis in some human and animal selected models[a]

Tumour system subline	Lung	Liver	Brain	Ovary	Spleen	Lymph node
Murine B16 melanoma (*i.v. or i.c.*)						
B16–F1	+	+/−	+/−	+/−	−	+/−
B16–F10	++++	+/−	−	+	+/−	+/−
B16–F15b	+++	−	+++	+	+/−	+/−
Murine RAW117 large cell lymphoma (*i.v. or s.c.*)						
Raw117-P	+/−	+/−	−	−	+/−	−
Raw117-H10	+/−	++++	−	−	++	−
Murine MTI mammary carcinoma (*i.v. or s.c.*)						
TC3	+++	+	−	+/−	−	+
Chicken MD lymphoma (*i.v.*)						
AL-2	−	++++	−	+/−	−	−
AL-3	−	+	−	++++	−	−
Human A375 melanoma (*i.v. in nude mice*)						
A375-P	+/−	+/−	−	−	−	+/−
A375-SM	++	+/−	−	+/−	−	+
Human PC-3 prostatic carcinoma (*i.v. in athymic mice*)						
PC-3-125-lN	++++	−	−	−	−	−
PC-3-1-LN	++++	+	−	+	+/−	+++

[a]Metastasis: −, none; +/−, sometimes; +, few; ++, moderate; +++, many; ++++, large numbers and heavy tumour burden; i.v., intravenous; i.c., intracaecum. (Adapted from Nicolson, 1998.)

Table 2 Frequency of metastatic sites

Site	10%	10–30%	30–50%	50–70%	70%
Breast		Kidney, skin, brain	Adrenal	Liver, bone, lung	Lymph nodes
Bladder	Brain, skin	Kidney,bone	Adrenal, lung		
Cervix	Brain, skin	Kidney, bone	Adrenal, lung		
Colorectum	Skin	Brain, kidney, lung	Bone, adrenal, liver	Lymph nodes	
Kidney	Skin, bone	Brain, kidney	Liver	Lung	
Lung	Lung	Kidney, distant nodes	Adrenal, brain	Bone	Liver, local lymph nodes
Melanoma		Kidney	Adrenal, brain, bone, skin	Lung, liver nodes	
Ovary	Brain, skin, kidney	Bone, adrenal	Lung, liver nodes		
Prostate	Brain, skin	Kidney, adrenal, liver, lung	Bone, nodes		

(Adapted from Weiss, 1992, and references therein.)

organ location of distant metastases appears to be the first capillary bed encountered. Lung metastases from sarcoma or colorectal cancer dissemination to the liver can be considered examples of this kind of metastasis. In the gynaecological tumours, distant metastases are seen in two forms: serosal dissemination, such as liver capsule metastases from ovarian cancer, and capillary-associated dissemination, such as lung parenchymal disease. However, not all metastases can be explained by anatomical considerations alone, such as metastasis to the ovary from breast carcinoma or dissemination to the liver from ocular melanomas, and hence must be considered as organ tropism.

The organ preference for metastatic colonization is heavily influenced by communications between the circulating tumour cell and the target host tissue. Various molecular mechanisms attempt to explain preferential organ distribution during metastasis. First, cancers shed equal numbers of tumour cells into the vascular system, and thus tumour cells disseminate equally to all organs, but only grow preferentially in some specific organs. For example, the insulin-like-growth factors are present in liver and lung and have been implicated in growth and motility for breast and lung carcinoma. Second, circulating tumour cells may adhere preferentially to the endothelial luminal surface. Nicolson *et al*. have identified endothelial surface antigens that may mediate preferential adhesion of circulating tumour cells. Lastly, circulating tumour cells may respond to soluble factors diffusing locally out of target organs. Chemokines are growth factor-like molecules which bind to G-protein coupled receptors. Circulating leukocytes and stem cells are known to use chemokine mechanisms to home in on specific organs. They induce leucocytes to adhere tightly to endothelial cells and migrate toward the highest concentration of chemokine. Since this behaviour seemed identical with that required for metastatic tumour cells, Mueller *et al*. hypothesized that tumour cells may co-opt the same chemokines to direct metastatic organ preference. They conducted a comprehensive survey of known chemokines and found a receptor–ligand pair (CCR4 and CXCL12) which fit the profile expected for breast cancer metastasis homing to bone, lung and liver. *In vitro*, the CXCL12 ligand stimulated breast cancer cells to carry out the basics of invasion: pseudopodial protrusion, directed migration and penetration of extracellular matrix barriers. *In vivo*, using animal models, the authors blocked metastasis to CXCL12-rich lung tissue by treatment with a neutralizing antihuman CXCR4 monoclonal antibody.

ANGIOGENESIS

The transition from normal epithelium to invasive carcinoma is preceded by, or is concomitant with, activation of local host vascular channels and stromal fibroblasts. Stromal cell activation and recruitment by the tumour cell promotes premalignant cell transformation and malignant invasion. For example, during the transition from *in situ* to invasive carcinoma, disorganisation and disruption of the periglandular basement membrane and a local neovascular 'blush' can precede frank malignant conversion (Guidi *et al*., 1997). Neovascularization offers a portal for dissemination. Locally activated vascular channels at the invasive edge of the tumour are highly permeable and offer a reduced barrier for intravasation (Dvorak *et al*., 1995). A variety of molecules have been found to mediate angiogenesis *in vitro* and *in vivo*. Among these are basic fibroblast growth factor (bFGF), angiogenin, vascular permeability factor (VPF) and tumour necrosis factor α and β (TNF-α, TNF-β) (Folkman and Klagsbrun, 1987).

Neovascularization is a form of physiological invasion (Fidler and Ellis, 1994; Folkman, 1995). Endothelial cells migrate, elaborate degradative enzymes and traverse extracellular matrix barriers along a chemotropic gradient emanating from the tumour cells. Physiological and malignant invasion employ similar molecular mechanisms.

The difference is that malignant invasion persists. Neovascularization, wound healing and neurite outgrowth during embryogenesis are examples of physiological invasion. In response to trophic signals, vascular cells, wounded epithelial sheets or neurites will migrate, penetrate tissue barriers and establish appropriate new anastomoses (Kohn and Liotta, 1995; Carmeliet and Jain, 2000). However, when the trophic signal is removed or the injury is repaired, physiological invasion ceases. Malignant cells perpetually stimulate host stromal and vascular cells to conduct physiological invasion. Promotion of the local invasive environment creates a permissive field for the malignant cell.

Rapid-growing tumours are capable of shedding up to millions of tumour cells into the vascular circulation by angiogenesis and invasion alone (Liotta *et al*., 1974). Furthermore, experimental studies show that less than 0.05% of circulating tumour cells are successful in initiation of metastatic colonies, making metastatic dissemination a highly inefficient process (Liotta *et al*., 1974; Nicolson, 1991). These studies are also clinically validated by the observation that circulating tumour cells are detected in patients who never form a metastasis.

INVASION

Invasion is the active translocation of neoplastic cells across tissue boundaries and through host cellular and extracellular matrix barriers. Invasion is dependent on the coordinated activity of a series of interacting proteins extending from the inside of the cell to the cell surface and the adjacent host cellular and extracellular microenvironment (**Figure 3**). Cellular adhesion, local proteolysis and motility are the triad of necessary functions that mediate invasion. While invasion is not directly caused by growth pressure, nevertheless the genetic and proteomic deregulation that causes invasion can effect neoplastic proliferation indirectly by promoting cancer cell survival.

CELLULAR ADHESION

Normal tissue morphology and organ architecture are tightly regulated by a communication reciprocity between the tissue cells and the extracellular matrix (ECM) and/or basement membrane. Adhesion is more than just anchoring – it is dynamic solid phase signal transduction (Fashena and Thomas, 2000). Receptors sensing changes in the cell-ECM state provide extracellular signals, which trigger corresponding intracellular signal transduction pathways that regulate proliferation, differentiation and migration. Receptors involved in sensing the ECM include growth factor and hormone receptors which recognize ligands solubilized from the ECM and receptors which directly bind to the solid-phase molecules of the ECM. The latter include the integrins, the cell adhesion molecules (CAMs) and cadherins (**Figure 3**).

The integrin family of cell surface extracellular matrix proteins consists of heterodimeric units, designated α (140 kDa) and β (95 kDa). An important aspect of this family is that integrins can exist in a binary 'on' or 'off' state, thereby selectively changing affinity for corresponding ligands (Juliano and Haskill, 1993). Activation of integrins has been shown to be involved in cell migration, cell proliferation and metastatic dissemination. Loss of sustained integrin stimulation (Frisch and Ruoslahti, 1997) has been associated with apoptosis. Integrin interactions with intracellular and extracellular molecules determines function. Such interactions are dictated by the context of the intergrins in the cell function, not the absolute levels of integrins. For example, increased expression of integrin receptors on cell surfaces has been associated with an invasive phenotype of melanoma and squamous carcinoma of the head and neck, whereas loss of integrin expression status has been shown in cancers of breast, prostate and colon (Chammas and Brentani, 1991).

A wide variety of cell–cell adhesion receptors (CAMs) have been studied for their role in cancer invasion. These include, but are not limited to, intercellular adhesion molecules (ICAMs), L-, E- and P-selectins, vascular cell adhesion molecules (VCAMs), neural cell adhesion molecules (NCAMs) and neuroglial cell adhesion molecules (NG-CAMs). Unlike other receptors that bind proteins, selectins bind carbohydrate ligands on endothelial cells. The cadherins comprise a family of transmembrane glycoproteins that mediate Ca^{2+}-dependent cell–cell adhesion (Takeichi, 1991). Special intracellular proteins, the catenins, form zipper-like structures constituting extracellular cell–cell bonds with the cell cytoplasm. These interactions are regulated by tyrosine phosphorylation as well as additional cell to cell communications. Down-regulation of epithelial cadherin transcription, E-cadherin, has been shown to correlate with an aggressive cancer cell phenotype (Frixen *et al*., 1991; Vlemininckx *et al*., 1991). For example, transforming Madin–Darbey canine kidney (MDCK) cells by H-*ras* not only diminshed E-cadherin expression, but also increased the invasive behaviour of these cells. This effect could be reversed by transfecting E-cadherins back into the transformed cells (Vleminckx *et al*., 1991). Furthermore, overexpression of E-cadherin in highly invasive tumour types (bladder, breast, lung and pancreas) caused loss of invasiveness. Surface receptors that participate in cell–cell adhesion and interaction can activate signalling pathways responsible for maintaining normal cell and tissue architecture. Deregulation of these receptors in cancer can promote invasion by (a) reducing cell–cell adhesion which prevents shedding of tumour cells and (b) failing to suppress the inappropriate mixing of tumour cell and host cell populations during invasion.

MOTILITY

Translocation of individual cells across tissue boundaries is a necessary component of invasion. Cell motility and migration are not unique to tumour cells. This process is essential for normal immune cell function and for embryological development, organogenesis and gastrulation. Deregulation and persistence of motility may distinguish carcinoma cells from their normal epithelial counterparts (Nabeshima *et al*., 1997). The direction of tumour cell motility is controlled by a multitude of chemoattractants, including cytokines (hepatocyte growth factor), collagen peptides, formyl peptides and autocrine growth factors (e.g. Autotaxin) (Anzano *et al*., 1983). These agents may stimulate both the initiation and maintenance of tumour cell motility and the directness of that migration. Chemoattractants can be secreted by host stromal cells or the tumour cells themselves, or be released from the extracellular matrix.

An early event in motility is cytoskeletal remodelling causing extension of a dominant pseudopod toward the direction of movement. This is followed by translocation of the whole cell body (Stossel, 1993; You *et al*., 1996). The pseudopod of the invading cell has been renamed an 'invadopodia' because it may direct local proteolytic machinery (Bowden *et al*., 1999) literally to create a tunnel in the extracellular matrix in front of the cell. Protruding invadopodia, in response to chemoattractants, may serve to sense organs for the migrating cell to locate directional clues, to secrete motility-stimulating factors, to promote propulsive traction for locomotion and induce matrix proteolysis. The complexity of tumour cell migration requires that more than one agent is involved in the direction, location, and magnitude of the migratory response. During the course of invasion, the tumour cell must interact with the extracellular matrix components and be exposed to host-derived factors. Tumour cells have receptors for many of these potential attractants. Therefore, the response of tumour cells to autocrine motility stimulation and also endo- or paracrine stimulation by matrix components and host-derived factors is important to tumour motility.

EXTRACELLULAR MATRIX DEGRADATION DURING INVASION

Proteolytic modification of the cell surface and the extracellular matrix is believed to be an essential component of invasion (Liotta *et al*., 1980), both neoplastic and physiological. The major enzymes that degrade the ECM and cell-associated proteins are (1) the matrix metalloproteinases (MMPs), a family of secreted and membrane anchored proteinases, (2) the adamalysin-related membrane proteinases, (3) the bone morphogenetic protein 1 type metalloproteinases and (4) tissue serine proteinases including tissue plasminogen activator, urokinase, thrombin and plasmin (Werb, 1997). Major ECM barrier substrates for degradative enzymes include collagens (more than 13 types), proteoglycans, laminin, fibronectin and vitronectin. Each compartment of the ECM contains a different complement of matrix molecules. Collagens I and III are examples of collagens preferentially localized to stroma, while collagens IV and V are predominant in the basement membrane, which forms the border between epithelium and stroma. Proteolysis of the ECM is observed in trophoblast implantation, embryo morphogenesis, wound healing, tissue remodelling and angiogenesis. An imbalance in the ratio of proteinases to protease inhibitor can regulate vascular morphogenesis and invasion (Ura *et al*., 1989). All classes (serine, aspartyl, cysteinyl and metallo) of matrix-degrading proteinases participate, and coactivate each other, in the tumour–host invasion field (Nakajima *et al*., 1987; Ostrowski *et al*., 1988; Reich *et al*., 1988). Evidence also exists that proteases inside the cell may also be involved during invasion (Koblinski *et al*., 2000).

A large body of literature exists correlating degradative enzyme activity with cancer invasion and metastasis. The most studied proteases include tissue-type plasminogen activator (tPA), plasmin, cathepsin-D, -B, -L and -G, the urokinase plasminogen activator (uPA), metalloproteinases and the heparanases. Urokinase plasminogen activator, a serine protease, has been shown to correlate with a metastatic phenotype of cells. Antibodies against uPA block human HEP-3 cell invasion and murine B16–F10 melanoma cell metastasis after tail vain injection (Ossowski and Reich 1983; Esheicher *et al*., 1989). Moreover, overexpression of uPA in H-*ras* transformed cell lines enhance lung metastases (Axelrod *et al*., 1989). Inhibition of metalloproteinases has been demonstrated to inhibit cell invasion (DeClerck *et al*., 1991). MMPs can be divided into three general classes: (1) interstitial collagenases, (2) stromelysins and (3) gelatinases. Interstitial collagenase degrades type I, II, III and VII collagens. Stromelysins degrade type I, III, IV, V and IX collagens, laminin, fibronectin, and gelatin. The third group of the MMP family, the gelatinases (MMP-2 and MMP-9), can degrade collagen type I, II, III, IV, V, VII, IX and X and fibronectin (Emonard and Grimaud, 1990). Association of MMP-2 and MMP-9 with the invasive phenotypes is abundant in the literature. Inhibition of MMP-2 by TIMP-1 reduces cellular invasion *in vitro* and *in vivo*. Induction of H-*ras* oncogene enhances expression of MMP-2 and MMP-9. Invasive colonic, gastric, ovarian and thyroid adenocarcinomas showed positive immunoreactivity for MMP-2, whereas normal colorectal, gastric mucosa and benign ovarian cysts showed reduced or negative staining (Monteagudo *et al*., 1990; Levy *et al*., 1991). A delicate balance between TIMPs and MMPs may act as a positive and negative feedback control regulating vascular morphogenesis and invasion (Mignatti *et al*., 1986). MMPs and TIMPs have direct, and indirect, effects on angiogenesis, which are separate from their proteolytic functions (Chambers *et al*., 1997; Hoegy *et al*., 2001). Heparan sulfate

proteoglycans (HSPGs), major and ubiquitous components of the ECM, are substrates for heparanases, which cleave heparan sulfate glycosaminoglycan side chains. Augmentation of heparanase activity has been associated with tumour aggressiveness (Nakajima *et al*., 1988; Vlodavsky *et al*., 1995). Heparin and similar polysaccharides inhibit metastasis (Parish *et al*., 1987). Transfection of nonmetastatic murine T-lymphoma Eb cell lines with full-length human heparanase cDNA (Vlodavsky *et al*., 1999) enhances the metastatic phenotype in animal models. (See the chapters on *Models for Tumour Cell Adhesion and Invasion* and *Tumour Metastasis Models*.)

COORDINATION OF THE MACHINERY OF INVASION AT THE CELL SURFACE

Significant progress has been made in our understanding of the molecular cross-talk between tumour cells and host cells at the invasion front. A cascade of cytokines, motility factors, matrix receptors, enzymes and enzyme inhibitors simultaneously carries out the regulation, steering, proteolysis, traction and locomotion required for invasion (**Figure 3**).

Remodelling of the extracellular matrix, within the immediate pericellular environment of the cell, appears to be a necessary step in local invasion (Liotta *et al*., 1991; Werb, 1997). The complement of enzyme classes is tightly and exquisitely regulated by a series of activation steps and specific inhibitors. In a striking demonstration of host–tumour interdependence, a majority of the enzymes and inhibitors complexed at the invasion front are contributed by host cells, not by the invading tumour cells (Nakahara *et al*., 1997; Bowden *et al*., 1999; Coussens *et al*., 2000).

The enzyme machinery is confined to the cell surface at the point of invading pseudopodia 'invadopodia' by binding the enzymes to adhesion sites, cell surface receptors and adjacent ECM molecules (Nakahara *et al*., 1997; Bowden *et al*., 1999; Hoegy *et al*., 2001). MT1-MMP, a membrane-anchored ECM-degrading enzyme, contains a transmembrane–cytoplasmic sequence that confines it to microinvasion sites on the tumour cell invadopodia surface (**Figure 3**). In complex with one of the tissue inhibitors of metalloproteinases (TIMP-2) it becomes a receptor and activator of MMP-2 (Nakahara *et al*., 1997), a soluble MMP produced by stromal fibroblasts and endothelial cells. The serine proteinase uPA is confined to the invading pseudopodia through a cooperation between integrins and the uPA receptor (uPA-R) (Andreasen *et al*., 1997). uPA-R is an adhesion receptor for vitronectin, and also interacts laterally with integrin β chains. Proteolysis of ECM proteins modifies integrin mediated anchorage, focal adhesions and cytoskeletal architecture and triggers signalling molecules such as focal adhesion kinase (FAK) (Braga, 2000; Fashena and Thomas, 2000). Such heterotypic complexes direct and confine the enzymatic field at the forward edge of the invading cell, leaving intact the peripheral and distal attachment sites required for traction. As the invading cell moves forward through ECM barriers, the leading edge complex of enzymes, inhibitors and receptors molecules cycle through adhesion, deadhesion and proteolysis. The direction of tumour cell invasion and migration can be influence by chemoattractants and by marking of preferred adhesion pathways. Local attractants include (1) hepatocyte growth factor/scatter factor, which binds to the Met (c-Met) receptor (Wernert, 1997; Jo *et al*., 2000), (2) proteolysed matrix fragments which are recognized by integrins (Varner and Cheresh, 1996) or (3) cytokines and growth factors, such as EGF and TGF-β released from the degraded matrix (Roberts *et al*., 1992). Cryptic RGD sites exposed by proteolysis (Davis, 1992; Fukai *et al*., 1995; Varner *et al*., 1995; Varner and Cheresh, 1996) may guide the path in front of the invading tumour cells.

The combination of microdissection and protein microarrays has been successfully applied to the microworld of early stage cancer (Clark *et al*., 2000; Paweletz *et al*., 2001). Protein lysate microarrays consist of very small mass quantities (picograms) of protein lysates from cell lines, whole lysed tissue or microdissected subpopulations of lysed tissue cells immobilized and arrayed on a solid phase. The array can be probed with antibodies recognizing phosphorylated forms of signal proteins. Detection is highly sensitive, quantitative and precise, so that the state of signal pathways may be profiled. Individual subpopulations of host and tumour tissue cells within a microscopic field of invasion or premalignant transition can be microdissected and individually studied.

In conclusion, the process of cancer invasion is a coordinated effort by tumour cells and host cells within a microinvasion field. Within this field the tumour cells exchange cytokines, enzymes, inhibitors and growth factors which promote invasion by all cells involved. Pericellular remodelling of the ECM is commensurate with invasion. The different events of the metastatic cascade of angiogenesis, adhesion, proteolysis, motility and proliferation may provide useful and novel therapeutic targets. Investigators have identified some of the critical molecules involved in the extracellular cross-talk taking place among and between cells in the invasion field. This synthesis provides strategies for a new therapy concept 'stromal therapy' which targets the tumour–host communication interface. (See the chapter *Targeting the Extracellular Matrix*.)

REFERENCES

Aboseif, S., *et al*. (1999). Mesenchymal reprogramming of adult human epithlelial differentiation. *Differentiation*, **65**, 113–118.

Andreasen, P., *et al*. (1997). The urokinase type plasminogen activator system in cancer metastasis: a review. *International Journal of Cancer*, **71**, 1–22.

Anzano, M. A., *et al*. (1983). Sarcoma growth factor from conditioned medium of virally transformed cells is composed of both type alpha and type beta transforming growth factors. *Proceedings of the National Academy of Sciences of the USA*, **80**, 6264–6268.

Astrow, A. B. (1994). Commentary: rethinking cancer. *Lancet*, **343**, 494–495.

Axelrod, J. H., *et al*. (1989). Expression of human recombinant plasminogen activators enhance invasion and experimental metastasis of Ha-RAS-transformed NIH 3T3 cells. *Molecular Cell Biology*, **9**, 2133–2141.

Bowden, E., *et al*. (1999). An invasion related complex of cortactin, paxillin, and PKC u associates with invadopodia at sites of extracellular matrix degradation. *Oncogene*, **18**, 4440–4449.

Braga, V. (2000). The crossroads between cell–cell adhesion and motility. *Nature Cell Biology*, **2**, E182–E184.

Brown, P. D. and Giavazzi, R. (1995). Matrix metalloproteinase inhibition: a review of anti-tumour activity. *Annals of Oncology*, **6**, 967–974.

Carmeliet, P. and Jain, R. K. (2000). Angiogenesis in cancer and other diseases. *Nature*, **407**, 249–257.

Chambers, A. F. and Matrisian, L. M. (1997). Changing views of the role of matrix metalloproteinases in metastasis. *Journal of the National Cancer Institute*, **89**, 1260–1270.

Chammas, R. and Brentani, R. (1991). Integrins and metastases: an overview. *Tumour Biology*, **12**, 309–320.

Clark, E. A., *et al*. (2000). Genomic analysis of metastases reveals and essential role for RhoC. *Nature Cell Biology*, **406**, 532–535.

Coussens, L. M., *et al*. (2000). MMP-9 supplied by bone marrow-derived cells contributes to skin carcinogenesis. *Cell*, **103**, 481–490.

Davis, G. E. (1992). Affinity of integrins for damaged extracellular matrix: alpha v beta 3 binds to denatured collagen type I through RGD sites. *Biochemical and Biophysical Research Communications*, **182**, 1025–1031.

DeClerck, Y. A., *et al*. (1991). Inhibition of tumor cell invasion of smooth muscle cell layers by recombinant human metalloproteinase inhibitor. *Cancer Research*, **51**, 2151–2157.

Dvorak, H. F., *et al*. (1995). Vascular permeability factor/vascular endothelial growth factor, microvascular hyperpermeability, and angiogenesis. *American Journal of Pathology*, **146**, 1029–1039.

Emonard, H. and Grimaud, J. A. (1990). Matrix metalloproteinases: a review. *Cell Molecular Biology*, **36**, 131–153.

Esheicher, A., *et al*. (1989). Characterization of the cellular binding site for the urokinase type plasminogen activator. *Journal of Biological Chemsitry*, **264**, 1180.

Fashena, S. and Thomas, S. M. (2000). Signaling by adhesion receptors. *Nature Cell Biology*, **2**, E225–E229.

Fidler, I. J. and Ellis, L. M. (1994). The implication of angiogenesis for the biology and therapy of cancer metastasis. *Cell*, **79**, 185–188.

Folkman, J. (1995). Angiogenesis in cancer, vascular, rheumatoid and other disease. *Nature Medicine*, **1**, 27.

Folkman, J. and Klagsbrun, M. (1987). Angiogenic factors. *Science*, **235**, 442–447.

Frisch, S. and Francis, H. (1994). Disruption of epithelial cell–matrix interactions induces apoptosis. *Journal of Cell Biology*, **124**, 619–626.

Frisch, S. M. and Ruoslahti, E. (1997). Integrins and anoikis. *Current Opinions in Cell Biology*, **9**, 701–706.

Frixen, U. H., *et al*. (1991). E-cadherin mediated cell–cell adhesion prevents invasiveness of human carcinoma cells. *Journal of Cell Biology*, **113**, 173–185.

Fukai, F., *et al*. (1995). Release of biologically activities from quiescent fibronectin by conformational change and limited proteolysis by matrix metalloproteinases. *Biochemistry*, **34**, 11453–11459.

Gallager, H. S. and Martin, J. E. (1969). The study of mammary carcinoma by mammography and whole organ sectioning, early observation. *Cancer*, **23**, 855–873.

Guidi, A. J., *et al*. (1997). Vascular permeability factor (vascular endothelial growth factor) expression and angiogenesis in patients with ductal carcinoma in situ of the breast. *Cancer*, **80**, 1945–1953.

Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, **100**, 57–70.

Hoegy, S., *et al*. (2001). Tissue inhibitor of metalloproteinases-2 (TIMP-2) suppresses TKR-growth factor signaling independent of metalloproteinase inhibition. *Journal of Biological Chemistry*, **276**, 3203–3214.

Jo, M., *et al*. (2000). Cross talk between epidermal growth factor receptor and c-Met signal pathways in transformed cells. *Journal of Biological Chemistry*, **275**, 8806–8811.

Juliano, R. L. and Haskill, S. (1993). Signal transduction from the extracellular matrix. *Journal of Cell Biology*, **120**, 577–585.

Koblinski, J. E., *et al*. (2000). Unraveling the role of proteases in cancer. *Clinica Chimica Acta*, **291**, 113–135.

Kohn, E. C. and Liotta, L. A. (1995). Molecular insights into cancer invasion: strategies for prevention and intervention. *Cancer Research*, **55**, 1856–1862.

Levy, A., *et al*. (1991). Increased expression of the 72 kDa type IV collagenase in human colonic adenocarcinoma. *Cancer Research*, **51**, 439–444.

Liotta, L. A., *et al*. (1974). Quantitative relationships of intravascular tumor cells: tumor vessels and pulmonary metastases following tumor implantation. *Cancer Research*, **34**, 997.

Liotta, L. A., *et al*. (1980). Metastatic potential correlates with enzymatic degradation of basement membranes. *Nature*, **284**, 67–68.

Liotta, L. A., *et al*. (1991). Cancer metastasis and angiogenesis: an imbalance of positive and negative regulation. *Cell*, **64**, 327–336.

Mignatti, P., *et al*. (1986). In vitro angiogenesis on the human amniotic membrane: requirement for basic fibroblast growth factor-induced proteinases. *Journal of Cell Biology*, **108**, 671–682.

Monteagudo, C., *et al*. (1990). Immunohistologic distribution of type IV collagenases in normal, benign, and malignant breast tissue. *American Journal of Pathology*, **136**, 585–592.

Mueller, A., *et al*. (2001). Involvement of chemokine receptors in breast cancer metastasis. *Nature*, **403**, 50–56.

Nabeshima, K., *et al*. (1997). Cohort migration of cancer cells. *Connective Tissue*, **29**, 199.

Nakahara, H., *et al*. (1997). Transmembrane/cytoplasmic domain mediated membrane type 1-matrix metalloproteinase docking to invadopodia is required for cell invasion. *Proceedings of the National Academy of Sciences of the USA*, **94**, 7959–7964.

Nakajima, M., *et al*. (1987). Degradation of basement membrane type IV collagen and lung subendothelial matrix by rat mammary adenocarcinoma cell clones of differing metastatic potentials. *Cancer Research*, **47**, 4869–4876.

Nakajima, M., *et al*. (1988). Heparanase and tumor metastasis. *Journal of Cell Biochemistry*, **36**, 157–167.

Nicolson, G. L. (1991). Gene expression, cellular diversification and tumor progression to the metastatic phenotype. *Bioessays*, **13**, 337–342.

Ossowski, L. and Reich, E. (1983). Antibodies to plasminogen activator inhibit human tumor metastasis. *Cell*, **35**, 611–619.

Ostrowski, L. E., *et al*. (1988). Expression pattern of a gene for a secreted metalloproteinase during late stages of tumor progression. *Molecular Carcinogenesis*, **1**, 13–19.

Parish, C. R., *et al*. (1987). Evidence that sulphated polysaccharides inhibit tumor metastasis by blocking tumor cell-derived heparanase. *International Journal of Cancer*, **40**, 511–518.

Paweletz, C. P., *et al*. (2001). Reverse phase protein microarrays which capture disease progression shown activation of pro-survival pathways at the cancer invasion front. *Oncogene*, **20**, 1981–1989.

Reich, R., *et al*. (1988). Effects of inhibitors of plasminogen activator, serine proteases, and collagenase IV on the invasion of basement membranes by metastatic cells. *Cancer Research*, **48**, 3307–3312.

Roberts, A. B., *et al*. (1992). TGF-beta: regulation of extracellular matrix. *Kidney International*, **41**, 557–559.

Stossel, T. P. (1993). On the crawling of animal cells. *Science*, **260**, 1086–1094.

Takeichi, M. (1991). Cadherin cell adhesion receptors as a morphogenetic regulator. *Science*, **251**, 1451–1455.

Ura, H., *et al*. (1989). Expression of type IV collagenase and procollagen genes and its correlation with tumorigenic, invasive, and metastatic abilities of oncogene transformed human bronchial cells. *Cancer Research*, **49**, 4615–4621.

Varner, J. A. and Cheresh, D. A. (1996). Integrins and cancer. *Current Opinions in Cell Biology*, **8**, 724–730.

Varner, J. A., *et al*. (1995). The integrin α v β3: angiogenesis and apoptosis. *Cell Adhesion Communications*, **3**, 367–374.

Vleminckx, K., *et al*. (1991). Genetic manipulation of E-cadherin by epithelial tumor cells reveals an invasion suppressor role. *Cell*, **66**, 107–119.

Vlodavsky, I., *et al*. (1995). Inhibition of tumor metastasis by heparanase inhibiting species of heparin. *Invasion and Metastasis*, **14**, 290–302.

Vlodavsky, I., *et al*. (1999). Mammalian heparanase: gene cloning, expression and function in tumor progression and metastasis. *Nature Medicine*, **5**, 793–802.

Weiss, L. (1992). Comments on hematogenous metastatic patterns in humans as revealed by autopsy. *Clinical and Experimental Metastasis*, **10**, 191–199.

Werb, Z. (1997). ECM and cell surface proteolysis: regulating cellular ecology. *Cell*, **91**, 439–442.

Wernert, N. (1997). The multiple roles of tumor stroma. *Virchows Archives*, **430**, 433–443.

You, J., *et al*. (1996). Responses of tumor cell pseudopod protrusion to changes in medium osmolality. *Journal of Cell Physiology*, **167**, 156–163.

Zhuang, Z., *et al*. (1995). Identical allelic loss on chromosome 11q13 in microdissected *in situ* and invasive human breast cancer. *Cancer Research*, **55**, 467–471.

FURTHER READING

Carmeliet, P. and Jain, R. K. (2000). Angiogenesis in cancer and other diseases. *Nature*, **407**, 249–257.

Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, **100**, 57–70.

Kohn, E. C. and Liotta, L. A. (1995). Molecular insights into cancer invasion: strategies for prevention and intervention. *Cancer Research*, **55**, 1856.

Nicolson, G. L. (1998). Organ specificity of tumour metastasis: role of preferential adhesion, invasion and growth of malignant cells at specific secondary sites. *Cancer and Metastasis Reviews*, **7**, 173–188.

Chapter 17

Angiogenesis

Ute Modlich and Roy Bicknell
University of Oxford, John Radcliffe Hospital, Oxford, UK

CONTENTS

WHAT IS ANGIOGENESIS?

Like all tissues, tumours depend on a continuous supply of oxygen and nutrients for their survival. The ability to recruit a functional blood supply is therefore central to tumorigenesis. Angiogenesis is the growth of new blood vessels from the pre-exisiting vasculature by budding and sprouting of endothelial cells. This is in contrast to vasculogenesis, which is the *de novo* formation of blood vessels from endothelial precursor cells (called angioblasts) (Rissau, 1997). Vasculogenesis occurs mainly during embryogenesis, in particular the development of tissues of endodermal origin, although some ectodermal and mesodermal derived tissues acquire their vessels through angiogenesis, e.g. the kidney and the brain (Beck and D'Amore, 1997).

In an adult, the vasculature is remarkably quiescent and angiogenesis occurs only very rarely. Diffusion of oxygen in tissues is limited to a distance of about 150 μm. Therefore, tissue growth is restricted to a few cubic millimetres if no new vasculature is formed. Under physiological conditions, angiogenesis is involved in the turnover of tissues in the female reproductive system (endometrium, placenta, follicle maturation and corpus luteum formation in the ovaries) and in wound healing. Angiogenesis is, on the other hand, an essential component of many pathologies such as diabetic retinopathy, rheumatoid arthritis, psoriasis and tumour growth.

'TUMOUR GROWTH IS ANGIOGENESIS DEPENDENT'

Around 30 years ago, it was recognized that the growth of solid tumours is angiogenesis dependent (Folkman, 1971, 1990). Thus, tumours remain in a dormant state of a few millimetres in diameter (prevascular phase) unless they are able to recruit their own vascular bed. This does not mean that the tumour cells are unable to proliferate, but an inadequate supply of oxygen and nutrients results in a high rate of apoptosis (cell death). The tumour exists in a balance of proliferation and regression which precludes tumour expansion (tumour dormancy). Only after a tumour has recruited its own blood supply can it expand in size. This involves the production of angiogenic factors secreted into the tissue by the tumour cells and is known as the 'angiogenic switch' (Hanahan and Folkman, 1996). This induces angiogenesis in the adjacent quiescent vasculature, allowing the growth of new vessels into the tumour.

Why and how new vessels are formed is of great interest in terms of anticancer therapy because inhibition of new vessel formation could restrict tumour growth. The tumour is essentially 'starved'. Such strategies could also include the destruction of tumour vessels which are already formed. Increased understanding of the mechanism of angiogenesis could allow subsequent design of therapeutics which interfere with the process. The aim of this chapter is to describe the sequence of events during angiogenesis and to

outline its importance in tumour biology. (See the chapter *Antiangiogenic Therapy*.)

MORPHOLOGY OF BLOOD VESSELS

The entire vasculature is lined by a single layer of specialized cells, the endothelial cells. They form a simple squamous epithelium which rests on a basement membrane and surrounds a lumen in which blood flows (**Figure 1**). The basement membrane contains collagen types I and IV, fibronectin, laminin, entactin and other non-collagen glycoproteins. Endothelial cells, together with their basement membrane, constitute the lamina intima of a blood vessel. The lamina intima is surrounded by pericytes, smooth muscle cells or cardiomyocytes, collectively described as mural cells. Pericytes lie within the basement membrane of the endothelial cells whereas the others surround larger vessels and the endocardium as an additional layer outside the basement membrane (as a part of the lamina media). Pericytes are recruited by endothelial cells during development and play an important role in vessel maturation and stabilization. They appear to suppress the turnover of endothelial cells. During neovascularization pericytes are selectively lost from angiogenic vessels, including the tumour vasculature.

The endothelium in an adult has a surface area of $>1000\,m^2$. In addition to forming a static physiological barrier, it possesses secretory, synthetic, metabolic and immunological functions. The endothelium is highly heterogeneous; its precise nature varies depending upon the function of the vascular bed in different tissues.

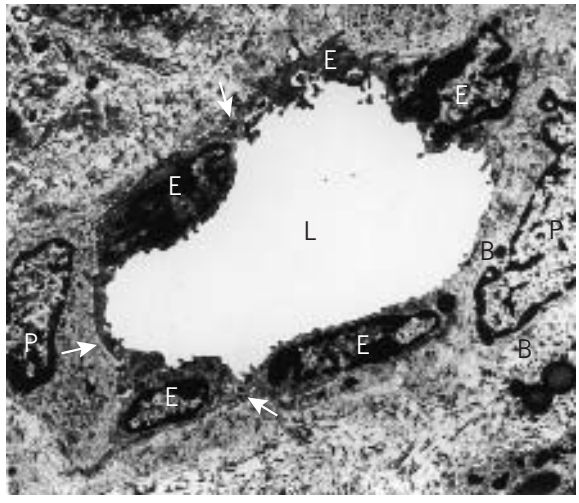


Figure 1 Ultrastructure of a capillary. The vessel lumen (L) is surrounded by five endothelial cells (E). The cells rest on a continuous basement membrane (B). Two pericytes (P) lie within the basement membrane. The endothelial cell contacts (tight junctions) are visible as electron-dense areas of the endothelial cell membranes (arrows).

Histologically there are three main taxonomic classes of endothelia. For example, blood vessels of the brain and the retina are especially tight (continuous endothelium), whereas the endothelium of the sinusoidal vessels in the liver, spleen and bone marrow contains intracellular gaps not covered by a basement membrane and so allows cellular trafficking (discontinuous endothelium); the endothelium in endocrine glands, the kidney and small intestine shows intracellular holes (fenestrated endothelium), which allows extensive exchange of substances.

At the molecular level, the endothelium of different tissues varies in its surface phenotype and protein expression. The distinct differences are induced by the extracellular matrix which adapts the endothelium to its functional needs in a tissue. This observation has led to the concept of 'vascular addresses.' It explains the homing of inflammatory cells in specific tissues because they adhere to the vessel wall by attaching to specific endothelial cell surface molecules. The unique vascular addresses of organs are also the reason for tissue-directed metastasis of tumour cells. It has been shown that tumour cells recognize specific adhesion molecules on the endothelium. They adhere there and are subsequently able to extravasate and to invade the tissue. In the tissue stroma they then form micrometastases.

THE TUMOUR VASCULATURE IS DISORGANIZED

The vascular tree of tumour vessels appears 'chaotic' and 'disorganized.' Tumour vessels are also hyperpermeable ('leaky') owing to a discontinuous endothelium and lack of pericytes. Morphologically the vessels appear highly tortuous. The tumour vasculature is formed in two ways: by co-opting pre-existing vessels and by induction of new vessels by angiogenesis. In addition, every tumour induces its own characteristic vascular bed which is different for each tumour type. As a result of persistent growth, the tumour vasculature constantly changes its shape and is dynamic rather than static. In tumours, up to 30% of the vasculature consists of arterio-venous shunts where the blood bypasses the capillaries, precluding exchange of nutrients. Tumour blood vessels lack mural cells as well as appropriate innervation and therefore blood pressure is poorly controlled (only by pre-existing arterioles). The blood flow in the tumour is slow but higher in the tumour periphery than in the centre. Different regions of perfusion can be seen within tumours: necrotic, seminecrotic and well-vascularized regions, in order from the centre of the tumour to the periphery. The blood flow in the periphery of a tumour is higher than that in the surrounding normal tissue (advancing front), whereas in the centre, blood flow can be interrupted transiently resulting in complete stasis; blood flow may then be re-established in the opposite direction.

Extravasation of molecules from the blood occurs through diffusion, convection and partially through transcytosis. Despite the fact that blood flow within tumours is slow, and the vessels are leaky, the delivery of therapeutics into tumours is inefficient. This is due to a high interstitial fluid pressure in the tumour tissue compromising convective movements of therapeutics. In regions of highest pressure in a tumour, the hydrostatic and osmotic pressure in the vessels is opposed by that in the tissue, inhibiting the exchange of substances. This results in hypoxic and hypoglycaemic regions within a tumour. In addition, there are totally anoxic areas. Because the vasculature is dynamic, and perfusion rates in different regions of a tumour change frequently, hypoxic regions can become reoxygenated again. Similarly, normoxic regions can experience sudden oxygen deprivation and become hypoxic. Owing to the poor perfusion of vessels, endothelial cells also can become hypoxic. It has been shown that hypoxia is a potent inducer of many molecules involved in angiogenesis.

These observations highlight the fact that the tumour vasculature is not an extension of 'normal' blood vessels but that it is itself distinct and characteristic. This renders the tumour vasculature an excellent target for antitumour therapy. In addition, targeting the vessels is associated with a number of advantages over current therapies (Augustin, 1998). Targeting its uniqueness should minimize interference with other blood vessels in the body. In addition, endothelial cells are much more accessible than tumour cells to systemically administered therapeutics because of their intimate contact with the blood. Finally, destruction of only a few endothelial cells can have detrimental effects on the tumour. The destruction of a single blood vessel will lead to an amplification of tumour destruction because about 10 layers of tumour cells are dependent on the supply by one blood vessel.

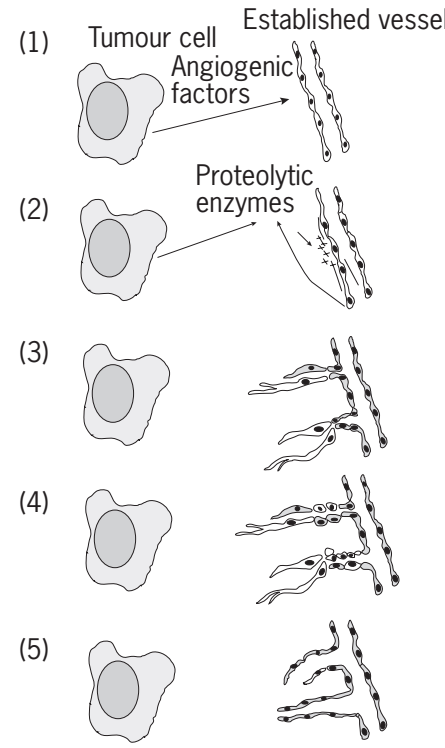


Figure 2 The angiogenic cascade. (1) Angiogenic stimulus; (2) Degradation of the basement membrane and ECM by proteases released from tumour and activated endothelial cells; (3) Migration of EC towards angiogenic stimulus; (4) Endothelial cell proliferation; (5) Tube formation and vessel maturation.

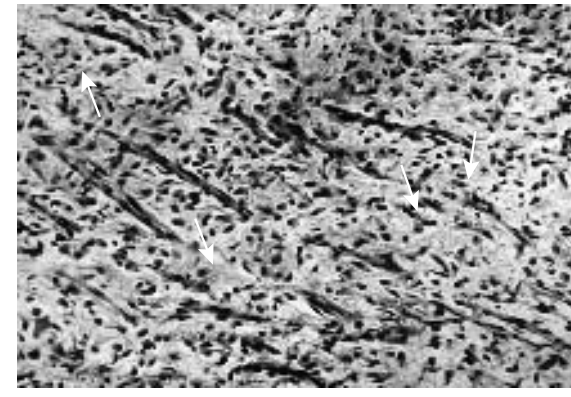


Figure 3 Angiogenic blood vessels in the rat sponge assay (paraffin section, haematoxylin/eosin stained). In this assay inert sponge is implanted subcutaneously. Endothelial cells migrate from the surrounding tissue into the sponge and form new vessels. Some vessels are indicated by arrows.

MECHANISMS OF ANGIOGENESIS

A tumour can only continue to expand in size if it induces a blood supply. When a tumour starts to produce angiogenic factors it activates endothelial cells in the vasculature of the surrounding tissue to initiate angiogenesis (**Figure 2**). The angiogenic stimulus induces the endothelial cells of the 'mother' vessels to change from a quiescent to an activated phenotype (Auerbach and Auerbach, 1994). These endothelial cells produce proteolytic enzymes which break down their basement membrane. This is the prerequisite for endothelial cells to migrate into the surrounding tissue towards the angiogenic stimulus. The migrating endothelial cell changes shape to an elongated phenotype (**Figure 3**). They start expressing typical cell surface molecules which allow the cells to migrate along the extracellular matrix. These endothelial cells also start to proliferate and to form new tubes. Finally, these new tubes anastomose into loops that allow blood flow. The recruitment of pericytes and smooth muscle cells completes angiogenesis through blood vessel maturation. The endothelial cells in the new vessels then revert to a quiescent phenotype.

Table 1 Angiogenic factors

Growth factors
Vascular endothelial growth factor (VEGF)
Placenta growth factor (PlGF)
Basic fibroblast growth factor (bFGF or FGF-2)
Acidic fibroblast growth factor (aFGF or FGF-1)
Transforming growth factor-α (TGF-α)
Transforming growth factor-β TGF-β)
Platelet-derived growth factor (PDGF)
Hepatocyte growth factor (HGF)
Granulocyte colony-stimulating factor (G-CSF)
Cytokines
Tumour necrosis factor-α (TNF-α)
Interleukin 1 (IL-1)
Chemokines
Interleukin 8 (IL-8)
Enzymes
Platelet-derived endothelial cell growth factor (PD-ECGF)
Angiogenin
Prostaglandins
PGE_1
PGE_2

ANGIOGENIC FACTORS

In the last 20 years, many factors that have angiogenic activity have been identified. The first to be identified were the fibroblast growth factors (aFGF or FGF-1 and bFGF or FGF-2), which are now known to be pleiotrophic growth factors. Others include the transforming growth factors (TGF-α and TGF-β), platelet-derived growth factor (PDGF), hepatocyte growth factor (HGF), granulocyte colony-stimulating factor (G-CSF), tumour necrosis factor-alpha (TNF-α), platelet-derived endothelial cell growth factor (PD-ECGF, thymidine phosphorylase), interleukin-8 and prostaglandins (PGE_1, PGE_2) (**Table 1**). All of these have been studied extensively and have been shown to be angiogenic (Folkman and Shing, 1992), but none of these are endothelial cell specific. The only growth factor known to be specific for endothelial cells is vascular endothelial growth factor (VEGF).

Vascular Endothelial Growth Factor (VEGF)

VEGF was the first member to be identified of a growing family of vascular endothelial growth factors and is now referred to as VEGF-A. Other members include VEGF-B to E and placenta growth factor (PIGF-1 and 2). VEGFs are homodimeric proteins and mediate their activity through tyrosine–kinase receptors (VEGF-receptors, VEGFR1-3) which are almost exclusively expressed on endothelial cells. VEGFR-1 (flt-1, fms-like kinase-1) binds VEGF-A, VEGF-B and PIGF with strong affinity whereas VEGFR-2

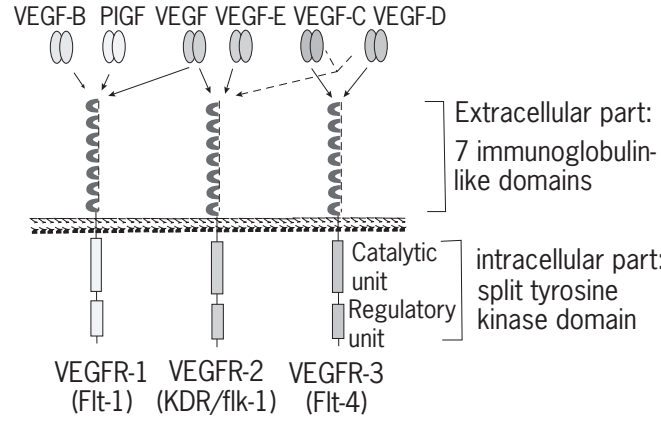


Figure 4 VEGF receptors and their ligands.

(flk-1, foetal liver kinase-1 or the human homologue KDR, kinase-insert domain receptor) binds VEGF-A, VEGF-C and VEGF-D with lower affinity. VEGFR-3 (flt-4) is only expressed on lymphatic endothelium and is the ligand for VEGF-C which induces lymphangiogenesis. VEGF-D has been discovered to be an additional ligand for the VEGFR-3. VEGF-E seems to signal mainly through VEGFR-2 (**Figure 4**).

Of all the VEGFs, VEGF-A has been studied most intensively (for a review, see Neufeld *et al*., 1999). It exists in at least five different splice variants ($VEGF_{121,145,165,189,206}$) encoding isoforms of differing length derived from a single gene. With the exception of $VEGF_{121}$, which is secreted, all VEGFs are heparin binding. Therefore, they accumulate in the extracellular matrix and can be released from there by proteolytic enzymes. There is, in addition, a sixth 110 amino acid isoform of VEGF arising from proteolytic cleavage of $VEGF_{189}$, when bound to the cell surface. Heparan sulfate proteo glycans are known to modulate growth factor signalling of many heparin-binding growth factors from their respective receptors, and the same might apply to VEGF signalling.

VEGF is an endothelial cell mitogen and chemoattractant, it promotes cell migration, inhibits apoptosis and modulates the permeability of the endothelial cell layer (it was first identified as vascular permeability factor (VPF) in 1983). Hence it has major roles as a key regulator in angiogenesis and vasculogenesis. Mice deficient in one allele of VEGF (VEGF +/− mice) show early embryonic death (embryonic day 11–12) due to cardiovascular defects in most sites of early blood vessel formation (embryo and yolk sac). This suggests that a minimal dosage requirement for the growth factor exists because a single allele could not rescue the phenotype. Microinjection of VEGF into quail embryos during development induces uncontrolled and unlimited vascularization at sites that are normally avascular. Together these observations show that the level of VEGF expression is tightly regulated and that small variations can have fatal effects.

Figure 5 Rat aortic ring angiogenesis assay. A section of a rat aorta is placed in a fibrin gel in cell culture. Endothelial cells from the inner lining of the aortic ring (AR) sprout into the fibrin gel. Some of the newly formed tubes are seen to anastomose.

VEGFR-2-deficient mice (*flk-1* $-/-$ mice) die *in utero* between embryonic days 8.5 and 9.5. Endothelial cells in these animals fail to differentiate from their angioblastic precursors and the animals die as a result of a total lack of vascular structures. VEGFR-2 is the earliest marker for differentiation towards endothelial cells in development. Similarly, VEGFR-1-deficient mice (*flt-1* $-/-$) show abnormal and disorganized vascular structures and die between embryonic days 9.5 and 10.5. These two gene-targeted mice, despite both developing vascular defects and being embryonic lethal, show very different phenotypes. Furthermore the VEGF receptors do not show redundancy in their function.

VEGF has been established as an angiogenic factor *in vitro* in two- and three-dimensional cell culture systems (**Figure 5**). VEGF also proved its angiogenic properties in the rabbit cornea assay and the chick chorioallantoic assay. Tumour cells overexpressing VEGF form faster-growing tumours and contain significantly more blood vessels than tumours formed from untransfected control cells. VEGF also synergizes with bFGF and the angiogenic effects of each growth factor are potentiated.

VEGF expression is regulated by hypoxia and hypoglycaemia, and especially high levels are present within hypoxic regions in tumours. Oxygen tension in tumours is about 50 times lower than in normal tissues and hypoxia could be a major factor driving tumour angiogenesis. The cellular response to hypoxia is mediated by the hypoxia-inducible factor Hif-1. Hif is a dimer comprised of two subunits by Hif-1α and Hif-1β. This dimeric complex recognizes a specific DNA sequence, known as the hypoxia-response element (HRE), which is found within the promoters of hypoxia regulated genes such as the VEGF promoter. The promoter of VEGFR-1 also contains hypoxia response elements whereas the promoter of VEGFR-2 does not. Hence the expression of VEGFR-1 is enhanced by hypoxia directly but not the VEGFR-2. The expression of VEGFR-2 can be induced by binding of VEGF to VEGFR-2 and as a result the VEGFR-2 is indirectly upregulated by hypoxia.

Macrophages in Tumour Angiogenesis

Unlike physiological angiogenesis, tumour growth is usually accompanied by an inflammatory response. Widespread infiltration of inflammatory cells during angiogenesis is a feature of wound healing. Thus, tumours have been described as 'wounds which do not heal.' Tumours recruit inflammatory cells, especially macrophages (tumour-associated macrophages (TAMs)). In some tumours macrophages can represent up to 50% of the cell population and are often the major source of angiogenic molecules in a tumour (see **Figure 6**). Tumour cells express the very potent monotactic chemokine MCP-1 (monocyte chemotactic protein-1) which attracts macrophages into the tumour. TAMs have two important functions within the tumour. First, they confer antitumour effects by their immunological functions and have been implicated as part of the host immune defence mechanism against tumours. Second, those same cells can promote angiogenesis through secretion of cytokines, such as interleukin 1 and TNF-α, chemokines and growth factors (VEGF, bFGF, EGF, PD-ECGF and HGF).

Chemokines act as chemoattractants for haematopoietic cells and confer similar effects on endothelial cells (Moore *et al*., 1998). Chemokines are grouped into two main subfamilies, the CC-chemokines and the CXC-chemokines, based on their structure. In CXC-chemokines the two cysteines near the N-termini are separated by a single amino acid whereas in CC-chemokines these cysteines are adjacent. Endothelial cells express all known receptors for CXC-chemokines (CXCR-1 to CXCR-4 and DARC) but none of the receptors for CC-chemokine (there are to date eight identified receptors). This corresponds well with the observation that CC-chemokines have no direct effects on endothelial cells (members of the CC-chemokine subfamily are MCP-1 and RANTES). The subfamily of CXC-chemokines includes interleukin 8 (IL-8), platelet factor-4, growth-related antigen (GRO-β) and interferon-γ-inducible protein (IP-10). In addition, Il-8 was shown to induce endothelial cell proliferation and to induce angiogenesis *in vivo*. This effect can be potentiated by IL-1 and TNF-α. Interestingly, a number of chemokines from this sub-family have an antiangiogenic function (platelet factor-4, GRO-β, IP-10). These CXC-chemokines have in common that they lack the ELR-motif (glutamine–leucine–arginine) between the first cysteines. The pro- or anti-angiogenic properties of a chemokine might be dependent on this motif.

Macrophages secrete TNF-α, a multifunctional cytokine which has both angiogenic and antiangiogenic properties. These conflicting reports are a result of different actions of TNF-α at high and low doses. Whereas high doses of TNF-α

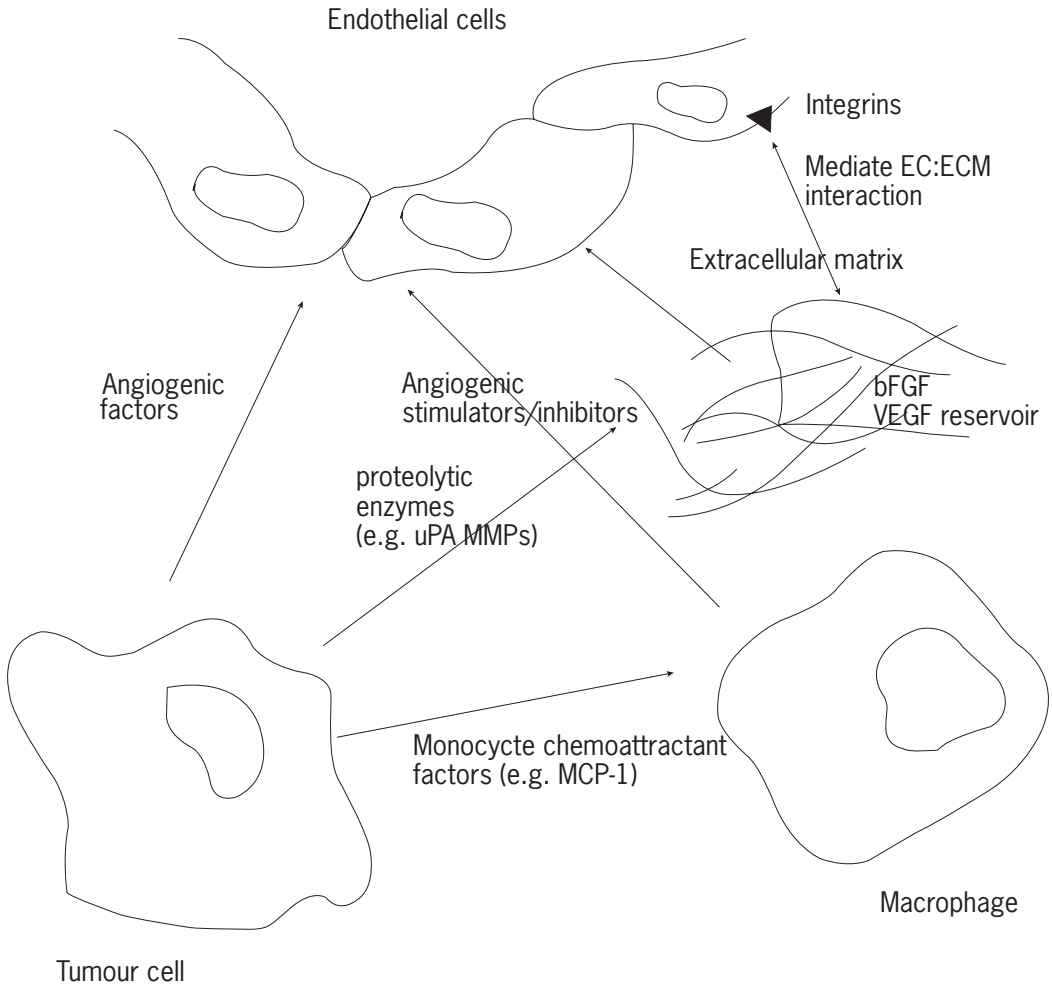


Figure 6 Interaction between tumour cells, endothelial cells and macrophages. Tumour cells produce angiogenic factors, which induce endothelial cells to initiate angiogenesis. Tumour cells also secrete chemoattractants, that recruit macrophages, and proteolytic enzymes, which can release growth factors from the extracellular matrix. Endothelial cells interact with the extracellular matrix by adhesion molecules (e.g. integrins), which mediate endothelial cell migration.

are directly cytotoxic to tumours, it is angiogenic at low doses. Because of its direct cytotoxic effects, TNF-α was studied following systemic administration *in vivo*. Unfortunately, the side effects were severe and may be related to the fact that TNF-α is the major mediator of septic shock. At high doses, TNF-α exerts its antitumour effects mainly through its cytotoxicity on endothelial cells and inhibition of angiogenesis rather than through actions on tumour cells. These findings are consistent with the fact that tumour cells are often insensitive to TNF-α alpha-mediated cytotoxicity.

At low doses, TNF-α induces endothelial cell migration and tube formation *in vitro*, but it also inhibits endothelial cell proliferation. It is chemotactic for leukocytes and induces IL-1 and GM-CSF expression. It also induces the expression of E-selectin, ICAM-1 and VCAM-1 on endothelial cells and can mediate leukocyte recruitment into the tissue stroma.

Fibroblast Growth Factors

aFGF and bFGF are endothelial cell mitogens and stimulate endothelial cell migration. They induce the production of proteases by endothelial cells and stimulate tube formation in three-dimensional cell culture systems. They are angiogenic in all *in vivo* assays and are today often used as a positive control in angiogenesis assays. Because of their heparin-binding abilities, FGFs are bound to the extracellular matrix and are released by proteases involved in the angiogenic cascade. Endothelial cells are not only reactive to FGFs through two FGF receptors but they also produce bFGF themselves. In this way, endothelial cells respond to FGF in both paracrine and autocrine manner. The FGFs interact with almost all molecules involved in angiogenesis. A complete discussion of this growth factor is beyond

the scope of this chapter (for a detailed review, see Christofori, 1997).

THE BREAKDOWN OF THE BASEMENT MEMBRANE

Angiogenesis is an invasive process involving migration of endothelial cells into the surrounding tissue. In the first instance, this requires the degradation of the basement membrane of the pre-existing vasculature, to enable endothelial cells to leave the organized structure of the vessel wall. Several proteolytic enzymes are involved in the degradation of the extracellular matrix (ECM). They include the plasminogen activator system and the matrix metalloproteinases (MMPs).

The Plasminogen Activator System

The plasminogen activator (PA) system includes the urokinase plasminogen activator (uPA) and the tissue-type plasminogen activator (tPA), both of which convert plasminogen into plasmin. The major physiological function of plasmin is to degrade fibrin; however, it also degrades components of the ECM such as laminin and collagen. In addition, it also activates metalloproteinases and elastase. Whereas tPA functions during fibrinolysis, uPA functions mainly in angiogenesis. uPA is secreted as a proenzyme and binds to its receptor on the cell surface, where it becomes activated. By binding to the uPA receptor it also activates downstream signals, resulting in induction of cell migration and invasion. In fact, the presence of the uPA receptor seems to be critical for cell migration because very little migration could be observed in uPA receptor negative cells. Endothelial cells express uPA as well as an uPA inhibitor (PAI). PAI binds to the active uPA and stimulates its internalization into cells following binding to the uPA receptor, resulting in its breakdown. The expression of uPA and its receptor can be induced by FGF-2. This effect is mediated via the FGF receptors. Hence it seems likely that the FGF-2-induced migration in endothelial cells is mediated by uPA. Because of its heparin-binding ability, FGF accumulates in the ECM and is released by proteases including plasmin. Therefore, uPA and FGF increase their function in an autocrine manner because uPA is expected to release FGF-2 and at the same time FGF-2 induces migration through uPA.

The Matrix Metalloproteinases

The MMPs are a family of extracellular endopeptidases which are secreted by a variety of cells including epithelial cells, fibroblasts and inflammatory cells. Endothelial cells express MMP-1, MMP-2, MMP-9 and the membrane-associated MT-1-MMP (for a review, see Stetler-Stevenson, 1999). These endopeptidases are secreted as inactive proenzymes and need to be activated by cleavage of a proteolytic fragment. The substrates of the different endopeptidases include all known components of the ECM but the specific substrate of each enzyme has not yet been identified. Four different endogenous inhibitors of MMPs have been identified; they are the tissue inhibitors of metalloproteinases TIMP-1 to -4. Endothelial cells studied in a two-dimensional culture produce only very low levels of MMPs, whereas expression can be induced in endothelial cells by growing them in three-dimensional collagen gel systems. In addition, changes in cell shape also alter MMP expression (especially mechanical stress) and it has been suggested that integrins, as the mediators of cell–matrix interactions, are involved in the regulation of MMP expression. This has been shown for the $\alpha 2\beta 1$ and the $\alpha v\beta 3$ integrins. The $\alpha v\beta 3$ integrin interacts with MMP-2 through the C-terminal hemopexin-like (PEX) domain of MMP-2. This is interesting because the PEX domain does not contain the integrin binding RGD motif. The PEX domain is also the binding site for TIMP-2. TIMP-2 binding of MMP-2 is necessary for the MT-1-MMP-mediated activation of MMP-2 from the proenzyme. So PEX can inhibit angiogenesis by direct competition to MMP-2 in two different ways. First, PEX inhibits activation of MMP-2 by binding to TIMP-2. Second, PEX binds to the integrin on the cell surface, which would be a prerequisite for MMP-2 activation. Natural breakdown products of MMP-2 are generated and PEX accumulates in tissues where neovascularization occurred. These findings suggest that endogenous PEX acts as a natural inhibitor of MMP-2 function and as a result is an endogenous inhibitor of angiogenesis.

Angiogenesis Inhibitors Encrypted within Larger Proteins

Similarly to PEX, several other endogenous angiogenesis inhibitors have been characterized which are all fragments of larger molecules. These inhibitors are formed by proteolytic breakdown of their parent molecules. The first of these inhibitors was identified as early as 1985 and is a 29-kDa fragment of fibronectin. It has been shown to inhibit endothelial cell proliferation. Fibronectin itself is an abundant molecule which has no such inhibitory function.

More recently discovered endogenous angiogenesis inhibitors are angiostatin and endostatin (for reviews, see Kim Lee Sim, 1998; Oehler and Bicknell, 2000). Angiostatin is a 36-kDa fragment of plasminogen which is cleaved by macrophage-derived MMPs or proteases derived from tumours (e.g. prostate carcinomas). Endostatin is a 20-kDa fragment of collagen XVIII. Systemic administration of both angiostatin and endostatin induces tumour regression by destruction of the tumour vasculature. They also inhibit the formation of new blood vessels in a tumour. They therefore maintain a tumour in a state of

dormancy. Both molecules have been identified in the urine or serum of tumour-bearing mice. In these animals a primary tumour existed but metastases did not occur. Based on these findings, and on the clinical observation that primary tumours suppress the growth of metastasis, it was concluded that inhibitory factors are produced by the primary tumour. The inhibition of metastatic spread was a direct result of angiogenesis inhibition, and angiostatin and endostatin act specifically on endothelial cells. In addition, both molecules are endogenous factors which are generated by proteolytic cleavage of larger molecules. These 'mother' molecules have no inhibitory function but have angiogenesis promoting effects.

N-Terminal fragments of the human prolactin/growth hormone family are antiangiogenic whereas the hormones themselves are angiogenic. They have also been shown to stimulate the expression of PAI. During the maturation of type I collagen, a ~90-kDa N-terminal fragment is cleaved which is homologous to a domain of the antiangiogenic matrix molecule thrombospondin. The fragment has antiangiogenic effects both *in vitro* and *in vivo*. Restin is a 22-kDa fragment of collagen XV that has been found by homology search with endostatin. It inhibits endothelial cell migration *in vitro* but does not effect proliferation. Vasostatin inhibits bFGF-induced angiogenesis *in vitro* and *in vivo*. It is a 180-kDa fragment of calreticullin. Calreticullin has been shown to exert similar effects.

It seems that endogenous antiangiogenic factors are released under physiological conditions. It might be true that the vasculature is kept in a quiescent state by a balance between angiogenesis-promoting factors and angiogenesis inhibitors and that a deregulation of this balance induces angiogenesis (endogenous angiogenesis inhibitors; see **Table 2**).

Table 2 Naturally occurring angiogenesis inhibitors

Enzyme inhibitors
Tissue metalloproteinase inhibitors (TIMP1-4)
Plasminogen activator inhibitor (uPAI)
Angiogenesis inhibitors encrypted in larger proteins
29-kDa fragment of fibronectin
16-kDa fragment of prolactin
Angiostatin: 36-kDa fragment of plasminogen
Endostatin: 20-kDa fragment of collagen type XVIII
90-kDa N-terminal fragment of collagen type I, homologue to a thrombospondin domain
Cleaved conformation of antithrombin
PEX: hemopexin-like domain of MMP-2
Vasostatin: N-terminal domain of calreticulin
Restin: 22-kDa fragment of collagen XV
Cytokines
Tumour necrosis factor-alpha (TNF-α)
Interferon
Chemokines
Platelet factor-4
Growth-related antigen (GRO-β)
Interferon-γ-inducible protein (IP-10)

MIGRATION

During the formation of new vascular sprouts, endothelial cells migrate into the surrounding tissue following chemoattractant stimuli. Migrating endothelial cells adapt their shape and become elongated with multiple pseudopodia. They cover a larger surface and are therefore in more intimate contact with components of the ECM. Consistent with this, they upregulate the expression of receptors which mediate cell–ECM contacts. Integrins are the major ECM receptors. They are transmembrane heterodimeric cell adhesion molecules which are composed of an α and a β subunit. There are at present 20 members of this family derived from 15 α and eight β chains. Integrins mediate adhesion to all known components of the ECM while one of them can often recognize more than one ECM molecule. Integrins also bind to cell surface molecules of the immunoglobulin superfamily (ICAM-1, ICAM-2 and VCAM-1) and in this situation are involved in cell–cell adhesion processes.

Some integrins recognize ECM molecules through the so-called RGD motif (Arg-Gly-Asp), which is present in fibronectin, vitronectin, proteolysed collagen, fibrinogen, von Willebrand factor, osteopontin and thrombospondin. The $\alpha_5\beta_1$, $\alpha_{IIb}\beta_3$ and most $\alpha_v\beta$ integrins bind the RGD motif. Some integrins, such as $\alpha_2\beta_1$, bind to an Asp-Gly-Gly-Ala sequence in type I collagen. Both $\alpha_2\beta_1$ and $\alpha_1\beta_1$ integrins are upregulated on migrating endothelial cells following treatment with VEGF and the migration of endothelial cells on a collagen matrix can be inhibited by antibodies against the α chains of these two integrins. An *in vivo* model of angiogenesis was inhibited by combinations of the same antibodies without effects on non-angiogenic vessels.

The most extensively studied integrin involved in angiogenesis is the $\alpha_v\beta_3$ integrin (Eliceiri and Cheresh, 1999). It is upregulated on tumour blood vessels, during wound healing and retinal neovascularization. Angiogenesis is inhibited by a specific monoclonal antibody (LM609) which blocks binding of this integrin and thus disrupts endothelial cell–ECM adhesion. In the chick chorioallantoic membrane assay and in the rabbit cornea eye pocket assay LM609 blocked bFGF-induced angiogenesis whilst VEGF-induced angiogenesis was inhibited by antibodies against the $\alpha_v\beta_5$ integrin. A humanized form of the antibody LM609 (Vitaxin) is in clinical trials.

The RGD peptide itself is able to induce blood vessel regression *in vitro* and *in vivo* through competition with matrix proteins for their integrin binding site. Adhesion is required for endothelial cell survival and these cells undergo apoptosis when adhesion is disrupted. Endothelial

cells of quiescent vessels are less sensitive to antiadhesive treatment than angiogenic and migrating cells. Experiments with endothelial cells grown on RGD-coated beads showed that binding to the ligand (in this case the RGD peptide on the bead) is not sufficient to prevent endothelial cells from undergoing apoptosis. It was suggested that spreading of endothelial cells is necessary for survival. It has also been postulated that the mechanical force generated by cell–ECM interactions mediates gene expression by changes in the cytoskeleton and nuclear morphology.

Integrins interact with the cytoskeleton through their cytoplasmic domains. This interaction is initiated by binding of integrins to the ECM. The mechanical properties of the ECM modulate the strength of this interaction. In fibroblasts the binding of integrins to the cytoskeleton enables the cells to pull themselves forward over a stationary substrate by retrograde movement of the cytoskeleton. This mechanism allows them to migrate. Similarly, integrins are involved in endothelial cell migration.

Integrins also act as signalling receptors. Binding to ECM has been shown to result in elevated intracellular calcium levels, elevated pH, activation of the inositol and DAG pathways and tyrosine phosphorylation. Furthermore, ligand binding to $\alpha_v\beta_3$ is essential for the sustained activation of MAP kinases by angiogenic factors. Tyrosine phosphorylation of VEGFR-2 by VEGF was enhanced by growing endothelial cells on vitronectin, a ligand for $\alpha_v\beta_3$. This was inhibited by antibodies directed against the β_3 integrin subunit. As mentioned earlier, $\alpha_v\beta_3$ integrin binds to MMP-2 and is involved in proteolysis.

PROLIFERATION

Another important step in angiogenesis is the proliferation of endothelial cells. This would provide new vasculature with additional endothelial cells to cover the inner surface of the new vessels. Indeed, many angiogenic factors have mitogenic activity, e.g. bFGF and VEGF.

The turnover rate of endothelial cells in the adult human body is extremely low. Incorporation experiments with radioactively labelled thymidine (H^3TdR) in rabbits *in vivo* estimated a proliferative index for endothelial cells in retinal vessels of about 0.01% (1 h after injection). Following two days of treatment with H^3TdR, about 0.2% of the endothelial cells showed incorporation of thymidine and a turnover time of ~1000 days was calculated. Interestingly, pericytes were discovered to have an even lower proliferation index (0.06%). In contrast, rapidly renewing tissues were shown to have much higher labelling indices, such as the epithelium of the cornea (3%) and the epithelium of the duodenum (14%). In a separate study, endothelial cell proliferation indices of different tissues were compared. Endothelium of the brain and muscle was especially quiescent (0.8% and 0.5% of cells incorporated H^3TdR after 3 days), whereas higher indices were measured in the liver endothelium (4.4% after 3 days).

In contrast to endothelial cells within normal tissues, endothelial cell proliferation indices in tumours is several orders of magnitude higher. Labelling indices in experimental tumours in animals have been reported as up to 32% (anaplastic sarcoma of Wistar rats). Others have estimated proliferation indices between 4.5 and 20%. There was, however, no correlation between the tumour growth rate and the endothelial cell proliferation index. Some studies have reported higher proliferation indices for tumour cells than for endothelial cells (35–11.4%), whereas others have shown the opposite. In general, however, the endothelial cell proliferation is higher in the tumour periphery than in the centre. Artefactual results caused by inadequate perfusion of the tumour centre can be excluded because tumour cells were efficiently labelled.

Endothelial cell proliferation indices in animal tumour models differ significantly from those obtained from human tumours. Endothelial cell proliferation is considerably slower in human tumours than in experimental tumours in animals (2.2–2.8% in breast carcinomas, 5% in gliomas, 0.1–0.6% in prostatic carcinomas). However, these endothelial cell proliferation indices are still significantly higher than in benign human tissues (benign breast tissue, 0.06%; benign hyperplasia in the prostate, 0.023%). These observations argue for active proliferation of endothelial cells during tumour-induced angiogenesis.

The discrepancy between human and experimental tumours becomes obvious if one looks at the rate of growth of experimental tumours. Within 1 week these tumours grow to about 20% of the body weight of the animals, a rate unthinkable in human tumours. Nonetheless, the proliferation indices for endothelial cells in human tumours are 30 times higher than the proliferation indices for normal tissues.

VESSEL FORMATION AND MATURATION

In the final stage of angiogenesis, vessel assembly and maturation result in a vessel that is a stable conduit for blood flow. This is achieved by two processes, namely anastomosis of the developing sprouts and recruitment of pericytes.

Subsequent to the degradation of the basement membrane and outgrowth of vessel sprouts, the sprouts develop into hollow tubes and two sprouts may then join to form a tube through which blood flows. Once formed, the endothelial tubes recruit a layer of pericytes to surround and stabilize the vessel. Pericytes are recruited as local mesenchymal cells, and induced to differentiate into pericytes by the endothelial cells. These cells are large with a prominent nucleus and multiple long processes that embrace the vessel. They also secrete factors, such as

TGF-β, which stabilize the endothelial cells and prevent their proliferation. Furthermore, pericyte density is negatively correlated with the turnover time of the endothelial cells in various tissues, consistent with the stabilizing actions of these cells.

In addition to the pericyte-derived vessel-stabilizing factors, angiopoietin-1 (Ang-1) is a local tissue-derived vessel-stabilizing factor that acts on the receptor Tie-2. Tie-2 is a receptor–tyrosine kinase and is mainly expressed on endothelial cells. The closely related ligand angiopoietin-2 (Ang-2) acts as an antagonist to Ang-1 on this receptor. The related receptor–tyrosine kinase Tie-1 is mainly expressed on endothelium during embryogenesis but downregulated in adulthood. Its ligands have so far not been identified. Targeted disruption of the *Tie-2* gene in mice results in embryonic death between days 9.5 and 10.5 and these animals show heart defects and disorganized vessels on sites where vessels form by vasculogenesis. The endothelial cells appear rounded and have only weak connections with their mural cells. A similar but milder phenotype was developed in *Ang-1* −/− mice which die on embryonic day 12.5. In addition mice overexpressing Ang-2 show very similar phenotypes to *Tie-2* and *Ang-1* −/− mice. These observations show that the angiopoietins and Tie-2 are factors controlling the 'tightness' of vessels and that indeed the angiopoietins have antagonistic effects. *Tie-1* −/− mice die between embryonic day 13.5 and birth as result of the loss of vascular integrity. These animals show 'electron light' endothelial cells because of numerous intra- and intercellular holes which allows blood and plasma extravasation.

During development Ang-1 and Ang-2 are expressed throughout the vasculature, both by the developing endothelium and by its supporting cells, with Ang-2 antagonizing the vessel-stabilizing and maturation function of Tie-2, allowing vasculogenesis and angiogenesis. Ang-1 continues to be expressed into adulthood, but Ang-2 in only expressed in areas that undergo vascular remodelling, such as the female reproductive system. Indeed, expression of Ang-2 appears to be permissive of angiogenesis, suggesting a requirement for 'dematuration' to initiate vessel sprouting (Lauren *et al*., 1998). Vessels which are destabilized by Ang-2 can respond to VEGF and initiate angiogenesis. Without an angiogenic stimulus blood vessels influenced by Ang-2 will regress (Holash *et al*., 1999). Only vessels where no pericytes support the endothelial cells can undergo regression whereas the contact of endothelial cells with pericytes prevents regression.

THE PROGNOSTIC SIGNIFICANCE OF TUMOUR ANGIOGENESIS

This chapter has presented evidence that tumour growth is angiogenesis dependent and in the chapter on *Antiangiogenic Therapy* it is shown that inhibition of tumour angiogenesis is currently one of the most exciting avenues of anti-cancer therapy. To conclude this chapter, we will review the prognostic significance of tumour angiogenesis. This encompasses primarily three areas of interest: (1) analysis of the primary tumour microvessel density, (2) expression of angiogenic factors within tumours and (3) analysis of systemic markers of active angiogenesis.

Intratumoural Microvessel Density

In 1991 it was shown that the presence of vascular hotspots, that is, areas of high vascular density, within primary human breast tumours correlated with poor patient survival. This was the first such study linking tumour vascularization, or the end result of angiogenesis, with prognosis. Subsequently, many similar studies have appeared showing a correlation between the presence of vascular hotspots and prognosis in a wide range of different tumour types. Making use of this information in the clinic has, however, proved difficult. To date no simple, quick and reliable way to quantitate the vascular density has appeared. All methods require examination of slides by an experienced pathologist, which is time consuming, expensive and potentially subjective. Nevertheless, proof of the correlation between angiogenesis and prognosis encourages the search for less subjectively quantitated molecular markers of tumour angiogenesis that could substitute for the assessment of vascular density. Such markers will no doubt be found in time.

Systemic Markers of Active Angiogenesis

Angiogenesis is an active process occurring within the tumour, but (excluding women in the menstrual cycle) in the healthy individual not elsewhere in the body, thus metabolic changes arising as a result of the angiogenesis could in principle provide a diagnostic test of tumorigenesis. Potential markers include plasma levels of angiogenic factors and markers of activated endothelium. Amongst angiogenic factors, VEGF has received the most attention as a potential prognostic marker. Thus, studies have suggested that serum VEGF could be a predictor of relapse-free survival in primary human breast cancer. Recently, high urinary VEGF has been shown to correlate with recurrence in bladder cancer, providing an easily accessible marker with which to monitor the disease. Thymidine phosphorylase (platelet-derived endothelial cell growth factor) was shown to be elevated in cancer patients some 15 years before it was recognized to have angiogenic activity and to be elevated in virtually all primary tumours. No doubt more studies will follow.

Other markers of vascular activity are the endothelial leukocyte adhesion molecules VCAM and E-selectin. VCAM and E-selectin mediate leukocyte exit from the

blood through the endothelium. Both molecules are upregulated on endothelium in inflammatory sites. This is also widespread in tumour endothelium where tumours are frequently awash with inflammatory cytokines. Several studies are now examining whether the presence of VCAM or E-selectin in the plasma are useful markers of tumour angiogenesis.

REFERENCES

Auerbach, W. and Auerbach, R. (1994). Angiogenesis inhibition: a review. *Pharmacology and Therapeutics*, **63**, 265–311.

Augustin, H. G. (1998). Antiangiogenic tumour therapy: will it work? *Trends in Pharmacological Science*, **19**, 216–222.

Beck, L. and D'Amore, P. A. (1997). Vascular development: cellular and molecular regulation. *FASEB Journal*, **11**, 365–373.

Christofori, G. (1997). The role of fibroblast growth factors in tumor progression and angiogenesis. In: Bicknell, R., *et al*. (eds), *Tumour Angiogenesis*, 201–237 (Oxford, Oxford University Press).

Eliceiri, B. P. and Cheresh, D. A. (1999). The role of the αv integrins during angiogenesis: insight into potential mechanisms of action and clinical development. *Journal of Clinical Investigation*, **103**, 1227–1230.

Folkman, J. (1971). Tumor angiogenesis: therapeutic implications. *New England Journal of Medicine*, **285**, 1182–1186.

Folkman, J. (1990). What is the evidence that tumor growth is angiogenesis dependent? *Journal of the National Cancer Institute*, **82**, 4–6.

Folkman, J. and Shing, Y. (1992). Angiogenesis. *Journal of Biological Chemistry*, **267**, 10931–10934.

Hanahan, D. and Folkman, J. (1996). Patterns and emerging mechanisms of the angiogenic switch during tumorigenesis. *Cell*, **86**, 353–364.

Holash, J., *et al*. (1999). New model of tumor angiogenesis: dynamic balance between vessel regression and growth mediated by angiopoietins and VEGF. *Oncogene*, **18**, 5356–5362.

Kim Lee Sim, B. (1998). Angiostatin and endostatin: endothelial cell-specific endogenous inhibitors of angiogenesis and tumour growth. *Angiogenesis*, **2**, 37–48.

Lauren, J., *et al*. (1998). Is angiopoietin-2 necessary for the initiation of tumor angiogenesis? *American Journal of Pathology*, **153**, 1333–1339.

Moore, B. B., *et al*. (1998). CXC chemokines mechanism of action in regulating tumor angiogenesis. *Angiogenesis*, **2**, 123–134.

Neufeld, G., *et al*. (1999). Vascular endothelial growth factor (VEGF) and its receptors. *FASEB Journal*, **13**, 9–22.

Oehler, M. K. and Bicknell, R. (2000). The promise of anti-angiogenic cancer therapy. *British Journal of Cancer*, **82**, 749–752.

Rissau, W. (1997). Mechanisms of angiogenesis. *Nature*, **386**, 671–674.

Stetler-Stevenson, W. G. (1999). Matrix metalloproteinases in angiogenesis: a moving target for therapeutic intervention. *Journal of Clinical Investigation*, **103**, 1237–1241.

FURTHER READING

Bicknell, R., *et al*. (eds) (1997). *Tumour Angiogenesis*. (Oxford, Oxford University Press).

Chapter 18

Cell Proliferation in Carcinogenesis

Nicholas A. Wright
Imperial Cancer Research Fund and University of London, London, UK

CONTENTS

INTRODUCTION

Investigators in the field of cancer research have always regarded cell proliferation, *a fortiori*, as of paramount importance. The reasons for this are not difficult to identify. It is generally accepted that tumours arise as a result of a series of mutations occurring in a cell, often said to be a stem cell. In the colorectal epithelium, for example, several of the mutations required for malignant transformation have been identified; such a series of mutations accumulate in a single cell and its progeny, and this single cell, having acquired properties which endow it with characteristics ensuring its growth and survival, undergoes a series of divisions which eventually result in the development of a neoplasm. A similar series of molecular events is envisioned for the development of other tumours such as the lung, gastric carcinoma and skin. These mutations are thought to confer upon the transformed cell an advantage which enables it to survive – as a mutated clone – and replace the normal cells in the tissue, eventually establishing itself as a neoplasm.

Thus neoplasms are clonally derived, that is, the single cell transformed is the ancestor of all cells which compose the neoplasm. Once established, a mechanism must ensure the growth and propagation of the mutated clone within the epithelium. It is often said that the mechanism lies in an inherent ability of the mutated clone to outgrow its normal counterparts – the 'carcinogenic advantage' is thus a proliferative advantage.

A second reason for the putative importance of cell proliferation in this context is the spread of the mutated clone as it evolves and establishes itself within the host tissue. Third, when a mutant clone has indeed established itself, we observe the phenomenon of tumour growth as the lesion becomes macroscopically evident. During this phase, cell proliferation, in the context of cell birth and cell loss, defines the rate of growth of the tumour.

Finally, when such a lesion is treated, there have been attempts to predict the outcome by the components of cell proliferation.

We shall examine these proposals in turn. Central to our initial proposal is the concept that tumours are indeed clonally derived.

HOW DO TUMOURS BEGIN IN HUMANS?

We often say that tumours arise from a single cell, and therefore form clonal populations. 'If a proliferation is clonal it is a neoplasm' is a statement often heard. Tumour clonality is an important concept in our attempts to understand malignant transformation, and it is worthwhile reminding ourselves of the experimental basis of this proposal. Why do we believe that epithelial tumours, such as the early adenoma in the colonic mucosa, are clonal proliferations? The main methods which have been used for the analysis of clonality in human tumours have been based on X-chromosome inactivation analysis and the detection of somatic mutations. Viral integration, e.g. by Southern blotting in Epstein–Barr virus (EBV)-associated tumours or in hepatitis B- or C-associated liver tumours, are also useful; there is excellent agreement between X-inactivation and EBV integration in nasophayrngeal carcinoma, for example.

In early X-chromosome inactivation studies, the haplotypes of glucose-6-phosphate dehydrogenase (G6PD) were used (Beutler *et al*., 1967; Fialkow, 1976), replaced more recently by methods based on restriction length polymorphisms of X chromosome-linked genes such as glycerophosphate kinase (PGK), the androgen receptor gene (HUMARA), hypoxanthine phosphoribosyltransferase (HPRT), the M27β probe for *DXS285* and p55 and glucose-6-phosphate dehydrogenase. Early on in embryogenesis, genes on one of the two X chromosomes are

randomly inactivated by methylation of cytosine residues within promoter regions; once methylated, such CpG islands are functionally and heritably inactive and it is usually believed that this inactivation is stable, even during malignant change. Thus in approximately half of the cells of the embryo the paternal X chromosome is active, and in the rest it is the X chromosome from the mother. The pattern of fragments produced by DNA digestion with a methylation-sensitive enzyme such as SnaB1 and a further endonuclease corresponding to a restriction fragment length polymorphism, in PGK–BstCX1, for example, can be used to investigate the clonality of any tissue specimen. Informative cases in woman using these markers are reported to vary from 45% with PGK and HUMARA to over 90% with M27β/DXS255.

But can we be certain that the results from such studies will be reliable? Well, the methylation pattern of DNA can be abnormal in malignancy, with both increases and reductions in methylation, and the possibility exists that X-chromosome inactivation may not be valid as an indicator of clonality because of such abnormalities in DNA methylation (Jones, 1996). Moreover, it is possible that X inactivation might be nonrandom, being either constitutive or cell-type specific. Studies in normal haematopoietic and lymphoid tissues have shown skewed X inactivation, possibly favouring the paternal or the maternal X chromosome, which could indicate a nonrandom X-chromosome inactivation pattern. Although there are claims that extremely unbalanced inactivation of the X chromosome is an uncommon phenomenon, skewed inactivation is seen in 23% of women with HPRT and PGK and 22–33% with M27β, in peripheral blood and in bone marrow and skin, indicating tissue specificity, and perhaps related to the number of stem cells in the tissue at the time of X chromosome inactivation. If this number is small, it will result in skewing, with increased probability as the stem cell pool size diminishes (Fialkow, 1973). Moreover, in some embryonic tumours such as retinoblastoma and Wilms' tumour, with LOH on 13q and 11p, respectively, show preferential loss of maternal and paternal alleles. This is also seen in sporadic osteosarcoma. Hence X inactivation analysis is not without its problems, and conclusions drawn from it must be viewed critically.

In mutation analysis, finding the same mutation in key genes, such as k-*ras* or *p53* in multiple tissue sample from the same tumour, or from unconnected tumours, has also been said to indicate a clonal origin, but we should note the possibility that the same mutation is induced in separate precursor cells by a single carcinogen, e.g. aflatoxin causes specific *p53* mutations in hepatocellular carcinomas.

The demonstration of heterogeneity of microsatellite instabilities, e.g. in multiple gastric carcinomas, has also been proposed as a marker for polyclonality – origin from more than one cell – but if genome alterations continue to occur at microsatellite loci with evolution of the tumour, with resulting genetic diversity within the same clone, then microsatellite instability will not be an appropriate molecular marker of clonality. The presence of cytogenetically unrelated abnormal clones demonstrated by karyotypic analysis has also been used as evidence of polyclonality, but the existence of such clones might, however, be due to chromosomal rearrangement in non-neoplastic epithelial or stromal cells: cytogenetically abnormal clones are present in apparently non-neoplastic breast lesions such as fibrocystic disease (from which breast carcinoma probably arises). Mutated clones have been reported even in histologically normal breast epithelium, with several at potential tumour-suppressor gene sites, indicating that genetic abnormalities accumulate before pathological changes can be detected.

We should note that these observations involve the biochemical or molecular examination of homogenized tissues, and in human tissues there have been few opportunities to observe directly the clonal development of very early tumours, and this introduces other problems. In development, a clone is a family of cells which derive from a common progenitor, and these remain more or less contiguous throughout the growth of the embryo. A patch, on the other hand, is defined as a group of cells which share the same genotype, contiguous at the moment of consideration, which, for the current argument, share the same X-chromosome inactivation pattern. Clone size and patch size are not strictly equivalent, since multiple clones of the same genotype could contribute to a single patch; similarly, a single clone could be anatomically separated into different patches.

Figure 1 illustrates this concept as it applies to the clonal origin of tumours. A tumour arising from the centre of a patch will be of clonal origin when assessed by the pattern of X inactivation. The only chance of detecting a polyclonal proliferation would be when such a lesion arises from the margin of a patch boundary (Schmidt and Mead, 1990). This is seen in normal mouse epidermis at patch boundaries, where hair follicles appear polyclonal, but of course are clonal within patches. As X inactivation occurs at about the time of implantation, the pool of stem cells is small, ~15 cells for the skin. Skewing towards one parental allele is therefore very possible, and indeed this is seen in human skin specimens, indicating a large patch size. Polyclonal tumours would be commonest at patch boundaries, the incidence being dependent on the size of the patch, and the incidence of such tumours could be small. And if such rare polyclonal lesions are found, because of a large patch size, it is usually explained away by the lesion consisting of more than one tumour, of clonal origin, that have mixed or collided (Farber, 1990, 1997). However, the rarity of such lesions would indeed be expected if as seems likely, X-linked patches are fairly large.

This is well illustrated by an example: for many years, based in X-chromosome inactivation analysis, it has been believed that atheromatous plaques, occurring in the walls of large- and medium-sized arteries, are monoclonal, and

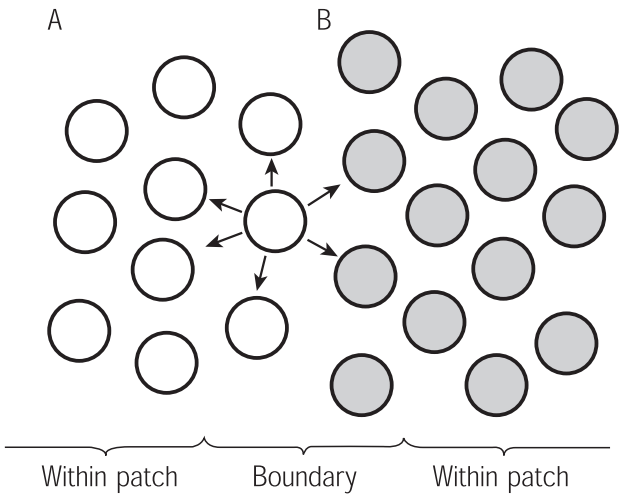


Figure 1 Demonstrating the importance of patch size in the clonal histogenesis of tumours. Different patches are shown, for example, illustrating a different pattern of X-inactivation. In the patches are tissues units, e.g. colonic crypts. A tumour arising within a patch will show a single pattern of X-inactivation of the other marker, and the only chance of detecting a polyclonal tumour is if such a lesion arose from the border of the patch, i.e. between A and B. (From Schmidt and Mead, 1990, *Bioessays*, **12**, 37–40.)

such plaques have even been regarded as some sort of benign neoplasm (Benditt and Benditt, 1973). This clonal proliferation is supported by X-inactivation studies, and was proposed to be due to a somatic mutation, induced perhaps by genotoxic chemicals, or an infection, giving a hit in a single cell, which then develops into an atheromatous plaque (Murry *et al*., 1997). However, by mapping X-chromosome inactivation patterns in human aortic smooth muscle, using the HUMARA method, the patch size in aortic media and intima was found often to exceed 4 mm, indeed a large area in terms of numbers of cells (Chung *et al*., 1998). Because of this large patch size, X-chromosome inactivation analysis cannot distinguish between a monoclonal and a polyclonal origin for atheromatous plaques. This underlines the importance of knowledge of the patch size in such studies.

The study of multiple lesions in the same patient does provide a way of avoiding the problem of patch size, especially if more than one clonal marker is used. Thus the probability that all the tumours examined would have the same X-chromosome inactivation pattern is 0.5^{n-1}, where n equals the tumour number. For example, if allele loss on 9q, an independent event, was also present, then the probability that the same pattern in each tumour is due to chance is $(0.5^{n-1})(0.5^{n-1})$, and so on for each independent marker used.

Similarly, when methods based on the analysis of homogenized tissues give results suggesting polyclonality, this is usually attributed to contamination with underlying stromal cells of different clonal derivation and some results are attributed to normal tissue contamination even when the incidence of polyclonal tumours amounts to 40% of the total. However, it is clear that even the fibroblasts or myofibroblasts very closely applied to the epithelium in the colonic crypt are of different clonal derivation, and there may be invading inflammatory cells, which, if numerous enough, can give rise to disparate results in clonality analyses involving PCR techniques. Normal tissue can be trapped and enclosed by surrounding neoplastic tissue, and again give discordant results. Finally, the problem of poor sampling may give a false impression of monoclonality in a polyclonal tumour.

It is also important to know when, during the life history of the development of the tumour, it is examined. Mature lesions are inappropriate, since tumours of polyclonal origin may become clonal because of clonal evolution (Nowell, 1976, 1986). All clones, except one, are eliminated or reduced to the point of being undetectable, seen in chemically induced mouse fibrosarcomas, initially polyclonal, which evolve to a clonally derived population in time, because of the later selection of a dominant clone (Woodruff *et al*., 1982). This is also seen in colorectal adenomas, where heterogeneity of k-*ras* mutations is observed, which is lost after the lesion has evolved to become an invasive carcinoma (Ajiki *et al*., 1994). As noted above, genetic heterogeneity in a tumour does not necessarily indicate a polyclonal origin, since genetic instability is a major feature of malignant tissues and many new clones may arise during tumour development.

Having critically reviewed the available methods, let us see what they say about how tumours arise in humans.

Non-neoplastic and Preneoplastic Lesions

Many tumours arise from preneoplastic lesions, not themselves neoplastic, which were previously regarded as hyperplasias, which involve changes in many cells and therefore by definition polyclonal. However, such lesions are often clonal proliferations themselves. Extremely relevant is the growing recognition that mutations in important genes such as *p53* are found in tissues such as squamous epithelium preceding any dysplastic change, such as sun-exposed normal epidermis. Clones of keratinocytes with immunoreactive p53 and p53 mutations (exons 5–8) are seen in sun-exposed but otherwise normal skin. This has also been reported in morphologically normal mucosa from individuals with upper aerodigestive tract tumours. Moreover, microsatellite instability has been found in normal mucosa from patients with ulcerative colitis, reflecting the increased risk of malignancy in these patients, since microallelotyping shows no allelic loss in transitional mucosa adjacent to colorectal neoplasms.

However, such losses have been reported in normal breast tissue.

In the gastric mucosa of a single patient, three separate hyperplastic polyps of the fundus have been found which harbour foci of dysplasia showed the same k-*ras* codon 12-point mutation, present in both hyperplastic and dysplastic areas. An explanation is that the progeny of a single transformed cell spreads through the mucosa: either surface spreading and surface implantation has occurred or a single cell could populate a gastric gland, which then spreads through the mucosa by gland fission.

In the female ovary, malignant tumours, endometrioid carcinomas, are thought to arise from foci of endometriosis and the majority of such endometriotic foci are clonal (Jimbo *et al*., 1997). Endometriosis arises either from implantation of shed endometrial cells, or from metaplasia of the pelvic peritoneum: if patch size considerations can be excluded, this means from a single endometrial or mesothelial cell. Endometriotic foci can show aneuploidy and loss of heterozygosity at candidate tumour-suppressor loci in 9p, 11q and 22q. The derivative tumours, the endometrioid carcinomas, are clonally derived, arising from a monoclonal proliferation that itself can show genetic defects.

The preneoplastic lesions from which breast cancers develop – proliferative breast disease and similar lesions – show cytogenetic abnormalities indicating the presence of multiple clones. X-chromosome inactivation studies and detection of microsatellite alterations both show that atypical duct hyperplasia and intraduct papillomas appear clonal proliferations and consequently cytogenetic alterations have already occurred at this stage (Lakhani *et al*., 1995).

Naevocellular naevi, either congenital or acquired, are, in some cases, the precursor lesions of malignant melanomas, are apparently polyclonal lesions on X-chromosome inactivation analysis (HUMARA and PGK combined), but malignant melanomas are clonal. So such naevi are presumably hamartomas – an abnormal proliferation of cells in an organ or tissue where these cell types would normally be found. However, hamartomas in tuberous sclerosis, where multiple cell types are seen, show clonal 9q34 or 16p13.3 LOH and clonal X-inactivation patterns, while pulmonary chondroid hamartomas also contain clonal cytogenetic abnormalities in the chondroid component.

In the human liver opinion is divided concerning the clonality of lesions often regarded as preneoplastic: some maintain that while lesions such as benign adenomatous hyperplasia and focal nodular hyperplasia are polyclonal, but hepatic adenomas and even small (<25 mm) hepatocellular carcinomas are clonal. Others, examining the integration patterns of hepatitis B virus, claim that atypical adenomatous hyperplasisa and focal nodular hyperplasias are clonal. Hepatocellular carcinoma is frequently multifocal and whether these arise from a single clone or independently is controversial, but most authors agree that an independent origin is more likely. Thus, after liver damage, clonal selection occurs during regeneration, leading to the genesis of persistent benign focal proliferations, which may be themselves clonal. This is followed by the development of clonal hepatocellular cancer from one or more such nodules. There is also substantial evidence from hepatitis B and C virus integration that between 0.5 and 43% of regenerative nodules in the resulting cirrhosis are monoclonal (Aihara *et al*., 1994), whereas in other types of cirrhosis, some 54% of regenerative liver nodules are clonal, but that the associated hepatocellular carcinomas are clonal by X-inactivation analysis, and differ from the nodules by 18q loss. This suggests that regenerative liver nodules showing a polyclonal pattern evolve into a clonal population, developing into hepatocellular carcinomas, which are also clonal.

Preinvasive Lesions

The field cancerization hypothesis states that multiple cells form independent tumours, since carcinogenic exposure affects multiple cells in the field (Slaughter *et al*., 1953), and predicts that second primary or synchronous tumours arise from independent genetic events. Thus 11% of individuals with oral cancer had multiple upper aerodigestive tract tumours, and multiple invasive foci are associated with overlying areas of *in situ* squamous carcinoma in these lesions. There is now genetic evidence for such an independent origin: in the upper aerodigestive tract, multiple synchronous squamous tumours appear independent and multicentric in origin. However, the concept of clonal origin and expansion is problematic in organs where several metachronous tumours appear; such a synchronous or second primary tumour may indicate recurrence or indeed lateral spread from a single tumour. Thus, although the field cancerization hypothesis would predict a multicentric, polyclonal origin, with the demonstration of a clonal origin for these tumours, lateral migration from the original clone would be a distinct possibility.

There is conflicting evidence for the nature of field cancerization from the study of tumours of the upper aerodigestive tract and adjacent mucosa. In laryngeal and pharyngeal tumours, multiple samples taken at tumour-distant sites show different and independent mutations in the *p53* gene, favouring a discontinuous, multifocal and polyclonal process, rather than migration of premalignant basal keratinocytes giving a clonal development of multiple primary, secondary or recurrent tumours.

In the stomach, discordant mutation patterns of *APC*, *MCC* and *p53* are found in many cases of multiple gastric carcinomas, again in accord with a multicentric origin, and this is also seen in multiple colorectal tumours, parathyroid adenomas and in separate (Emmert-Buch *et al*., 1995) intraepithelial neoplastic lesions, which show different clonal patterns of allele loss at 8p12-21, suggesting an

independent origin. However, multiple synchronous carcinomas in the bladder and other pelvic organs apparently show a common clonal origin – X-chromosome inactivation and allele loss at 9q and 17p are identical, as are c-*erB2* and *p53* mutations in multiple synchronous urothelial tumours (Lunec *et al*., 1992). Multiple serous adenocarcinomas in the ovaries, peritoneum and endometrium show the same p53 mutation and clonal cytogenetic abnormalities, findings confirmed by X-chromosome inactivation. Multiple sites of occurrence of sporadic ovarian cancer on the ovarian surface and pelvic peritoneum are clonal, as assessed by LOH, p53 mutations and X-chromosome inactivation analysis in the same patient (Jacobs *et al*., 1992).

In multifocal breast carcinomas, an increasingly common finding, karyotypically identical clones are detectable, indicating intramammary spread from a single carcinoma either by focal lymphatic spread or by intraductal spread. Moreover, in phyllodes tumours, widely separated deposits show the same monoclonal stromal component.

If lesions some distance away from each other are clonal, how can this be explained? Of course, there is always the possibility, always very difficult to exclude, that the disease process which causes the tumour to develop has a characteristic genetic fingerprint, therefore seen in all examples of the tumour. The other possibility is that of a common mutated progenitor cell, and the expansion of this mutant clone in some way, at an early stage in tumour development. Examination of *p53* mutations in the nonmalignant but dysplastic tracheobronchial mucosa of individuals who smoke shows the same G:C to T:A transversion in codon 245 at multiple sites in both lungs (Franklin *et al*., 1998). Thus the expression and dispersion of a single mutant progenitor bronchial epithelial cell clone throughout the airways is possible, aided possibly by a proliferative advantage. p53 mutations are early events in upper aerodigestive tract carcinogenesis, prior to the development of invasive lesions, being found in premalignant lesions of the head and neck, lung and oesophagus. p53 mutations apparently do not show an increased incidence with cancer progression, but do show clonal fidelity in a variety of tumours. An early event, prior to p53 mutation, might establish a mutated clone, which migrates laterally, possibly aided by a mutation in a cell cycle control gene, or a cell adhesion gene. In multiple bladder cancer, all tumours lose the same 9q allele as an early event (Chung *et al*., 1995, 1996), possibly encoding for a growth control or adhesion molecule, and cells repopulate the urothelium by lateral migration or mucosal seeding. In most discontinuous foci of CIN3 in the cervix, individual lesions show the same X-chromosome inactivation pattern, suggesting intraepithelial spread. In the skin, it is not uncommon to see migration of morphologically abnormal cells laterally from a lesion such as Bowen's disease.

The multiple deposits of Kaposi sarcoma, a widely disseminated malignancy, appear clonal in any one patient, indicating a clonal origin and wide intravascular dissemination, or an initial vascular hyperplasia, with later clonal evolution. Diffusely infiltrating gliomas are clonal, and multiple discrete meningiomas share clonal neurofibromatosis 2 (*NF2*) mutations, while most individual meningiomas appear clonally derived.

Sporadic, multinodular goitres contain nodules which are regarded as being hyperplastic and therefore polyclonal, but there is substantial X-chromosome inactivation evidence that many of these nodules are clonal populations. The presence of a TSH receptor mutation may be pivotal; most cases showed monoclonality on HUMARA analysis, raising the possibility that during thyroid hyperplasia a cell with a mutation at this locus leads to the initiation of autonomous clonal growth. Multiple nodules in the same patient are mostly clonal, with activation of the same allele, indicating intraglandular spread by follicular budding, although clonal nodules with different X-chromosome inactivation patterns can be seen in the same gland. This could mean a different pathogenetic mechanism for clonal and polyclonal nodules, or indeed evolution of clonal from polyclonal nodules, as we have seen above in the liver. Most follicular, papillary and anaplastic carcinomas are clonal. Parathyroid adenomas are monoclonal, as indicated by X-chromosome inactivation and the identification of clonal abnormalities of the parathyroid hormone gene. In MEN1, where all four parathyroids are enlarged and appear hyperplastic, allelic loss on 11q indicates that proliferation is clonal, although when LOH on 11q is combined with X-chromosome inactivation studies, the parathyroid lesions in MEN1 were shown to be polyclonal, suggesting that multiple neoplastic clones grow, coalesce and replace the parathyroid gland. Sporadic parathyroid hyperplasias, which are either a primary phenomenon or secondary to such conditions as chronic renal failure, were previously regarded as polyclonal proliferations, but X inactivation and allelic loss of 11q shows that 38% of primary parathyroid hyperplasias and 64% of hyperplasias secondary to renal failure harbour clonal proliferations. Again we can interpret this as clonal neoplastic evolution from a pre-existing polyclonal hyperplasia. Diffuse parathyroid hyperplasias in uraemia are polyclonal, but the individual nodules in nodular hyperplasia are clonal, indeed showing different clonal patterns of X inactivation in the same gland, once more favouring the hypothesis that monoclonal proliferations evolve from hyperplasias. These findings favour the view that hyperplasia begins in endocrine glands as a polyclonal process, but then becomes a clonal hyperplasia, and the borderline between this phase and that of a benign clonal neoplasm is difficult to delineate.

Invasive carcinomas of the cervix, which are clonal in origin, arise in areas of cervical intraepithelial neoplasia (CIN). Severe dysplasia or CIN3 is also a clonal

proliferation, although lesser degrees of dysplasia (CIN2) appear more commonly polyclonal in X-inactivation studies. In this respect, some vulvar hyperplasia, considered to be preneoplastic lesions in this tissue, appear clonal: the derivative VIN and invasive carcinomas are also clonal proliferations. In nasopharyngeal carcinoma (NPC), combined X-inactivation studies (PGK), X-linked RPLPs, and EBV integration show that carcinomas are mainly clonal, but hyperplastic epithelia, and early atypical hyperplastic epithelia, are polyclonal, and clonality emerges only at the moderate/severe dysplastic stage.

A relationship between intestinal metaplasia and carcinoma of the stomach and lower oesophagus has long been suspected: the non-dysplastic metaplasia adjacent to carcinomas is clonal on X-inactivation analysis (HUMARA), and also shows LOH for *APC*, changes also seen in the dysplastic and neoplastic tissues of Barrett oesophagus. Although it is not yet clear if the *APC* change is clonal, microsatellite analysis shows allelic imbalance not only on 5q but at multiple other sites, in both premalignant and malignant Barrett epithelium, supporting the concept of clonal expansion from metaplasia through dysplasia to carcinoma (Zhuang *et al*., 1996).

In most, but not all, of the above discussion the commonality has been the need to homogenize the tissue, albeit, in some cases, after microdissection. Evidence from studies where direct observation is possible gives disparate results. In a patient with familial adenomatous polyposis, who was also a sex chromosome chimaera (XO/XY), the colon contained hundreds of adenomas, ranging in size from monocryptal adenomas to microadenomas 2.5 mm in diameter; no larger adenomatous polyps were seen (Novelli *et al*., 1996). If an adenoma was of clonal origin, all dysplastic crypts within it would be expected to be entirely XO or entirely XY. Localization of the Y chromosome in tissue sections showed that monocryptal adenomas were entirely XO or XY, with no mixed pattern. However, many adenomas (76%) were polyclonal. Isoenzyme studies of G6PD in black females have shown that colonic adenomas from patients with Gardner's syndrome were polyclonal (Hsu *et al*., 1983), while studies using X-linked RFLPs show that both spontaneous and familial adenomas are clonal (Fearon *et al*., 1987). The minimum size of the adenomas in this might explain the discrepancy, through monoclonal conversion as size increases. Polyclonal derivation of adenomas has also been found in chimaeric mice – made between a *Min* mouse (which has an *APC* mutation) and a *Min*/ROSA mouse (Merrit *et al*., 1997), and in $Apc+/^{min}/+/-$ (De Wind *et al*., 1998) mouse chimaeras. Possible mechanisms of polyclonality in these lesions include 'field' effects causing adenomas to cluster (nonrandom collision), a passive process involving fusion of two or more *APC*-negative clones early in tumour development, but the high frequency of mixed adenomas found in both these studies is inconsistent with a random appearance of *APC*-negative clones in the mucosa, and suggests that some regions of the intestine have an increased potential for initiation. Multiple clones may be required for early adenoma growth, or perhaps early adenomas may induce adenomatous growth in surrounding crypts, especially in FAP since all cells already have a single *APC* mutation and perhaps some derangement of APC function. These latter scenarios imply active cooperation between clones. Indeed, conversion of normal crypts to adenomatous crypts apparently occurs at the margins of FAP adenomas.

What can we conclude? In many tissues the bulk of the evidence indicates that preneoplastic changes occur which are hyperplastic, but polyclonal, indicating increased rates of cell production among many cells. After some time, a genetic event occurs in which a clone of cells appears, which expands and establishes itself. This clone has a 'carcinogenic advantage,' which allows such expansion. Further growth of the clone is associated with genetic evolution and the appearance of mutations, which lead to the development of an early clonal neoplasm. Further clonal evolution develops the invasive phenotype. However, we must concede that further studies, with methods which allow the direct examination of clonality, might provide different conclusions, and we cannot disregard the concept that the earliest lesions are polyclonal, needing the cooperation of several clones, and that monoclonality is itself the result of further clonal evolution.

HOW DO CLONES ESTABLISH THEMSELVES?

The simplest way in which we can conceive of a single mutated cell establishing itself among a population of normal cells is if the mutation selects for an advantage that involves factors which modulate cell proliferation – the ability to divide faster – or the ability to survive longer. While conceptually these are significant advantages to being able to survive longer, or indefinitely, in the genesis of a neoplasm, it is, in practice, very difficult to demonstrate. While some early neoplastic lesions do show abnormal expression of genes involved in apoptosis, this is a far cry from demonstrating a aggregation of cells brought about by reduced cell death. Moreover, simple modelling approaches have suggested that mechanisms involving apoptosis are unlikely to be involved at this stage.

Is there any evidence that very early neoplastic lesions show an increased cell production rate? Using the colon as a paradigm, the earliest lesions detectable, the so-called 'aberrant crypt foci,' show increased labelling indices with a number of markers of cell proliferation, whereas in established adenomas, the rate of cell production also increases as assessed by these markers. The importance of even a minor increase in cell proliferation

can be assessed by examination of how a mutated stem cell – of which there are several in each intestinal crypt – competes with its colleagues for ascendancy, and in some cases succeeds, populating the crypt with its progeny. Intestinal crypts, in both animals and humans, are clonal populations, ultimately derived from a single cell. The evidence from this comes from the analysis of chimaeric mice and mice heterozygous for a defective G6PD gene carried on the X chromosome and randomly expressed after Lyonization. In these chimaeras, crypts are always composed wholly of cells from one component mouse strain (Ponder *et al*., 1985), and with similar findings in the G6PD heterozygote crypts are indeed clonal populations and the stem cell repertoire includes all intestinal crypt and gastric gland lineages. **Figure 2 (see colour plate section)** shows the results of an experiment in which mice showing uniform staining for G6PD are given a single dose of mutagen: in the weeks that follow, crypts appear which are composed of cells with a different, mutated phenotype. There is an induction of a rapid, but transient, increase in crypts which show a partial, or segmented, mutated phenotype (**Figure 2a**) (Park *et al*., 1995). Later, there is an increase in crypts showing a wholly mutated phenotype. The emergence of the partially mutated crypts, and their replacement by wholly mutated crypts, can be explained by a mutation at the G6PD locus in a single stem cell from which all lineages are derived. The partially mutated crypts are crypts in the process of being colonized by progeny from the mutated stem cell, and this crypt will ultimately develop into a wholly mutated crypt.

Approximately 9% of the Caucasian population secrete sialic acid lacking in *O*-acetyl substituents, readily recognized by histochemical staining. This is explained by genetic variability in the expression of the enzyme *O*-acetyl transferase (OAT) and this 9% of the population is homozygous for inactive *OAT* genes, *OAT*−/*OAT*−. Some 42% of the population are heterozygous, *OAT*−/*OAT*+, and loss of this gene converts the genotype to *OAT*−/−*OAT*−. This is indeed seen in about 42% of the population, and again is explained by a mutation or loss of the gene by nondisjunction in a single crypt stem cell and the colonization of this crypt by the clonal progeny of the mutated stem cell.

If we again imagine that a stem cell carries a mutation, not in its *G6PD* or its *OAT* gene, but in a cell cycle gene, increasing its cell production rate, then the monoclonal conversion process will not be stochastic, with equal competition between stem cells, but the mutant stem cell would have an advantage, and would succeed in colonizing the crypt. **Figure 2c** shows a further method in which mutated clones can be propagated – the mutated clone has expanded by crypt fission, where crypts divide to make two daughter crypts, and thus spread through the mucosa. If a gene controlling crypt fission became mutated, then of course the crypt would spread further through the mucosa. Indeed, in conditions such as ulcerative colitis, large areas of the mucosa become occupied by the same aneuploid stem line, conceptually, at any rate, the result of increased crypt fission.

In squamous epithelium we have seen above that tiny p53-positive clones, containing mutated p53, occur early in the carcinogenic process. These clones may owe their existence to prolonged survival, given the presence of a mutated *p53* gene. However, there is substantial evidence that rates of cell production, in preneoplastic lesions such as actinic keratosis, are elevated. In fact, it is possible to predict the shape of preinvasive skin lesions such as actinic keratosis, Bowen disease and lentigo maligna, by assuming different values for the carcinogenic advantage, in terms of increased cell production. Similar observations of elevated rates of cell production have been recorded in other situations where there is a defined morphological sequence of carcinogenesis – in the liver, bladder, oesophagus and lung, for example.

We have seen above that, in many tissues, clonal evolution occurs on a background of hyperplasia, implicitly involving increased cell production. It is further evident that the establishment of the mutated clone is, in many tissues, associated with an increase in cell production rate. It is easy to conceive now such increases in cell proliferation, in cells such as stem cells, can lead to the establishment and development of a mutated clone (**Figure 2**). However attractive the concept seems, we cannot say definitively whether such increased cell proliferation of stem cells is involved. For example, in **Figure 2**, increased longevity, or changes in the adhesion molecule status of the progeny, could give much the same result. An ability to study cell proliferation, and then expression of cell cycle-related genes, in stem cells during carcinogenesis, would be a very useful, if experimentally difficult, prospect.

HOW DO ESTABLISHED TUMOURS GROW?

The basic concepts of cell proliferation in tumour growth have been extensively and well reviewed.

Tumour Growth Curves

There has been extensive analysis of tumour growth curves – obtained by plotting the volume or weight of the tumour against time – in the past, in the (forlorn) expectation that basic truths about the nature of tumour growth might be discovered. Although this has not been the case, some interesting facts have emerged. Many tumours, usually transplantable, growing in experimental animals, initially grow in an exponential manner, that is, after an initial lag phase, there is logarithmic growth (**Figure 3**). However, as growth proceeds, and the

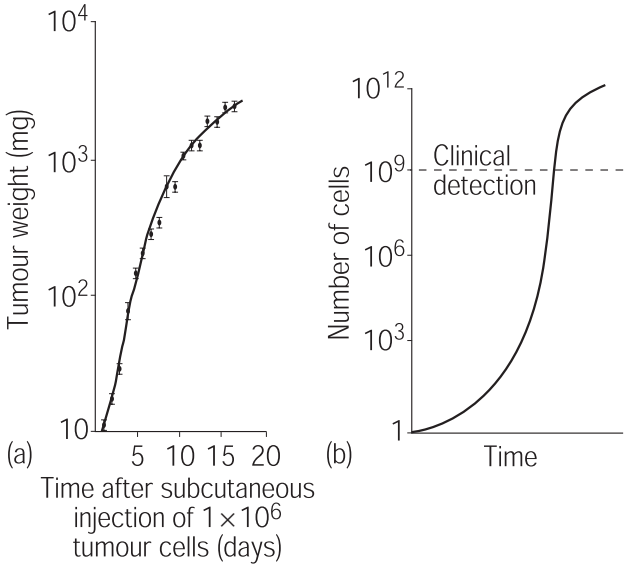


Figure 3 (a) A typical growth curve for a rapidly growing mouse tumour, with the curve fitted by eye; (b) idealized growth curve for a human tumour, showing the initial lag phase, a period of exponential growth and a retardation in the rate of increase of tumour cell number at large size. Note the long latent period before clinical detection.

tumour increases in volume, growth progressively slows. This is the so-called Gompertzian growth curve, where growth retardation is itself exponential. Although such tumours rarely obey such a growth curve in its entirety, it is usually a good approximation. Note that, at the asymptote, growth is barely discernible. Most experimental tumours in rodents grow very rapidly, particularly the transplantable ones, since they have been selected for a rapid growth rate.

In humans, tumours grow much more slowly; most of the doubling in cell number has occurred before the tumour becomes clinically evident. In circumstances where it has been possible to measure the rates of tumour growth, e.g. by serial observation of pulmonary primary or metastatic lesions, a number of tumours do indeed appear to grow exponentially. However, in other cases, growth is extremely variable, with periods of quiescence followed by rapid growth. Some human tumours grow extremely slowly, e.g. colorectal carcinoma has a reported doubling time (the time taken for the tumour to double in size) of over 600 days. Others, such as childhood or embryonal tumours, show doubling times of 10–20 days. There is some evidence that tumour growth rates vary with age; for example, breast carcinomas in women below the age of 50 years show a median doubling time of 80 days, whereas women over the age of 50 years show a doubling time of 140 days.

Kinetic Parameters During Tumour Growth

Why does tumour growth slow down? Well, the net rate of growth is defined by the relationship between the rate of cell production – the birth rate – and the rate of cell death or other modes of loss, the cell loss rate. In the earlier phases of growth, as seen in the early exponential part of the curve in **Figure 3**, most cells are growing exponentially; thus every cell which divides contributes two cells to the population, with effectively no cell loss. What defines the rate at which the exponential portion of the curve ascends?: two parameters – the cell cycle time (the interval between two divisions of the same cell) and the growth fraction, or proliferating population (the proportion of cells in the tumour which are dividing). There is no evidence in solid experimental tumours, at any rate, that the cell cycle time changes at all during tumour growth, but as growth advances, there are substantial studies showing that the growth fraction declines with time. In experiments where the growth fraction has been measured in different parts of the tumour, it is the areas furthest from the blood supply that show the largest reductions in growth fraction – the centre of the tumour in most lesions, or when the tumour is 'corded,' or shows cords of surviving tumour cells surrounding blood vessels – in the distal portions of the cord.

While the decline in the growth fraction is certainly responsible for a proportion of the decline in the rate of tumour growth, by far the largest factor is the considerable increase in the cell loss rate. This is due, of course, to the tumour outgrowing its blood supply and to large numbers of cells in the lesion dying. In fact, at the asymptote in the curve, where the curve levels off, the rate of cell loss almost equals the rate of cell birth, with little resulting in the way of net growth. It is sometimes useful to define a cell loss factor, ϕ, which is the ratio of the cell loss rate to the birth rate. At the asymptote in **Figure 3**, this will be close to 1.

Where data are available, experience culled from animal tumours is also applicable to humans. However, measurements are usually confined to the later phases of tumour growth, since most lesions present, biologically speaking, very late. Thus many carcinomas appear to be low-growth fraction, high-cell loss lesions. In clinically detected colorectal carcinomas, for example, the cell cycle time is 48 h, the growth fraction is 14% and the cell loss factor 98%, i.e. for every 100 cells which are born, two survive. At least this gives us some insight into why such tumours are difficult to treat with modalities such as radiation and cytotoxic chemotherapy. Of the surviving cells in the tumour, some 86% are not in the cell cycle, mainly because they are hypoxic. Such cells would be resistant to irradiation, and even many fractions might not recruit many cells into cycle. Similarly, 5-fluorouracil, commonly used in these lesions, is, in the main, an S-phase poison. With so few cells in cycle, a brisk response would

not be expected, unless the growth fraction is increased by debulking and redistribution.

CELL PROLIFERATION IN PROGNOSIS AND CANCER THERAPY

There has been a great deal of effort and energy expended in attempts to relate measurements of cell proliferation to both prognosis and the prediction of tumour response, and to define subgroups requiring additional or adjuvant therapy. It is probably true to say that such attempts have failed.

It is now some 80 years since measurements of mitotic activity were first routinely made in human tumours, and some 40 years since tritiated thymidine, the first real S-phase-specific marker, became available. Since then, measurements using techniques such as flow cytometry, bromodeoxyuridine labelling and the several antibodies against the Ki67 antigen, have followed. to say nothing of numerous attempts to clone human tumour cells *in vitro* and predict their levels of chemosensitivity against a panel of agents, a method recently resurrected. There have also been attempts to plan chemotherapeutic regimens on the basis of measurements of cell proliferation, e.g. the so-called 'synchronization therapy' for head and neck cancers and acute leukaemias, where doses of cytotoxics are given when the tumour cells are 'synchronized' in a cell cycle phase most sensitive to the agent in question.

It is perhaps salient to ask which of these measurements are generally and routinely used in cancer treatment outside of research protocols. Precious few, is the sobering answer. In tumour diagnosis, pathologists routinely do a formal mitotic count (rather than merely observe the presence of mitoses) in smooth muscle cell tumours of the uterus, some central nervous system tumours and stromal tumours of the gastrointestinal tract. In the grading of tumours, and therefore in the prediction of prognosis, mitotic counts are routinely carried out in a number of circumstances, as in the assessment of grade of invasive ductal carcinoma of the breast as a component of the Nottingham Prognostic Index or sometimes in ovarian carcinomas, either as a raw count as a component of a morphometric index. And have more sophisticated measurements reached routine status? Some advocate the use of Ki67 in the differential diagnosis of difficult melanocytic lesions, but there has been little use of such methods elsewhere. Perhaps the advent of new markers, such as antibodies against the minichromosome maintenance proteins, said to be more sensitive than Ki67, may make a difference.

CONCLUSIONS

In human cancer, it is clear that cell proliferation is involved in the origin and evolution of mutated clones, although the details at present elude us. Cell proliferation studies give us insight into how tumours grow, and in some instances why treatment is so difficult. Measurements of cell proliferation are at present of little routine value in diagnosis and prognosis. It is clear that a great deal of work needs to be done if we are to understand fully the role of cell proliferation in tumour development.

REFERENCES

Aihara, T., *et al*. (1994). Clonal analysis of regenerative nodules in hepatitis C virus-induced liver cirrhosis. *Gastroenterology*, **107**, 1805–1811.

Ajiki, T., *et al*. (1994). K-*ras* gene mutation related to histological atypias in human colorectal adenomas. *Biotechnical Histochemistry*, **70**, 90–94.

Benditt, E. P. and Benditt, J. M. (1973). Evidence for a monoclonal origin of human atherosclerotic plaques. *Proceedings of the National Academy of Sciences of the USA*, **70**, 1753–1756.

Beutler, E., *et al*. (1967). Value of genetic variants of glucose-6-phosphate dehydrogenase in tracing the origin of malignant tumours. *New England Journal of Medicine*, **276**, 389–391.

Chung, G. T. Y., *et al*. (1995). Sequential molecular changes in lung cancer development. *Oncogene*, **11**, 2591–2598.

Chung, G. T. Y., *et al*. (1996). Clonal evolution of lung tumours. *Cancer Research*, **56**, 1609–1614.

Chung, I.-M., *et al*. (1998). Clonal architecture of normal and atherosclerotic aorta. Implications for atherogenesis and vascular development. *American Journal of Pathology*, **152**, 913–923.

De Wind, N., *et al*. (1998). Mouse models for hereditary nonpolyposis colorectal cancer. *Cancer Research*, **58**, 248–255.

Emmert-Buch, M. R., *et al*. (1995). Allelic loss on chromosome 8p12–21 in microdissected prostatic intraepithelial neoplasia. *Cancer Research*, **55**, 2959–2962.

Farber, E. (1990). Clonal adaptation during carcinogenesis. *Biochemical Pharmacology*, **39**, 1837–1846.

Farber, E. (1997). Monoclonal and polyclonal development of digestive tract tumors in chimeric mice. *Japanese Journal of Cancer Research*, **88**, 663.

Fearon, E. R., *et al*. (1987). Clonal analysis of human colorectal tumours. *Science*, **238**, 193–197.

Fialkow, P. (1973). Primordial cell pool size and lineage relationships of five human cell types. *Annals of Human Genetics*, **37**, 39–48.

Fialkow, P. J. (1976). Clonal origin of human tumours. *Biochimica Biophysica Acta*, **458**, 283–321.

Franklin, W. A., *et al*. (1997). Widely dispersed p53 mutation in respiratory epithelium. *Journal of Clinical Investigation*, **100**, 2133–2137.

Hsu, S. H., *et al*. (1983). Multiclonal origin of polyps in Gardner syndrome. *Science*, **251**, 951–953.

Jacobs, I. J., *et al*. (1992). Clonal origin of epithelial ovarian carcinoma; analysis by loss of heterozygosity, p53 mutation,

and X-chromosome inactivation. *Journal of the National Cancer Institute*, **84**, 1793–1798.

Jimbo, H., *et al*. (1997). Evidence for monoclonal expansion of epithelial cells in ovarian endometrial cysts. *American Journal of Pathology*, **150**, 1173–1178.

Jones, P. A. (1996). DNA methylation errors and cancer. *Cancer Research*, **56**, 2463–2467.

Lakhani, S. R., *et al*. (1995). Atypical ductal hyperplasia of the breast: clonal proliferation with loss of heterozygosity on chromosomes 16q and 17p. *Journal of Clinical Pathology*, **48**, 611–615.

Lunec, J., Challen, C., Wright, C., Mellon, K. and Neal, D. (1992). *c-erB2* amplification and identical p53 mutations in comcomitant transitional carcinomas of renal pelvis and urinary bladder. *Lancet*, **339**, 439–440.

Merrit, A. J., *et al*. (1997). Polyclonal structure of intestinal adenomas in ApcMin/+ mice with concomitant loss of Apc+ from all tumor lineages. *Proceedings of the National Academy of Sciences of the USA*, **94**, 13927–13931.

Murry, C. E., *et al*. (1997). Monoclonality of smooth muscle cells in human atherosclerosis. *American Journal of Pathology*, **151**, 697–706.

Novelli, M. R., *et al*. (1996). Polyclonal origin of colonic adenomas in an XO/XY patient with FAP. *Science*, **272**, 1187–1190.

Nowell, P. C. (1976). The clonal evolution of tumour cell populations. *Science*, **194**, 23–28.

Nowell, P. C. (1986). Mechanisms of tumor progression. *Cancer Research*, **46**, 2203–2207.

Park, H.-S., *et al*. (1995). Crypt fission in the small intestine and colon: a mechansism for the emergence of glucose-6-phosphate locus-mutated crypts after mutagen treatment. *American Journal of Pathology*, **147**, 1416–1427.

Ponder, B. A. J., *et al*. (1985). Derivation of mouse intestinal crypts from single progenitor cells. *Nature*, **313**, 689–691.

Schmidt, G. H. and Mead, R. (1990). On the clonal origin of tumours – lessons from studies of intestinal epithelium. *Bioessays*, **12**, 37–40.

Slaughter, D. P., *et al*. (1953). 'Field cancerisation' in oral stratified squamous epithelium. *Cancer*, **6**, 963–968.

Woodruff, M. F. A., *et al*. (1982). Clonal interaction in tumours. *Nature*, **299**, 822–824.

Zhuang, Z., *et al*. (1996). Barrett's esophagus: metaplastic cells with loss of hetrozygosity at the *APC* gene locus are clonal precursors to invasive adenocarcinoma. *Cancer Research*, **56**, 1961–1964.

FURTHER READING

Garcia, S., *et al*. (1999). Field cancerization, stem cells and clonality: the spread of mutated clones in epithelial sheets. *Journal of Pathology*, **187**, 61–81.

Garcia, S., *et al*. (2001). The clonal origin and clonal evolution of human epithelial tumours. *International Journal of Experimental Pathology*, **81**, 117–143.

Hall, P. A., *et al*. (eds) (1992). *Clinical Aspects of Cell Proliferation* (Springer, Heidelberg).

Kinzler, K. W. and Vogelstein, B. (1997). Gatekeepers and caretakers. *Nature*, **386**, 761–763.

Pretlow, T. P., *et al*. (1994). Aberrant crypt foci and colon tumours in F344 rats have similar increases in proliferative activity. *International Journal of Cancer*, **56**, 599–602.

Quinn, C. and Wright, N. A. (1990). The clinical assessment of proliferation and growth in human tumours; evaluation of methods and applications as diagnostic variables. *Journal of Pathology*, **160**, 93–102.

Shibata, D. (1998). The dynamics of early intestinal tumour proliferation: to be or not to be. In Precancer: Biology, Importance and Possible Prevention. *Cancer Surveys*, **32**, 181–200.

Steele, G. G. (1977). *Growth Kinetics of Tumours* (Clarendon Press, Oxford).

Wright, N. A. and Alison, M. R. (1984). *The Biology of Epithelial Cell Populations*, Vols 1 and 2 (Oxford University Press, Oxford).

Wright, N. A. (2001). Stem cell repertoire in the intestine: the origins of self-renewal, cell lineage and cancer. *International Journal of Experimental Pathology*, **81**, 89–116.

Chapter 19

Identifying Cancer Causes Through Epidemiology

Hans-Olov Adami
Karolinska Institutet, Stockholm, Sweden

Pagona Lagiou
University of Athens Medical School, Goudi, Athens, Greece

Dimitrios Trichopoulos
Harvard School of Public Health, Boston, MA, USA

CONTENTS

INTRODUCTION

Cancer control is an important objective of contemporary medicine. It may be achieved through primary prevention (e.g., reduction of exposure to carcinogenic agents or increase in resistance towards them by immunization or other procedures), down-staging (earlier recognition of clinical symptoms or signs), screening for preclinical diagnosis of asymptomatic cancer or preneoplastic conditions, curative and palliative treatment and rehabilitation schemes. Although these approaches are complementary, there are strong theoretical arguments and overwhelming empirical evidence suggesting that primary prevention represents the most promising strategy for effective cancer control. Implementation of primary prevention requires identification of carcinogenic agents and of the conditions that favour the exposure of susceptible individuals to these agents. An agent is considered to be carcinogenic when a change in the frequency or intensity of exposure to this agent is accompanied by a predictable change in the frequency of occurrence of cancer of particular type(s) at a later time.

The issue of causal inference has generated intense debates among both philosophers and scientists (Rothman, 1988). In a simplified scheme that reflects, to a considerable extent, the ideas of Sir Karl Popper (Buck, 1975), causal inference follows a pattern of interconnected cycles. Each cycle includes an examination of the existing data in the light of the prevailing views and questions, formulation of an aetiological hypothesis that adequately addresses these questions and evaluation of the compatibility of the hypothesis with new sets of data specifically generated or assembled for this purpose. Compatibility cannot be equated to proof, but a hypothesis gains credibility when it repeatedly resists refute.

In 1969, the World Health Organisation (WHO), through its International Agency for Research on Cancer (IARC), initiated a programme to evaluate carcinogenic risks to humans and to produce monographs on individual agents, groups or mixtures (IARC, 1987). For the evaluation, evidence for carcinogenicity in humans and experimental animals, and also other relevant data in experimental systems and humans, are taken into account. However, in the final overall evaluation, an agent is judged to be 'conclusively' carcinogenic in humans when there is sufficient evidence of carcinogenicity in humans derived from relevant and valid epidemiological studies. Evidence for carcinogenicity in experimental animals and *in vitro* experimental systems is given less weight, because of well-known problems related to species specificity. Series of clinical case reports are also considered, since they can be thought of as incomplete epidemiological processes, with 'control' information based implicitly on background clinical experience and information.

THE 'NATURAL HISTORY' OF THE EPIDEMIOLOGICAL IDENTIFICATION OF A CANCER CAUSE

The formulation of aetiological hypotheses is usually based on the examination of existing data. These data may represent the results of studies in experimental animals, e.g. the occurrence of papillary carcinoma in the bladder of mice after exposure to tobacco tar encouraged investigators to examine whether an association between tobacco smoking and bladder cancer also existed in humans. In other instances, the data may refer to 'unusual' or 'interesting' cases reported in the clinical literature, e.g. the hypothesis linking inorganic trivalent arsenic compounds to skin cancer (IARC, 1980a) and phenacetin-containing analgesics to renal pelvic carcinoma (IARC, 1980b) have been based, to a large extent, on clinical observations and pathophysiological considerations. There have also been situations where hypotheses were developed and subsequently tested on the basis of biological and theoretical arguments. For example, it was hypothesized that passive smoking may cause lung cancer, because sidestream smoke is not qualitatively different from mainstream smoke, and there is no threshold in the dose-dependent relation between active smoking and risk of lung cancer (Trichopoulos, 1994).

In most instances, however, aetiological hypotheses are developed on the basis of statistical associations between the cancer under consideration on the one hand, and various personal characteristics of the affected individuals, in addition to the time and place occurrence pattern of their disease, on the other. By revealing who, when and where are affected by a particular cancer, one has already gone a long way towards discovering why the particular cancer has occurred. Such observations, collectively considered under the term 'descriptive epidemiology,' represent either the products of routinely recorded information or the byproducts of analytical epidemiological studies designed to address other specific aetiological hypotheses.

An aetiological relation presupposes the existence of a statistical association, but for diseases such as cancer, which are defined according to criteria at the histological, cytological or subcellular level, a cause does not have to be, and usually is not, either necessary or sufficient. For example, not all hepatitis B virus carriers develop hepatocellular carcinoma, and this cancer can develop without the presence of, or even exposure to, the hepatitis B virus (Stuver, 1998). Furthermore, the existence of a statistical association between a particular agent and a particular form of cancer does not necessarily imply the existence of an underlying causal relation. It is possible, in fact common, that the association reflects coexistence of the particular agent with another agent (the confounding factor) which represents the real cause of the particular cancer. Thus, lung cancer patients may report excessive use of alcoholic beverages, simply because in several cultures tobacco smoking and alcohol drinking tend to be positively correlated (**Figure 1**). Even in the absence of confounding, a statistical association is not an infallible indication of a causal relation. Non-smoking lung cancer patients, for example, may report higher alcohol intakes compared with healthy individuals, because they provide more truthful and accurate histories of habits for which there is real or perceived social disapproval (information bias).

Analytical epidemiological studies are designed to explore whether an association between a particular agent or characteristic and a particular cancer actually exists, and what is its real strength after eliminating, as far as possible, all recognizable effects of confounding and bias. The most commonly used measures of strength of the association between a particular agent or characteristic and a particular cancer is the relative risk; this generic term covers the rate ratio, the risk ratio and the odds ratio. The relative risk indicates how many times higher (or lower) is the frequency of occurrence of the particular cancer among individuals exposed to the agent (or possessing the characteristic), compared with individuals not exposed to the agent (or not possessing the characteristic).

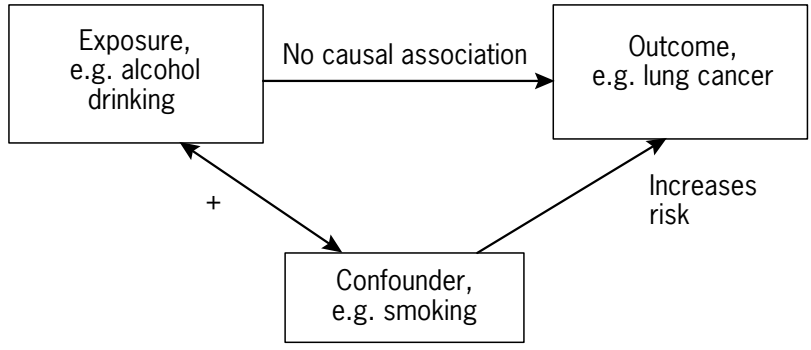


Figure 1 Confounding of the association between consumption of alcoholic beverages and lung cancer by tobacco smoking.

Analytical epidemiological studies have been traditionally designated as cohort or case control. In cohort studies, exposed and nonexposed individuals are followed over time and the frequency of occurrence of the cancer under investigation in the two (or more, if several exposure levels can be ascertained) groups is calculated, allowing the direct estimation of the incidence rate ratio, which is a variant of the relative risk (**Figure 2**). In case-control

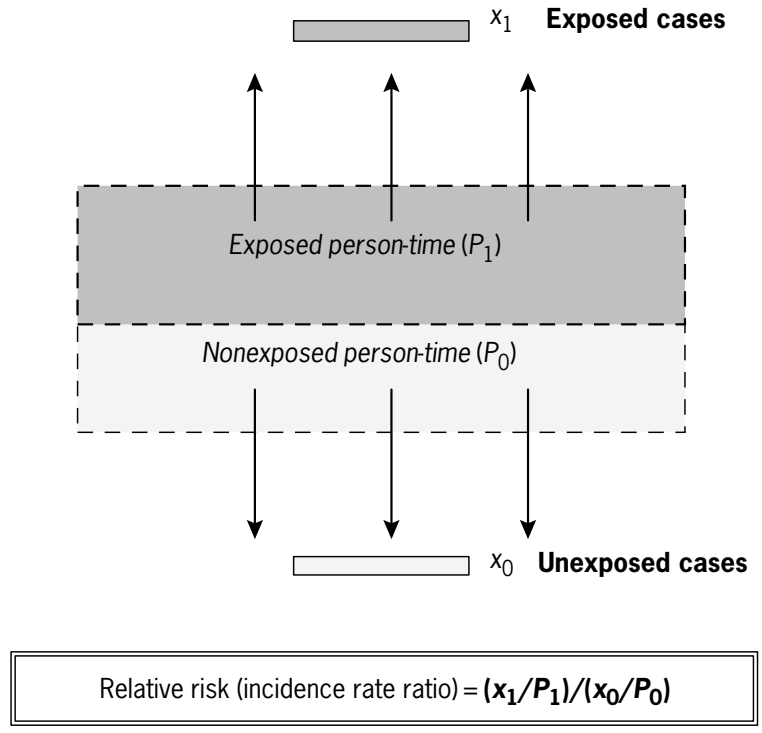


Figure 2 A cohort study. (Adapted from Walker, 1991.)

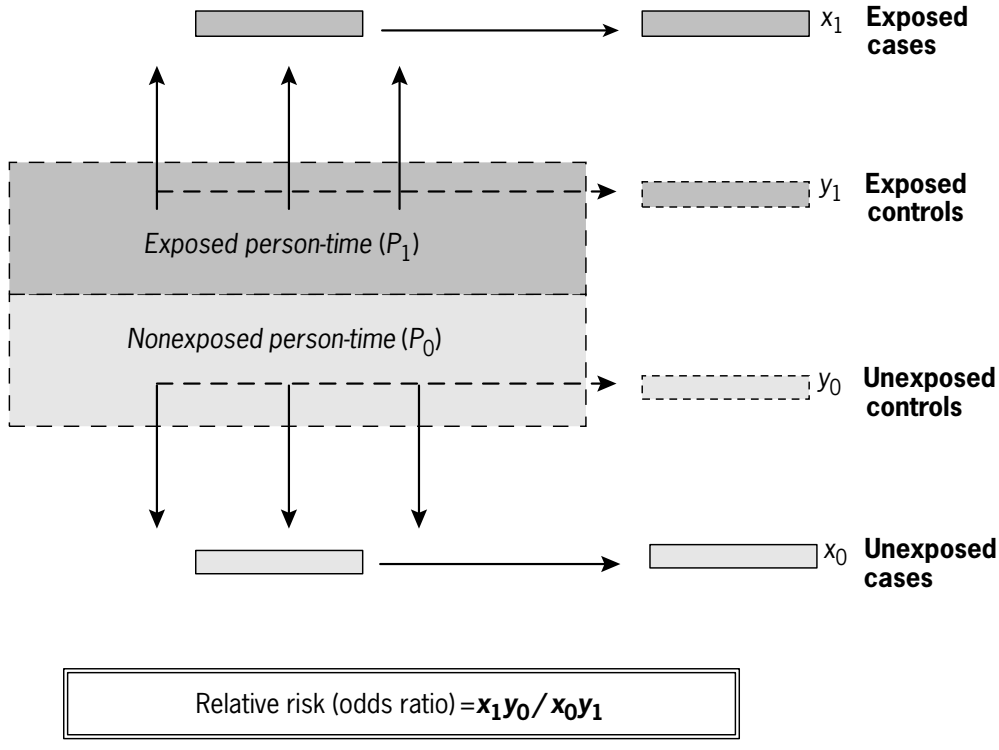


Figure 3 A case-control study. (Adapted from Walker, 1991.)

studies, patients with a recent diagnosis of the cancer under investigation are compared with individuals free of this cancer. These individuals (controls) should be representative of the population which gave rise to the cases with respect to the antecedent frequency of exposure to the agent under study. It can be shown that this design permits the calculation of the odds ratio, which is another variant of the relative risk. The actual frequency of occurrence of the cancer under consideration among exposed and non-exposed individuals usually cannot, however, be calculated in a case-control study (**Figure 3**).

In both cohort and case-control studies, measurable confounding factors and identifiable biases can be controlled for in the analysis, either by stratification or by multivariate modelling (Breslow and Day, 1980, 1987). Contemporary authors have indicated that a sharp distinction between cohort and case-control studies is artificial and unjustifiable (Miettinen, 1985; Rothman and Greenland, 1998). However, most epidemiological studies are still reported under these headings and their distinction may be useful for readers who are not methodologically orientated.

DESCRIPTIVE CANCER EPIDEMIOLOGY

Descriptive cancer epidemiology examines the distribution of individual cancer types by a number of characteristics. These characteristics may refer to individuals themselves (personal characteristics, including age, gender, occupation, education, marital status, etc.), to the place of disease initiation or occurrence (place characteristics, including the country and region of residence, the urban or rural nature of the area, the altitude, the latitude, etc.) or the time pattern of occurrence (time characteristcs, including long-term trends, time elapsing between exposure to a certain agent and the appearance of a particular cancer, etc.). Certain characteristics are inherently multidimensional, e.g. migration refers to individuals, in addition to the country of origin and the time of migration.

In most instances, descriptive epidemiological associations can only generate aetiological hypotheses, because observed associations are frequently confounded by many factors and are also subject to a number of potential biases. For example, the high incidence of stomach cancer in Japan (**Figure 4**) can be attributed to a number of intercorrelated genetic, nutritional and socio-cultural characteristics of the Japanese people. However, on rare occasions, descriptive epidemiological data can be revealing by themselves, particularly when an otherwise rare tumour occurs frequently among persons with a certain defining characteristic, usually their occupation or medical treatment for an earlier condition. For example, the high incidence of cancer of the nasal cavity among workers in nickel refineries strongly suggests that the agents or the processes involved in nickel refining are carcinogenic.

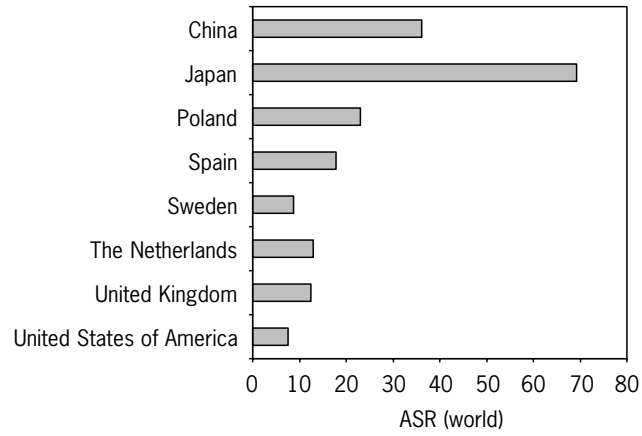


Figure 4 Age-standardized incidence rates of stomach cancer among males. (Based on data from Ferlay *et al.*, GLOBOCAN 2000: Cancer Incidence, Mortality and Prevalence Worldwide, Version 1.0. IARC CancerBase No. 5. IARC Press, Lyon, 2001.)

Personal Characteristics

Among the many personal characteristics, those most frequently explored in descriptive epidemiology are age, gender, religion, marital status, occupation and socio-economic class. These characteristics are rarely themselves aetiologically relevant, but they are commonly correlated with exposure to carcinogenic factors, and for this reason they are frequently considered risk indicators or risk factors. Furthermore, several of these characteristics are readily available because they are routinely recorded for administrative reasons.

Because the incidence of most cancers increases rapidly with age, exceptions to this pattern may have aetiological significance. Acute leukaemia, for example, shows both an early and a late peak, which led investigators to consider intrauterine carcinogenesis. The subsequent discovery that intrauterine exposure to ionizing radiation may increase the risk of childhood leukaemia provided empirical support to this hypothesis. Similarly, the bimodal incidence pattern of Hodgkin disease in many countries (**Figure 5**) led to the prevailing hypothesis that the early peak is due to an infectious agent. Finally, the slowing of the rate of increase of the incidence of breast cancer after menopause has underlined the importance of ovarian oestrogens in the aetiology of this disease.

Many cancers are more common among men than among women. This has been attributed, in many instances, to the higher past exposure of men to tobacco smoking and alcohol drinking. However, some intriguing gender differences remain unexplained. The incidence of hepatitis B-positive hepatocellular carcinoma is three times higher among men, even though the prevalence of chronic hepatitis B virus infection is only two times higher among them.

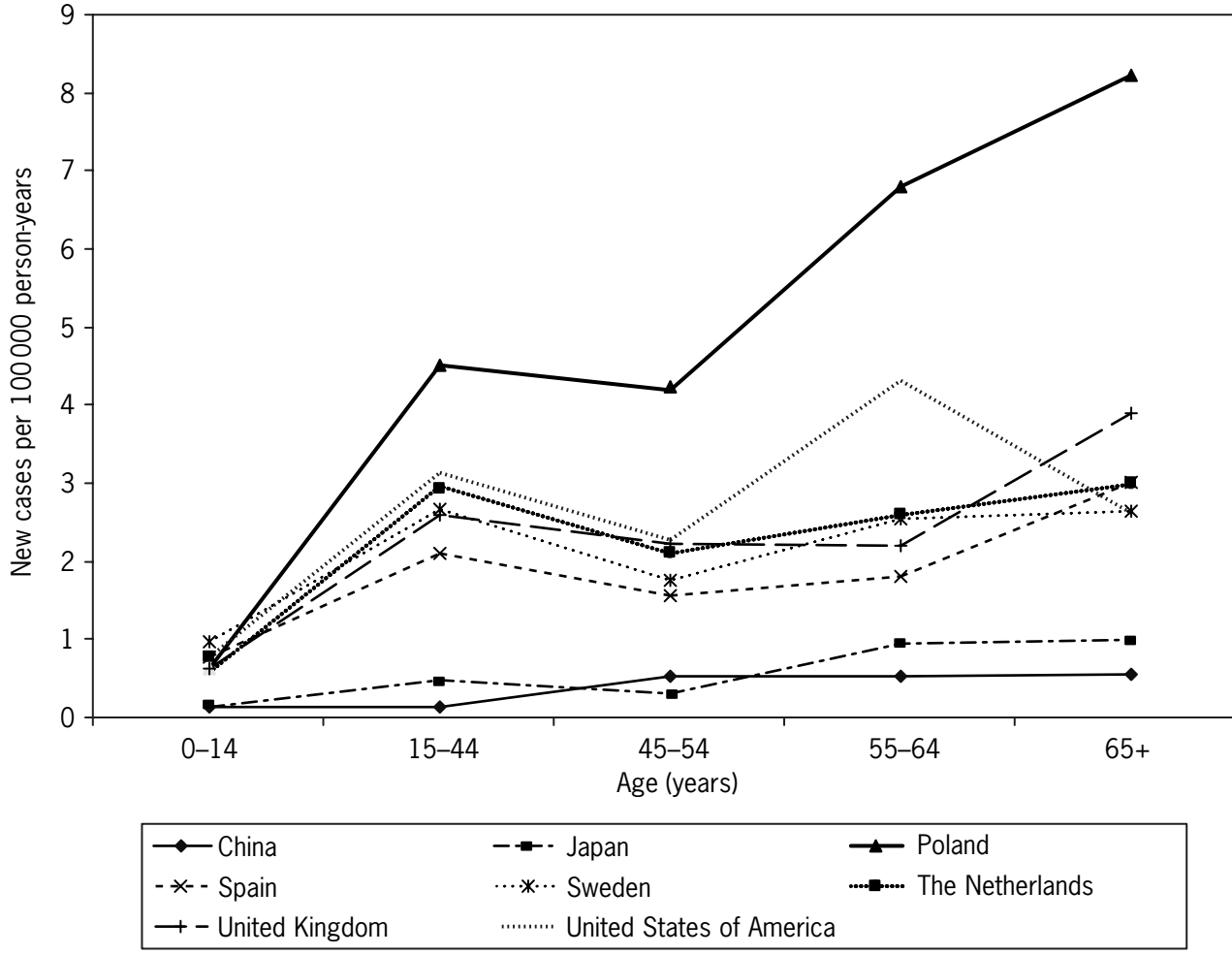


Figure 5 Age-specific incidence rates of Hodgkin disease among males. (Based on data from Ferlay *et al.*, GLOBOCAN 2000: Cancer Incidence, Mortality and Prevalence Worldwide, Version 1.0. IARC Cancer Base No. 5. IARC Press, Lyon, 2001.)

In contrast, the incidence of gallbladder cancer (and also cholelithiasis) is considerably more common among women.

Religion, ethnic origin and race are frequently interrelated. The realization that environmental factors, including lifestyles, are more important than hereditary factors in explaining cancer variation among population groups (Lichtenstein *et al*., 2000) has suggested that those groupings should be viewed in the socio-cultural, rather than genetic, context. Several hypotheses concerning cancer causation have been generated by the unusual lifestyles and cancer occurrence patterns among Mormons and Seventh Day Adventists in the United States. Also, the low incidence of cervical cancer among nuns has led to the hypothesis that sexual activity is an important determinant of this cancer.

Marital status has been an important parameter for the formulation of aetiological hypotheses concerning cancer occurrence in the female reproductive organs. Marital status associations, most likely reflecting underlying causal links of nulliparity with breast cancer, sexual activity with cervical cancer and low parity with ovarian cancer, are stronger in socio-cultural settings in which marital status is strongly correlated with sexual activity and reproduction.

Several forms of cancer, notably cancers of the stomach and uterine cervix, are more frequent in the lower socioeconomic classes, and these observations have led to the development of aetiological hypotheses implicating, respectively, particular dietary patterns (salty foods, inadequate intake of vegetables and fruits), or inadequate hygienic conditions and infrequent utilization of barrier contraceptive methods. Furthermore, occupational mortality statistics have been extremely valuable in pointing out occupations that increase the risk of particular types of cancer.

Place and Time Characteristics

Cancer mortality statistics by site, age and gender are published, on an annual basis, by the WHO for most countries of the world. Furthermore, the IARC publishes, every 5 years, Cancer Incidence in the Five Continents, containing data – by site, age and gender – generated by a worldwide network of cancer registries (Parkin *et al*., 1992). Correlations between the incidence of, or mortality from, particular cancers in various countries on the one hand, and per capita consumption in these countries of various food groups or nutrients on the other, generated hypotheses about the nutritional aetiology of several cancers (Armstrong and Doll, 1975).

National cancer rates have also been used extensively in studies of migrant populations, to distinguish the relative importance of heredity and environment (including lifestyle) in the generation of the large international variability of cancer occurrence at most sites. In migrant populations, e.g. Japanese migrating to the United States, the incidence of large bowel cancer approaches that of the host country within a few decades, whereas for stomach cancer a whole lifetime is needed (Haenszel and Kurihara, 1968) and for breast cancer several generations may be required for incidence assimilation (Ziegler *et al*., 1993). Lastly, large-scale geographic patterns have been used to generate hypotheses linking malaria endemicity with the occurrence of Burkitt lymphoma (Burkitt, 1970) and sunlight exposure with the incidence of skin cancer (Scotto *et al*., 1982).

Long-term temporal trends of lung and stomach cancer in developed countries (**Figure 6**) have led to hypotheses implicating tobacco smoking and food preservation modalities in their aetiology. Specifically, the rapid rise of lung cancer mortality among men during the first half of the twentieth century, and the much slower rise of the corresponding mortality among women, indicated that a novel strong carcinogen became widespread among men but much less so among women. Few agents, besides tobacco, would meet these criteria, and many analytical epidemiological studies have demonstrated that smoking was indeed the responsible factor. Whilst no comparable clear explanation exists for the remarkable decline in stomach cancer mortality, many investigators speculate that the large-scale

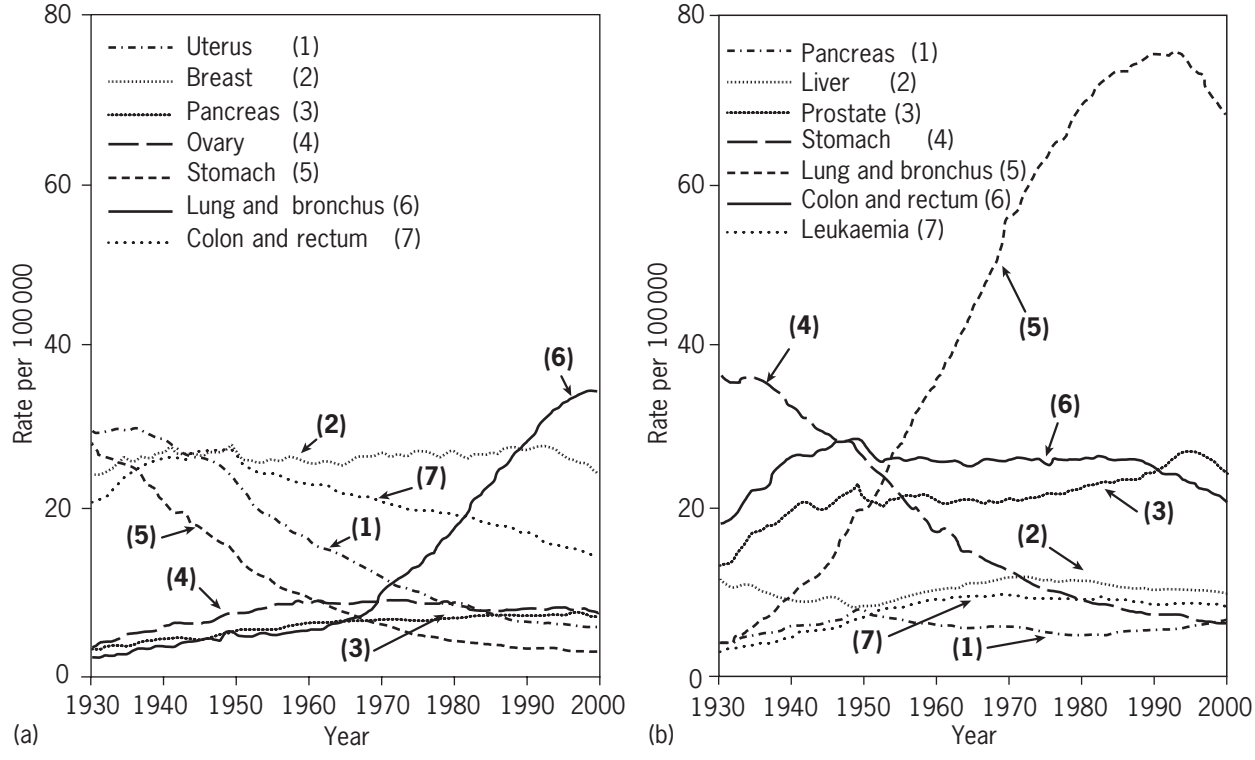


Figure 6 Age-adjusted (to the 1970 US standard population) cancer death rates by site among (a) Females and (b) males in the USA from 1930 to 1996, expressed per 100 000 person-years. (a) Uterus cancer death rates are for uterine cervix and uterine corpus combined. Note: Owing to changes in ICD coding, numerator information has changed over time. Rates for cancers of the uterus, ovary, lung and bronchus, and colon and rectum are affected by these coding changes. (b) Note: Owing to changes in ICD coding, numerator information has changed over time. Rates for cancers of the liver, lung and bronchus, and colon and rectum are affected by these coding changes. (From Greenlee *et al*., 2000, *Cancer Journal for Clinicians*, **50**, 7–33.)

introduction of refrigeration for food preservation may have been responsible.

Clustering in both time and place, a characteristic of diseases caused by infectious agents, has been investigated using an ingenious procedure developed by Knox (1963) for cases of acute childhood leukaemia. Notwithstanding inconclusive results, the hypothesis postulating involvement of infectious agent(s) is still widely entertained.

FORMULATION OF AETIOLOGICAL HYPOTHESES

The reasoning leading to the generation of aetiological hypotheses follows, explicitly or implicitly, a set of rules which focus on difference, agreement, concomitant variation and analogy (MacMahon and Trichopoulos, 1996).

According to the rule of difference, when both a particular set of factors and the incidence of a particular cancer differ between two populations, one or more of these factors will likely contribute to the occurrence of the disease. When the difference in incidence is large, in relative terms, and only one or two factors form the differentiating set, then the hypothesis is more likely to be correct, and vice versa. Thus, the relatively high incidence of Kaposi sarcoma among homosexual men in the United States in the early 1980s indicated that sexual preference was involved in the occurrence of the disease, since few other, if any, characteristics distinguish homosexual from heterosexual men. Also, the lower incidence of cervical cancer among nuns points to sexual activity as an important causal factor, since nuns differ from other women with respect to only a few characteristics and practices. By contrast, the usually higher incidence of lung cancer in large cities cannot be attributed confidently to air pollution, since city dwellers differ from rural inhabitants with respect to a number of exposures, including those related to occupation, diet, smoking, environmental tobacco smoke and radon.

The rule of agreement focuses on common factors identifiable in different settings characterized by a high incidence of particular cancers. Thus, exposure to ionizing radiation was the common factor characterizing patients with ankylosing spondylitis therapeutically irradiated and children diagnostically X-rayed *in utero*, when both groups were found to have increased incidence of leukaemia. Also, multiplicity of sexual partners is the likely common factor characterizing women with squamous carcinoma of the uterine cervix and male homosexuals with anal cancer.

Concomitant variation represents a quantitative expression of the first two rules, particularly the first. The concomitant variation of per capita fat consumption and the incidence of, or mortality from, colorectal cancer, in international correlations (Armstrong and Doll, 1975) was important for the hypothesis linking fat consumption to colorectal cancer – a hypothesis supported by the results of many, although not all, analytical epidemiological studies. Correlations based on countries, regions or other population groups (as distinct from correlations based on individuals) are termed population, group or ecological correlations. From the epidemiological point of view, they may be considered half-way between descriptive and analytical studies, and their aetiological importance is, as a rule, limited.

Inference by analogy represents the fourth and arguably the softest rule. The paradigm of poliomyelitis, in which early infection is harmless, whereas late infection causes the paralytic form of the disease, has been invoked by several authors in attempts to explore the aetiology of childhood leukaemia.

The credibility of an aetiological hypothesis developed on the basis of descriptive epidemiological data increases substantially if the suspected agent causes cancer in experimental animals, preferably in more than one species and in the same organ as in humans. In the absence of a satisfactory experimental model, demonstration of *in vitro* mutagenicity provides supportive evidence, as does the identification of a likely pathogenetic mechanism. Nevertheless, the ultimate criterion of the validity of an aetiological hypothesis – whether derived from descriptive epidemiological findings or formulated on the basis of experimental data, clinical information or theoretical considerations, or any combination of the above – is in most cases epidemiological. Specifically, an aetiological hypothesis is refuted when it generates predictions that appear incompatible with the collective evidence of analytical epidemiological studies. In contrast, it remains provisionally valid when it generates predictions compatible with the accumulated empirical evidence from human studies.

EPIDEMIOLOGICAL STUDIES EVALUATING CANCER CAUSATION

In theory, the best empirical evidence regarding causation should come from randomized trials in humans. Studies in humans do not face the pitfalls inherent in the reasoning by analogy from animal to human data, whereas randomization allows control for distorting influences by both known and unsuspected confounding factors; in addition, double-blind designs minimize the potential for several types of bias. Indeed, therapeutic clinical trials are also aetiological studies exploring the causation of a better clinical outcome by a certain process, the study treatment. However, experimental studies of cancer causation cannot be undertaken in humans for ethical reasons, except when there is evidence that a particular factor may actually protect from cancer (in which instance the absence of the factor can be thought of as the carcinogenic agent). Even under these conditions, experimental studies in humans

are still impractical. Among cancer patients, the outcome under investigation (death, metastasis or recurrence) is a relatively frequent effect, and the corresponding study size can be manageable. Among healthy individuals, the frequency of occurrence of any particular cancer is low and the corresponding study size must be very large, making the follow-up and compliance problems exceedingly difficult. Nevertheless, a few such studies have been undertaken, either by investing large resources (e.g. the IARC study targetting the prevention of hepatocellular carcinoma in the Gambia through active immunization against hepatitis B virus) or by reducing the required sample size by focusing on preneoplastic conditions that identify high risk individuals.

Epidemiological studies specifically designed to address a particular aetiological hypothesis are usually called analytical. The objective of analytical studies is to document causation between exposures and a certain disease. In analytical investigations, measurements and categorical assignments apply to individuals, whereas this is not necessary in descriptive epidemiological studies. Thus, in order to examine in an analytical study whether chronic carriers of hepatitis B virus (HBV) are more likely to develop liver cancer than noncarriers, it is necessary to classify the individuals under study according to their HBV carrier state and the development or not of liver cancer during a specified time period (Hennekens and Buring, 1987; MacMahon and Trichopoulos, 1996; Rothman and Greenland, 1998).

Ecological studies in epidemiology, as opposed to individual-based studies, occupy an intermediate position between descriptive and analytical investigations, in that they share many characteristics with descriptive studies, but may also serve aetiological objectives. In ecological studies, the exposure and the disease under investigation are ascertained not for individuals but for groups or even whole populations (Morgenstern, 1982). Thus the prevalence of HBV in several populations could be correlated with the incidence of liver cancer in these populations, even though no information could be obtained as to whether any particular individual in these populations was or was not an HBV carrier and has or has not developed liver cancer.

When an exposure is fairly common (e.g. smoking, sunlight, poverty, even prevalence of HBV carriers), ecological studies should be able to reveal the effects of these exposures. Thus, following the increase in tobacco consumption, the incidence of lung cancer increased sharply over time; skin melanoma is more common in geographic latitudes with more sunshine exposure; stomach cancer is generally more common in low-income social strata; and the incidence of primary liver cancer is higher in populations with higher prevalence of HBV (Tomatis, 1990). As a corollary, the inability of ecological studies to demonstrate an association between a widespread exposure that has rapidly increased over time (e.g. extremely low-frequency magnetic fields) and the incidence of a disease allegedly caused by these fields (e.g. childhood leukaemia) challenges the causal nature of the positive association reported from some analytical investigations.

In analytical epidemiological studies, there are several ways through which an association, or lack thereof, is assessed, but the most common measure is the relative risk. A value equal to 1 implies that the exposure under study does not affect the incidence of the disease under consideration. In contrast, values <1 and >1 indicate, respectively, an inverse or a positive association (MacMahon and Trichopoulos, 1996; Rothman and Greenland, 1998). When the relative risk is >2 and the associated unbiased, unconfounded, precise and causal, an exposed-case patient is more likely than not to have developed the disease because of the exposure. When the relative risk is >1 but <2 the exposed patient is more likely than not to have developed the disease for reasons other than the exposure. This is because a relative risk of, say, 1.5 has a baseline component equal to 1 that characterizes the unexposed, and a component equal to 0.5 that applies only to the exposed (MacMahon and Trichopoulos, 1996). For instance, if the risk of a nonsmoking 55-year-old man developing lung cancer in the next 10 years is 1%, and that of a same age and gender smoker is 5% (relative risk 5), only 4% in the smoker's risk (i.e. 4/5 of the total 5%) can be attributed to his smoking.

Cohort Studies

In cohort studies (**Figure 2**), individuals are classified according to their exposure and are followed over time for ascertainment of the frequency of disease occurrence in the various exposure-defined categories. In each category the frequency of occurrence is calculated either as risk (proportion of those who developed the disease under study among all individuals in this category) or as incidence rate (number of those who developed the disease divided by the sum of the time each of the individuals in this category has been under observation). Defining characteristics of cohort studies are that they are exposure based and are patently or conceptually longitudinal.

In cohort studies, the groups to be studied are selected on the basis of exposure. In some studies the groups are chosen on the basis of a particular exposure (special exposure cohorts), whereas in others, groups offering special resources for follow-up are initially chosen and the individuals are subsequently allocated according to exposure status (general population cohorts). The first approach may be necessary when rare exposures need to be studied, while the second approach is appropriate when the exposure under consideration is fairly common in general populations (e.g. smoking or major dietary components).

The cohort study groups are observed over a period of time to determine the frequency of occurrence of disease among them. The distinction between retrospective and

prospective cohort studies depends on whether or not the cases of disease have occurred in the cohort at the time the study began. In a retrospective cohort study, all the relevant exposures and health outcomes have already occurred when the study is initiated. In a prospective cohort study, the relevant causes may or may not have occurred at the time the study began, but the cases of disease have not yet occurred and, following selection of the study cohort, the investigator must wait for the disease to appear among cohort members.

From a methodological point of view, there are two types of cohort studies: closed or fixed cohorts and open or dynamic cohorts. Closed cohorts are frequent in occupational epidemiology and study of outbreaks, whereas open cohorts dominate cancer epidemiology and are the basis for most case-control studies. The key distinction between open and closed cohorts is how membership in the cohort is determined. In a closed cohort, membership is determined by a membership-defining event which occurs at a point in time. Studies based on follow-up of closed cohorts may be analysed using either cumulative incidence (risk) measures or by counting person-time and calculating incidence rate measures. Analyses based on cumulative incidence measures are useful only under certain conditions (that is, no loss to follow-up; no competing risks; exposure status unchanged throughout follow-up; study subjects followed for the same period of time). Whether or not these assumptions are met, it is always valid to conduct analyses based on person-time, using incidence rate measures in the setting of a closed cohort.

Open cohorts are composed of individuals who contribute person-time to the cohort, while meeting criteria for a membership-defining state. Once individuals can no longer be characterized by this state, they cease to contribute person-time to the corresponding cohort. In studies based on open cohorts it is not possible to measure cumulative incidence (risk) directly. Instead, analyses are based on person-time using incidence rate measures. Thus, if among 5000 non-smoking men followed for an average period of 10 years ($P_0 = 50\,000$ person-years) $x_0 = 25$ were diagnosed with lung cancer, whereas among 10 000 smoking men followed for an average period of 8 years ($P_1 = 80\,000$ person-years) $x_1 = 600$ were diagnosed with lung cancer, the incidence rate among nonexposed would be 50 per 10^5 person-years, that among exposed 750 per 10^5 person-years and the incidence rate ratio would be $(600/80\,000)/(50/50\,000) = 15$.

Case-control Studies

In case-control studies (**Figure 3**), patients recently diagnosed with the disease under consideration form the case series, and their exposure to the factor under investigation is ascertained through questionnaires, interviews, examination of records, undertaking of laboratory tests in biological samples and other means. The pattern of exposure to the study factor in the population that generated the case series is then evaluated through a properly selected control series. If only two categories of exposure are relevant (exposed and unexposed), the relative risk can be estimated by multiplying exposed cases with unexposed controls and dividing this product by the product of unexposed cases and exposed controls (the so-called odds ratio, that adequately estimates rate ratio and risk ratio). Thus, if among 200 male patients diagnosed with lung cancer (cases), $x_1 = 150$ were smokers and $x_0 = 50$ nonsmokers, whereas among 300 men of similar age as the cases but without lung cancer (controls), $y_1 = 50$ were smokers and $y_0 = 250$ were nonsmokers, the odds ratio x_1y_0/x_0y_1, which is a good approximation to the relative risk, is $(150 \times 250)/(50 \times 50) = 15$.

Two features of case-control studies make them susceptible to bias: the ascertainment of exposure after the occurrence of disease (information bias), and suboptimal processes for control selection (selection bias) (Hennekens and Buring, 1987; MacMahon and Trichopoulos, 1996; Rothman and Greenland, 1998).

Some case-control designs, in particular the population-based ones, are methodologically superior to other case-control variants. A case-control study is called population-based when controls are chosen from the clearly defined population from which all cases have arisen – in other words, had one of the controls developed the disease under study it would have definitely been included among the cases. Cohort studies, however, generally produce more credible results than those emerging from case-control investigations, because, if properly designed and implemented, they are, as a rule, free from information and selection bias.

Frequently, case-control studies are matched, in the sense that controls are chosen so as to match particular cases with respect to gender, age, race or any other factor that is probably related to the disease under study but not intended to be analysed in the particular study. Matching is not strictly necessary, nor does it confirm the validity of results, but it improves statistical efficiency, i.e. statistical power, or the ability to substantiate a true association (Rothman and Greenland, 1998). If matching has been used in the enrollment of cases and controls, the statistical analysis should accommodate the matching process, through either a matched analysis (e.g. conditional modelling) or unmatched analysis with explicit control of the matching factors (proper application of unconditional modelling).

The advantages and disadvantages of cohort and case-control studies in exploring the causes of cancer in humans are described in several textbooks on epidemiology. Briefly, cohort (or prospective or follow-up) studies trace in a more natural way the time sequence of causal phenomena. However, the case-control approach is, in essence, an extension of a practice familiar to physicians, that of case history taking. Confounding is just as common

in cohort as in case-control studies, but information bias concerning exposure to suspected agent(s) is more common in case-control investigations, since the existence of a serious disease, such as cancer, can affect recollection and reporting, as well as several directly ascertainable biochemical and immunological variables. By contrast, case-control studies are usually more powerful, in the statistical sense, since the frequency of exposure to many suspected agents is more common than the frequency of occurrence of most cancers during a not-too-long period of follow-up.

CHANCE, CONFOUNDING AND BIAS

Three issues need to be resolved before an epidemiological association could be considered true and therefore deserve interpretation in causal terms: chance, confounding and bias.

Probabilistic processes always have a built-in uncertainty, but we can reduce the chance-related uncertainty by using progressively larger numbers in a study, and we can assess its possible influence by utilizing statistical procedures that generate what has become known as the *P* value. This value indicates how likely one would have been to observe an association as extreme as, or more extreme than, that found between a particular exposure and a certain disease, if there were in fact no association. The true meaning of the *P* value, however, is poorly understood and the concept itself is widely misused.

P values are traditionally expressed as numerical fractions of unity. For example, a *P* value of 0.1 for a particular positive association indicates that there is a 10% chance that such an association or more extreme (or a symmetrically opposite one, i.e. an inverse association) would appear by chance, even if there were in reality no association at all. In essence, the *P* value is interpretable as such when only one comparison or one test is performed. When multiple comparisons or multiple tests are carried out, the set of the respective *P* values lose their collective interpretability. Procedures for adjusting *P* values according to the number of comparisons undertaken or tests performed have been proposed, but they are not universally accepted (Rothman, 1990) and are rarely utilized. However, other things being equal, the real significance of a certain *P* value is weaker when the number of tests performed is larger. *P* values can be more confidently interpreted as suggestive of real phenomena or genuine associations when there is independent evidence in support of the process that generated the *P* value.

A *P* value of 0.05 or smaller is traditionally treated in medical research as evidence that an observed association may not have arisen by chance. However, small *P* values, including values considerably smaller than 0.05, do not guarantee that an association (or difference) is genuine, let alone causal. Even when the *P* value is very small and was generated from a randomized trial, it could still be dismissed when the relevant result makes no sense (Miettinen, 1985). As a corollary, our daily lives are full of highly unlikely events and coincidences and, at the extremes, thousands of people have become wealthy from lotteries and many more have died in strange accidents, even though the probabilities of the respective events are extremely small. The lesson is simple: highly unlikely events that would have been associated with extremely small *P* values do happen by chance. In addition to small *P* values, science requires judgement relying on sound substantive knowledge in order to discard chance.

The *P* value itself does not convey information about the strength of the respective association: a weak association may be statistically highly significant (very small *P*) when the study is large, and a strong association may be statistically nonsignificant (large *P*) when the study is small (Rothman and Geenland, 1998). In order to integrate information about the strength of an association (as reflected in the relative risk effect measure) and its statistical significance, the concept of confidence interval has been developed. The confidence interval describes the range of values that the true relative risk is likely to take with, for example, 95% confidence (95% confidence interval) on the basis of the data through which the association was evaluated. The confidence interval is intimately linked to the *P* value, so that for a statistically significant ($P < 0.05$) association, the 95% confidence interval of the measure of this association (e.g. the relative risk) does not include ('rejects') 95 times out of 100 the value of this measure which would have indicated no association (the no association value, also called the 'null value,' is 1 for relative risk).

Random variation *per se* in epidemiological studies is not an insurmountable problem. Larger studies and eventually combined analyses, through systematic statistical evaluations of results of several independent investigations, can effectively address genuine chance-related concerns. Such combined analyses have been termed meta-analyses and pooled analyses. There is no real distinction between the two terms, although meta-analysis has been more frequently used when published data are combined, whereas in pooled analysis primary data from different studies may be made available to the investigator who undertakes the task to combine them. Meta-analyses and pooled analyses have been widely and effectively used for randomized clinical trials and intervention studies, in which confounding and bias are of limited concern (Sacks *et al*., 1987). For observational epidemiological studies, however, the role of meta-analysis is not universally accepted (Shapiro, 1994; Feinstein, 1995), because no statistical summarization can effectively address problems generated by residual confounding, unidentified bias and the different ways investigators choose to present their results.

Confounding is the phenomenon generated when another factor that causes the disease under study, or is

otherwise strongly related with it, is also independently related to the exposure under investigation (**Figure 1**). Thus, if one wishes to examine whether hepatitis C virus (HCV) causes liver cancer, hepatitis B virus (HBV) would be a likely confounder because HBV causes liver cancer and carriers of HBV are more likely to also be carriers of HCV (because these two viruses are largely transmitted by the same route). Hence, if the confounding influence of hepatitis B is not accounted for in the design (by limiting the study to hepatitits B-negative subjects) or in analyses of the data, then the strength of the association between hepatitis C and liver cancer would be overestimated.

A more trivial example is the strong association between carrying matches (or a lighter) on the one hand and developing lung cancer on the other. Neither matches nor lighters cause lung cancer and their association to the disease is due entirely to confounding by cigarette smoking. The confounding factor, cigarette smoking, is the true cause of lung cancer and the dependence of cigarette lighting on matches or lighters generates the confounded entirely spurious association of the latter two factors with the disease.

There are several ways to deal with confounding, some simple, some more complicated. They all assume that two conditions are satisfied: (1) that all the confounders have been identified or at least suspected and (2) that the identified or suspected confounders can be adequately measured (Greenland, 1980). When the study is fairly large, it is always possible to evaluate in the analysis all suspected confounders, but the ability to conceptualize and measure all of them accurately frequently goes beyond the control of the investigator. The result is what has been termed residual confounding (Greenland, 1980; MacMahon and Trichopoulos, 1996; Rothman and Greenland, 1998).

Compounding the problems of epidemiological studies is that the data are never of optimal quality. Data collection relies on the memory of and accurate reporting by study participants, laboratory procedures or existing records. None of these sources is perfect, and frequently the errors are unequal between the compared groups and generate biased results. A reasonable concern is that cases, or their relatives, are inclined to link the disease under study to particular exposures for conscious or subconscious reasons. Cases may also try harder than controls or their relatives do to recall or identify past exposures (information bias).

A well thought-out protocol, standardized procedures and built-in quality control measures can reduce bias and allow quantification of its potential impact. However, complete assurance that bias has been eliminated can never be achieved. In addition, the reliance of case-control studies on a control series that simultaneously has to meet criteria of compliance, comparability to the case series, statistical efficiency and general practicality makes them susceptible to selection bias of unpredictable direction and magnitude (Wacholder *et al*., 1992a–c). Hospital controls, neighbourhood controls and controls enrolled through searches of telephone lists have their own problems, and these have been extensively discussed (MacMahon and Trichopoulos, 1996).

When results of an observational epidemiological study designed to address a specific hypothesis are striking, the study is large and there is no evidence of overt confounding or major biases, it is legitimate to attempt aetiological inferences. The rationale is that powerful confounding presupposes strong risk factors that are unlikely to be missed, and that major biases can be traced to gross and easily identifiable protocol violations. Interpretation becomes particularly problematic when a weak association turns out to be statistically significant perhaps in a large but imperfect data set. Although that association could reflect a weak but genuine causal association, it might more likely be the result of residual confounding, subtle unidentifiable bias or chance following a multiple testing process.

Repeated demonstration of an association of similar direction and magnitude in several studies, undertaken by different investigators in different population groups, increases confidence in a genuine causal basis but cannot conclusively establish this. Nor do meta-analyses prove causality; these techniques essentially address the issue of chance and provide no guarantee that a particular bias, unrecognized confounding or selective reporting have not operated in the constituent studies. It is at this stage that biology confronts epidemiology and the ability to reconcile the two perspectives should be the guiding principle in interpreting epidemiological results.

INFERENCE OF CANCER CAUSATION IN EPIDEMIOLOGY

Criteria for inferring causation from epidemiological investigations have been proposed, over the years, by several authors (MacMahon *et al*., 1960; US Department of Health, Education and Welfare, 1964; Hill, 1965; IARC, 1987; Evans, 1993). The philosophy of causation in epidemiology and medicine has also been examined in various essays (Rothman, 1988). In spite of differences in emphasis, a similar set of principles have been invoked by most authors. Sir Austin Bradford Hill (Hill, 1965) advocated the nine widely used criteria listed in **Table 1** to distinguish causal from noncausal associations. These criteria, although sensible and useful, do not separately address the inherently different issues that are posed by the results of a single study, the results of several studies and the likelihood of causation in a certain individual (Cole, 1997). In reality, the perceived likelihood of an association between a particular exposure and a certain disease being causal moves forward or backward in a continuous

Table 1 The Hill criteria for inferring causation (Hill, 1965)

Strength	A strong association is more likely to be causal. The measure of strength of an association is relative risk and not statistical significance
Consistency	An association is more likely to be causal when it is observed repeatedly and in different populations
Specificity	When an exposure is associated with a specific outcome only (e.g. a cancer site or even better a particular histological type of this cancer), then it is more likely to be causal. There are exceptions, however (e.g. smoking causing bronchitis, lung cancer, pancreatic cancer)
Temporality	A cause should not only precede the outcome (disease), but also be compatible with its latency period (in noninfectious diseases) or its incubation period (infectious disease)
Biological gradient	This criterion refers to the presence of a dose–response relationship. If the frequency of the outcome increases when an exposure is more intense or lasts longer, then it is more likely that the association is causal
Plausibility	An association is more likely to be causal when it is biologically plausible
Coherence	This criterion implies that a cause and effect interpretation of an association should not conflict with what is known of the natural history and biology of the disease
Experimental evidence	If experimental evidence exists, then the association is more likely to be causal. Such evidence, however, is seldom available in human populations
Analogy	The existence of an analogy (e.g. if a drug causes birth defects, then another drug could also have the same effect) could enhance the credibility that an association is causal

spectrum as research results accumulate. The evidence for causality is declared as sufficient when a particular threshold has been reached (MacMahon and Trichopoulos, 1996; Cole, 1997).

Criteria for causality can be invoked, explicitly or implicitly, in evaluating the results of a single epidemiological study, although, in this instance, a firm conclusion is all but impossible (single study level, or level I, according to the classification introduced by Cole and used here (Cole, 1997)). Criteria for causality are more frequently used for the assessment of evidence accumulated from several epidemiological studies and other biomedical investigations. At this stage, the intellectual process is inductive, moving away from the specifics towards generalization (several studies level, or level II).

Regulatory agencies and policy makers may recommend standards, set limits or authorize action even when the scientific evidence is weak, particularly at levels surrounding the proposed standards or limits. These procedures serve public health objectives by introducing a wide safety margin, but should never be confused with the establishment of causation based on scientific considerations alone. Finally, when causation has been established at level II, then, and only then, can the cause of the disease in a particular individual be considered (specific person level, or level III). At this level, the intellectual process is deductive and deterministic, moving from the general concept of disease causation to the examination of what has caused disease in a particular individual.

Whereas causality can be conclusively established between a particular exposure as an entity and a particular disease as an entity, it is not possible to conclusively establish such a link between an individual exposure and a particular disease of a certain individual, e.g. smoking and development of lung cancer. It is possible, however, to infer deductively that the specific individual's illness more likely than not was caused by the specified exposure. For this conclusion all the following criteria must be met (Cole, 1997):

- The exposure under consideration, as an entity, must be an established cause of the disease under consideration, as an entity (level II).
- The relevant exposure of the particular individual must have properties comparable (in terms of intensity, duration, associated latency, etc.) to those of the exposure that, as an entity, has been shown to cause the disease under consideration, again as an entity.
- The disease of the specified person must be identical with or within the symptomatological spectrum of the disease which, as an entity, has been aetiologically linked to the exposure.
- The patient must not have been exposed to another established or likely cause of this disease. If the patient has been exposed to both the factor under consideration (e.g. smoking) and to another causal factor (e.g. asbestos), individual attribution becomes a function of several relative risks, all versus the completely unexposed: relative risk of those who only had the exposure under consideration, relative risk of those who had only been exposed to the other causal factor(s) and relative risk of those who have had a combination of these exposures.
- The scientifically defensible, properly adjusted, relative risk of the disease among individuals who have had the exposure under consideration at the given magnitude should be >2. Only then (as previously discussed) is the diseased individual more likely than not to have developed the disease because of the specified exposure (US Department of Health, Education and Welfare, 1964; MacMahon and Trichopoulos, 1996). If, for

example, the relative risk is 2.5, the component 1 of the background risk is less than the component 1.5 of the excess risk and the individual is more likely than not to have developed the disease because of the exposure.

CONCLUSION

Modern epidemiology has become a rich and powerful toolbox for the study of biological phenomena in humans. Because manipulation of exposures, many of which may be harmful, is usually neither feasible nor ethical, epidemiologists have to base their inferences on experiments that human subjects undertake intentionally, naturally or even unconsciously. The study of risk of lung cancer among smokers compared with nonsmokers is one classical example of a 'natural' experiment. Because human life is characterized by a myriad of complex, often interrelated behaviours and exposures – ranging from genetic traits and features of the intrauterine environment to growth rate, physical activity, sexual practices, use of tobacco, alcohol and pharmaceutical compounds, dietary intake, exposure to infections, environmental pollutants and occupational hazards, etc. – epidemiologic investigation is difficult and challenging. Given this complexity, it is not surprising that from time to time epidemiological studies generate results that appear confusing, biologically absurd or even contradictory. However, it is reassuring that a wealth of new knowledge has been generated by epidemiological studies over the last few decades, and that this knowledge now lays the scientific ground for primary prevention of many major cancers and other chronic diseases among humans globally.

Given the multitude of problems with which epidemiological studies have to deal, the variability of the conditions under which they are undertaken and the abundance of potential sources of confounding and bias, it is remarkable that their results are as consistent as they are. However, it should be stressed that, in epidemiology, more than in any other field of biomedical research, it is the collective evidence that matters, rather than the results of a particular study, however large and well done the latter may be. Only the quality-adjusted collective evidence can address the issues of hidden bias and confounding that may find their ways into any particular study, and only the collective evidence can provide an approximation to generalizability.

REFERENCES

Armstrong, B. and Doll, R. (1975). Environmental factors and cancer incidence and mortality in different countries, with special reference to dietary practices. *International Journal of Cancer*, **15**, 617–631.

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research. Vol. I. The Analysis of Case-control Studies.* IARC Scientific Publication No. **32**, 5–338 (International Agency for Research on Cancer, Lyon).

Breslow, N. E. and Day, N. E. (1987). *Statistical Methods in Cancer Research. Vol. II. The Design and Analysis of Cohort Studies.* IARC Scientific Publication No. **82**, 1–406 (International Agency for Research on Cancer, Lyon).

Buck, C. (1975). Popper's philosophy for epidemiologists. *International Journal of Epidemiology*, **4**, 159–168.

Burkitt, D. P. (1970). Geographical distribution. In: Burkitt, E. P. and Wright, D. H. (eds), *Burkitt's Lymphoma*. 186–197 (Livingstone, Edinburgh).

Cole, P. (1997). Causality in epidemiology, health policy and law. *Environmental Law Reporter*, **27**, 10279–10285.

Evans, A. (1993). *Causation and Disease: A Chronological Journey* (Plenum, New York).

Feinstein, A. R. (1995). Meta-analysis: statistical alchemy for the 21st century. *Journal of Clinical Epidemiology*, **48**, 71–79.

Greenland, S. (1980). The effect of misclassification in the presence of covariates. *American Journal of Epidemiology*, **112**, 564–569.

Haenszel, W. and Kurihara, M. (1968). Studies of Japanese migrants. I. Mortality from cancer and other diseases among Japanese in the United States. *Journal of the National Cancer Institute*, **40**, 43–68.

Hennekens, C. H. and Buring, J. E. (1987). *Epidemiology in Medicine* (Little, Brown, Boston).

Hill, A. B. (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine*, **58**, 295–300.

IARC (1980a). *IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans. Vol. 23. Some Metals and Metallic Compounds*. 39–142 (International Agency for Research on Cancer, Lyon).

IARC (1980b) *IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans.* Vol. 24, Some Pharmaceutical Drugs. 135–62. (International Agency for Research on Cancer, Lyon).

IARC (1987). *IARC Monographs on Carcinogenic Risks to Humans. Overall Evaluations of Carcinogenicity: an Updating of IARC Monographs Volumes 1 to 42*, Supplement 7. 440 (International Agency for Research on Cancer, Lyon).

Knox, G. (1963). Detection of low intensity epidemicity. Application to cleft lip and palate. *British Journal of Preventive and Social Medicine*, **17**, 121–127.

Lichtenstein, P., *et al*. (2000). Environmental and heritable factors in the causation of cancer – analyses of cohorts of twins from Sweden, Denmark, and Finland. *New England Journal of Medicine*, **343**, 78–85.

MacMahon, B. and Trichopoulos, D. (1996). *Epidemiology: Principles and Methods* (Little, Brown, Boston).

MacMahon, B. *et al*. (1960). *Epidemiologic Methods* (Little, Brown, Boston).

Miettinen, O. S. (1985). *Theoretical Epidemiology: Principles of Occurrence Research in Medicine.* (Wiley, New York).

Morgenstern, H. (1982). Uses of ecologic analysis in epidemiologic research. *American Journal of Public Health*, **72**, 1336–1344.

Parkin, D. M., *et al*. (eds) (1992). *Cancer Incidence in Five Continents*, Vol. VI. IARC Scientific Publication No. 120 (International Agency for Research on Cancer, Lyon).

Rothman, K. J. (ed.) (1988). *Causal Inference* (Epidemiology Resources, Chestnut Hill, MA).

Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology*, **1**, 43–56.

Rothman, K. J. and Greenland, S. (1998). *Modern Epidemiology*, 2nd edn (Lippincott-Raven, Philadelphia, PA).

Sacks, H. S., *et al*. (1987). Meta-analysis of randomized controlled trials. *New England Journal of Medicine*, **316**, 450–455.

Scotto, J., *et al*. (1982). Solar radiation. In: Schottenfeld, D. and Fraumeni, J. F. (eds), *Cancer Epidemiology and Prevention*. 254–276 (W. B. Saunders, Philadelphia, PA).

Shapiro, S. (1994). Meta-analysis/Shmeta-analysis. *American Journal of Epidemiology*, **140**, 771–778.

Stuver, S. O. (1998). Towards global control of liver cancer? *Seminars in Cancer Biology*, **8**, 299–306.

Tomatis, L. (ed.) (1990). *Cancer: Causes, Occurrence and Control*. IARC Scientific Publication No. 100 (International Agency for Research on Cancer, Lyon).

Trichopoulos, D. (1994). Risk of lung cancer from passive smoking. *Principles and Practice of Oncology: PPO Updates*, **8**, 1–8.

US Department of Health, Education and Welfare (1964). *Smoking and Health. Report of the Advisory Committee to the Surgeon General of the Public Health Service*. Publication 1103 (US Government Printing Office, Washington, DC).

Wacholder, S., *et al*. (1992a). Selection of controls in case-control studies: I. Principles. *American Journal of Epidemiology*, **135**, 1019–1028.

Wacholder, S., *et al*. (1992b). Selection of controls in case-control studies: II. Types of controls. *American Journal of Epidemiology*, **135**, 1029–1041.

Wacholder, S., *et al*. (1992c). Selection of controls in case-control studies: III. Design options. *American Journal of Epidemiology*, **135**, 1042–1050.

Ziegler, R. G., *et al*. (1993). Migration patterns and breast cancer risk in Asian-American women. *Journal of the National Cancer Institute*, **85**, 1819–1827.

Adami, H.-O. and Trichopoulos, D. (1999). Epidemiology, medicine and public health. *International Journal of Epidemiology*, **28**, s1005–s1008.

MacMahon, B. and Trichopoulos, D. (1996). *Epidemiology: Principles and Methods*, 2nd edn (Little, Brown, Boston).

Saracci, R. and Trichopoulos, D. (1996). Aetiological leads. In: Peckham, M., *et al*. (eds), *Oxford Textbook of Oncology*. 167–174 (Oxford University Press, Oxford).

Trichopoulos, D. and Adami, H.-O. (2001). Concepts in cancer epidemiology and etiology. In: Adami, H.-O., *et al*. (eds), *Cancer Epidemiology* (Oxford University Press, Oxford).

Trichopoulos, D., *et al*. (1997). Epidemiology of cancer. In: DeVita, V. T., Jr, *et al*. (eds), *Cancer: Principles and Practice of Oncology*, 5th edn. 231–257 (Lippincott-Raven, Philadelphia, PA).

FURTHER READING

Authors' Work

Adami, H.-O. and Trichopoulos, D. (eds) (1998). Progress and enigmas in cancer epidemiology. *Seminars in Cancer Biology*, **8**, 215–314.

Other Publications

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research. Vol. I. The Analysis of Case-control Studies*. IARC Scientific Publication No. **32**, 5–338 (International Agency for Research on Cancer, Lyon).

Breslow, N. E. and Day, N. E. (1987). *Statistical Methods in Cancer Research. Vol. II. The Design and Analysis of Cohort Studies*. IARC Scientific Publication No. **82**, 1–406 (International Agency for Research on Cancer, Lyon).

Hennekens, C. H. and Buring, J. E. (1987). *Epidemiology in Medicine* (Little, Brown, Boston).

Rothman, K. J. and Greenland, S. (1998). *Modern Epidemiology*, 2nd edn (Lippincott-Raven, Philadelphia, PA).

Schottenfeld, D. and Fraumeni, J. F. (eds) (1996). *Cancer Epidemiology and Prevention*, 2nd edn (Oxford University Press, New York).

dos Santos Silva, I. (1999). *Cancer Epidemiology: Principles and Methods* (International Agency for Research on Cancer, Lyon).

Szklo, M. and Javier Nieto, F. (1999). *Epidemiology: Beyond the Basics* (Aspen, Gaithersburg, MD).

Tomatis, L. (ed.) (1990). *Cancer: Causes, Occurrence and Control*. IARC Scientific Publication No. 100 (International Agency for Research on Cancer, Lyon).

Walker, A. M. (1991). *Observation and Inference: An Introduction to the Methods of Epidemiology* (Epidemiology Resources, Boston).

Chapter 20

Mechanisms of Chemical Carcinogenesis

Nigel J. Gooderham and Paul L. Carmichael
Imperial College School of Medicine, London, UK

CONTENTS

INTRODUCTION

The concept that chemicals can cause cancer (chemical carcinogenesis) has been accepted for some time. As long ago as 1775, the English physician Sir Percival Pott noted the incidence of scrotal cancer in chimney sweeps and perceptively suggested that the disease was related to their occupation. He further suggested that it was soot that was the cause of their disease. Since this observation, the number of chemicals strongly associated with the development of cancer has substantially increased. Other notable historical examples include the development of skin tumours associated with oils and bladder cancers due to exposure to dyes and pigments. A more detailed account of the history of chemicals and cancer can be found in Lawley (1994).

The list of chemicals that can induce cancer is extensive; they can show high specificity for the organ in which the tumour is induced and in the molecular mechanisms through which they operate. Early observations of chemical carcinogenicity were often made using crude mixtures such as coal tars and subsequent studies have shown that these mixtures comprise a complex range of chemical entities. Chemical carcinogens include organics and inorganics, fibres and particulates, and biologically active materials such as hormones. The organic chemicals probably comprise the largest group.

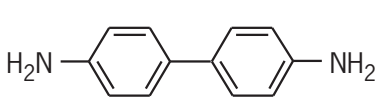
CHEMICALS CAN CAUSE CANCER

Many chemicals are direct-acting carcinogens yet many more require metabolic activation in order to exert their carcinogenicity. In most cases it is chemical reactivity that dictates this carcinogenicity and therefore factors and environments, which influence this reactivity, are important in the nature and site of the carcinogenic action.

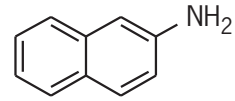
Organic Chemical Carcinogens

The diversity of organic chemicals that have been shown to cause cancer is considerable. Notable examples are shown in **Figure 1**. They range from low molecular mass simple halogenated hydrocarbons to very complex multiheterocyclic molecules. All possess, or have the potential to possess, key functionalities that are intimately involved in their carcinogenic action. Some of the first chemicals found to be carcinogens were the polycyclic aromatic hydrocarbons (PAHs). These chemicals are major components of coal tars and soots and application to the skin of rodents showed them to be powerful carcinogens. Many of these PAHs have been identified (**Figure 2**) and include benzo[*a*]pyrene, dibenz[*a*,*c*]anthracene, 3-methylcholanthrene, 7,12-dimethylbenz[*a*]anthracene and chrysene. All are multi-ring aromatic chemicals composed of carbon and hydrogen. Substituted PAHs in which a nitro or amino or azo function is incorporated into the structure are also carcinogenic; examples include 4-dimethylaminoazobenzene, 2-naphthylamine, benzidine and 1-nitropyrene (**Figures 1** and **2**). However, aromaticity is not an obligatory feature for chemicals that cause cancer and a group of lower molecular mass chemicals, the nitrosoamines and nitrosoamides, are equally potent carcinogens.

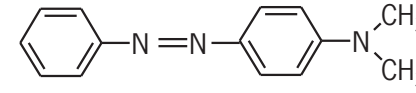
Other examples of cancer-causing organic chemicals include aflatoxins, which are generated by the mould *Aspergillus flavus*. Indeed, aflatoxin B_1 is one of the most potent hepatocarcinogens known, capable of inducing tumours in rodents, fish, birds and primates, including



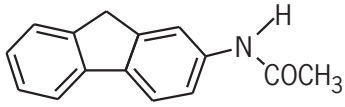
Benzidine



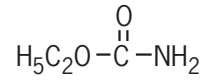
2-Naphthylamine



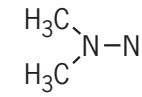
4-Dimethylaminoazobenzene



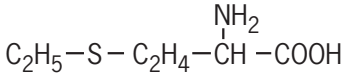
2-Acetylaminofluorene



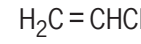
Ethyl carbamate



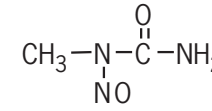
Dimethylnitrosamine



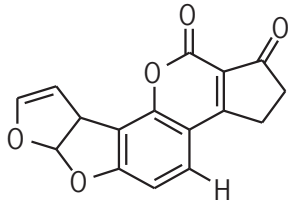
Ethionine



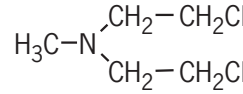
Vinyl chloride



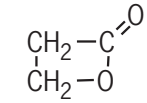
Methylnitrosourea



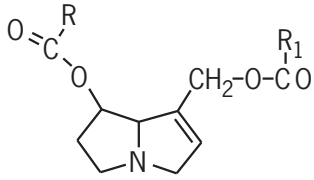
Aflatoxin B_1



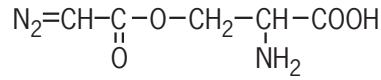
Nitrogen mustard



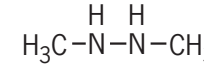
β-Propiolactone



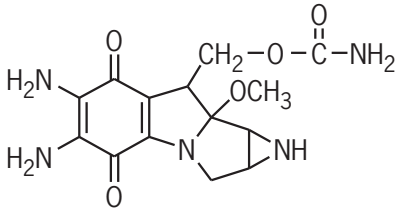
Pyrrolizidine alkaloids



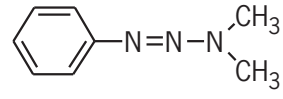
Azaserine



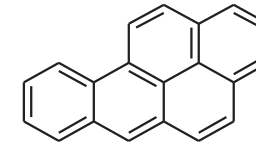
Dimethylhydrazine



Mitomycin C



3,3-Dimethyl-1-phenyltriazine



Benzo[*a*]pyrene

Figure 1 Examples of chemicals that can cause cancer.

humans. The mould is a contaminant of many food crops (e.g. grains and peanuts), especially when stored under warm, humid conditions, and is a significant hazard in many parts of the world, especially Africa and Asia (Wogan *et al*., 1992). However, the aflatoxins are not the only naturally produced organic carcinogens that present a hazard to humans; other notorious examples include hydrazine derivatives in certain mushrooms, cycasin from

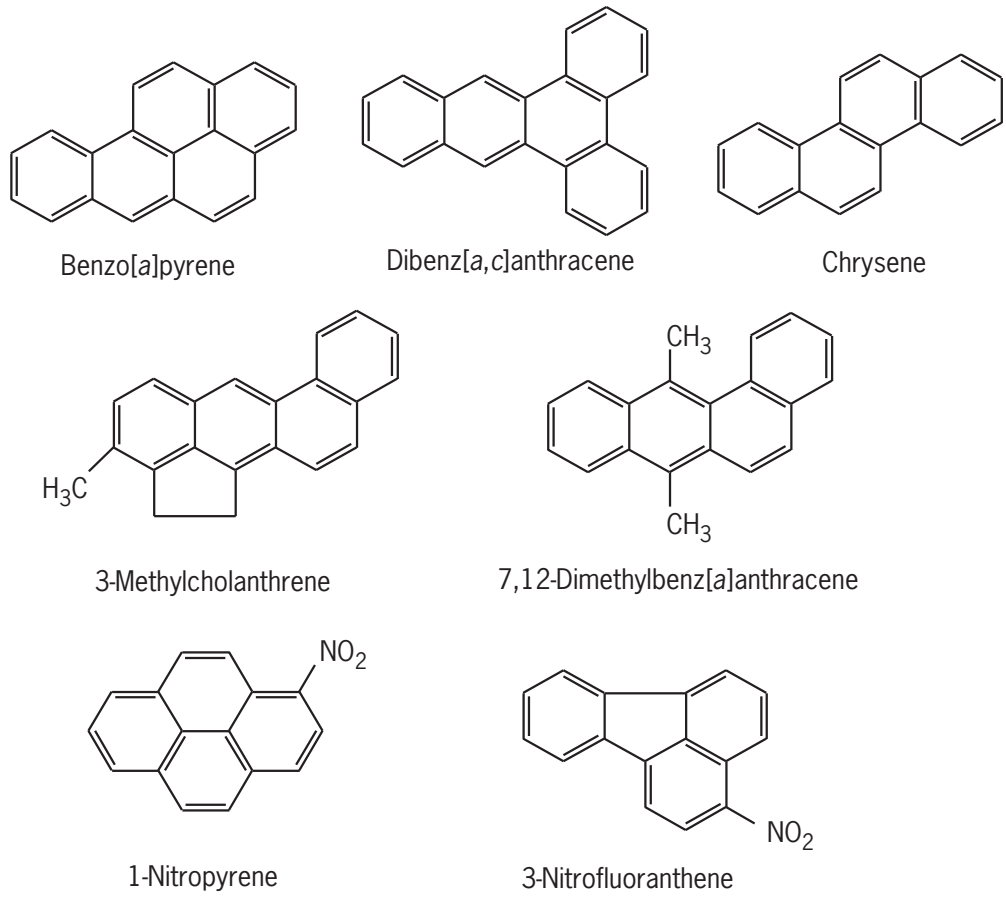


Figure 2 Examples of carcinogenic polycyclic aromatic hydrocarbons.

cycad nuts, pyrrolizidine alkaloids and ptaquiloside in various plants (see the chapter *Dietary Genotoxins and Cancer*).

Although most of the aforementioned chemicals require metabolic activation in order to exert their carcinogenicity, there are many examples of chemical carcinogens whose structure incorporates inherently reactive functional groups. Such chemicals are direct-acting carcinogens (see **Figure 3**). In each case, the reactivity of the key functional group enables the chemical to damage directly DNA, proteins and other cellular macromolecules.

Inorganic Chemical Carcinogens

Although the carcinogenic potential of many elements has not been adequately evaluated, several are known to be carcinogenic in laboratory animals and good epidemiological data support their potential as human carcinogens. Notable examples include compounds of cadmium, chromium and nickel that have all been shown to cause cancer in animals and are established human carcinogens with documented industrial exposure. In contrast, in the case of arsenic, there is significant evidence for it being a human carcinogen despite negative animal data. Other elements that are clearly carcinogenic in animals and therefore are suspect human carcinogens include beryllium, cobalt, iron, lead, titanium and zinc (Sky-Peck, 1986).

Inert Chemical Carcinogens

Some chemicals are carcinogenic even in the absence of chemical reactivity; their physical presence in tissues can be enough to form a tumour. Implantation of certain plastics and fibres into animals can induce sarcomas, usually at the site of implantation. It is the physical size and nature of the material that appear to be important for the development of cancer and not the chemical composition (Brand *et al*., 1975). Although rodents are susceptible to these agents, other species are resistant, e.g. the guinea pig, and interestingly, implantation of plastic and other inert materials into humans, in the form of prostheses, rarely generates a sarcoma. Of more importance to humans are the fibres that are known to cause cancer; asbestos is notorious in this respect and human exposure results in mesothelioma and bronchiogenic carcinoma. The development of asbestos-induced neoplastic disease is related to the crystal structure and dimensions of the fibres rather than the chemical composition of the material (Lippmann, 1993). Fibres that are about 5 μm in length with a diameter of <0.5 μm induce mesothelioma, whereas

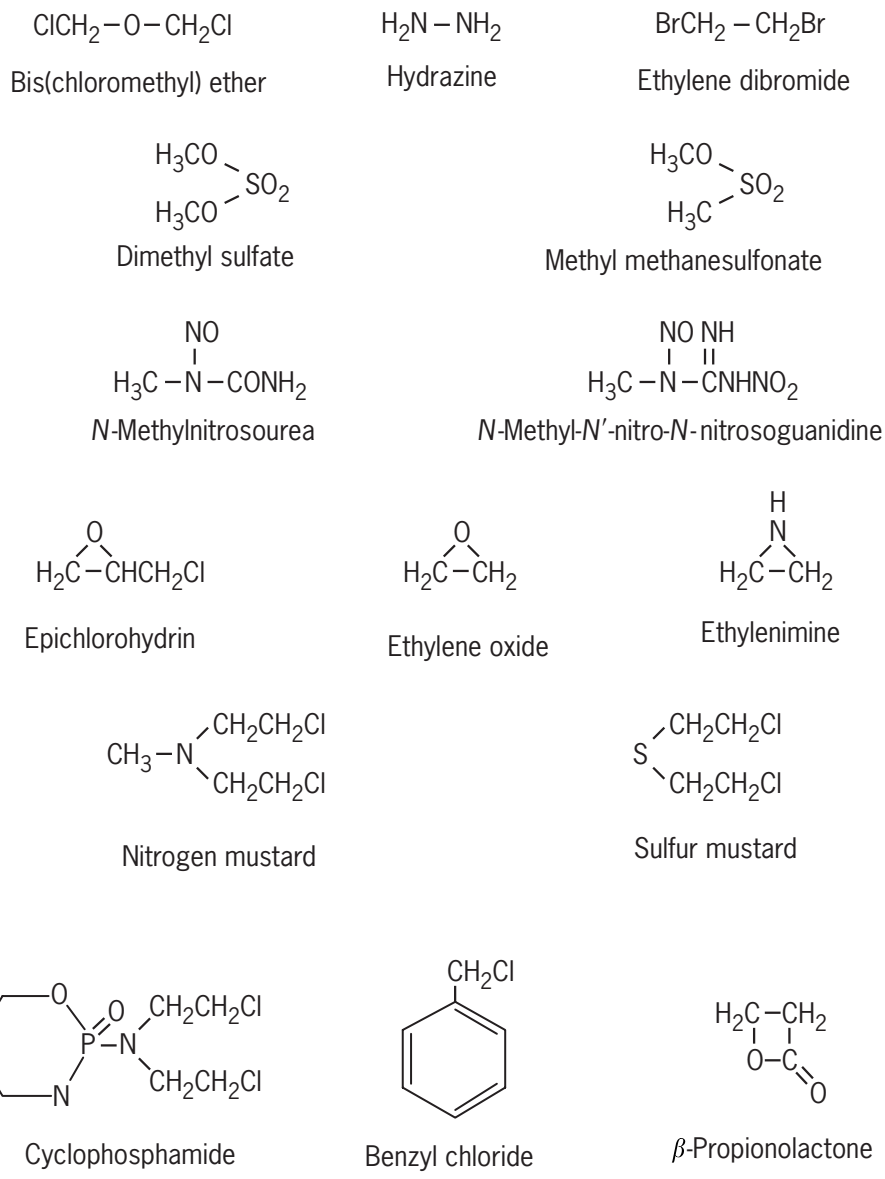


Figure 3 Examples of direct-acting chemical carcinogens.

fibres that are $\geq 10\,\mu m$ in length induce carcinoma. Since not all asbestos fibres conform to these dimensions, different types and sources of asbestos vary in their carcinogenic potency.

Hormonal Chemical Carcinogens

The first link between hormones and carcinogenesis can probably be ascribed to the Italian physician Ramazzini, who in the eighteenth century observed an increased incidence of breast cancer among nuns. It is now well established that never having children is associated with an increased breast cancer risk and, over a century after Ramazzini's observations, George Thomas Beatson (1896) pointed out that a relationship existed between breast cancer and the ovary, the major site for the production of oestrogen. Indeed, the case for endogenous oestrogens in cancer promotion is well established but increasingly concern has arisen regarding external phyto- and xenooestrogens in our environment and the role they may play in carcinogenesis. Furthermore, the administration of chemicals that alter the synthesis or secretion of hormones can lead to neoplastic disease. For example, chemical modulation of thyroid and pituitary growth hormone can lead to neoplasms and changes in human growth hormone, transforming or insulin-like growth factors and testosterone (**Figure 4**) are all associated with carcinogenicity under circumstances where their normal function is interfered with. In each case there is an interruption to the normal hormonal relationship experienced by the target organ. An increasing number of synthetic compounds that possess steroid hormone and antisteroid hormone activity

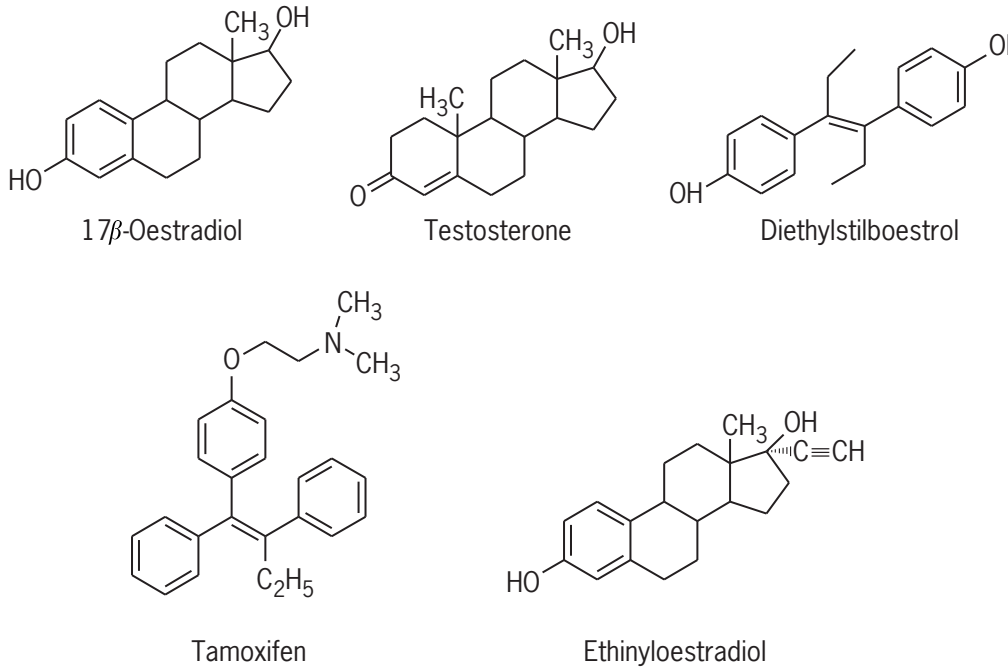


Figure 4 Examples of hormonal carcinogens.

have been found to be chemical carcinogens (**Figure 4**). One of particular interest has been the highly effective anti-breast cancer agent tamoxifen, which also possesses carcinogenic activity in the uterus in a small proportion of women using the drug (Carmichael, 1998).

METABOLISM OF CHEMICAL CARCINOGENS

Many chemicals require metabolic activation in order to exert their carcinogenic potential. The pioneering studies of Elizabeth and James Miller showed that metabolic activation of azo dyes led to their covalent binding to cellular macromolecules. They went on to show with the model carcinogen 2-acetylaminofluorene that hydroxylation of the amide nitrogen generated a metabolite that was more carcinogenic than the parent molecule. Subsequently it was found that these primary products of metabolism, although activated, could be further metabolized to even more reactive derivatives (Miller, J. A., 1970 and Miller, E. C., 1978). It was the Millers who understood that these products were potent electrophiles and comprehensively described their rapid covalent interactions with cellular macromolecules (Miller and Miller, 1981). This led to their proposal that chemical carcinogens that require such metabolic conversion in order to exert their carcinogenic effect should be called procarcinogens and that their highly reactive electrophilic metabolites were ultimate carcinogens. This further led to the concept of proximal carcinogens (e.g. *N*-hydroxy-2-acetylaminofluorene), which were intermediates between the parental procarcinogen and the ultimate carcinogenic metabolite. Although this concept has now been with us for more than two decades, the structure of many ultimate carcinogens is still not thoroughly understood and in many cases may comprise a number of different metabolites of the same parent compound.

There are a number of these metabolic pathways that together are part of a more extensive defence system, the overall role of which is ideally to process and detoxify noxious chemicals. Enzyme-catalysed and diverse in nature, these reactions have been defined and split into what are called phase I and phase II metabolism (Williams, 1971). Phase I can be separated into oxidation, reduction and hydrolytic reactions and phase II comprises a series of conjugation reactions in which a polar endogenous group is added to the xenobiotic chemical. The overall effect of this biochemistry is to convert xenobiotics, which are often lipophilic molecules, into more polar water-soluble and therefore more readily excreted products. Generally, phase I reactions unmask or introduce a functional group into the molecule and phase II metabolism conjugates the derivative with a polar water-soluble endogenous molecule, that is often acidic in nature. However, it is these same pathways of detoxification metabolism that can inadvertently bioactivate chemical carcinogens. For a more detailed description of phase I and phase II metabolism reactions and associated enzymology, the reader is directed to Jakoby *et al*. (1982), Parkinson (1996) and Gonzalez (1989).

The majority of procarcinogens are activated by mechanisms involving two-electron-mediated metabolic reactions primarily catalysed by the mixed function oxidase enzyme systems, often involving cytochrome P-450 enzymes. However, a number of one-electron reactions are known to be capable of activating xenobiotics in co-oxidation processes. For example, PAHs have been found to be bioactivated during the synthesis of prostaglandins from arachidonic acid. A key enzyme in this process is prostaglandin H synthetase, which catalyses the oxygenation of arachidonic acid to the endoperoxide prostaglandin G_2 and also has peroxidase activity, whereby it reduces the hydroperoxide prostaglandin G_2 to the alcohol prostaglandin H_2. In these reactions the peroxidase activity of the enzyme yields a free radical product that can donate electrons to xenobiotics (Eling *et al*., 1990). Other enzyme systems which can participate in these one-electron activation reactions include constitutive peroxidases such as myeloperoxidase and lactoperoxidase, both of which are capable of activating xenobiotics. Although these co-oxidation pathways are not as quantitatively important as the mixed-function oxidase activities, their presence in tissues that lack mixed function oxidase activity can be an important contributor to xenobiotic activation.

CHEMICALS CAN DAMAGE DNA

As discussed above, the metabolism of chemical agents to reactive species is a common feature of carcinogenicity. Once bioactivated (often via proximal carcinogens or intermediate chemicals formed on the way to the creation of the ultimate carcinogen), for most classical chemical carcinogens, some form of DNA damage is the norm. Because of this DNA-damaging activity, such agents are known as genetic or genotoxic carcinogens. However, this is not the case for all chemical carcinogens and some agents bring about carcinogenicity through no direct alteration or damage to the DNA. These agents can be classified as a separate group known as epigenetic carcinogens and their effects are commonly mediated through other changes involving growth factor expression or complex effects on signal transduction mechanisms (see the chapter *Non-genotoxic Causes of Cancer*). Some of the common ways in which chemical agents may be genotoxic are summarized in **Figure 5**.

The nature of carcinogen damage to DNA is dependent upon the chemical agent and its metabolism, but can often include simple changes to the DNA such as the hydroxylation of the bases. Products of such damage

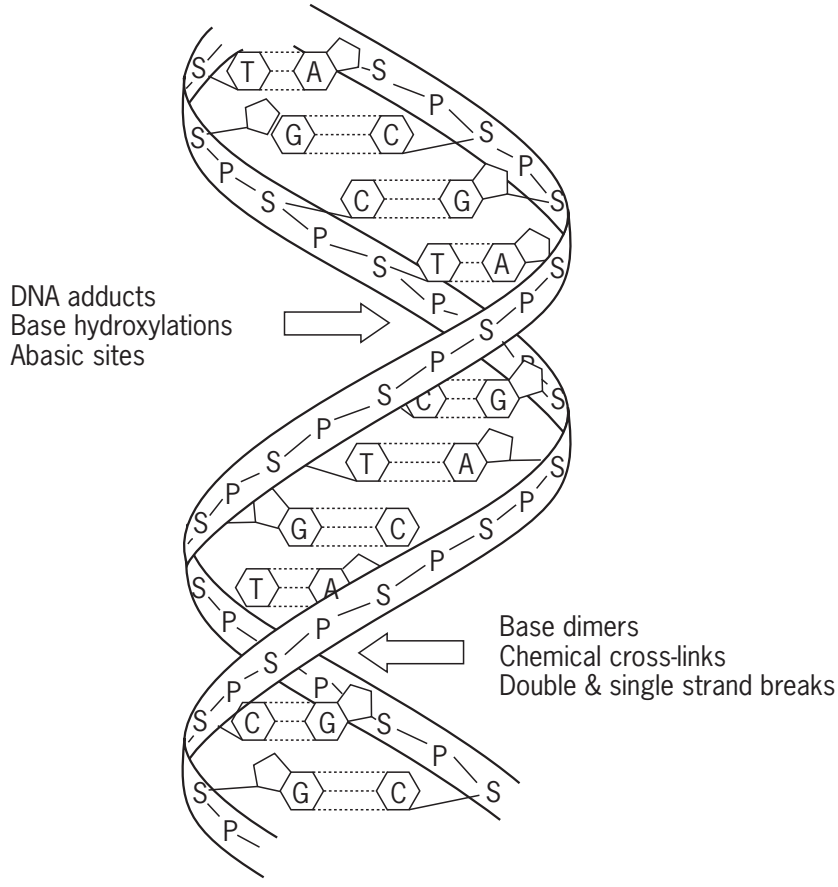


Figure 5 Carcinogen damage to DNA.

include 8-hydroxydeoxyguanosine (formed by the hydroxylation of guanine bases), which is considered to be a mutagenic lesion in DNA, and believed to be formed by attack of the DNA by highly reactive free radicals. Free radicals are a common product of a number of chemical carcinogens, which either carry an unpaired (and so-called radical) electron on the molecule, or are formed from oxygen (oxygen free radicals) via metabolism. Hydroxylations and other more extensive radical damage may result in the loss of bases and the creation of abasic sites, either apurinic or apyrimidinic. Free radical attack of DNA is also responsible for the formation of genotoxic strand breaks in the DNA. These can be formed at either one strand or may span both sides of the sugar–phosphate backbone resulting in a double-strand break.

One of the most important ways, however, in which genotoxic carcinogens may bring about DNA damage is through a chemical binding directly to the DNA. For some small molecules this binding may be considered an alkylation, although some longer chain molecules are capable of forming cross-links across the bases. For many chemical carcinogens the product of their bioactivation (via the creation of an electron- and DNA-seeking electrophile) is the formation of what are known as 'DNA adducts.' These covalent modifications of the DNA, normally at the bases, are generally considered to be one of the primary initiating events (see below) in chemical carcinogenesis and they are formed by many of the agents mentioned above, including PAHs, aflatoxins and aromatic amines. With chronic exposure to such genotoxic agents, DNA adducts may reach steady-state levels in target tissues. During cell replication the DNA adducts can result directly in mutations in genes that control cell growth and thus lead to neoplasia. The levels of these DNA adducts appear to be dose related and are generally predictive of tumour incidence across species. Thus, the accurate estimation and identification of human carcinogen-induced DNA adducts is one of the most important predictive tools or biomarker for the assessment of human cancer risk (discussed in the chapter The Formation of DNA Adducts).

The position and nature of carcinogen-induced DNA adducts dictate the type of mutation that can result (Dogliotti, 1996). For example, small alkylating agents will adduct to guanine at the N7 position owing to the highly nucleophilic nature of the site. In contrast, other more bulky aromatic amine agents will attack the purine ring preferentially, such as the C8 position of guanine and others such as diol epoxides of PAHs bind to the N2 and N6 positions. The adduct products of these reactions are converted to mutations when the cell attempts to repair the damage or replicate itself. These include point mutations, frameshifts involving loss or gain of either a single or multiple bases, chromosomal aberrations, anueploidy or polyploidy. Once introduced, the mutation becomes fixed within the DNA sequence and therefore heritable. The type of mutation that a chemical induces is dependent upon the way in which the chemical interacts with the DNA. The site of chemical attack is important but so is the influence of the bulk of the chemical and the way in which it can influence the structure of the DNA. For example, the amide 2-acetylaminofluorene differs from the amine 2-aminofluorene only by the presence of a carbonyl group. Both chemicals are metabolically activated via their *N*-hydroxy derivatives, yet the bulk of the amide–DNA adduct intercalates within the DNA, distorting the helix whereas the bulk of the amine–DNA adduct remains on the helix exterior. The consequence of these different arrangements is that the amide adduct results in frameshift mutations whereas the amine causes transversions.

The Fate of Carcinogen DNA Damage

Regardless of the specific nature of the carcinogen damage to DNA, it is important to note that aberrations at the primary sequence may not necessarily result in cancer. Indeed, there are at least three possible fates for such carcinogen damage. These are summarized in **Figure 6**.

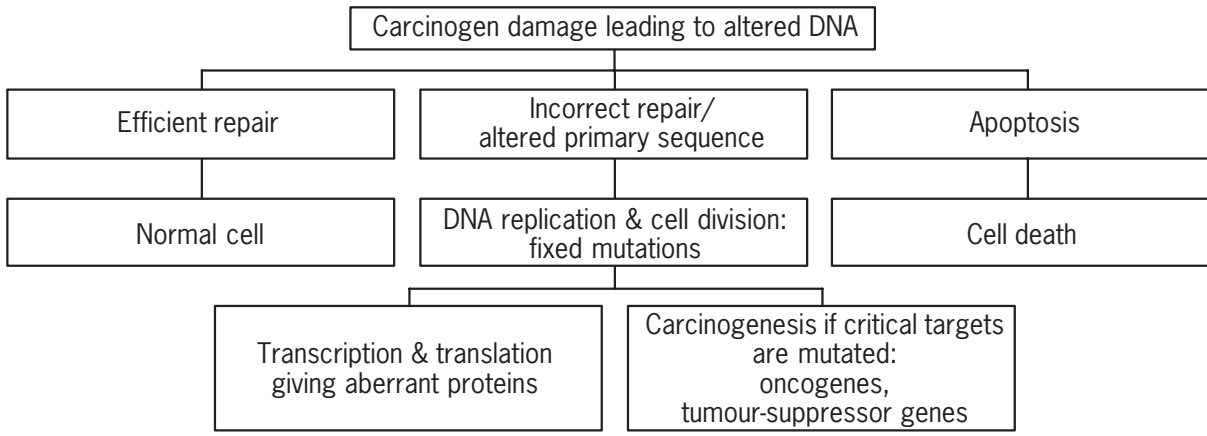


Figure 6 Possible fates for carcinogen-damaged DNA.

In some circumstances, the damage to the genetic machinery by a potential chemical carcinogen may be so extensive or be recognized as so crucial that cells initiate the process of apoptosis or programmed cell death. This behaviour and the involvement of such important proteins as p53 are discussed in detail elsewhere in this book, but through this activity, affected cells may effectively commit suicide and thus prevent the formation of potentially cancerous clones. However, if the carcinogen-altered DNA remains unchanged or if repair of the DNA is incorrect, then following DNA replication and cell division, daughter cells may contain fixed mutations. Transcription and translation of mutated genes in daughter cells will give rise to incorrect or inappropriate proteins being produced but carcinogenesis will not occur unless critical gene targets in the DNA are mutated (see later).

It follows, therefore, that repair of DNA-carcinogen damage in cells should ideally be fast, efficient and highly accurate. Indeed, mammalian cells invest a great deal of effort in order to achieve this and there are many genes and enzymes involved in maintaining the fidelity of the DNA sequence. These include enzymes that can bring about the reversal of DNA damage or repair the DNA through either base excision or nucleotide excision.

The Repair of Chemical-damaged DNA

Clearly the most straightforward way for a cell to repair a gene lesion is to remove it directly and thus regenerate the normal base at its correct position. One example of such direct removal involves the photodamage to DNA caused by ultraviolet (UV) radiation. UV radiation creates a number of photoproducts in DNA including the formation of mutagenic pyrimidine dimers between neighbouring thymine bases. Of these, the cyclobutane–pyrimidine photodimer can be repaired by a photoreactivating enzyme, although a 6–4-photoproduct cannot be repaired. The enzyme operates by binding to the photodimer and splits it back to the original bases, although the enzyme requires light energy and hence cannot work in the dark. Another enzyme involved in the direct removal of DNA damage is alkyltransferase, which can remove certain alkyl groups from altered guanine bases, although if the levels of alkylation are high enough then the enzyme activity can become saturated.

There are several pathways for excision of altered bases, often with the concomitant removal of neighbouring bases. Some systems recognize any lesion that causes a significant distortion of the DNA double helix and creates endonucleolytic cuts, several bases away and on either side of the lesion. The simple loss of a purine or pyrimidine may also initiate endonuclease cutting. The single-stranded DNA containing the damaged or missing base can then be removed and the short gap is filled in by DNA polymerases. Through this form of excision repair, bulky DNA adducts formed from the binding of PAHs and aflatoxins can be removed, as can certain UV-damaged bases. For some forms of damage (e.g. alkylation), DNA glycosylases remove just the altered base, leaving an apurinic or apyrimidinic (AP) site. The AP site is subsequently excised by endonucleases as above. Some repair pathways are capable of recognizing errors even after DNA replication has taken place. One of these systems, termed the mismatch repair system, can detect mismatches that occur during DNA replication (Friedberg, 1985). (See also chapter on *Genomic Instability and DNA Repair*.)

CHEMICALS CAN INDUCE MUTATIONS

The interaction of a chemical with DNA is not, in itself, a mutagenic event. The interaction of a xenobiotic with DNA can lead to mutagenesis due to attempts by the cell to repair the damaged DNA or during replication of the damaged DNA. Either way, a mistake may be made in which an inappropriate base is inserted or is lost (or a series of bases are inserted or are lost) from the region of DNA, with the consequence that a mutation is acquired, which is heritable. If the mutation occurs in a crucial piece of DNA, i.e. within a structural sequence that codes for an important gene controlling cell growth or a sequence that regulates the expression of such a gene, then the consequences may be disastrous. However, despite the human genome (some three billion bases) coding for approximately 28 000 genes, the majority of the DNA in a cell does not code for crucial information and mutations are usually not significant. However, the threat to the cell by DNA-damaging influences is substantial and it is not surprising that numerous proteins are involved in the detection and repair of damaged DNA.

Since mutation is evidence that a chemical is capable of inducing heritable genetic change, and since tumours often contain characteristic mutational patterns or mutational spectra in critical genes, the notion arises that the carcinogenic process is heavily dependent upon the acquisition of mutation in key genes (Harris, 1991). Furthermore, the observation that acquisition of specific types of mutation in such key genes is consistently associated with the formation of tumours, strengthens the argument that these changes are a driver of tumorigenesis. Although it is often difficult to detect and evaluate cancer-causing mutations in primary target cells and tissues, it is relatively straightforward to detect and identify mutations in surrogate reporter genes, and from information such as mutation frequency, nature and sequence context, the potential for a particular chemical to participate in the cancer process can be evaluated. This realization has driven the development of assays designed to evaluate mutagenic potential (and therefore carcinogenic potential). Currently systems are available to assess not only crude mutagenic activity, but also the likelihood of specific transitions, transversions (point mutations) or

deletions or insertions (frameshift mutations) or clastogenicity or aneuploidy. Since each of these types of DNA damage are detectable in tumours, the ability of a xenobiotic to induce such damage and therefore the relevance of the chemical to the neoplastic process can be assessed. For a more detailed description, see the chapter *Short-Term Testing for Genotoxicity*.

In vivo chemical-induced mutation is often subtle and insidious and it is usually difficult to ascribe particular mutations with exposure to a specific chemical, because evidence of the chemical is removed (loss of the DNA–chemical adduct) by introduction of the mutation. However, different classes of reactive chemicals appear to have a preference for individual DNA bases. For example, activated aromatic amines and PAHs appear to target guanine bases preferentially, probably owing to the fact that guanine is the most basic of the four nucleic acid bases. Some activated xenobiotics tend to induce transversions whereas others cause transitions or deletions. This has led to the concept of mutational fingerprints in which a specific activated xenobiotic is thought to induce a particular type of mutagenic response with a frequency which is greater than that expected by chance alone (Thilly, 1990; Aguilar *et al*., 1993; Yadollahi-Farsani *et al*., 1996). The concept of mutational fingerprints raises the prospect of identifying chemical causality in the occurrence of a particular type of mutational event, despite the fact that direct evidence for involvement (presence of the chemical) is not available.

The cooked food-borne carcinogen 2-amino-1-methyl-6-phenylimidazo[4,5-*b*]pyridine (PhIP) offers an example of a chemical whose mutagenic effects are characteristic and predictable (see **Table 1**). Numerous *in vitro* studies have shown this chemical to be an effective bacterial and mammalian cell mutagen (Gooderham *et al*., 1997). DNA sequence analysis of PhIP-induced mutation indicates that activation of this promutagen to damaging electrophilic species results in predominantly guanine-based mutation events. Several studies have shown the spectrum of these mutations to comprise mainly G to T transversions, with a few guanine transitions, and a significant percentage of frameshift mutations involving loss of G–C base pairs but not A–T base pairs (**Table 1**). Interestingly, PhIP-induced mutation appears to involve little or no mutation at A–T base pairs. It has also been commonly observed that PhIP-induced guanine mutation occurs primarily on the non-transcribed DNA strand, implying that damage induced on the transcribed strand is preferentially and successfully repaired. Additionally, detailed examination of the sequence context of PhIP induced mutation, shows that there appears to be a preferred motif (5′ GGGA 3′) within which the mutations are detected. It has been suggested that collectively, these observations describe a PhIP 'mutation fingerprint' and that the appearance of such a combination of mutational changes could be evidence for involvement of the chemical in their causation (Gooderham *et al*., 1996). In the case of PhIP, the nature and preferences of its DNA reactive species (PhIP nitrenium ion) tend to support the type of mutagenic outcome associated with the chemical. PhIP activated by cytochrome P-450-1 family enzymes is oxidised to the *N*-hydroxy derivative. Subsequent esterification of the *N*-hydroxy metabolite generates the acetoxy or sulfoxy derivatives, which spontaneously decompose, forming the highly reactive nitrenium ion (see **Figure 7**). The nitrenium ion shows preference for attacking guanine bases leading to the deoxyguanine-C8-PhIP (dG-C8-PhIP) and dG-N2-PhIP products. The inherent basicity of the purine molecule is a likely driver of the reaction and therefore it is no surprise that monotonous runs of guanine present as highly susceptible targets.

Table 1 PhIP induces a mutation fingerprint *in vitro* and *in vivo*

Mutation type	Percentage of mutations					
	hprt[a]	*hprt*[b]	*dhfr*[c]	*lacZ*[d]	*lacI*[e]	*lacI*[f]
Base substitutions	83	89	80	66	71	71
At GC pairs						
GC→TA	63	56	65	33	49	43
GC→AT	15	27	5	25	10	10
GC→CG	5	6	5	5	9	15
At AT pairs	0	0	5	3	3	3
Single bp frameshifts	13	9	5	22	26	22
(GC bp)	13	9	5	20	26	21
(AT bp)	0	0	0	2	0	1
Other mutations	4	2	15	12	3	8

[a]*hprt* gene in V79 Chinese hamster fibroblast cells (Yadollahi-Farsani *et al.*, 1996)
[b]*hprt* gene in TK6 human lymphoblastoid cells (Morganthaler and Holzhauser, 1995).
[c]*dhfr* gene in CHO Chinese hamster ovary cells (Carothers *et al.*, 1994).
[d]*lacI* gene in Muta™mouse mice large intestine (Lynch *et al.*, 1998).
[e]*lacZ* gene in Big Blue mice large intestine (Okonogi *et al.*, 1997).
[f]*lacZ* gene in Big Blue rat mammary gland (Okochi *et al.*, 1999).

As discussed, the mechanisms whereby mutations arise and their chemical nature can vary considerably. In some instances, adducted DNA can base pair with its correct complementary base, apparently unaffected by the adduct whereas in other cases base pairing becomes degenerate and inappropriate bases are used during repair or replication. If the adduct is not correctly repaired, then replication can result in a mutation being fixed. In other instances, the presence of the adduct physically blocks DNA synthesis, effectively terminating the process at that point. Since such blocks would be lethal, the cell uses bypass mechanisms to overcome the block and adenine is frequently used to pair with the damaged (adducted) base. If adenine is not the original complementary base, then a mutation has been introduced into the sequence.

The ability of chemicals to induce mutation is an important aspect of their involvement in carcinogenesis. This is particularly true for the initiation process but it also

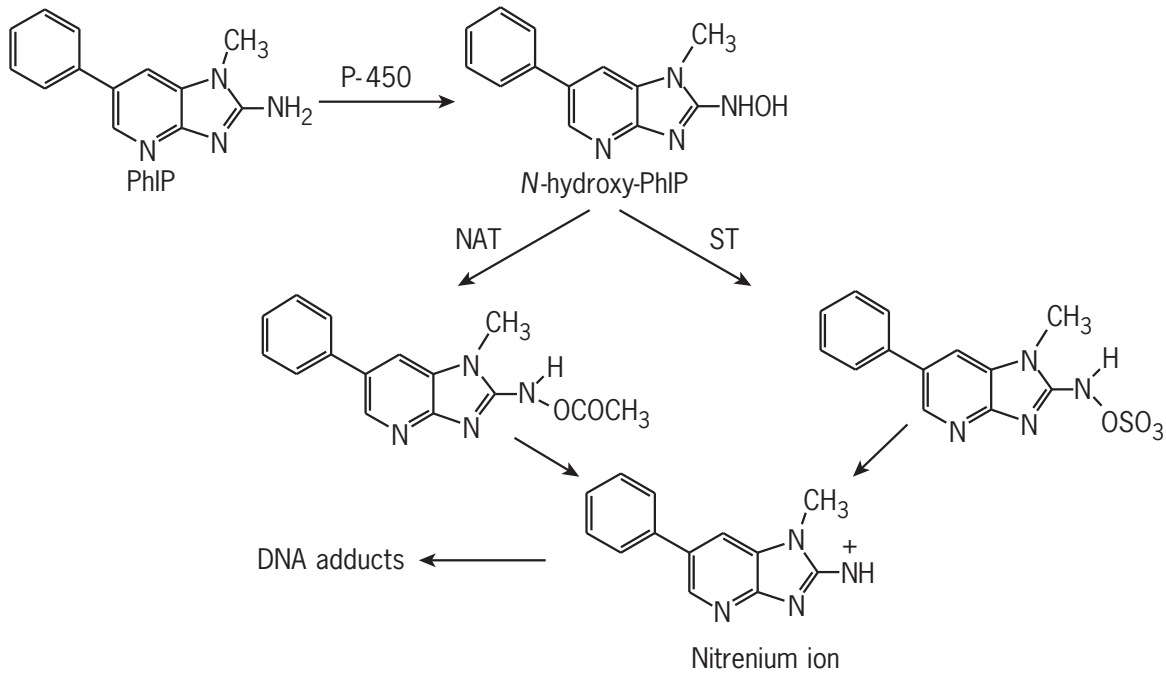


Figure 7 The metabolic activation of 2-amino-1-methyl-6-phenylimidazo[4,5-*b*]pyridine (PhIP). P-450, cytochrome P-450; NAT, *N*-acetyltransferase; ST, sulfur transferase.

contributes to promotion and progression stages during which cells continue to accumulate genetic damage at key gene sites (see later).

EXAMPLES OF CHEMICAL CARCINOGENS AND THEIR METABOLISM

2-Acetylaminofluorene

2-Acetylaminofluorene (2-AAF) is a potent mutagen and carcinogen, which induces tumours in a number of species in the liver, bladder and kidney. As previously mentioned, metabolism of this compound is the key to its carcinogenicity. 2-AAF is a substrate for cytochrome P-450 enzyme, of which there is a superfamily in mammals (Gonzalez, 1989; Porter and Coon, 1991). The product of this metabolism (**Figure 8**) is *N*-hydroxy-2-acetylaminofluorene (*N*-OH-2-AAF), which is a more potent carcinogen that the parent molecule. The formation of *N*-hydroxy metabolites has been found for many aromatic amine, amide, nitro and nitroso compounds, many of which exert toxicity through this type of derivative. For nitro and nitroso compounds, the *N*-hydroxy metabolite is formed by reduction rather than oxidation.

In the case of 2-AAF, the *N*-hydroxy metabolite is not the ultimate carcinogen but a proximal carcinogen, and this compound undergoes several enzymic and non-enzymic rearrangements. The compound can be *O*-acetylated by cytosolic *N*-acetyltransferase enzyme to yield the *N*-acetyl-*N*-acetoxy derivative, which can spontaneously rearrange to form the arylamidonium ion and a carbonium ion and interact directly with DNA to produce DNA adducts. In addition to esterification by acetylation, the *N*-OH-2-AAF can be *O*-sulfated by cytosolic sulfur transferase enzyme giving rise to the *N*-acetyl-*N*-sulfoxy product. This is again unstable and spontaneously generates the arylamidonium ion and carbonium ion, which can adduct to DNA. In addition, the cytosolic *N*,*O*-arylhydroxamic acid acyltransferase enzyme catalyses the transfer of the acetyl group from the N atom of the *N*-OH-2-AAF to the O atom of the *N*-OH group to produce *N*-acetoxy-2-aminofluorene (*N*-OH-2-AF). This reactive metabolite spontaneously decomposes to form a nitrenium ion, which will also react with DNA. However, the product of this latter reaction is the deacetylated aminofluorene adduct (see **Figure 8**). The interconversion of amide and amine metabolites of 2-AAF can further occur via the microsomal enzyme deacetylase, producing the *N*-hydroxy metabolite of the amine derivative. Subsequent esterification of the aryl hydroxylamine by sulfur transferase yields the reactive sulfate ester, which also spontaneously decomposes to form the reactive nitrenium ion. The reactive nitrenium, carbonium and arylamidonium ion metabolites of 2-AAF react with the nucleophilic groups in DNA, proteins and endogenous thiols such as glutathione (Miller and Miller, 1981). These interactions can be demonstrated *in vitro* and *in vivo*. Other metabolites such as the *N*,*O*-glucuronide, although not directly activated products, can be important in the carcinogenic process due to the fact that they are capable of degradation to proximal

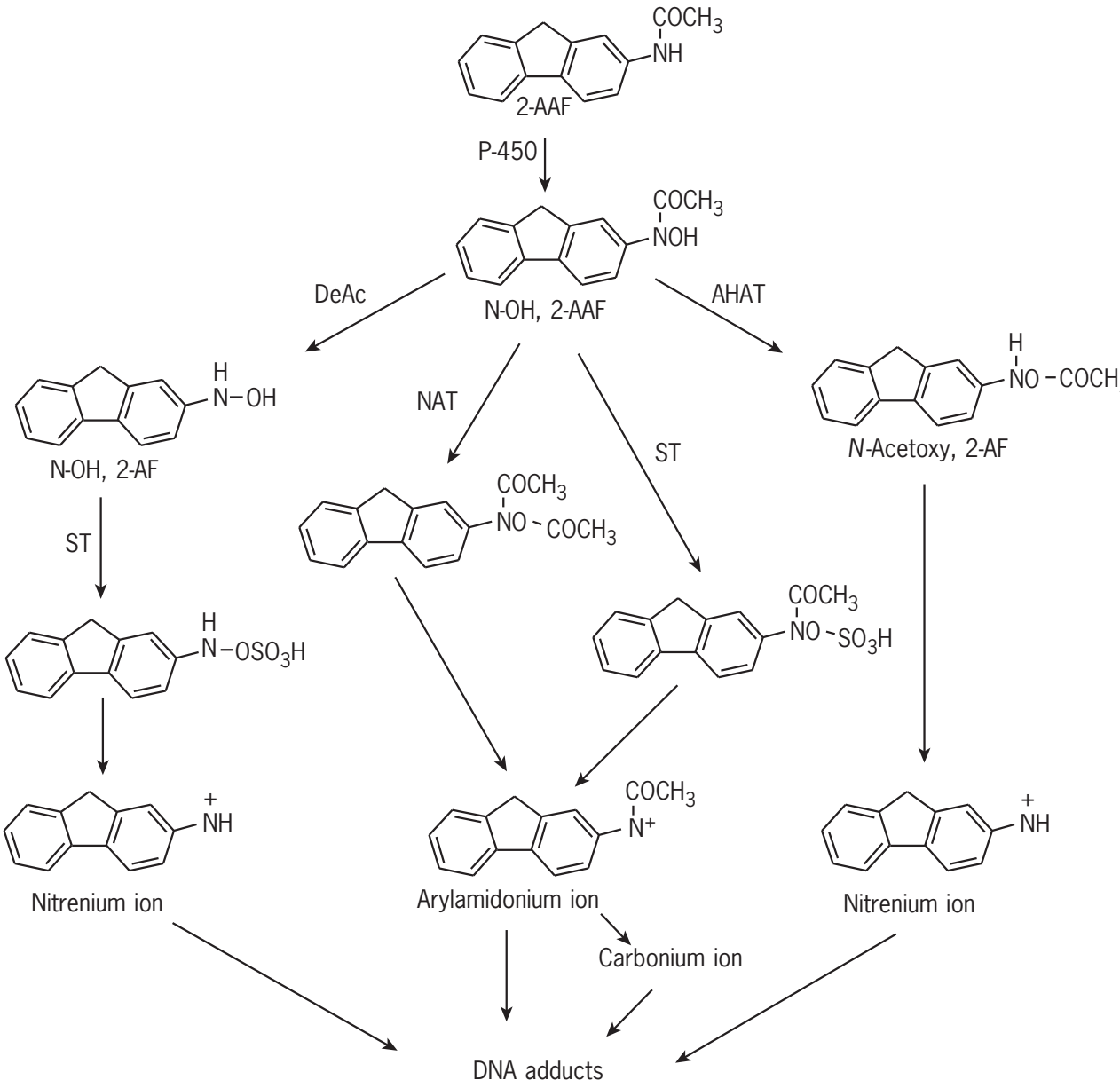


Figure 8 The metabolic activation of 2-acetylaminofluorene. P-450, cytochrome P-450; DeAc, deacetylase; AHAT, *N,O*-arylhydroxamic acid acyltransferase; NAT, *N*-acetyltransferase; ST, sulfur transferase.

N-hydroxy metabolites. Thus, the *N,O*-glucuronide of *N*-OH-2-AAF is less reactive than the acetoxy and sulfate esters, yet it may be involved in the formation of bladder tumours. The mechanism for this is thought to involve degradation of the glucuronide in the bladder due to the acidic pH of urine giving rise to the *N*-hydroxy proximal carcinogen, which is then a substrate for esterification.

N-Hydroxylation of 2-AAF is not the only route of metabolism. Ring hydroxylation, catalysed by cytochrome P-450 enzymes, can also occur which generates metabolites that are not carcinogenic *per se*. These ring-hydroxylated products can be further metabolized to glucuronidated products that are readily excreted.

Benzidine

As can be seen from the example of 2-AAF, the reactivity and carcinogenicity of aromatic amides can involve their conversion to aromatic amines. Another example of this class of chemical carcinogen is benzidine. Benzidine (**Figure 9**) is a carcinogenic bifunctional aromatic amine. It can undergo several routes of metabolism, but with regard to its carcinogenicity, *N*-hydroxylation and *N*-esterification are important. The amine function at both ends of the molecule is subjected to *N*-acetylation to the corresponding amide and they can also be *N*-hydroxylated by cytochrome P-450 enzymes. The resulting aryl

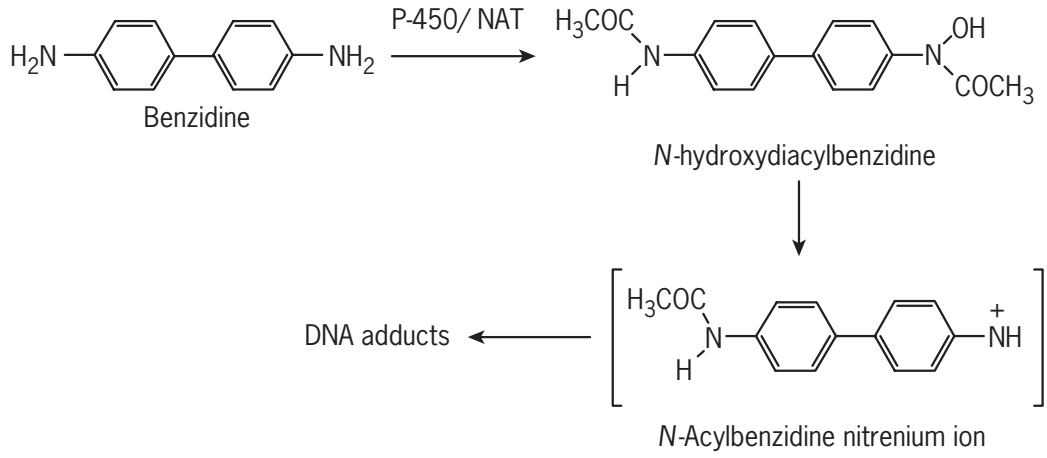


Figure 9 The metabolic activation of benzidine. P-450, cytochrome P-450; NAT, *N-acetyltransferase*.

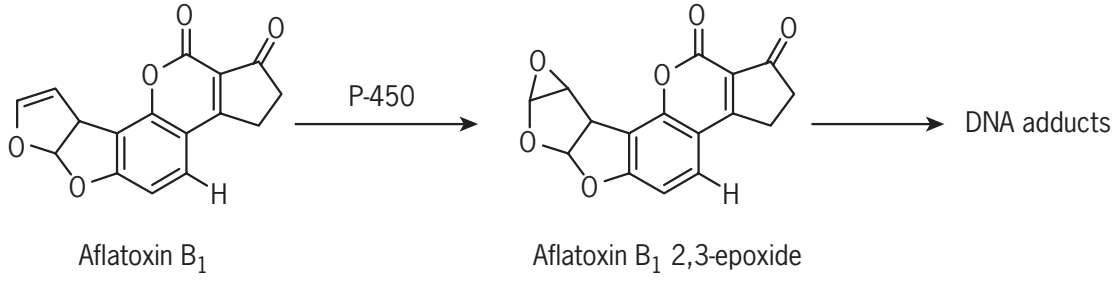


Figure 10 The metabolic activation of aflatoxin B_1. P-450, cytochrome P-450.

hydroxamic acid is unstable and rearranges to form electrophilic nitrenium ion derivatives which will rapidly interact with cellular nucleophiles such as DNA, forming DNA adducts (Searle, 1984).

Aflatoxin B_1

Aflatoxin B_1 is one of a family of mycotoxin contaminants of food crops such as grain and groundnuts. Produced by *Aspergillus flavus*, especially in hot and humid conditions, there are four main types of aflatoxin, B_1, B_2, G_1 and G_2. Aflatoxin B_1 is not only the most toxic, but is also the most carcinogenic. Contamination of food with aflatoxins is a significant problem in parts of Africa and Asia where conditions are particularly favourable for the growth of the *Aspergillus* organism. Indeed, epidemiological studies show that in areas of the world where contamination of food supplies with aflatoxins is high, there is an associated incidence of human hepatocellular carcinoma. Although contamination of grain crops with aflatoxins is not restricted to these parts of the world, in the developed countries food surveillance programmes backed up by tight legislative control regulating maximum permissible levels of aflatoxins in food crops control human exposure to these potent chemical carcinogens.

The basis of the carcinogenicity of aflatoxin B_1 centres on the carbon–carbon double bond (**Figure 10**) in the terminal furan ring of the molecule (Wogan *et al*., 1971). The chemical is a substrate for cytochrome P-450 enzymes that oxidize the carbon–carbon double bond. The epoxide product is an electrophile that rapidly reacts with cellular nucleophiles such as DNA and the resulting DNA–aflatoxin B_1 adduct is powerfully promutagenic. This and the fact that it is highly hepatotoxic makes aflatoxin B_1 one of the most powerful carcinogens known.

Benzo[*a*]pyrene

The PAHs are a large group of environmentally important chemical carcinogens. Benzo[*a*]pyrene is a prominent example of these compounds, being ubiquitous in our environment since it is present in the smoke and fumes from many diverse sources, including cigarettes, exhaust fumes and the burning of many different organic materials including wood and fossil fuels. It is one of the most extensively studied PAHs. It is a substrate for the mixed-function oxidases, being extensively converted to a variety

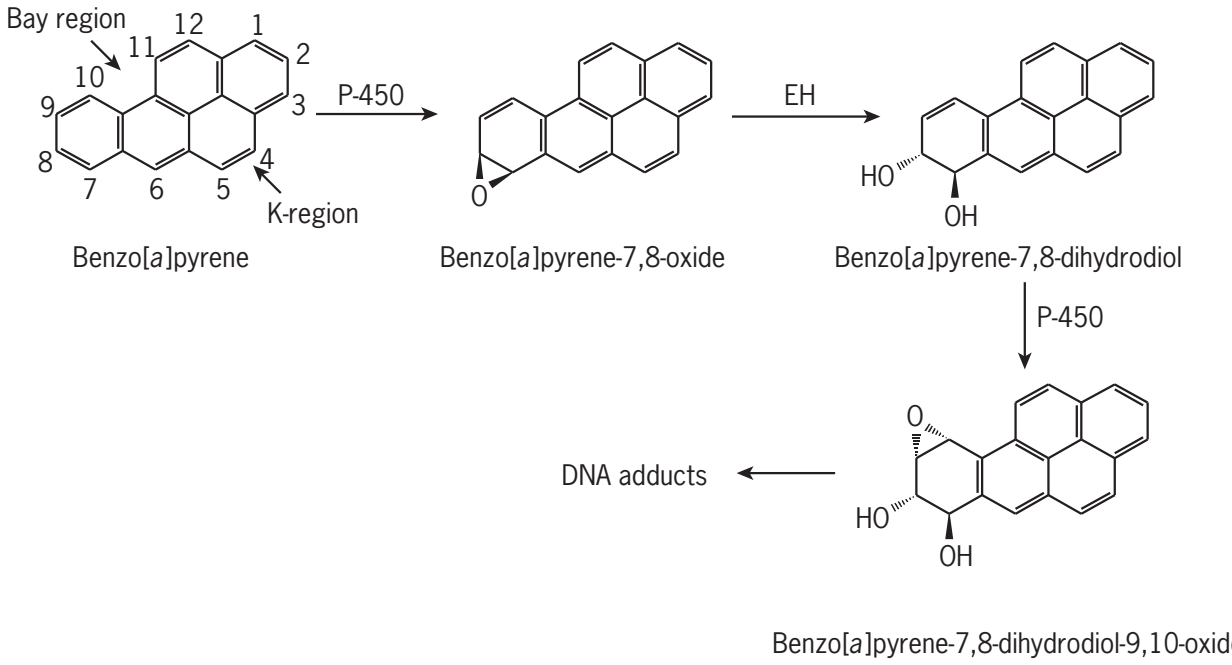


Figure 11 The metabolic activation of benzo[*a*]pyrene. P-450, cytochrome P-450; EH, epoxide hydrolase.

of different oxidized metabolites including epoxides, phenols, diols, dihydrodiols and their conjugated products, particularly with glutathione, glucuronic acid and sulfate. The genetic toxicity of the PAHs is based around the formation of their epoxides (Sims *et al*., 1974). For example, with benzo[*a*]pyrene, cytochrome P-450-1 family enzymes can generate a series of epoxides around the different rings of the molecule, some of which are known to be more carcinogenic than others. Electronically, the most reactive portion of the benzo[*a*]pyrene molecule is the so-called 'K region' (see **Figure 11**), yet it is epoxides of the 'bay region,' which are thought to be the most tumorigenic. The formation of the ultimate carcinogen of benzo[*a*]pyrene involves cytochrome P-450-mediated epoxidation at the 7,8-position of the molecule. This epoxide is hydrolysed by the enzyme epoxide hydrolase to form the 7,8-dihydrodiol. The dihydrodiol subsequently undergoes cytochrome P-450-mediated oxidation, forming the 7,8-dihydrodiol-9,10-epoxide, which is thought to be the ultimate carcinogen. However, diastereoisomers of these metabolites are formed metabolically and the (+)-benzo[*a*]pyrene, (7*R*,8*S*)-dihydrodiol-(9*S*,10*R*)-epoxide formed by cytochrome P-450 1A1 and epoxide hydrolase generates a species that is more mutagenic than other isomers. Interestingly, although the K region 4,5-epoxide is highly mutagenic, the bay region 7,8-dihydrodiol-9,10-epoxide is carcinogenic whereas the 4,5-epoxide and the 9,10-epoxide are not, indicating the requirement for the specific dihydrodiol formation. Other metabolites of benzo[*a*]pyrene are known to be cytotoxic and mutagenic, e.g. the 3,6-quinone derivative, but have poorer carcinogenic potential (Conney, 1982).

Dimethylnitrosamine

The nitrosamines are another extensively studied family of chemical carcinogens. Dimethylnitrosamine, representative of this family, is hepatotoxic, mutagenic and carcinogenic, causing kidney tumours with acute exposure and liver tumours after chronic exposure (Magee *et al*., 1976). Additionally, tumours of the stomach, oesophagus and central nervous system are found. Again, cytochrome P450 enzymes are central to the metabolic activation of dimethylnitrosamine, involving *N*-demethylation to *N*-methylnitrosamine. This metabolite rearranges to form methyldiazohydroxide, then methyldiazonium ion and ultimately methyl carbonium ion. It is the methyl carbonium ion that is the DNA-damaging species since it is a highly reactive alkylating agent (see **Figure 12**). The degree of DNA methylation that occurs after exposure to dimethylnitrosamine correlates with susceptibility to tumour formation. The DNA sites that appear to be most sensitive to alkylation are the N^7 position of guanine and to a lesser extent the O^6 position of guanine, but it is the latter site that appears to correlate best with mutagenicity and carcinogenicity. Alkylation at guanine O^6 leads to guanine–thymidine mispairing, causing a GC to AT transition. Paradoxically, acute doses of dimethylnitrosamine cause greater DNA methylation in the liver than the kidney, yet tumours are preferentially found in the kidney. Clearly, although the liver has more metabolic activation capacity, it also has better protective mechanisms. This balance between activation and protective mechanisms is also likely to be a significant factor in the susceptibility

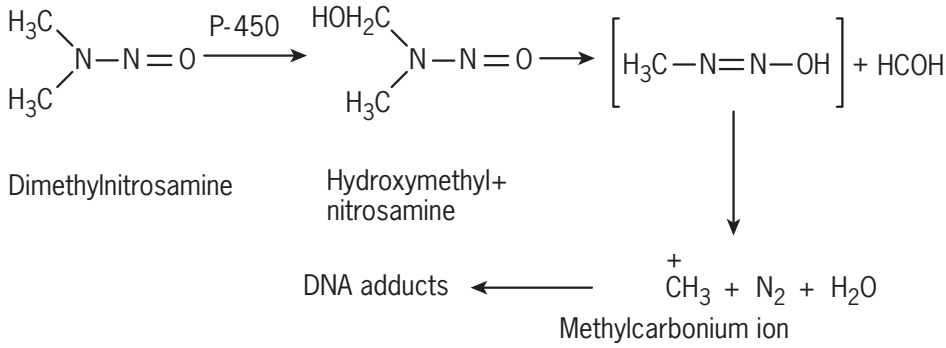


Figure 12 The metabolic activation of dimethylnitrosamine. P-450, cytochrome P-450.

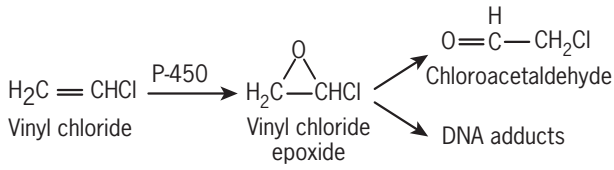


Figure 13 The metabolic activation of vinyl chloride. P-450, cytochrome P-450.

of other organs and tissues to dimethylnitrosamine-induced carcinogenesis.

Vinyl chloride

Vinyl chloride is a simple halogenated allyl compound that is extensively used in the plastics industry, being the starting point for a number of polymer syntheses, particularly the manufacture of poly(vinyl chloride). A gas at room temperature and therefore usually stored as a liquified gas under pressure, the use of the material in an industrial setting is substantial and numerous workers have suffered well-documented accidental occupational exposure, for example being overcome by vinyl chloride solvent narcosis during the cleaning of reaction vessels. Acute exposure to the material is associated with an unusual form of liver tumour, known as haemangiosarcoma, a tumour of reticuloendothelial cells, and not hepatocytes, giving rise to tumours of the hepatic vasculature. This very rare type of liver cancer has only been observed in workers who have been exposed to vinyl chloride.

For tumorigenicity, vinyl chloride requires metabolic activation by cytochrome P-450 enzymes (Bolt, 1988). Oxidation to the epoxide intermediate (**Figure 13**) is the first step resulting in the subsequent formation of chloroacetaldehyde. Although a number of other metabolites are known to be generated, the tumorigenicity of the compound is likely to be dependent upon the epoxide and chloroacetaldehyde metabolites, both of which react with cellular nucleophiles. Subsequently glutathione is depleted and the excess reactive metabolites can then react with nucleic acids and protein. The available evidence suggests that the epoxide is the predominant species binding to DNA where it reacts with deoxyguanosine at the N^7 position and with RNA to give 1,N^6-ethenoadenosine and 3,N^4-ethenocytidine, whereas the chloroacetaldehyde probably binds to protein.

THE BIOLOGY OF CHEMICAL CARCINOGENESIS

The involvement of chemicals in damaging DNA is only one part of their potential role in carcinogenesis. In progressing to the neoplastic state, cells must undergo fundamental changes in their biology and many of these changes can be driven by chemical intervention of essential cellular activity. Current perception of the neoplastic process has been shaped by fundamental studies performed in the 1940s using PAHs with the mouse skin carcinogenesis model (Berenblum and Shubik, 1947). From such studies emerged the concept of the multistep nature of carcinogenesis and the defining of these steps as initiation, promotion and subsequently progression (see **Table 2**).

Initiation

A key feature of the initiation process (**Table 2**) is the requirement for cell replication (Pitot and Dragan, 1996). Once a chemical has damaged DNA by forming an adduct, or inducing a strand break, etc., and provided that the cell recognizes the damage, repair processes will intervene. Should the damage be misrepaired, or not recognized by the cell, then the outcome can be (1) aberrant transcription of the affected DNA if it is a structural gene lesion and on the transcribed strand, (2) altered expression if the damage is located within a regulatory sequence or (3) no effect if the damage is within noncoding or nonregulatory DNA. If the damage is at a critical gene site or extensive, then the cell may opt to apoptose. However, since the cell has two copies of each autosomal gene loci, compensation for the

Table 2 The stages of initiation, promotion and progression

Initiation	Promotion	Progression
Irreversible	Reversible	Irreversible
Cell not morphologically recognizable		Morphologically discernable
Single exposure to chemical is sufficient	Requires multiple exposure to chemical	Driven by multiple and varied exposure
Can occur spontaneously	Endogenous and exogenous chemicals can promote	Cells show altered sensitivity to endogenous chemicals
'Fixed' by cell division	Expanded through cell division	Driven by cell division and migration
No obvious dose–response threshold	Measurable dose–response threshold and maximum effect	Effect of individual chemicals difficult to discern
Involves simple mutation (transitions, transversions, frameshifts)	Does not necessarily involve mutation	Frequently involves complex mutation
Can involve single genes	Often involves multiple genes	Usually involves chromosomal alterations (clastogenicity, aneuploidy, polyploidy)
Proto-oncogenes and tumour-suppressor genes can be mutated	Altered levels of gene expression rather than mutation	Substantial alteration in gene expression due to multiple mutational events
Simple genotypic changes with corresponding phenotype change	Epigenetic phenotype changes	Complex phenotypic change
Limited growth advantage over surrounding cells	Chemical mediated selective growth advantage	Significant growth advantage due to genetic change and environment

(Adapted from Pitot and Dragan, 1996.)

damaged gene is often possible. The only cell directly affected is the recipient of the damage, the initiated cell. However, if division occurs, a daughter cell will inherit the mutation, which becomes fixed irrespective of whether it was misrepaired or remained as an unrepaired adduct in the parent cell, and will be inherited in all future generations of this lineage. If the cell type is generally quiescent, then the effect of the DNA lesion is unlikely to become apparent unless the cell experiences a proliferative stimulus (e.g. by promotion). However, if the cell is normally proliferative (e.g. a stem cell), then there is a much greater chance of neoplastic development, even in the absence of concerted promotion.

Promotion

Once the initiated cell has been encouraged to replicate, the initiated genetic damage is irreversibly fixed. Yet the phenotypic characteristics of the initiated cell remain insidious because it is surrounded by (and perhaps compensated by) normal cells. When the initiated cell is subjected to promotional influences, the effect is to encourage clonal expansion of the initiated cell. A key feature of the promotion response is its reversibility, i.e. if the promoting agent is removed, the initiated cell population is no longer encouraged to proliferate (Boutwell, 1964). The promotional stimulus is usually not a direct interaction of the promoting agent on DNA but acts to encourage growth and proliferation of the cell by exaggerating favourable environmental conditions, especially for initiated cells (see **Table 2**). This can be achieved by chemical interference with normal cellular signal transduction mechanisms with the consequence of enhanced transcription and translation.

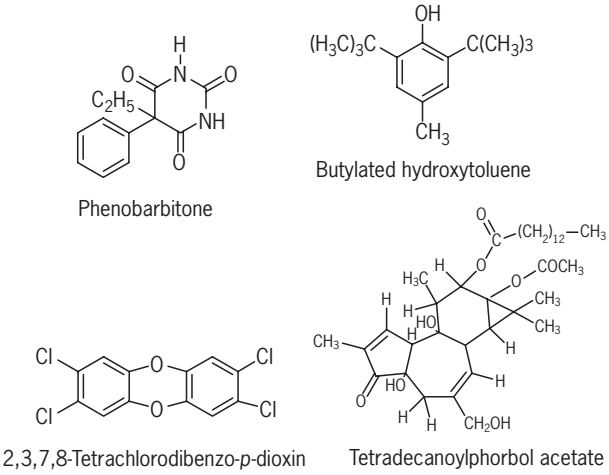


Figure 14 Examples of chemical promoters.

A number of well-characterized promoting agents, both natural and xenobiotic, are shown in **Figure 14**. A good example is tetradecanoylphorbol acetate (TPA), a naturally occurring component of croton oil. This chemical interacts with the key protein, protein kinase c, causing enhanced signalling via inositol triphosphate and diacylglycerol pathways, resulting in increased transcription and strong stimulation of growth. 2,3,7,8-Tetrachlorodibenzo-*p*-dioxin (TCDD) is another example of a xenobiotic with exceptional promoting activity, being particularly effective in the liver, lung and skin. There are also a number of very powerful naturally occurring promoting agents including androgens and oestrogens (Taper, 1978). In all cases removal of the agent can rapidly lead to regression of the preneoplastic lesion.

Another important characteristic of promoting agents is the existence of concentration thresholds. Thus a minimum dose of agent is required before effective promotion can be detected, after which there is a dose–response relationship before a maximum response is reached. This latter effect is probably due to saturation of physiological targets, usually endogenous receptors. Other physiological factors can alter the impact of promoting chemicals, e.g. diet and hormonal status will influence the promotion of numerous preneoplastic lesions including those in the mammary and hepatic tissues.

Progression

As the initiated cell population is expanded during promotion, individual cells acquire further genetic damage, which can be either agent-mediated or spontaneous, thereby introducing genetic heterogeneity into the promoted population. Acquisition of these changes is irreversible and progressive (**Table 2**). Ultimately, some promoted cells will acquire karyotype instability, which can bestow growth advantage over surrounding cells. In some cases this will remove environmental constraints over such cells, allowing them to escape proximal cellular and humoural influences. In many cases, such cells will have acquired clastogenic changes that bestow significant phenotypic change (Foulds, 1965; Welch and Tomasovic, 1985).

CRITICAL GENE TARGETS

The reaction between an activated carcinogen and DNA is concentration dependent. Since DNA comprises only four bases then, in theory, chemicals should damage DNA at any site within the genome. In practice there appears to be preferred sites at which attack occurs. The factors which govern the chance of attack at such sites would be expected to include accessibility, the presence of modifying influences (extent of methylation or association with nuclear proteins), etc. Furthermore, factors such as (1) DNA repair, (2) the presence of transcriptionally active or silent genes, (3) the role of the gene in normal cell function, (4) whether the chemical attacks a structural domain of the gene or (5) a regulatory domain and (6) whether the DNA target has any function whatsoever, will affect outcome. Although DNA is chemically the same throughout the genome, some sections (containing critical genes) are clearly more important to the correct functioning of the cell than are others. The number of currently recognized critical gene targets is only a fraction of the estimated 28 000–30 000 genes of the human genome, but is growing. These critical genes are involved in controlling the growth and differentiation of cells, damage or mutation of which is highly correlated with the neoplastic process; such genes are described as oncogenes or tumour-suppressor genes.

Oncogenes

Under normal circumstances oncogenes exist as proto-oncogenes (Garrett, 1986); they code for proteins that are involved in important pathways within the cell such as growth, signal transduction, cell cycle and nuclear transcription. Chemical-induced damage to key sites within the proto-oncogene can convert the sequence to one that codes for an aberrant protein. The product of this genetic change (conversion to an oncogenic variant of the gene) is a protein that has an altered sequence due to loss, gain or replacement of amino acids, and no longer responds to appropriate controlling stimuli. The activation of proto-oncogenes to oncogenes can be achieved by several well-recognized processes (Pitot, 1986). These may include (1) mutation of the structural gene, (2) mutation of regulatory sequences, (3) amplification of the structural gene, (4) translocation of the gene to a site of inappropriate gene regulation (e.g. within the control of a powerful promoter sequence) after clastogenic events or (5) epigenetic alteration of either structural or regulatory sequences leading to altered regulation of gene expression (e.g. altered gene methylation). Each of these mechanisms generates a functionally dominant gene and can be induced by chemical interaction at key sites within DNA. Examples of specific oncogenes are described in **Table 3** and can be grouped into functional biochemical classes, which code for key regulatory activities involved in correct functioning of the cell. These genes tend to be involved in pathways that mediate cell growth, differentiation and signalling. The ability of mutagenic chemicals to activate oncogenes is well established. A good example is the codon 12 K-*ras* mutations found in aberrant crypt foci in the colon of rats treated with azoxymethane (Shivapurkar *et al*., 1994).

Tumour-suppressor Genes

Tumour-suppressor genes normally act to suppress the neoplastic phenotype. As with oncogenic activation, chemical damage to a tumour-suppressor gene can give rise to a protein with altered activity, in this case making it unable to operate as a suppressor of the cancer process (Knudson, 1993). Since these genes tend to be functionally recessive, full loss of phenotypic expression often requires mutation or loss of both alleles and chemicals tend to destroy or alter the normal cellular activity of the tumour-suppressor gene. Many of these genes have a nuclear location/function and are directly involved in maintaining genome integrity. Examples of tumour-suppressor genes are given in **Table 3**. Chemical-mediated damage of the *p53* gene is a frequently encountered event in chemical-induced neoplastic disease. For example, aflatoxin B_1 exposure is associated with a high incidence of hepatocellular carcinoma in parts of Africa. Examination of the *p53* gene in these tumours reveals characteristic mutations at specific

Table 3 Examples of critical genes involved in carcinogenesis

Gene	Function	Localization
Oncogenes		
SIS, *FGF*, *INT2*, *WNT1*	Growth factors	Extracellular
MET, *NEU*, *EPH*, *EGRF*, *FMS*, *KIT*, *HER2*, *RET*, *ROS*,	Receptor/protein tyrosine kinases	Extracellular/cell membrane
SRC, *ABL1*, *FPS*, *FGR*, *FYN*, *HCK*, *LCK*, *YES*	Nonreceptor tyrosine kinases	Cell membrane/cytoplasmic
MAS	Receptors lacking protein kinase activity	Cell membrane/cytoplasmic
RAS, *GIP2*, *GSP*	Membrane-associated G proteins	Cell membrane/cytoplasmic
BCR, *DBL*, *ECT2*	RHO/RAC-binding proteins	Cytoplasmic
RAF, *PIM1*, *BCR*, *EST*, *MOS*,	Cytoplasmic protein serine kinases	Cytoplasmic
STY	Protein serine, threonine and tyrosine kinase	Nuclear
BCL1, *CRK*, *ODC*, *NCK*	Cytoplasmic regulators	Cytoplasmic
MYC, *FOS*, *JUN*, *BCL3*, *CBL* *ERBA*, *ETS*, *HOX*, *MYB*, *MYCL*, *REL*, *TAL1*, *SKI*	Nuclear transcription factors	Nuclear
BCL-2	Mitochondrial membrane factor	Mitochondrial/cytoplasmic
Tumour-suppressor genes		
NF1	GTPase activation	Cell membrane/cytoplasmic
RB-1	Cell cycle-regulated nuclear transcription repressor	Nuclear
P53	Cell cycle-regulated nuclear transcription repressor	Nuclear
WT1	Zinc finger transcription factor	Nuclear
HMLH1	Mismatch DNA repair	Nuclear
BRCA1	DNA repair enzyme	Nuclear
APC	Regulates cytoskeletal networks	Cytoplasmic
DCC	Cell adhesion molecule	Plasma membrane
VHL	Signal transduction or cell–cell contact	Plasma membrane
NME	Cell receptor	Plasma membrane
CMAR/CAR	Cell attachment	Plasma membrane
WNT	Growth factor	Extracellular matrix
YES1	Tyrosine kinase	Plasma membrane

(Adapted from Hesketh, 1994.)

hot spots (e.g. codons 248/249). These tumours and p53 mutations can be reproduced *in vivo* in animals and *in vitro* in cells exposed to pure aflatoxin B_1 (Aguilar *et al*., 1993).

CONCLUSION

Cancer is often perceived as largely due to inherited defective genes, a view fuelled by recent discoveries in this area. However, the biggest cause of human cancer is the influence of chemical carcinogens. Indeed, the biggest contributors to the burden of human cancer (estimated to be about 90% of all cancers) are carcinogens contained within tobacco smoke, diet and the environment (Doll and Peto, 1981).

Carcinogenesis is a multistep process and many of these steps can be influenced by chemicals. Clearly a chemical with carcinogenic or mutagenic properties should be used with caution, especially in cases where it would come into contact with humans. Yet there will be some instances where the benefit of using the chemical will outweigh its immediate risk, e.g. as a chemotherapeutic agent. Under such circumstances its use must be carefully regulated and monitored.

As discussed, many chemicals are carcinogenic through metabolic bioactivation and genotoxic activity at the level of the genome. However, this primary activity is only one factor in a sequence of complex events. Of thousands of chemicals tested, a few hundred have been identified as carcinogenic in rodents and around 50 are human carcinogens. With the development of new pharmaceuticals and industrial chemicals, carcinogenicity testing has become an essential activity and has led to a clearer understanding of the mechanistic basis of the carcinogenicity process.

The value of understanding the mechanisms whereby a chemical causes cancer cannot be underestimated. It provides a rational basis for risk assessment allows predictions of possible outcomes and informs the development of alternatives or strategies, which reduce risk. In the case of

therapeutically useful agents, it indicates how improved pharmaceuticals can be conceived and employed. It is also increasingly evident that each chemical with carcinogenic properties is unique and that grouping chemical carcinogens on the basis of their structural attributes is not always appropriate and thus prediction of genetic toxicity should always be confirmed by experimental data. In this respect, the development of new technologies such as transgenic animals that have altered expression of proto-oncogenes and tumour-suppressor genes greatly add to the battery of systems available for helping us understand cancer. Such models will continue to provide substantial information on the mechanisms through which chemicals and biological processes interact to drive the neoplastic process.

REFERENCES

Aguilar, F., *et al*. (1993). Aflatoxin B1 induces the transversion of G→T in codon 249 of the p53 tumour suppressor gene in human hepatocytes. *Proceedings of the National Academy of Sciences of the USA*, **90**, 8586–8590.

Beatson, G. T. (1896). On the treatment of inoperable cases of carcinoma of the mamma: suggestions for a new method of treatment with illustrative cases. *Lancet*, **2**, 104–107.

Berenblum, I. and Shubik, P. (1947). A new quantitative approach to the study of the stages of chemical carcinogenesis in the mouse's skin. *British Journal of Cancer*, **1**, 383–391.

Bolt, H. M. (1988) Roles of etheno–DNA adducts in tumorigenicity of olefins. *CRC Critical Reviews in Toxicology*, **18**, 299–309.

Boutwell, R. K. (1964). Some biological aspects of skin carcinogenesis. *Progress in Experimental Tumor Research*, **4**, 207–250.

Brand, K. G., *et al*. (1975). Etiological factors, stages, and the role of the foreign body in foreign body tumourigenesis. A review. *Cancer Research*, **35**, 279–286.

Carmichael, P. L. (1998). Mechanisms of action of antiestrogens: relevance to clinical benefits and risks. *Cancer Investigation*, **16**, 604–611.

Carothers, A. M., *et al*. (1994). Mutation and repair induced by the carcinogen 2-(hydroxylamino)-1-methyl-6-phenylimidazo[4,5-*b*]pyridine (N-OH-PhIP) in the dihydrofolate reductase gene of Chinese hamster ovary cells and conformational modelling of the dG-C8-PhIP adduct in DNA. *Chemical Research in Toxicology*, **7**, 209–218.

Conney, A. H. (1982). Induction of microsomal enzymes by foreign chemicals and carcinogenesis by polycyclic aromatic hydrocarbons: GHA Clowes Memorial Lecture. *Cancer Research*, **42**, 4875–4917.

Dogliotti, E. (1996). Mutational spectra: from model systems to cancer-related genes. *Carcinogenesis*, **17**, 2113–2118.

Doll, R. and Peto, R. (1981). *The Causes of Cancer* (Oxford University Press, Oxford).

Eling, T. E., *et al*. (1990). Prostaglandin H synthatase and xenobiotic oxidation. *Annual Review of Pharmacology and Toxicology*, **30**, 1–45.

Foulds, L. (1965). Multiple etiologic factors in neoplastic development. *Cancer Research*, **25**, 1339–1347.

Friedberg, E. C. (1985). *DNA Repair* (Freeman, New York).

Garrett, C. T. (1986). Oncogenes. *Clinica Chemica Acta*, **156**, 1–40.

Gonzalez, F. J. (1989). The molecular biology of cytochrome P450s. *Pharmacological Reviews*, **40**, 243–288.

Gooderham, N. J., *et al*. (1996). Heterocyclic amines: evaluation of their role in diet associated human cancer. *British Journal of Clinical Pharmacology*, **42**, 91–98.

Gooderham, N. J., *et al*. (1997). Assessing human risk to heterocyclic amines. *Mutation Research*, **376**, 53–60.

Harris, C. C. (1991). Chemical and physical carcinogenesis: advances and perspectives for the 1990s. *Cancer Research*, **51**, 5023s–5044s.

Hesketh, R. (1994). *The Oncogene Handbook* (Academic Press, London).

Jakoby, W. B., *et al*. (1982). *Metabolic Basis of Detoxification: Metabolism of Functional Groups*. 1–375 (Academic Press, New York).

Knudson, A. G. (1993). Antioncogenes and human cancer. *Proceedings of the National Academy of Sciences of the USA*, **90**, 10914–10921.

Lawley, P. D. (1994). Historical origins of current concepts of carcinogenesis. *Advances in Cancer Research*, **65**, 17–111.

Lipmann, M. (1993). Biophysical fibres affecting fibre toxicity. In: Wahrheit, D. B. (ed.), *Fibre Toxicology*. 259–303 (Academic Press, San Diego).

Lynch, A. M., *et al*. (1998). Genetic analysis of PhIP intestinal mutations in Muta^{TM} Mouse. *Mutagenesis*, **13**, 601–605.

Magee, P. N., *et al*. (1976). *N*-Nitroso compounds and related carcinogens. In: Searle, C. E. (ed.), *Chemical Carcinogens*. 491–625 (American Chemical Society, Washington, DC).

Miller, E. C. (1978). Some current perspectives on chemical carcinogenesis in humans and experimental animals: Presential Address. *Cancer Research*, **38**, 1479–1496.

Miller, E. C. and Miller, J. A. (1981). Searches for ultimate chemical carcinogens and their reactions with cellular macromolecules. *Cancer*, **47**, 2327–2345.

Miller, J. A. (1970). Carcinogenesis by chemicals: an overview – GHA Clowes Memorial Lecture. *Cancer Research*, **30**, 559–576.

Morgenthaler, P. M. and Holzhauser, D. (1995). Analysis of mutations induced by 2-amino-1-methyl-6-phenylimidazo[4,5-*b*]pyridine (PhIP) in human lymphoblastoid cells. *Carcinogenesis*, **16**, 713–718.

Okochi, E., *et al*. (1999). Preferential induction of guanine deletion at 5′-GGGA-3′ in the rat mammary glands by 2-amino-1-methyl-6-phenylimidazo[4,5-*b*]pyridine (PhIP). *Carcinogenesis*, **20**, 1933–1938.

Okonogi, H., *et al*. (1997). Agreement of mutational characteristics of heterocyclic amines in lacI of Big Blue mouse with

those in tumour related genes in rodents. *Carcinogenesis*, **18**, 745–748.

Parkinson, A. (1996). Biotransformation of xenobiotics. In: Klaassen, C. D., *et al*. (eds), *Casarett and Doull's Toxicology: the Basic Science of Poisons*. 113–186 (McGraw-Hill, New York).

Pitot, H. C. (1986). Oncogenes and human neoplasia. *Clinical Laboratory Medicine*, **6**, 167–179.

Pitot, H. C. and Dragan, Y. P. (1996). Chemical carcinogenesis. In: Klaassenm, C. D., *et al*. (eds), *Casarett and Doull's Toxicology: the Basic Science of Poisons*. 201–267 (McGraw-Hill, New York).

Porter, T. D. and Coon, M. J. (1991). Multiplicity of isoforms, substrates and catalytic and regulatory mechanisms. *Journal of Biological Chemistry*, **266**, 13469–13472.

Searle, C. E. (1984). *Chemical Carcinogens*, 2nd edn (American Chemical Society, Washington, DC).

Shivapurkar, N., *et al*. (1994). Sequence analysis of k-ras mutations in aberrant crypt foci and colonic tumours induced by azoyxmethane in Fisher-344 rats on high-risk diet. *Carcinogenesis*, **15**, 775–778.

Sims, P., *et al*. (1974). Metabolic activation of benzo[*a*]pyrene proceeds by a diol epoxide. *Nature*, **252**, 326–328.

Sky-Peck, H. H. (1986). Trace metals and neoplasia. *Clinical Physiology and Biochemistry*, **4**, 99–111.

Taper, H. S. (1978). The effect of estradiol-17-phenylpropionate and estradiol benzoate on *N*-nitrosomorpholine-induced liver carcinogenesis in ovariectomised female rats. *Cancer*, **42**, 462–467.

Thilly, W. G. (1990). Mutational spectrometry in animal toxicity testing. *Annual Review of Pharmacology and Toxicology*, **30**, 369–385.

Welch, D. R. and Tomasovic, S. P. (1985). Implications of tumour progression on clinical oncology. *Clinical and Experimental Metastasis*, **3**, 151–188.

Williams, R. T. (1971). *Detoxification Mechanisms*, 2nd edn (Wiley, New York).

Wogan, G. N. (1992). Aflatoxins as risk factors for hepatocellular carcinoma in humans. *Cancer Research*, **52**, 2114s–2118s.

Wogan, G. N., *et al*. (1971). Structure–activity relationships in toxicity and carcinogenicity of aflatoxins and analogues. *Cancer Research*, **31**, 1936–1942.

Yadollahi-Farsani, M., *et al*. (1996). Mutational spectra of the dietary carcinogen 2-amino-1-methyl-6-phenylimidazo [4,5-*b*]pyridine (PhIP) at the Chinese hamster *hprt* locus. *Carcinogenesis*, **17**, 617–624.

FURTHER READING

Balmain, A. and Brown, K. (1988). Oncogene activation in chemical carcinogenesis. *Advances in Cancer Research*, **51**, 147–182.

Dipple, A., *et al*. (1985). Metabolism of chemical carcinogens. *Pharmacology and Therapeutics*, **27**, 265–296.

Guengerich, F. P. (1992). Metabolic activation of carcinogens. *Pharmacology and Therapeutics*, **54**, 17–61.

Hodgson, E. and Levi, P. E. (1994). *Biochemical Toxicology*, 2nd edn (Appleton and Lange, Norwalk, CT).

Levine, A. J. (1993). Tumour suppressor genes. *Annual Review of Biochemistry*, **62**, 623–651.

Timbrell, J. (2001). *Principles of Biochemical Toxicology*, 3rd edn (Taylor and Francis, London).

Vainio, H. (1992). *Mechanisms of Carcinogenesis in Risk Identification*. *IARC Publications*, No. 116 (International Agency for Research on Cancer, Lyon).

Chapter 21

The Formation of DNA Adducts

David H. Phillips
Institute of Cancer Research, Sutton, UK

CONTENTS

DNA DAMAGE BY CARCINOGENS

A property common to many chemical carcinogens is that they, or one or more of their metabolites, are DNA reactive. Cellular responses to DNA damage in mammalian cells include DNA repair, cytotoxicity, apoptosis, mutagenesis and transformation to malignancy. These processes are either fundamental to maintaining the integrity of the cell or they set the cell on a path to mortality or malignancy. Thus the study of reactions between carcinogens and DNA, and the biological consequences of these reactions, is central to understanding the early stages of the carcinogenic process.

The identification of DNA as the genetic material and the solving of its structure occurred about 50 years ago, but it was not immediately appreciated that carcinogens exert their biological effects by damaging it. Although the discovery that carcinogens could form covalent bonds with cellular macromolecules dates from about the same time, for some years afterwards the prevailing hypothesis was that the deletion of key proteins was critical to the carcinogenic process. However, the demonstration in the 1960s that the potency of a series of carcinogens correlated with their ability to bind to DNA *in vivo*, and not with the extent of binding to protein or RNA (Brookes and Lawley, 1964), led to the acceptance of DNA as the critical target in carcinogenesis. Subsequently, the discovery of several classes of genes that control cellular function and maintain cellular integrity and which are commonly mutated in tumours (oncogenes, tumour-suppressor genes and mismatch repair genes) has made it evident that if certain critical genes are modified by carcinogens, the mutations that may ensue from erroneous replication of the damaged gene template will contribute to the transformation of a normal cell into a malignant one. Finally, the types of genetic alterations commonly found in tumour cells – point mutations, deletions, translocations, gene amplifications – can also be induced in cells by treatment with DNA adduct-forming chemicals. Given the long latency of most types of human and experimental cancer, it is not possible to observe directly the biological consequences of DNA adduct formation *in vivo*. Nevertheless, the very strong correlation between this early biochemical event and the subsequent biological manifestation of malignancy in many different studies places the cause-and-effect association beyond reasonable doubt.

Carcinogen-induced DNA damage can take several forms. It can result in breaks in the sugar–phosphate backbone of the molecule, either in one of the two strands of the double helix, or in both. Covalent binding of the carcinogen results in the formation of a chemically altered base (or, occasionally, phosphate group) in DNA that is termed an adduct. As DNA adducts are usually studied by fragmenting the DNA either chemically or enzymatically, the term nucleotide adduct describes a fragment consisting of carcinogen–base–deoxyribose–phosphate, nucleoside adduct consists of carcinogen–base–deoxyribose and a base adduct is the carcinogen-modified base only. Some carcinogens are bifunctional and can give rise to both monoadducts and cross-links in DNA, the latter being either intra- or interstrand cross-links. Many cancer chemotherapeutic agents have this property, and it is widely held

that interstrand cross-links are cytotoxic (accounting for the therapeutic properties of the drugs), whereas the monoadducts and intrastrand cross-links are potentially mutagenic and carcinogenic (Lawley and Phillips, 1996).

Most chemical carcinogens are not biologically active as such, but undergo metabolic activation in mammalian cells to reactive intermediates that react with DNA (see also the chapter *Mechanisms of Chemical Carcinogenesis*). Why do mammalian cells do this? The answer is that it is an aberration of the general mechanisms that cells employ to rid themselves of toxins, generally by making them water soluble. Metabolism of xenobiotic (foreign) compounds is carried out in mammalian cells by broad spectrum oxidative enzymes (Phase I metabolism) that introduce polar groups (e.g. hydroxyl groups) into molecules and render the molecules suitable substrates for conjugation (Phase II metabolism) with one of a variety of hydrophilic groups. The resulting conjugate is substantially more water soluble than the parent compound and thus more readily excreted from the organism. Phase I metabolites are often formed through transient generation of reactive compounds such as epoxides, but are rapidly converted to hydroxyl groups by further enzymatic conversion or by nonenzymatic reaction with water. However, if they are slow to convert, their presence in the cell may lead to their reaction with DNA. Similarly, most Phase II metabolites are water soluble and chemically stable, ideal properties for their efficient elimination from the cell and subsequent excretion from the organism; however, some carcinogens are converted to conjugates that are reactive, and the loss of the conjugated function generates a highly reactive carbocation that reacts with DNA.

There have been two approaches to determining the pathways of activation of carcinogens. In the first, metabolites of the carcinogen are isolated and identified, and their abilities to induce tumours and other genotoxicity endpoints are investigated. In the second approach, DNA adducts are detected and identified, and the pathways by which they are formed are deduced by determining the metabolites that can also give rise to them, and the cofactors necessary for their formation.

An example of a carcinogen that is metabolically activated to a reactive Phase I metabolite is benzo[*a*]-pyrene (BaP), a polycyclic aromatic hydrocarbon (PAH) that is widespread in the environment through its formation during the incomplete combustion of organic material (wood, coal, petrol, tobacco, etc.). Although it is converted to many different metabolites, one cytochrome P450(CYP)-dependent pathway results in metabolic activation (**Figure 1**). The initial epoxide formed is rapidly metabolized further by epoxide hydrolase to a dihydrodiol (called the proximate carcinogen), but this metabolite then undergoes further metabolism to a dihydrodiol epoxide, BPDE (the ultimate carcinogen). This time, the molecule is not a good substrate for epoxide hydrolase and although it is chemically unstable (i.e. it is reactive), it is sufficiently long-lived in mammalian cells to be able to migrate to the nucleus and react with DNA, via formation of a carbocation, to form chemically stable DNA adducts (Phillips, 1983).

An example of a carcinogen activated by Phase II metabolism is tamoxifen (**Figure 2**). In the liver of rats, where it causes hepatocellular carcinoma, it is converted first to α-hydroxytamoxifen (the proximate carcinogen) by a CYP-dependent Phase I step. This metabolite then undergoes Phase II metabolism by sulfotransferase to a sulfate ester conjugate. This compound dissociates to form

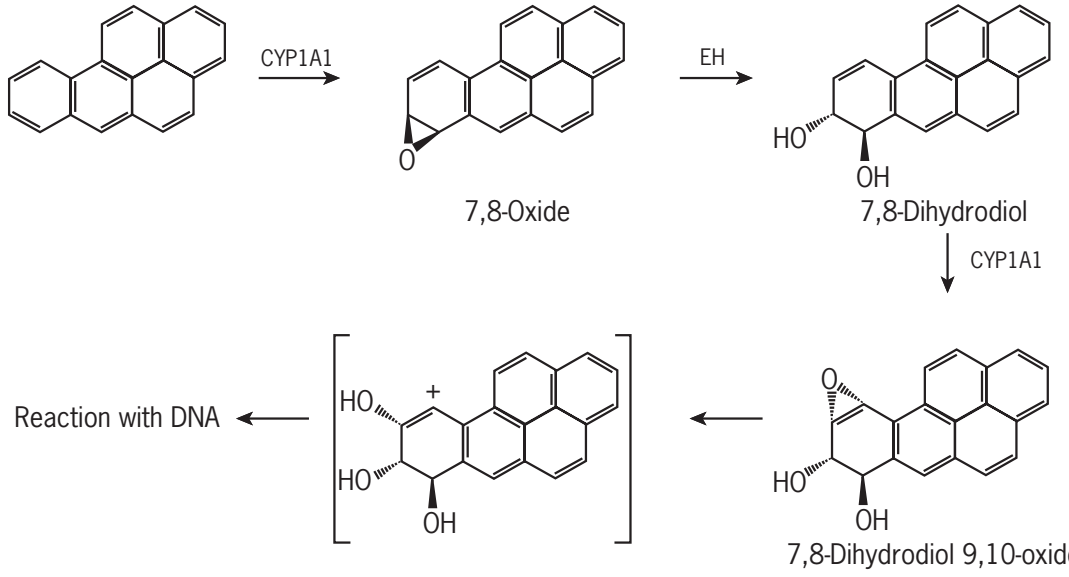


Figure 1 Major pathway of metabolic activation of benzo[*a*]pyrene. CYP1A1, cytochrome P450 1A1; EH, epoxide hydrolase.

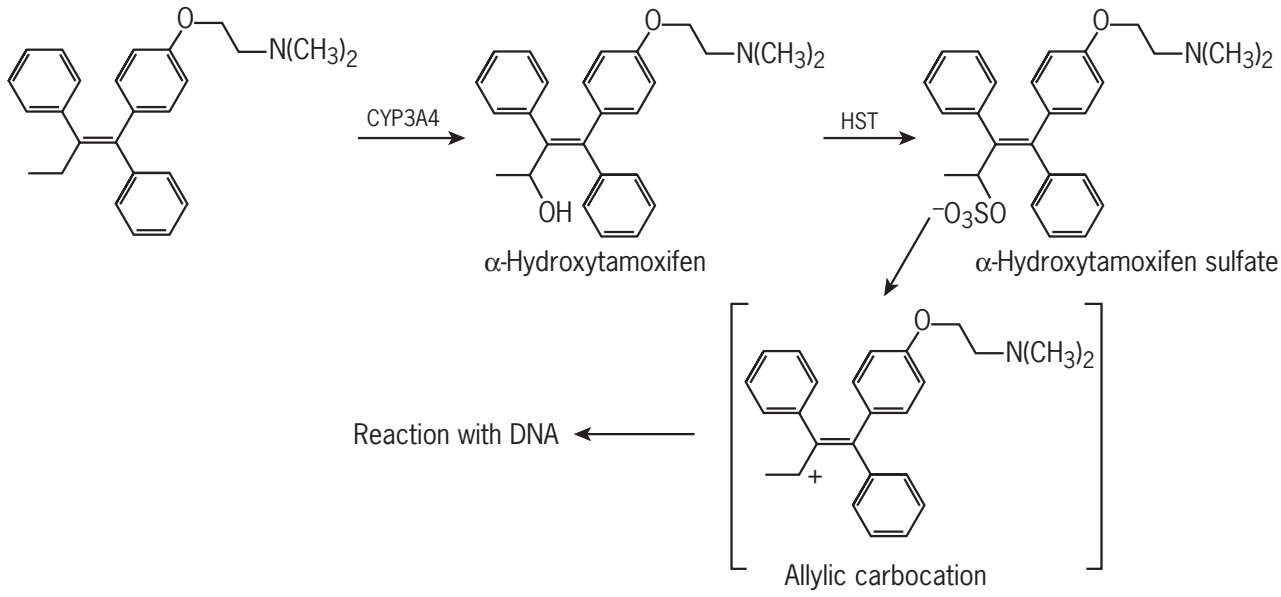


Figure 2 Major pathway of metabolic activation of tamoxifen. CYP3A4, cytochrome P450 3A4; HST, hydroxysteroid sulfotransferase (SULT2A family).

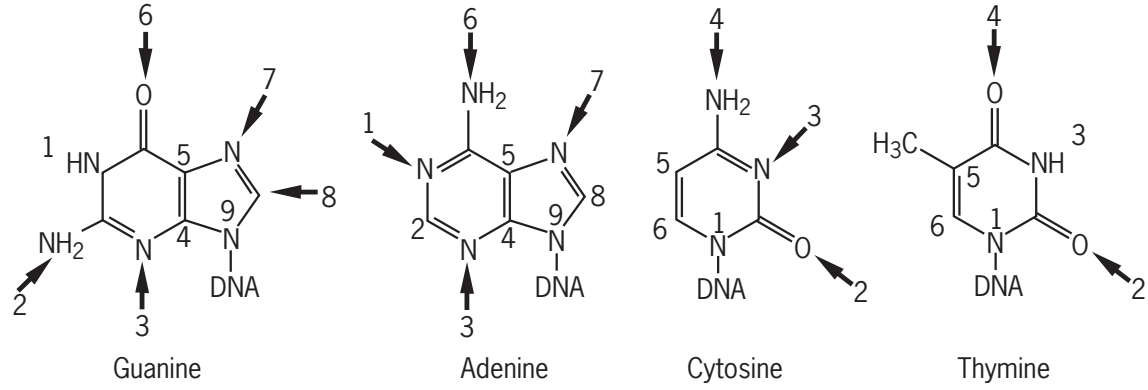


Figure 3 DNA bases and sites at which they are bonded to carcinogens.

a reactive carbocation that reacts with DNA to form DNA adducts (Davis *et al*., 1998).

This chapter describes the formation of DNA adducts by carcinogenic chemicals, with particular emphasis on their detection in human tissues, and the role of such studies in investigating the aetiology of cancer, in monitoring human exposure to environmental carcinogens and in determining cancer risk.

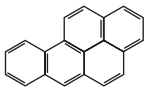
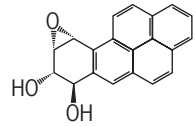
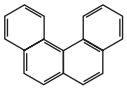
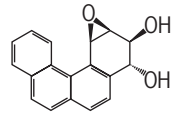
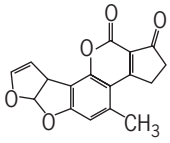
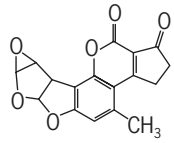
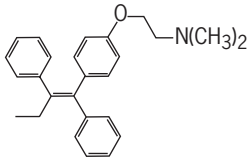
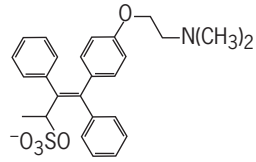
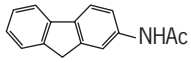
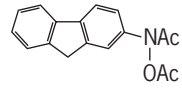
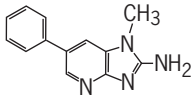
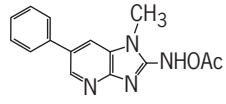
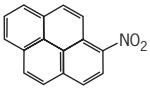
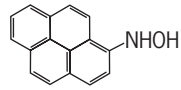
ADDUCTS FORMED BY CHEMICAL CARCINOGENS

Most of the covalent binding of carcinogens to DNA involves the modification of the purine and pyrimidine bases, although some agents also react with the phosphodiester linkages. Guanine is the most commonly and extensively modified base, with interactions occurring at N^2, N-3, O^6, N-7 and C-8 (**Figure 3**). Adducts with adenine are formed at the N-1, N-3, N^6 and N-7 atoms. Pyrimidine adducts are formed at the O^2, N-3, N^4 and C-5 of cytosine and at the O^2 and O^4 of thymine.

Alkylating agents that react by a predominantly S_N2 mechanism show a greater affinity for the ring nitrogens in DNA bases, these being the most nucleophilic sites in DNA (Osborne, 1984). As the S_N1 character of the reaction increases (i.e. via generation of an electrophilic carbocation that is the reactive species), so the proportion of the reaction that occurs at exocyclic groups increases (Lawley, 1984). Thus the ratio of O^6/N-7 alkylation of guanine in DNA is 0.7 for *N*-ethylnitrosurea (ENU), 0.1 for *N*-methylnitrosourea (MNU) and 0.004 for methyl methanesulfonate (MMS) (Lawley, 1984).

Representative examples of active metabolites of carcinogens and of their reaction sites with DNA are shown in **Table 1**. More comprehensive reviews of carcinogen–DNA interactions can be found elsewhere (Osborne, 1984; Cooper and Grover, 1990; Hemminki *et al*., 1994). PAHs, activated by diol epoxide formation, mainly form stable DNA adducts at the exocyclic amino groups of guanine and adenine. The more distorted from planarity the ultimate carcinogen, the greater is the proportional reactivity with adenine relative to guanine (Dipple, 1994). Tamoxifen also reacts at these sites. Nitroaromatic polycyclic hydrocarbons, activated at the nitro group, react with the C-8 position of guanine. Aromatic amines and heterocyclic amines also modify predominantly the C-8 position of

Table 1 Some representative carcinogens, their active metabolites and sites of modification of DNA

Carcinogen	Major active metabolite	Site of modification of DNA
Benzo[*a*]pyrene (BP) 	BP 7,8-diol 9,10-epoxide 	N^2-Guanine, N^6-adenine
Benzo[*c*]phenanthrene (BcPh) 	BcPh 4,3-diol 2,1-epoxide 	N^6-Adenine, N^2-guanine
Aflatoxin B_1 (AfB_1) 	AfB_1 8,9-epoxide 	*N7*-Guanine
Tamoxifen 	α-Hydroxytamoxifen sulfate 	N^2-Guanine, N^6-adenine
2-Acetylaminofluorene (AAF) 	*N*-Acetoxy-AAF 	*C8*-Guanine, N^2-guanine
Vinyl chloride 	Chloroethylene oxide 	3,N^4-Cytosine, 1,N^6-adenine, 3,N^2-guanine
2-Amino-1-methyl-6-phenylimidazo-[4,5-*b*]pyridine (PhIP) 	*N*-Acetoxy-PhIP 	*C8*-Guanine
1-Nitropyrene (1-NP) 	*N*-Hydroxy-1-aminopyrene 	*C8*-Guanine

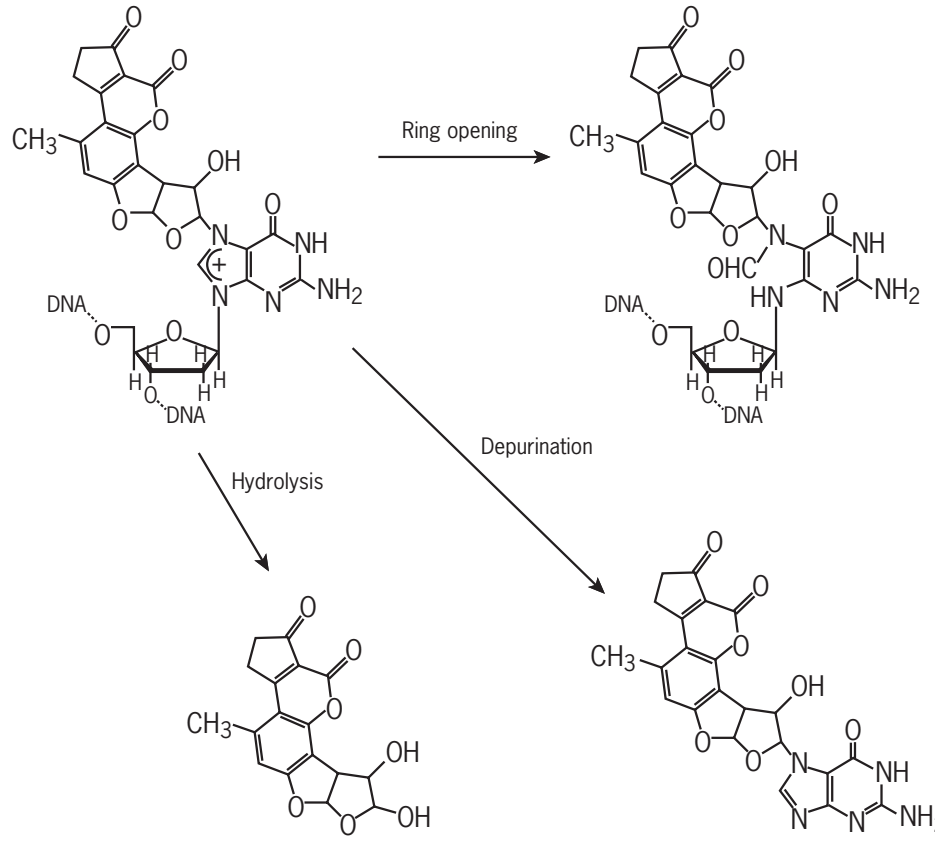
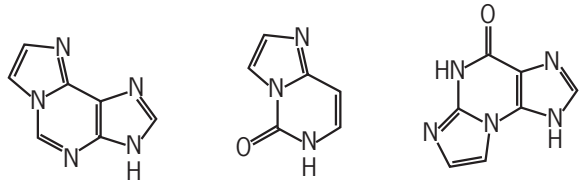


Figure 4 Conversion of the unstable aflatoxin-*N7*-guanine adduct in DNA to stable products.

guanine. Aflatoxin B_1, a mycotoxin, is metabolized to an epoxide that reacts at the N-7 position of guanine. This adduct structure is electronically charged and unstable. It can undergo one of three processes spontaneously: (1) hydrolysis to yield aflatoxin dihydrodiol and unmodified DNA; (2) depurination to yield an aflatoxin-base adduct, leaving an apurinic site in DNA; or (3) imidazole ring-opening to yield a stable adduct (**Figure 4**). Some compounds form cyclic adducts, in which two positions on the same base are modified by the same molecule. An example is vinyl chloride (and other vinyl halides), which forms etheno adducts with cytosine, adenine and guanine (**Figure 5**).

The large variety of DNA sites attacked by carcinogens leads to the question of whether some modifications are more biologically important than others. This is still a matter of debate, with some suggestions that adenine adducts formed by PAHs are more consequential than guanine adducts, even though benzo[*a*]pyrene forms very few of the former. Others have proposed that unstable, depurinating adducts are more important than stable adducts in causing mutations from which tumour initiation proceeds, although this theory has been challenged. Substitution at N-7 of guanine by simple alkylating agents appears to be ineffectual in causing mutations, whereas adducts at the O^6 position are highly promutagenic . On the other hand, aflatoxin B_1 appears to modify only N-7 in guanine, and is among the most potent carcinogens known (Osborne, 1984). Methylation of DNA at the O^6 position, a base-pairing position, will cause guanine to mispair with thymine, thereby causing a G → A transition after a round of replication (Lawley, 1984). The 7-position of guanine, however, is non-pairing; methylation produces little or no alteration to the tertiary structure of DNA, but substitution with a bulky molecule like aflatoxin B_1 would cause considerable distortion (and also leads to ring opening of the purine), which could in itself lead to replication errors by DNA polymerases.



1,N^6-Ethenoadenine 3,N^4-Ethenocytosine N^2,3-Ethenoguanine

Figure 5 Structures of etheno adducts formed by vinyl halides and by products of lipid peroxidation.

There is also some evidence that DNA modification by carcinogens does not occur at random, but is influenced to some extent by DNA sequence (Osborne, 1984).

Additionally, the potential for an adduct to give rise to a mutation may be dependent, to some extent, on its sequence context, giving rise to the concept of hotspots for DNA damage and mutation in some genes. Some of these possibilities are discussed later in this chapter.

METHODS OF DNA ADDUCT DETECTION

The last 25 years have seen the development of a number of sensitive methods for the detection and characterization of DNA adducts in mammalian cells and tissues (Phillips, 1990; Strickland *et al*., 1993; Phillips *et al*., 2000; Poirier *et al*., 2000), the most important of which are reviewed briefly here.

Radiolabelled Compounds

Because only a very small proportion of an applied dose of a carcinogen becomes bound to DNA in the exposed cells or tissue, very sensitive methods of detection are required to study DNA adduct formation *in vivo*. Most of the early work on adducts was done using radiolabelled carcinogens and, although other methods now provide comparable or greater sensitivity, the method still has its uses. With compounds labelled either with ^{3}H or ^{14}C, at a position of the molecule where the isotope will not be lost as a result of metabolism, detection of radioactivity in DNA isolated from the exposed animal or cultured cells is the starting point for the characterization of the DNA binding. Sensitivities of detection of one adduct in 10^{8} nucleotides are achievable with ^{3}H labelling, although ^{14}C labelling is less sensitive because of the much longer half-life of ^{14}C than ^{3}H (Phillips *et al*., 2000).

Limitations to the use of the method are the high costs of synthesizing radiolabelled compounds and the difficulty in doing chronic, multidose, exposure studies due to the hazards of the use of radioactive materials in these circumstances. Furthermore, it is seldom possible to use radioactive test compounds in studies involving human subjects. However, the recent adaptation of accelerator mass spectrometry (AMS) enables isotope ratios to be measured with great sensitivity in biological samples. With this method, the binding of a radiolabelled carcinogen to DNA is detected not by means of its decay (thereby linking the sensitivity to the half-life of the isotope) but by measuring the abundance of the radioisotope relative to that of the natural isotope. For ^{14}C it is possible to detect one part in 10^{15} parts total carbon, and in practice limits of adduct detection of greater than one in 10^{11} nucleotides have been achieved (Phillips *et al*., 2000). Because of this high sensitivity, it has been possible to obtain ethical approval to give minute amounts of a radioactive carcinogen, e.g. the mutagen formed in cooked food, 2-amino-1-methyl-6-phenylimidazo[4,5-*b*]pyridine (PhIP), to human subjects prior to surgery and to detect DNA adducts in the excised tissue. AMS does not give structural information on DNA adducts, and characterization requires chromatographic comparison with synthesized standards. Nevertheless, its ultra-sensitivity provides a means of establishing whether carcinogens thought to be nongenotoxic in their mechanism of action, such as 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD), are truly devoid of DNA binding activity.

^{32}P Postlabelling

The ^{32}P postlabelling method of analysis comprises a procedure for introducing a radiolabel into a DNA adduct after it has formed, by enzymatic phosphorylation of the deoxyribose group of the nucleotide adduct (Phillips, 1997). Chromatographic separation of the labelled adducts followed by detection and quantitation by measuring ^{32}P decay provides a highly sensitive assay, requiring only small (1–10 μg) quantities of DNA. Adducts from different classes of carcinogens with diverse structures can be detected by this method, including PAHs, aromatic amines, heterocyclic amines, unsaturated aldehydes, simple alkylating agents, reactive oxygen species and UV radiation (Beach and Gupta, 1992). It is also able to detect adducts formed from complex mixtures of chemicals, such as tobacco smoke and fossil fuel products. It is sensitive enough to detect adduct levels as low as one per mammalian cell. A limitation of the method is that it does not provide structural information; identification of adducts is reliant on co-chromatography with characterized synthetic standards. Adduct levels may be underestimated if the DNA is not completely digested or if the nucleotide adduct is not efficiently ^{32}P labelled by polynucleotide kinase (Phillips *et al*., 2000).

Immunoassays

Antisera elicited against carcinogen–DNA adducts can be used in immunoassays to detect adducts in human or animal tissues. Antibodies have been raised against a variety of carcinogen-modified DNAs, including those containing adducts of PAHs, aromatic amines, methylating agents, tamoxifen, UV radiation and oxidative damage. Immunoassays are highly sensitive but generally require more DNA for analysis than ^{32}P postlabelling. The assay is relatively inexpensive to perform and can be automated. Various sensitive methods (radioactive, colorimetric, fluorescent and chemiluminescent) have been developed for detecting bound antiserum. When combined with histochemistry it can be used to localize adducts within biological samples. Antibodies raised against a particular adduct can show cross-reactivity with adducts formed by other carcinogens of the same class, which can obscure both the nature of the adducts detected and the levels at which they are present. Nevertheless, this cross-reactivity does at least afford the opportunity to use immunoaffinity chromatography as a means of extracting adducts of a

specific class, such as PAHs, for further analysis by other methods.

Mass spectrometry and other physicochemical methods

Mass spectrometry is the most chemically selective method for DNA adduct detection and it can provide unequivocal identification of the nature of an adduct. This selectivity comes at a price, that of sensitivity, which has limited its application to human DNA adduct studies. However, it is a method in which technological advances are being made rapidly, and it can be predicted with confidence that in the near future it will provide much valuable data on the nature of DNA modifications in human tissues, as well as in experimental studies (Phillips *et al*., 2000; Poirier *et al*., 2000). Most studies to date have used mass spectrometry combined either with gas chromatography (GC-MS) or liquid chromatography (LC-MS). Earlier methods required volatility to be a property of a molecule for its characterization by mass spectrometry, necessitating derivatization of polar species such as adducts, but softer ionization conditions (e.g. fast atom bombardment (FAB)) have overcome this limitation to some extent (Weston, 1993; Phillips and Farmer, 1995). With the advent of newer methods, such as electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI), it will also become possible to investigate the presence of adducts in intact (high molecular mass) biomolecules including DNA and protein.

A number of carcinogens form adducts that are highly fluorescent, enabling their detection by fluorescence spectroscopy. These include adducts formed by PAHs and aflatoxins, cyclic (etheno) adducts and some methylated adducts. They can be analysed in intact DNA, DNA digests or as hydrolysis products. Low-temperature fluorescence spectroscopy (e.g. fluorescence line narrowing spectroscopy (FLNS) (Jankowiak and Small, 1991)) reveals considerable spectral fine structure that is lost at ambient temperature and that can be diagnostic for certain carcinogens, also revealing information about the conformation of the adducts.

Some adducts, notably the oxidative DNA lesion 8-hydroxy-2′-deoxyguanosine, are readily detected by high-performance liquid chromatography (HPLC) combined with electrochemical detection (ECD) (Halliwell and Dizdaroglu, 1992). The method has also been used to detect *N7*-methylguanine in combination with immunoaffinity purification.

Another specific method that can be used to detect adducts by means of their elemental content is atomic absorption spectrometry. An example of its use is the detection of platinum bound to DNA as a result of the treatment of cancer patients with platinum drugs (Weston, 1993). The method gives only the level of the element present, and no structural information.

EVIDENCE FOR THE BIOLOGICAL SIGNIFICANCE OF DNA ADDUCTS

While it is widely assumed that the formation of DNA adducts is an early and obligatory event in the process by which many carcinogens initiate tumours, it is by no means a sufficient event, and the long delay between carcinogen treatment and tumour appearance precludes a direct cause-and-effect demonstration. Nevertheless, it is the case that inhibition of DNA adduct formation will decrease the incidence of tumours formed subsequently, and increasing the adduct levels generally leads to a higher tumour yield.

That chemical modification of DNA can result in the same alterations as observed in mutated genes in tumours was observed with the H-*ras*-1 proto-oncogene transfected into NIH3T3 cells. Prior modification of the plasmid containing the gene resulted in mutations occurring in the DNA after transfection and replication of the host cells, manifested as the appearance of transformed foci. Mutations that activate *ras* genes occur in a few codons in the gene, so correlations between the sites of mutations in such experiments are less informative than in genes where there are many possible sites of DNA damage and mutation which can lead to altered function of the gene product. Such a gene is *p53*, where correlations can be usefully sought between the mutation spectra observed in different human tumours and clues sought to the nature of the initiating agent (Hainaut *et al*., 1998). This has led to strong evidence for the involvement of aflatoxin B_1 in the initiation of liver cancer in regions of high incidence in China. With respect to lung cancer, codons 157, 248 and 273 of the *p53* gene are frequently mutated in these tumours. G → T transversions are much more common in lung cancer of smokers than that of non-smokers, and these types of mutation are characteristic of bulky carcinogens, such as the PAHs which are present in tobacco smoke. When the sites of DNA adduct formation by benzo[*a*]pyrene diol epoxide, the reactive metabolite of benzo[*a*]pyrene, in the *p53* gene in HeLa cells and bronchial epithelial cells were determined, it was found that codons 157, 248 and 273 were preferentially modified (Denissenko *et al*., 1996).

Ultraviolet (UV) radiation causes DNA damage chiefly by dimerization of adjacent pyridines in the same DNA strand. The biological importance of these lesions is illustrated by the fact that sufferers of xerodema pigmentosum (XP), who have a deficiency in nucleotide excision repair mechanisms that remove pyrimidine dimers from DNA, are prone to sunlight-induced skin cancer. Moreover, mutations in the *p53* gene found commonly in such tumours, but rarely in tumours of internal organs, are tandem mutations occurring at pyrimidine pairs (CC → TT transitions), highly suggestive that they arose from UV-induced pyrimidine dimers (Dumaz *et al*., 1993).

Thus, there are examples of genetic changes in tumours that closely match the genetic changes that can be induced

experimentally in cellular DNA by specific genotoxic agents. These tumour-specific mutations in *p53* and the demonstration that chemically modified DNA transforms cells show that the mutations observed in human tumours could have arisen from the formation of carcinogen–DNA adducts *in vivo*. Clonal expansion of the mutated cells and the acquisition of further genetic alterations eventually leads to malignancy (Fearon and Vogelstein, 1990).

Another piece of evidence that strongly links DNA adduct formation to tumour initiation is the demonstration that XPA knockout mice, which are deficient in nucleotide excision repair, are highly sensitive to tumour induction by carcinogens that form stable adducts that would be removed from DNA in normal mice by this repair mechanism (van Steeg *et al*., 1998).

DNA ADDUCT DOSIMETRY

From a number of animal studies, it has been demonstrated that at chronic low doses, there is a linear relationship between the amount of carcinogen administered and the level of DNA adducts that results (Poirier and Beland, 1992). Adducts can be lost from DNA by depurination and by DNA repair, and can be diluted by DNA replication and cell division, so if the exposure is chronic, a steady-state level will be attained; typically this takes about 1 month of dosing. Where exposure is acute or of limited duration, it is clear from animal studies that a small proportion of the adducts persist in tissues for long periods, even in the presence of cell proliferation (e.g. in skin). Thus, the detection of DNA adducts can provide evidence of prior exposure to carcinogens even if the exposure was limited to a single dose.

It is also evident that DNA adducts may be formed in some tissues in the absence of tumour formation, indicating that adduct formation alone may not be sufficient for carcinogenesis (Poirier *et al*., 2000). Other tissue-specific events, such as cell proliferation, are required. Nevertheless, tumours do not form in the absence of adducts, and interventions that reduce adduct formation (e.g. co-administration of enzyme inhibitors or inducers) result in inhibition of carcinogenesis.

An analysis of adduct levels and tumour rates in experimental animals for 27 different chemicals has shown that the adduct levels required to produce a 50% incidence in liver tumours in rats or mice varies between 53 and 5543 adducts per 10^8 nucleotides (Otteneder and Lutz, 1999). This is a narrow range considering the diversity of the chemicals and their interactions with DNA. The analysis also shows that low levels of adducts are detectable in animal bioassays in which tumours were not observed. Thus it is accepted that there are uncertainties concerning the biological significance of low levels of DNA adduct formation, but there are not, as yet, sufficient data with which to define a threshold below which adduct levels can be deemed biologically irrelevant (Phillips *et al*., 2000).

ADDUCTS DETECTED IN HUMAN TISSUES

Studies of humans occupationally exposed to carcinogens have demonstrated the formation of DNA adducts in human tissues (see also the chapter *Occupational Causes of Cancer*). Many of these studies have involved exposure to PAHs with adduct formation being monitored in white blood cells or peripheral lymphocytes; for example, adduct levels are elevated in iron foundry workers, coke oven workers, aluminium plant workers, bus drivers exposed to diesel exhausts and roofers (Phillips, 1996). The same is true of residents of polluted regions of Poland, the Czech Republic and China (Perera, 2000). In some studies, dietary exposure to PAHs appears to be a stronger determinant of adducts in blood cells than does occupational exposure, so it is important that such factors be taken into consideration in interpreting the results of human biomonitoring studies.

Although DNA adducts of some sort have been detected in many studies of human tissues, there are still only a few studies in which the nature of the adduct has been unequivocally identified, and fewer still in which the origin of the DNA binding species can be defined. These limitations derive from the fact that many studies, particularly those employing ^{32}P postlabelling analysis, rely on the co-chromatography of the human adducts with synthetic standard adducts, rather than providing structural identification as such; although this can give reasonably reliable indications of the nature of the adducts formed *in vivo*, it cannot be considered sufficient evidence of identification. Thus those examples listed in **Table 2** all come from studies in which unambiguous physicochemical data (e.g. mass spectra or fluorescent spectra) on the properties of the adducts were obtained. In some cases, the nature of the DNA damaging agents can be deduced from the adduct, but in others, there is uncertainty or ambiguity because there is more than one potential source of the adduct.

As a general rule, several types of adduct have been detected in human DNA, at various levels. Thus 8-hydroxyguanine, originating from oxidative and free radical processes (see section on *Endogenous DNA Adducts*), is typically at levels of one in 10^4–10^5 nucleotides; etheno adducts, from lipid peroxidation or vinyl halides, are formed at between one in 10^7 and one in 10^8 nucleotides; O^6-methylguanine (formed by alkylating agents) is typically at one in 10^6–10^7 nucleotides; and bulky adducts (arising from smoking, pollution and diet) are often found at one in 10^7–10^8 nucleotides. With the exception of smoking-related adducts (see section on *Tobacco Exposure*) these can be regarded as approximate 'background' levels of DNA damage in human tissues due to environmental and/or endogenous DNA-damaging agents. Thus,

Table 2 Identified DNA adducts in human tissues

Adduct	Tissues
2-Amino-1-methyl-6-phenylimidazo[4,5-*b*]pyridine (PhIP, a food mutagen/heterocyclic amine) adducts	Colon
4-Aminobiphenyl (ABP, an aromatic amine) adducts	Bladder
4-(*N*-methyl-*N*-nitrosamino)-1-(3-pyridyl)-1-butanone (NNK, a tobacco-specific nitrosamine) adducts	Lung
Benzo[*a*]pyrene (BaP, a polycyclic aromatic hydrocarbon) adducts	Placenta, lung, leukocytes
Aflatoxin B_1 (AfB_1, a mycotoxin) adducts	Urine
Malondialdehyde-guanine	Liver, leukocytes
Thymine glycol	Placenta
5-Hydroxyethyluracil	Leukocytes
N7-(2-Hydroxyethyl)guanine	Liver
N^2,3-Ethenoguanine	Liver
8-Hydroxyguanine	Many tissues

when measuring adducts in a group of individuals suspected of being highly exposed to carcinogens and/or at elevated risk of cancer, it is important that such measurements be compared with the levels in a suitably selected control group of individuals.

DNA adducts have been detected in cancer patients undergoing chemotherapy (with, for example, platinum-based drugs, cyclophosphamide, melphalan, mitomycin C and the methylating agents dacarbazine and procarbazine) and studies have revealed considerable inter-individual variation in adduct levels among patients receiving similar dosage (Lawley and Phillips, 1996). It remains to be established whether such measurements may provide indications of therapeutic response.

ADDUCTS IN URINE

Because some adducts are chemically unstable and cause depurination of DNA, the excreted products, modified bases, can be detected in the urine (Shuker and Farmer, 1992). Examples of this include aflatoxin–guanine adducts where food contamination by this mycotoxin is endemic (e.g. in areas of Africa and China). Ethylated bases in urine can also serve as markers of exposure to ethylating agents. The method is of less use as a biomarker of exposure to methylating agents because there are instances in which methylated bases (e.g. 3-methylguanine) are ingested in the diet. Another caution is that carcinogen–purine adducts in urine may derive from both DNA and RNA. Urine contains significant numbers of exfoliated bladder epithelial cells, whose DNA can be isolated and analysed for the presence of adducts. An example is the case of a worker acutely exposed to MOCA (4,4′-methylenebis(2-chloroaniline)), whose urine samples yielded significant levels of adducts in the exfoliated urothelial cells for some weeks following the exposure.

TOBACCO EXPOSURE

The relationship between DNA adduct formation and tobacco smoke has been a fruitful area of research in which to explore the validity of the biomarker for a number of reasons (see also the chapter *Tobacco Use and Cancer*). First, a large proportion of the human population is regularly and habitually exposed to tobacco smoke because smoking is an addiction. Second, tobacco smoke contains at least 50 compounds that are known to be carcinogenic, including representatives of several distinct classes of compounds (PAHs, aromatic amines, *N*-nitrosoamines, azaarenes, aldehydes, other organic compounds and inorganic compounds). Most of these compounds are genotoxic carcinogens that form DNA adducts. Third, epidemiological studies have provided clear evidence that tobacco smoking causes not just lung cancer but also cancers in many other organs.

Many studies have compared DNA from smokers, ex-smokers and nonsmokers and have found that the levels of adducts in smokers are elevated in many target tissues: lung, bronchus, larynx, bladder, cervix and oral mucosa (Phillips, 1996). In some of these studies a linear correlation between estimated tobacco smoke exposure and adduct levels has been observed. In tissues of the respiratory tract adduct levels in ex-smokers tend to be intermediate between smokers and nonsmokers, indicating that adducts are removed through DNA repair and/or cell turnover. The half-life of adduct persistence appears to be between 1 and 2 years. This value is longer than would be predicted from adduct persistence studies in animals. A possible explanation is that the lungs of an ex-smoker continue to accumulate adducts after cessation of smoking owing to the continued presence in the lung of smoke and tar deposits.

For some of these studies specific adducts have been detected, but in others a more general measure of DNA damage has been made, namely aromatic/hydrophobic adducts detected by ^{32}P postlabelling, or PAH–DNA

adducts detected by immunoassay. Recent studies have found that when adduct levels are adjusted to take account of the level of tobacco smoke exposure, lung DNA from women smokers is more highly adducted that that of male smokers. This finding is interesting in view of the epidemiological evidence that women are at a 1.5–2-fold greater risk of lung cancer from smoking. It would appear that the adduct analysis provides biochemical, mechanistic evidence to support the morbidity data. The reason for the higher level of adduction in women could be related to levels of expression of metabolizing enzymes (e.g. CYP enzymes) that activate tobacco smoke carcinogens to DNA binding species (Mollerup *et al*., 1999).

Some, but not all, studies have shown elevated levels of lung adducts in cancer cases compared with controls. The relationship between adduct levels in target tissues (e.g. lung) and other tissues (e.g. blood) has been investigated to see whether the latter can serve as a more readily accessible surrogate source of DNA than the former. Results for smoking-related adducts have been inconsistent (Perera, 2000; Poirier *et al*., 2000), perhaps because other sources of exposure to some classes of carcinogens, such as the PAHs, which are also ingested as dietary contaminants, may contribute to the overall level of adducts in the blood but not to the same extent in the lung.

It should be emphasized that by measuring adducts in smokers at the time of cancer diagnosis, investigators are not looking at the biochemical events causal in the initiation of those tumours, as these would have occurred decades earlier. However, because smoking is addictive and habitual for most tobacco users, DNA adducts in tumour-adjacent tissue at the time of tumour manifestation can still serve as a useful biomarker that gives an indication of an individual's probable steady-state level of DNA damage maintained over a long period of time.

ENDOGENOUS DNA ADDUCTS

Thus far, the emphasis has been on exogenous, environmental sources of carcinogens, but there is a large body of evidence that shows that DNA is also subject to modification by a number of endogenous processes arising from normal metabolism, oxidative stress and chronic inflammation (Marnett and Burcham, 1993; Marnett, 2000). **Table 3** shows a list of endogenous DNA adducts that have been detected in human tissues.

The most abundant oxidized base is 7,8-dihydro-8-oxo-2′-deoxyguanosine (8-oxo-dGuo), but at least 30 other base modifications have been characterized in oxidized DNA. Some of these, including 8-oxo-dGuo, can be formed as a result of free radical attack on DNA (for example, by the hydroxyl radical) whereas others appear to a consequence of normal aerobic metabolism. Being the most abundant oxidative lesion, 8-oxo-dGuo is often used as a biomarker for oxidative DNA damage in humans.

Etheno adducts (see **Figure 5**) can be generated in DNA as a result of lipid peroxidation (Marnett and Burcham, 1993; Marnett, 2000), making them adducts of endogenous origin. Etheno bases are removed from DNA by a repair mechanism involving glycosylases, but the rate of removal appears to be slow. These lesions are efficient

Table 3 Endogenous DNA adducts detected in human tissues. (Adapted from Marnett and Burcham, 1993.)

Adduct	Tissue	Adduct levels (per 10^7 bases)	Method
7,8-Dihydro-8-oxo-2′-deoxyguanosine (8-oxo-dGuo)	Leukocytes	12 ± 7	HPLC-ECD
5-Hydroxy-2′-deoxycytidine (5-OH-dCyd)	Leukocytes	10 ± 5	HPLC-ECD
5-Hydroxy-2′-deoxyuridine (5-OH-dUrd)	Leukocytes	7 ± 6	HPLC-ECD
5,6-Dihydroxy-5,6-dihydro-2′-deoxyuridine (dUrdg)	Leukocytes	20 ± 15	HPLC-ECD
7,8-Dihydro-8-oxo-2′-deoxyadenosine (8-oxo-dAdo)	Leukocytes	230	GC-MS
5-(Hydroxymethyl)-2′-deoxyuridine (5-HmdUrd)	White blood cells	2300 ± 480	GC-MS
8-Hydroxy-6-methyl-1,N^2-propano-2′-deoxyguanosine (8-OH-6-Me-PdGuo)	Liver	6	^{32}P postlabelling
8-Hydroxy-1,N^2-propano-2′-deoxyguanosine (8-OH-PdGuo)	Liver	10	^{32}P postlabelling
1,N^6-Etheno-2′-deoxyadenosine (εdAdo)	Liver	0.7 ± 0.4	^{32}P postlabelling
3,N^4-Etheno-2′-deoxycytidine (εdCyd)	Liver	2.8 ± 0.9	^{32}P postlabelling
3-β-D-2′-Deoxyribofuranosylpyrimido [1,2-*a*]purin-10(3*H*)-one (M_1dGuo)	Liver, white blood cells, pancreas, breast	0.1–12	^{32}P postlabelling, MS, immunoassay

premutagenic lesions and, although they can also be formed by exogenous agents such as vinyl halides, they appear to have some credentials for consideration as endogenous origins of mutagenic and carcinogenic processes. Another promutagenic exocyclic base adduct that has been detected in human DNA is M_1dGuo (3-β-D-2′-deoxyribofuranosylpyrimido[1,2-*a*]purin-10(3*H*)-one), a product of the reaction of malondialdehyde (itself a product of lipid peroxidation and prostaglandin biosynthesis) with DNA (Marnett, 1999).

The use of ^{32}P postlabelling analysis has revealed the existence of a large variety of moderately polar lesions in DNA, termed I-compounds. The patterns observed are tissue- and species-specific and show, in experimental animals, significant increases with age. The levels of many of them are also dependent on the diets used to feed the animals. Paradoxically, their levels are actually lower in circumstances which result in tumour formation, such as feeding rats a choline-deficient diet or the administration of enzyme inducers, and they are lower in rat liver tumours than in normal liver. Furthermore, some I-compounds are subject to circadian rhythms.

There are many uncertainties about the role of endogenous DNA adducts in carcinogenesis. Some chemical carcinogens have been shown to cause oxidative damage to DNA as well as forming DNA adducts directly themselves. Other so-called nongenotoxic carcinogens, which do not directly damage DNA, are suspected of increasing levels of oxidative lesions in DNA, but it remains unclear whether this is a mechanism by which they induce tumour formation.

Several of the endogenous DNA adducts are found in human DNA at levels significantly higher than the levels of adducts from exogenous (environmental) carcinogens. Relatively little is known about the relationship between endogenous adduct levels and development of cancer. It might be expected that their levels would be increased in circumstances that ultimately lead to cancer. However, aside from evidence that oxidative DNA damage appears, in some instances, to be higher in tumour tissue than in normal adjacent tissue (e.g. in breast), other evidence for I-compounds suggests that their levels decrease with increasing cancer susceptibility. Until the nature of these compounds is more clearly understood, what their cellular function is remains a matter of speculation.

RISK ASSESSMENT

Elevated levels of adducts in human tissue are clearly biomarkers of exposure. Case-control studies, because they are retrospective, cannot establish causality (Perera, 2000). In order to determine the potential of DNA adducts as biomarkers of risk, it is necessary to examine their presence in subjects prior to the onset of disease, and then investigate at a later date whether those individuals in a cohort with higher adduct levels are the ones that subsequently develop tumours. To date only a few studies have done this. One study monitored aflatoxin B_1 exposure in Chinese men in Shanghai (Qian *et al*., 1994); 18 244 volunteers each gave a single urine sample, which was stored for future analysis. Fifty-five men subsequently developed liver cancer and their urine samples were analysed along with matched control samples. Levels of aflatoxin adducts in the samples were significantly higher in the cases than in the controls, indicating that this parameter was indeed a biomarker of risk. Interestingly, the more classical methods of measuring exposure, based on estimating dietary consumption from food analysis and questionnaires, failed to identify the cases as being more exposed to aflatoxin B_1 than the controls. Thus the use of a DNA adduct biomarker in this instance revealed a link between exposure and risk of cancer that more classical epidemiological methods of exposure assessment failed to demonstrate.

In another such nested case-control study, the predictive value of DNA adducts in white blood cells for lung cancer risk was determined amongst male smokers. Those smokers who subsequently developed lung cancer had significantly higher levels of aromatic DNA adducts, determined by ^{32}P postlabelling, than those who did not (Tang *et al*., 2000).

The limitations to the wider use of these methods for risk assessment are those which are generally encountered in prospective studies, namely that large numbers of subjects need to be recruited to the study and that it can take many years for the cases to appear and for the analyses on cases and matched controls to be conducted. In addition, the preservation and long-term storage of the biological samples of every individual recruited must be accomplished, adding considerably to the costs of the study.

In the field of genetic toxicology testing and regulatory affairs, the assessment of the potential carcinogenic and mutagenic properties of new compounds is of paramount importance. A number of *in vivo* and *in vitro* tests have been developed with different genetic endpoints, including bacterial mutagenicity, mammalian cell mutagenicity, micronucleus formation and aneuploidy (Phillips and Venitt, 1995), but no single test is reliably predictive (see also the chapter *Short Term Testing for Genotoxicity*). Many compounds are positive in some tests but negative in others, leaving uncertainties about the true characteristics of the compound. In these circumstances, investigating the compound for DNA adduct formation in suitable experimental systems may help to clarify the issue (Phillips *et al*., 2000). A good example of such a compound, already mentioned, is tamoxifen, which is a potent liver carcinogen in rats but negative in most regulatory short-term tests for genotoxicity. Despite this, it gives rise to DNA adducts in the target tissue and is considered to be carcinogenic to rats by a genotoxic mechanism.

GENE–ENVIRONMENT INTERACTIONS

From early studies in which short-term explant cultures of human tissue or primary cultures of human cells were treated with carcinogens, it was evident that there was a wide range of DNA adduct formation with samples from different individuals. This variability is also observed in studies of DNA adducts formed *in vivo* in individuals apparently exposed to similar levels of carcinogens and points to the influence of genetic differences in carcinogen metabolism and/or DNA repair.

Detailed discussion of the potential role of polymorphisms in carcinogen-metabolizing enzymes in determining cancer susceptibility is beyond the scope of this chapter, but what can be commented on here are those studies where correlations have been sought between DNA adduct levels and genotype.

Polymorphisms in the *CYP1A1* gene have been extensively studied, although there is still disagreement as to whether these have functional consequences for the activity or inducibility of the enzyme. Two of these are the Msp1 polymorphism (loss/gain of a restriction site) and an exon 7 polymorphism that results in the coding for valine in place of isoleucine (Perera and Weinstein, 2000). One study has found that US smokers with the exon 7 variant allele had higher levels of DNA adducts in their white blood cells than smokers with the normal allele (Perera and Weinstein, 2000). Newborn babies with the *CYP1A1* Msp1 restriction site had higher levels of adducts in placenta and cord blood than those without it (Whyatt *et al*., 2000). Polymorphisms in other *CYP* genes may also be important but have been less well studied.

A number of polymorphisms in Phase II metabolizing enzymes are also of interest. The *null* genotype of *GSTM1* (in which glutathione *S*-transferase μ1 is absent) is associated with a greater risk of lung cancer, although, curiously, the gene does not appear to be expressed in the lung. Individuals with the *GSTM1 null* genotype have higher levels of lung DNA adducts (Perera and Weinstein, 2000). Polymorphisms in other members of the GST family, such as *GSTP1* and *GSTT1*, are also suspected of influencing carcinogen–DNA adduct levels in the lung. Combinations of polymorphisms in Phase I and II enzymes may additionally refine susceptibility to DNA adduct formation and cancer risk from carcinogen (e.g. tobacco smoke) exposure.

In studies of breast cancer risk, there is evidence that possession of the *GSTM1 null* genotype results in higher levels of PAH–DNA adducts in breast tumour tissue (Perera and Weinstein, 2000), and in another study possession of 'slow' alleles of *N*-acetyltransferase 2 (*NAT2*), which detoxifies carcinogenic aromatic amines, resulted in higher levels of aromatic/hydrophobic adducts (Pfau *et al*., 1998).

Currently there is much interest in how combinations of polymorphisms in different Phase I and Phase II metabolizing enzymes may, together, confer greater risk on certain individuals than others of DNA adduct formation and, ultimately, cancer risk. As yet, however, no correlations have been found that can account for the very wide range of interindividual variability in carcinogen–DNA adduct formation among human subjects.

CONCLUSION

The formation of DNA adducts by many carcinogens is causally associated with their mechanism of tumour initiation. Mutations in key genes as a consequence of adduct formation are found in many tumours and the altered proteins that they encode have functions that accord with the phenotypic differences between normal and malignant cells. While the formation of adducts is a necessary, but not sufficient, event for malignant transformation, enhancement of adduct formation will increase tumour formation, and inhibition of adduct formation will restrict it (see also the chapter *Mechanisms of Chemical Carcinogenesis*).

The detection and characterization of DNA adducts in mammalian tissues are research fields that are still undergoing rapid development, and in recent years it has become possible to detect adducts with high levels of sensitivity and/or selectivity. Among the many uses for DNA adduct determination, those currently of greatest interest include monitoring human exposure to environmental carcinogens, investigating the mechanism of activation and tumour initiation of carcinogens, monitoring DNA repair and investigating endogenous DNA damage and oxidative processes. DNA adduct detection has an important role to play in the burgeoning field of molecular epidemiology. It is also a supplementary procedure for assessing new compounds for genotoxic activity and can, potentially, provide valuable data for assessing patient response to cytotoxic chemotherapy.

REFERENCES

Beach, A. C. and Gupta, R. C. (1992). Human biomonitoring and the ^{32}P-postlabelling assay. *Carcinogenesis*, **13**, 1053–1074.

Brookes, P. and Lawley, P. D. (1964). Evidence for the binding of polynuclear aromatic hydrocarbons to the nucleic acids of mouse skin: relation between carcinogenic power of hydrocarbons and their binding to DNA. *Nature*, **202**, 781–784.

Cooper, C. S. and Grover, P. L. (eds) (1990). *Chemical Carcinogenesis and Mutagenesis.* Vols I and II. *Handbook of Experimental Pharmacology*, Vol. 94. (Springer, Berlin).

Davis, W., *et al.* (1998). The metabolic activation of tamoxifen and α-hydroxytamoxifen to DNA binding species in rat hepatocytes proceeds via sulphation. *Carcinogenesis*, **19**, 861–866.

Denissenko, M. F., *et al*. (1996). Preferential formation of benzo[*a*]pyrene adducts at lung cancer mutational hotspots in P53. *Science*, **274**, 430–432.

Dipple, A. (1994). Reactions of polycyclic aromatic hydrocarbons with DNA. In: Hemminki, K., Dipple, A., *et al*. (eds), *DNA Adducts: Identification and Biological Significance*. 107–129 (IARC, Lyon).

Dumaz, N., *et al*. (1993). Specific UV-induced mutation spectrum in the *p53* gene of skin tumors from DNA-repair-deficient xeroderma pigmentosum patients. *Proceedings of the National Academy of Sciences of the USA*, **90**, 10529–10533.

Fearon, E. R. and Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. *Cell*, **61**, 759–767.

Hainaut, P., *et al*. (1998). IARC Database of *p53* gene mutations in human tumors and cell lines: updated compilation, revised formats and new visualisation tools. *Nucleic Acids Research*, **26**, 205–213.

Halliwell, B. and Dizdaroglu, M. (1992). The measurement of oxidative damage to DNA by HPLC and GC/MS techniques. *Free Radical Research Communications*, **16**, 75–87.

Hemminki, K., Dipple, A., Shuker, D. E. G., Kadlubar, F. F., Segerback, D. and Bartsch, H. (eds) (1994). *DNA Adducts: Identification and Biological Significance.* (IARC, Lyon).

Jankowiak, R. and Small, G. J. (1991). Fluorescence line narrowing: a high-resolution window on DNA and protein damage from chemical carcinogens. *Chemical Research in Toxicology*, **4**, 256–269.

Lawley, P. D. (1984). Carcinogenesis by alkylating agents. In: Searle, C. E. (ed.), *Chemical Carcinogens*, 2nd edn. Vol. 1, 325–484 (American Chemical Society, Washington, DC).

Lawley, P. D. and Phillips, D. H. (1996). DNA adducts from chemotherapeutic agents. *Mutation Research*, **355**, 13–40.

Marnett, L. J. (1999). Lipid peroxidation–DNA damage by malondialdehyde. *Mutation Research*, **424**, 83–95.

Marnett, L. J. (2000). Oxyradicals and DNA damage. *Carcinogenesis*, **21**, 361–370.

Marnett, L. J. and Burcham, P. C. (1993). Endogenous DNA adducts: potential and paradox. *Chemical Research Toxicology*, **6**, 771–785.

Mollerup, S., *et al*. (1999). Sex differences in lung CYP1A1 expression and DNA adduct levels among lung cancer patients. *Cancer Research*, **59**, 3317–3320.

Osborne, M. R. (1984). DNA interactions of reactive intermediates derived from carcinogens. In: Searle, C. E. (ed.), *Chemical Carcinogens*, 2nd edn. Vol. 1, 485–524 (American Chemical Society, Washington, DC).

Otteneder, M. and Lutz, W. K. (1999). Correlation of DNA adduct levels with tumor incidence: carcinogenic potency of DNA adducts. *Mutation Research*, **424**, 237–247.

Perera, F. P. (2000). Molecular epidemiology: on the path to prevention? *Journal of the National Cancer Institute*, **92**, 602–612.

Perera, F. P. and Weinstein, I. B. (2000). Molecular epidemiology: recent advances and future directions. *Carcinogenesis*, **21**, 517–524.

Pfau, W., *et al*. (1998). DNA adducts in human breast tissue: association with *N*-acetyltransferase-2 (NAT2) and NAT1 genotypes. *Cancer Epidemiology, Biomarkers and Prevention*, **7**, 1019–1025.

Phillips, D. H. (1983). Fifty years of benzo(*a*)pyrene. *Nature*, **303**, 468–472.

Phillips, D. H. (1990). Modern methods of DNA adduct determination. In: Cooper, C. S. and Grover, P. L. (eds), *Chemical Carcinogenesis and Mutagenesis*, Vol. 1. *Handbook of Experimental Pharmacology*, Vol. 94. 503–546 (Springer, Berlin).

Phillips, D. H. (1996). DNA adducts in human tissues: biomarkers of exposure to carcinogens in tobacco smoke. *Environmental Health Perspectives*, **104**, Supplement 3, 453–458.

Phillips, D. H. (1997). Detection of DNA modifications by the ^{32}P-postlabelling assay. *Mutation Research*, **378**, 1–12.

Phillips, D. H. and Farmer, P. B. (1995). Protein and DNA adducts as biomarkers of exposure to environmental mutagens. In: Phillips, D. H. and Venitt, S. (eds), *Environmental Mutagenesis.* 367–395 (Bios, Oxford).

Phillips, D. H., *et al*. (2000). Methods of DNA adduct determination and their application to testing compounds for genotoxicity. *Environmental Molecular Mutagenesis*, **35**, 222–233.

Phillips, D. H. and Venitt, S. (eds) (1995). *Environmental Mutagenesis*. (Bios, Oxford).

Poirier, M. C. and Beland, F. A. (1992). DNA adduct measurements and tumor incidence during chronic carcinogen exposure in animal models: implications for DNA adduct-based human cancer risk assessment. *Chemical Research in Toxicology*, **5**, 749–755.

Poirier, M. C., *et al*. (2000). Carcinogen macromolecular adducts and their measurement. *Carcinogenesis*, **21**, 353–359.

Qian, G. S., *et al*. (1994). A follow-up study of urinary markers of aflatoxin exposure and liver cancer risk in Shanghai, People's Republic of China. *Cancer Epidemiology, Biomarkers and Prevention*, **3**, 3–10.

Shuker, D. E. G. and Farmer, P. B. (1992). Relevance of urinary DNA adducts as markers of carcinogen exposure. *Chemical Research in Toxicology*, **5**, 450–460.

Strickland, P. T., *et al*. (1993). Methodologies for measuring carcinogen adducts in humans. *Cancer Epidemiology, Biomarkers and Prevention*, **2**, 607–619.

Tang, D., *et al*. (2000). Aromatic DNA adducts as a predictor of lung cancer risk in a case-control study. *Proceedings of the American Association for Cancer Research*, **41**, 221.

van Steeg, H., *et al*. (1998). Use of DNA repair-deficient XPA transgenic mice in short-term carcinogenicity testing. *Toxicology and Pathology*, **26**, 742–749.

Weston, A. (1993). Physical methods for the detection of carcinogen–DNA adducts in humans. *Mutation Research*, **288**, 19–29.

Whyatt, R. M., *et al*. (2000). Association between polycyclic aromatic hydrocarbon–DNA adduct levels in maternal and newborn white blood cells and glutathione *S*-transferase P1 and CYP1A1 polymorphisms. *Cancer Epidemiology, Biomarkers and Prevention*, **9**, 207–212.

FURTHER READING

Cooper, C. S. and Grover, P. L. (eds) (1990). *Chemical Carcinogenesis and Mutagenesis*, Vols I and II. *Handbook of Experimental Pharmacology*, Vol. 94 (Springer, Berlin).

Groopman, J. D. and Skipper, P. L. (1991). *Molecular Dosimetry and Human Cancer: Analytical, Epidemiological and Social Considerations.* (CRC Press, Boca Raton, FL).

Hemminki, K., *et al.* (eds) (1994). *DNA Adducts: Identification and Biological Significance.* (IARC, Lyon).

Phillips, D. H. and Venitt, S. (eds) (1995). *Environmental Mutagenesis.* (Bios, Oxford).

Schulte, P. A. and Perera, F. P. (1993). *Molecular Epidemiology: Principles and Practice.* (Academic Press, San Diego).

Chapter 22

Physical Causes of Cancer

Jeffrey L. Schwartz
University of Washington, Seattle, WA, USA

CONTENTS

INTRODUCTION

Unlike most chemical and viral carcinogens, physical carcinogens act by imparting energy into the biological material. If the energy imparted produces changes in the bonds holding molecules together, this will yield chemical changes and possibly biological effects. Radiation is the primary physical agent to which we are exposed in our environment. Radiation is a ubiquitous component of our environment. We are exposed to ultraviolet (UV) radiation from the sun, γ-rays from cosmic radiation and the decay of isotopes in building materials, air, water and food, and α-particles from radon and radon daughters that seep into our basements. We are exposed to X-rays and ultrasound from medical procedures, and both microwaves and radiofrequency (RF) radiation from various consumer products, including cellular telephones. Power generation leads to a host of different types of radiation exposure including electric and magnetic fields (EMFs). There is no way to avoid completely exposure to these potentially harmful levels of radiation. Understanding the risks associated with exposure can aid in developing appropriate protection standards and approaches to reducing risks.

Definitions

The amount of energy deposited in biological tissue, and therefore the types of changes seen in cells, depend on the nature of the radiation (**Table 1**). Ionizing radiation refers to those types of radiation that produce the ejection of an orbital electron from an atom or molecule and result in the formation of an ion pair. The ionization potential of most molecules in biological materials is 10–15 electronvolts (eV), so in order to be ionizing, the radiation must be able to impart at least that much energy. Ionizing radiation can be either electromagnetic (X-rays and γ-rays) or particulate (neutrons and α-particles). Energy loss varies with the energy of the incoming photon or particle, the charge of the particle and the character (atomic number, electron density) of the absorbing medium. The density of energy deposition along a track length has a profound influence on the subsequent biological effect. The spatial rate of energy loss along a track length is described by the term 'linear energy transfer' (LET). LET is defined as the energy lost (in kiloelectronvolts) per unit track length (in micrometres). X-rays and γ-rays are considered sparsely ionizing, low-LET radiation with ionization or ionization clusters being spaced relatively far apart. The LET for ^{60}Co γ-rays is 0.25 keV μm^{-1}, and that for 250-kV X-rays is about 3.0 keV μm^{-1}. Energetic particles tend to be more densely ionizing and have a higher LET than X-rays and γ-rays, although this is very dependent on the energy of the particle. The LET for a radon-derived α-particle is between 80 and 100 keV μm^{-1}. For energetic particles, the LET varies over the track length as the particle interacts and the energy spectrum changes. For high-energy particles, the density of ionization at the beginning of the track is fairly sparse and the LET is correspondingly low. As the particle loses energy, the density of ionization and the LET increases. At the end of the track, one may see a peak of ionization density (Bragg peak).

Table 1 Examples of ionizing and non-ionizing radiation

Ionizing radiation (> 10–15 eV)	Nonionizing radiation (<10 eV)
X-rays	UV
Gamma-rays	Microwaves
Alpha-particles	RF radiation
Neutrons	Ultrasound
	EMF

The major mode of energy loss for radiation having energy of 1–10 eV (nonionizing radiation) is excitation. Excitation refers to elevation of an electron to a higher ('excited') state. This state is transient, and when the electron returns to a ground state, the energy released can be in the form of visible light (fluorescence, phosphorescence) or chemical change (e.g. pyrimidine dimer formation). UV radiation is the major nonionizing radiation hazard in our environment. For UV radiation, both the wavelength and the fluence determine biological effects. UV radiation is subdivided into three wavelength bands, UVA (313–400 nm), UVB (290–315 nm) and UVC (220–290 nm). UVC is the most potent band for biological effects because DNA absorbs most strongly at the 254-nm wavelength. However, most UVC radiation is quickly absorbed in air. Thus UVB, which is also DNA damaging and potentially carcinogenic, is considered the greater environmental hazard.

For electromagnetic radiation that produces <1 eV, such as microwave and RF radiation, the energy deposited results primarily in molecular vibration and heat. These types of radiation are usually expressed by their frequency. Microwave radiation ranges from 300 MHz to 300 GHz, RF radiation from 300 Hz to 300 MHz and extremely low-frequency (ELF) radiation from 30 to 300 Hz.

Ultrasound consists of high-frequency acoustic waves too fast for us to hear. Human hearing cannot go beyond about 18 000 vibrations per second, or 18 kHz. The effects of ultrasound are usually classified into thermal, direct and cavitation effects. For nonthermal effects of ultrasound, cavitation is considered to be the most important. Under the right conditions, irradiation of a liquid with ultrasound leads to the formation and collapse of gas- and vapour-filled bubbles or cavities in the solution. The collapse of these bubbles can be violent enough to lead to chemical effects.

X- and γ-rays, UV radiation, microwaves and RF radiation are all forms of electromagnetic radiation. EMFs are not electromagnetic radiation, but mixtures of electric and magnetic fields. These fields emanate from electric power lines and all devices that use electricity. Electricity is usually delivered as alternating current. The resulting EMFs are of extremely low frequency and low energy. At the atomic level, weak electric and magnetic fields are too small to produce chemical changes by themselves. However, EMFs might act to modify biological processes by causing small changes in the frequency of events that trigger different signal transduction pathways. In this regard, all aspects of a field, including its frequency, amplitude and pattern, may be important.

Units of Dose and Activity

Radiation exposure is usually expressed either as energy incident on a surface or energy absorbed per gram of tissue. For ionizing radiation (**Table 2**), the gray (Gy) is the Système Internationale (SI) unit of dose that is most often used. It is equal to 1 J kg^{-1}. An older term still in use is the rad, which is equal to 0.01 Gy. For ionizing radiation, the unit of absorbed dose does not take into account the differences in efficiency with which one type of radiation might act. Therefore, to describe exposures to different types of radiation, a dose equivalent is used. The dose equivalent is calculated by multiplying the absorbed dose by a quality factor that takes into account the biological effectiveness of the radiation. The quality factor for ^{60}Co γ-rays is 1. For some energies of α-particles, the quality factor can be as high as 100. The original term used to compare radiation of different qualities was the rem (roentgen equivalent in man), which was equal to the dose in rad times a quality factor. The present SI term is the sievert (Sv), which is equal to the dose in gray times a quality factor.

Table 2 Ionizing radiation units

Type of unit	Unit	Definition
Dose	Gray (Gy)	1 J kg^{-1}
	Rad	1 rad = 0.01 Gy
Dose-equivalent	Sievert (Sv)	Dose $\times$ quality factor
	Rem	1 rem = 0.01 Sv
Activity	Bequerel (Bq)	1 dps[a]
	Curie (Ci)	1 Ci = 3.7×10^4 MBq

[a] Disintegration per second.

Radioactive isotopes decay, producing ionizing radiation at a rate specific for the type and concentration of the isotope. The intensity of the source (activity) is determined by the rate of nuclear transformations per unit time. The SI term used to describe activity is the becquerel (Bq), which is equal to one disintegration per second. The older term sometimes used is the curie (Ci), which was originally defined as the activity associated with 1 g of ^{222}Ra and later was defined as 3.7×10^{10} disintegrations per second; 1 Ci is equal to 3.7×10^4 MBq. Activity is a rate measure. It is not equivalent to dose and, without knowledge of the nature of the decay produced, it is not possible to assign dose or biological effect solely on the basis of activity measures.

Exposure to UV radiation is expressed in joules per square centimetre (J cm^{-2}) . Dose for UV raditation, is in joules. The UV index is another term used to describe UV dose rate. The UV index is a forecast of the amount of skin damaging UV radiation expected to reach the Earth's surface at the time when the sun is highest in the sky. The UV index ranges from 0 at night time to 15–16 in the tropics at high elevations under clear skies. The higher the UV index, the greater is the dose rate of skin-damaging UV radiation.

Incident microwave or RF energy is expressed as energy flux or power density in watts per square centimetre (W cm^{-2}). The specific absorption rate (in watts per

kilogram) is the measure of energy absorbed. Ultrasound exposure is similarly expressed in watts per square centimetre and also by their frequency in hertz. EMFs are described by frequency, amplitude and pattern. Magnetic field strength, a key component of EMF, is expressed in tesla or gauss ($1\,T = 10^4\,G$).

Sources of Exposure

Ionizing radiation exposure varies widely for different populations. In part, exposure levels depend on altitude and latitude. The atmosphere attenuates dose from extraterrestrial sources of radiation; thus higher elevations receive greater doses of γ-rays from cosmic radiation. In the USA, the highest background exposures in a heavily populated area occur in Denver, which experiences about 0.25 mSv per year more than the US average. Higher exposure levels also accompany high-altitude flight. For flights at around 40 000 feet, the average exposure is between 0.005 and 0.01 $mSv\,h^{-1}$.

The Earth's magnetic field acts to deflect the protons from cosmic radiation towards the poles. Thus the polar regions tend to receive greater doses of ionizing radiation. There are also areas in the world where the natural background is higher than the average. These areas include the monozite regions of India and Brazil. Background radiation levels in these areas may exceed 10 times the world average.

The average yearly dose of ionizing radiation in the USA is 3.6 mSv (360 mrem). To put this number in perspective, for a typical dental X-ray, the equivalent dose is 0.01–0.02 mSv and a chest X-ray is about 0.1 mSv. Cosmic rays and γ-rays from naturally occurring isotopes account for about 1 mSv per year (**Table 3**). Radon accounts for about 2 mSv per year. That is not to imply that we are exposed to large levels of radon. Instead, the α-radiation associated with radon decay has a high quality factor of around 20, hence the dose equivalent is high. As for anthropogenic radiation exposure, medical exposures account for about 0.5 mSv per year, while exposures from consumer products make up about 0.1 mSv per year. There are also exposures associated with a variety of other activities, including the use of coal-fired power plants, tobacco products and, as mentioned above, travel at high altitudes.

Table 3 Sources of ionizing radiation exposure (in the USA)

Source	Amount (mSv per year)
Natural	3.0
Radon	2.0
Cosmic	0.3
Terrestrial	0.3
Internal	0.4
Artificial	0.6
Medical	0.5
Consumer	0.1
Other	
Work, nuclear power generation, fallout	<0.01
Total	3.6

There are substantial daily and seasonal variations in UV exposure levels. The atmosphere also acts to attenuate dose of UV radiation from the Sun. Most UV radiation is absorbed by the ozone layer or reflected back into space. Atmospheric ozone, which is principally located in the stratosphere, is a strong absorber of UVB radiation. Ozone is formed when molecular oxygen is split by UV radiation. The singlet oxygen atom that results can combine with other molecules of molecular oxygen to form ozone. Ozone production is normally balanced by its destruction through similar photochemical events. In addition, ozone is depleted by the chlorine contained in chlorofluorocarbons (CFCs), a family of commonly used industrial compounds. In part due to these CFCs, the amount of ozone screening the Earth has decreased by an average of 3% per year over the last decade, and a polar ozone hole over the South Polar region has developed. The reduction in ozone has resulted in a corresponding increase in UV levels.

There are a myriad of consumer products that generate radiation. Microwave ovens and cellular telephones provide a common source of microwave and RF radiation. There are a number of medical and consumer devices that use ultrasound. All electrical devices also generate EMFs. Transmission power lines generate both strong electric fields and strong magnetic fields. Distribution lines generate weak electric fields but can generate strong magnetic fields. Exposure levels to EMFs around the home are in the range 0.1–2.5 mG. For homes near power lines, these levels may be as high as 5–10 mG. Immediately under the power line, magnetic field levels of 60–100 mG may be found. The Earth's magnetic field has a strength of about 500 mG, although, unlike EMFs, it is not changing directions 50–60 times per second and is therefore not comparable to electrically generated EMFs. EMFs also show intermittent spikes in frequency (transients) and harmonics, which are multiples of the standard frequency.

GENOTOXICITY

Cancer is a genetic disease and therefore, at some point, carcinogen exposure must result in genetic alterations (see also the chapter *Mechanisms of Chemical Carcinogenesis*). These can be produced directly through an interaction of the radiation with DNA, indirectly, through some intermediate molecule that in turn damages DNA, or radiation may produce epigenetic alterations in cell growth or metabolism that ultimately lead to the transformed phenotype.

Ionizing Radiation

Ionizing radiation induces a variety of different types of DNA alterations, including nucleotide-base alterations and breaks in the sugar–phosphate backbone (Schwartz, 1995). Normally, cells handle these base alterations and single-strand breaks very well using base excision repair processes. Base excision repair is an error-free process that removes damaged bases and then takes advantage of an intact complementary strand to fill in missing information. Defects in excision repair can lead to mutagen sensitivity and cancer susceptibility (see also the section *The Molecular Basis of Cell and Tissue Organisation*).

For X-rays, about one in 10–20 breaks span both strands of the DNA molecule. This ratio of double-to single-strand breaks increases with increasing LET. DNA double-strand breaks can form directly as a result of radiation exposure, or can develop during excision repair due to the formation of overlapping gaps. DNA double-strand breaks are repaired by homologous and nonhomologous recombination processes. These are inherently error-prone and can lead to large deletions and rearrangements. DNA double-strand breaks have been shown to be the primary lethal and mutagenic lesion induced by ionizing radiation. (See also the chapter *Genomic Instability and DNA Repair*.)

Most of the genetic alterations seen in irradiated cells are large deletions and rearrangements. This is mostly a reflection of the nonhomologous repair of DNA double-strand breaks. Because radiation involves the loss of so much genetic material, it is a relatively weak mutagen. Most of the genetic changes lead to cytotoxicity. The nature of the alterations induced by radiation is very dependent on gene locus. Some loci can tolerate the loss of large amounts of genetic material. These loci would tend to be more sensitive to mutation induction and would show primarily large deletion. Other loci, because of their proximity to essential genes, are less tolerant of large deletion formation. These loci will be less sensitive to mutation induction and show far smaller types of genetic changes.

The dose response for mutation induction by X-rays and γ-rays has a quadratic portion to it, suggesting that most mutations develop from the interaction of two breaks (**Figure 1**). Fractionating the dose or reducing the dose rate will in general lead to a reduction in the effectiveness of the radiation by allowing time for repair of these sublethal lesions. As LET increases, both the frequency and the shape of the dose response change. High-LET radiation such as α-radiation is more effective at inducing mutations and their dose response is linear, suggesting that a single α-track can produce enough damage for lesion interaction. Fractionation of high-LET radiation usually has little effect on radiation-induced genotoxicity. The lack of sparing at low dose rates is presumed to be due to the inability of cells to repair damage induced by high-LET radiation.

Recent studies have demonstrated that ionizing radiation can induce genetic instability in exposed cells.

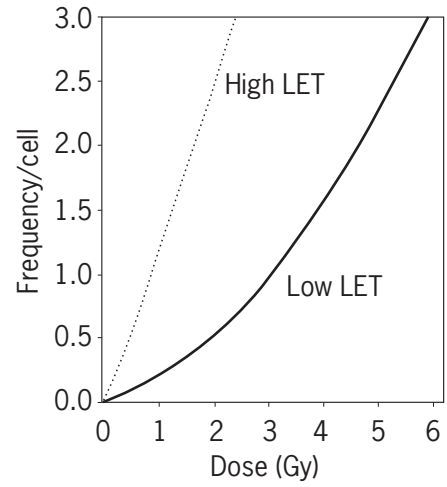


Figure 1 Examples of low- versus high-LET dose responses.

Delayed appearance of elevated frequencies of gene and chromosome mutations has been reported for cells exposed both *in vitro* and *in vivo*. The target for the effect is unknown. Instability is a frequent byproduct of radiation exposure. Rates as high 3% Gy^{-1} have been reported. (See also the chapter *Genomic Instability and DNA Repair*.)

Exposure to ionizing radiation can also induce alterations in gene expression as part of a generalized stress response. This stress response can affect DNA repair, cell toxicity and cell growth. As such, they may lead to more permanent genetic effects. While DNA damage can initiate these stress responses, other targets of radiation have also been identified. Alterations in redox levels or energy levels or specific molecular alterations in key protein molecules may trigger these responses. For example, direct activation of TGF-β by ionizing radiation can influence cell growth and apoptosis.

UV Radiation

The excitations induced by UV exposure will also produce DNA base changes and breaks in the sugar–phosphate backbone. The most important alterations appear to be cyclobutanepyrimidine dimers and pyrimidine-(6,4)-pyrimidone photoproducts. Replication of DNA containing these lesions leads to C to T transitions at dipyrimidine sites. Nucleotide excision repair processes normally repair these base alterations efficiently and without error. Like ionizing radiation, UV radiation can also induce alterations in gene expression as part of a generalized stress response. (See also the chapter *Genomic Instability and DNA Repair*.)

RF and Microwave Radiation

RF and microwave radiation do not produce enough energy to induce chemical change. They can lead to heat, however, and excessive heat may lead to more permanent

genotoxic changes (Moulder *et al*., 1999). Heat can also lead to the induction of a stress response, the spectrum of which overlaps with that seen with ionizing and UV radiation. It remains to be shown whether subthermal levels of microwave or RF exposure such as most of us experience can lead to permanent genetic effects. Most studies report no evidence that prolonged exposures to subthermal levels of microwave or RF radiation lead to genetic or epigenetic changes. The few positive studies suggest that the effects, if any, are small.

Ultrasound

Cavitation is considered to be the primary nonthermal effect of ultrasound that might lead to biological effects (Miller, 1987). Cavitation can lead to cell membrane damage by mechanical forces produced by cavitation. Ultimately, this can cause cell lysis. Cavitation can also lead to the formation of free radicals that can further damage cells. Alterations in cell morphology, membrane transport and cell growth following ultrasound exposure have all been reported, but the effects are small and transient. Ultrasound-induced free radicals have the potential to damage DNA. However, most reports on DNA damage or mutation following ultrasound exposure are negative. The few positive studies that demonstrate DNA strand breakage and gene mutations are for very high intensity exposures. Almost all of these studies are *in vitro* studies. It is less clear whether similar effects will be seen *in vivo*.

EMF

In vitro studies with EMF exposure have focused on endpoints commonly associated with cancer, including cell proliferation, signal transduction alterations and differentiation inhibition as well as on more traditional DNA damage endpoints (Moulder *et al*., 1999). Effects on gene expression, cell growth and signal transduction were reported for magnetic flux densities $>100\,\mu T$ or internal electric field strengths $>1\,mV\,m^{-1}$, but no consistent alterations have been reported for lower intensities. There are no reported direct effects on DNA. Disruption of normal circadian rhythm of melatonin was one postulated mechanism of EMF action. Studies on EMF effects suggest some effect on melatonin, but the significance of this observation for carcinogenesis is not obvious.

CARCINOGENESIS

It has clearly been established that ionizing and UV radiation are both carcinogens. In contrast, the epidemiological evidence for an association between cancer and microwave and RF radiation, ultrasound or EMF exposure is weak and inconsistent.

Ionizing Radiation

Information on cancer induction by human populations comes from epidemiological studies of exposed human populations (National Research Council, 1990, 1999; Schwartz, 1995). For ionizing radiation, the largest single group of exposed individuals are those Japanese exposed to the atomic bombs at Hiroshima and Nagasaki. About 280 000 individuals survived the immediate effects and about 80 000 have been followed for long-term effects. Other exposed populations include early radiation workers such as radium dial painters, uranium miners and populations exposed for therapeutic reasons. These and other smaller groups of exposed individuals serve as the human database for estimating risk for developing cancer following ionizing radiation exposure. (See also the chapter *Identifying Cancer Causes through Epidemiology*.)

At the molecular, cell and tissue levels, cancers induced by ionizing radiation are so far indistinguishable from those that occur spontaneously. Unlike with many chemical mutagens and UV radiation, there are no specific mutations associated with radiation exposure that are also seen in radiation-induced tumours. Recent studies suggest that the initiating event in radiation carcinogenesis may be the induction of genetic instability. (See also the chapter *Molecular Epidemiology of Cancer*.)

The types of tumours seen following radiation exposure are usually the same as seen spontaneously. The effect of radiation on cancer induction is usually inferred from the increase in frequency over background. Radiation-induced tumours appear in almost all tissues of the body, but sensitivity varies greatly for specific tissues and organs (**Table 4**). In general, the thyroid, female breast and certain blood forming organs are considered the most sensitive to induction by ionizing radiation in humans, while kidney, bone, skin, brain and salivary gland are considered the least sensitive. The lymphatic system, lung, colon, liver, and pancreas are among those tissues with moderate sensitivity. The differential sensitivity probably reflects a complex number of factors; it is not simply a reflection of spontaneous frequencies.

The types of cancers observed are also related to the nature of exposure. Many different types of cancer have been seen in the atomic bomb-exposed individuals who in general experienced total body irradiation. Early radiation workers exposed occupationally showed increases in skin

Table 4 Relative ionizing radiation sensitivity of tissues

Sensitive	Moderate	Resistant
Thyroid	Lung	Kidney
Breast	Colon	Bone
Blood-forming tissues	Liver	Skin
	Pancreas	Salivary gland
	Lymphatic system	Brain

cancer and leukaemias, reflecting the nature of their exposure. Many worked with radium, a bone-seeking isotope. Others studied relatively low-energy X-rays where skin would be the primary tissue exposed. Radium dial painters were a group (mostly women) who painted watch dials with a radioactive solution. In the course of their work, they ingested radium. The radium deposited in bone, resulting primarily in higher incidences of bone cancers. Uranium mineworkers inhaled radon gas. The α-radiation exposure resulting from that inhaled radon led to a higher incidence of lung cancer. Thorotrast was a contrast agent used in the late 1920s and early 1930s. It is a colloidal preparation of thorium-232 dioxide that tends to concentrate in the liver. Thorium is an α-emitter. Excess liver cancers and leukaemias were seen in these patients. During the late 1930s, patients in the UK with ankylosing spondylitis were treated with radiation to reduce bone pain. An excess of leukaemias has been reported in these individuals. In the past, radiation has been used to reduce thymus size in children and also for the treatment of tinea capitus. The critical organ exposed was the thyroid. These exposed children later showed excesses in the incidence of thyroid cancers.

The latent period for cancer induction following radiation varies with tumour type, radiation type, dose and dose rate. Leukaemias have the shortest latent periods (mean, 5–10 years), which no doubt accounts for their being seen with most radiation exposures. Solid tumours show latencies of between 20 and 30 years.

In general, the dose response for the induction of tumours by radiation follows a sigmoidal response. At low doses, there is little induction. A steep increase and then saturation or even decrease in tumour frequency at high doses follow. The dose response will vary depending on tissue type, dose rate and tumour latency time. As mentioned above, most of the available data on human carcinogenesis come from individuals exposed to relatively high doses of radiation. No human data exist for proven carcinogenic effects of radiation below 0.1 Sv. The dose response at low levels of exposure remains unknown. One predicts response in this region by extrapolation. Animal studies have suggested that the shape of this crucial portion of the curve may vary greatly from tissue to tissue and animal to animal. At present, linear extrapolation of high-dose effects is used to estimate risk for low doses of radiation. This is considered the most conservative approach. However, there is growing evidence for non-linearity at low doses (Pollycove, 1998). Some investigators have suggested that cells can adapt to low dose exposure and thus linear extrapolation overestimates risk at low doses.

As the LET of the radiation increases, the carcinogenic effects seen also increase in severity up to about $100\,\mathrm{keV\,\mu m^{-1}}$ (**Figure 2**). As LET increases beyond this point, the effect on carcinogenesis usually declines. This reflects 'overkill,' where individual ionization events are not less effective, but the increasing amounts of the energy released per ionization event are wasted.

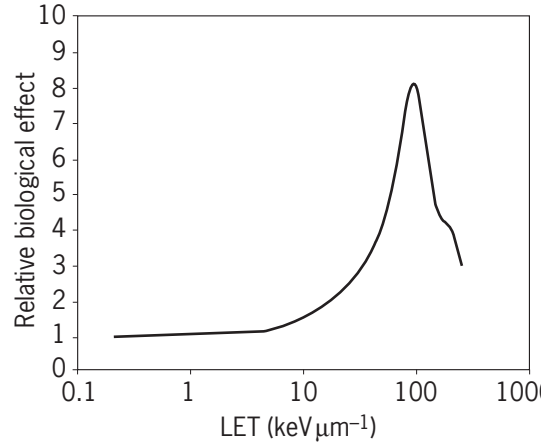


Figure 2 Variation of biological effect with LET. RBE is defined as the ratio of doses (standard: experimental) that yield the same biological effect.

As with other biological effects, protracting the dose over long periods of time reduces cancer incidence, and the effectiveness of fractionated or low-dose-rate irradiation in causing tumours is different for low- and high-LET radiation. Most studies on fractionated low-LET radiation have resulted in a reduction of tumour incidence for a given total dose. Presumably, fractionation of the dose allows for time to repair sublethal and subcarcinogenic damage. In contrast, fractionation of high-LET radiation or low-dose exposure usually has little effect on radiation-induced tumour formation. For some types of radiation, such as fission-spectrum neutrons, reducing the dose rate actually leads to increased transformation and more tumours. This is known as the inverse dose-rate effect. The largest inverse dose-rate effects are seen for fission-spectrum neutrons, with monoenergetic neutrons yielding reduced enhancements, and charged particles having LETs $> 120\,\mathrm{keV\,\mu m^{-1}}$ producing little or no enhancement. The inverse dose-rate effect is most prominent at low doses (< 20 cGy) and low dose rates ($< 0.5\,\mathrm{cGy\,min^{-1}}$).

Sensitivity to tumour induction varies for different species and strains of animals, is different for males and females and also shows interindividual variability (Schwartz, 1995). This variability suggests that the initial damage, which is presumed to be the same for a given dose and type of radiation, is subject to a number of host factors that modify response. Presumably these include repair capability, presence of endogenous viruses, cell proliferation status, endocrine levels, immune competence, age of irradiation and factors associated with genetic susceptibility. There are a number of genetic syndromes that show radiation sensitivity and cancer susceptibility (Murnane and Kapp, 1993).

Radiation risk is defined as the increase in the number of cancer deaths over that expected for an unirradiated population. It is expressed in units per person exposed

per gray of radiation. Estimates based on linear extrapolation of the atomic-bomb data and on other more limited data from pooled results of various partial body exposures give total cancer mortality risks for a general population exposed to whole body radiation of $(1–4) \times 10^{-2}$ per person-Gy.

UV Radiation

There is extensive epidemiological evidence supporting the direct role that sunlight plays in human skin cancer (van der Leun, 1984; de Gruijl, 1999; Green *et al*., 1999). (See also the section *Systemic Oncology*.) Patients who develop skin cancer generally have decreased melanin pigmentation. Melanin normally acts to protect skin from UV radiation by absorbing it. People with light complexions and who sunburn easily have a higher incidence of tumours. Basal cell carcinomas, the most common skin cancers in Caucasians, are found primarily on sun-exposed areas such as the head and neck where a dose–response relationship exists. There is even stronger evidence for the role of sunlight in causing squamous cell carcinomas of the skin. Although both types of tumours are more prevalent in geographical areas of high sun light exposure, there is a much greater increase in squamous cell carcinomas with decreasing latitude and increasing sunlight exposure. A reasonable correlation also exists between sunlight exposure and melanoma, but the relationship is not as clear as with basal and squamous cell carcinomas. Unlike basal and squamous cell carcinomas, melanomas occur most frequently on the upper back in males and lower extremities in females. Melanoma incidence does not follow a pattern of increased risk with cumulative UV exposure, whereas the incidences of basal and squamous cell carcinomas do increase with cumulative exposure. The risk of skin cancer is highly dependent on UV wavelength. (See also the chapter *Identifying Cancer Causes through Epidemiology*.)

Microwave and RF Radiation

Studies on microwave and RF radiation and cancer have included analyses of a wide variety of different populations, including radar laboratory workers, foreign service workers, military personnel and electrical workers (National Research Council, 1997; Moulder *et al*., 1999; National Institute of Environmental Health Sciences, 1999). In addition, there have been a number of animal studies that have looked at the effects of exposure. For many of the epidemiological studies, there is no precise information on dose. Often occupation is used to define exposure groups. In general these studies either find no effect of nonthermal levels of microwave or RF radiation on tumour induction, or show weak and inconsistent results. There is no evidence for any single type of tumour being induced by exposure, and no strong evidence for any dose–response relationship.

There have also been a number of long-term exposure studies with mice and rats. Too many of the animal studies suffer from poorly controlled exposures where heat stress is a component of exposure. Still, taken together, these studies do not support an effect of nonthermal levels of RF exposure on cancer induction. There is also mixed evidence for a tumour-promoting effect of RF exposure. Some groups have reported that RF exposure shortens tumour latency time and increases tumour frequency in carcinogen-exposed mice and in lymphoma-prone transgenic mice. Others see no effect on the promotion of spontaneous or chemically induced tumours. Hence it is not possible to conclude that nonthermal exposures to microwave or RF radiation have any effect on cancer induction or progression. (See also the chapter *Non-Genotoxic Causes of Cancer*.)

Ultrasound

There is no reported evidence for cancer induction by ultrasonic exposure. There are some suggestions that ultrasound might encourage neoplastic growth and promote metastases, but no strong evidence for either effect (Miller, 1987). (See also the chapter *Non-Genotoxic Causes of Cancer*.)

EMFs

The evidence for carcinogenic effects of EMFs is weak. The strongest evidence comes from epidemiological studies that observe associations between EMF exposure and leukaemia. The initial study by Wertheimer and Leeper (1979) suggested a causal association between risk of childhood leukaemia and exposure to magnetic fields. Wire code classifications were used to estimate exposure. Subsequent studies on this association have come to mixed conclusions. In general there appears to be at most a small increased risk of childhood leukaemia associated with EMF exposure (National Research Council, 1997; Moulder *et al*., 1999; National Institute of Environmental Health Sciences, 1999). There were also initial suggestions of an association between EMF exposure and brain cancers and lymphomas in children, but subsequent studies have not borne out this risk. Epidemiological reports of adult cancer induction by EMF were based primarily on occupational exposures. As with childhood leukaemia, there were mixed results for an association between EMF exposure and chronic lymphocytic leukaemia, suggesting at most a weak risk. The evidence for increased risk based on residential studies of adults is even weaker. (See also the chapter *Non-Genotoxic Causes of Cancer*.)

There have been numerous animal carcinogenicity studies of EMFs. The animal studies are all negative. There are no data to support any association between leukaemia and EMF exposure. Similarly, no evidence for any significantly increased frequency of tumours or changes in tumour latency or size was observed. The lack of any experimental data to back up the epidemiological data suggests that

there might be other factors that explain the increased risk of cancer. However, none has yet been identified.

RISK MODELS AND PROTECTION STANDARDS

There are a number of ways to model risk. Absolute risk refers to the number of cancers induced over spontaneous levels. Relative risk is a multiplicative increase over spontaneous. As cancer latency periods are generally long, time-dependent relative risk models are also used. With ionizing radiation, there are a number of other dependent variables that are considered in developing risk models. These include dose and $(\text{dose})^2$, age at exposure, time since exposure and gender.

Risk estimate development for ionizing radiation is handled by the BEIR (Biological Effects of Ionizing Radiation) Committee in the USA and by UNSCEAR (United Nations Scientific Committee on the Effects of Atomic Radiation) for the United Nations. The NCRP (National Council for Radiation Protection) in the USA and the ICRP (International Commission of Radiation Protection) for the United Nations use BEIR/UNSCEAR information to develop appropriate radiation protection standards. Total cancer mortality risks for a general population exposed to whole-body radiation is currently based on linear extrapolation of the atomic-bomb data and on other more limited data from pooled results of various partial body exposures. Risks are 5×10^{-2} per person-Sv for low dose and low dose rates and 10×10^{-2} per person-Sv for high dose and high dose rates. The values for the working (adult) population are about 80% of those for the general (adult and child) population. Exposures to the general public are limited to 1 mSv per year while occupational standards are 50 mSv per year. Embryo exposure is limited to 0.5 mSv per month.

There are no corresponding limits for UV or ultrasound exposures, although protective measures are usually required for working with UV- or ultrasonic-producing equipment. For microwave radiation, the recommended exposure limit is $10\,\text{mW}\,\text{cm}^{-2}$. It is based on thermal effects of microwaves. RF limits are one-tenth of the microwave standard. The International Commission on Non-Ionizing Radiation Protection has set up guidelines for EMF exposure. Magnetic field exposures are limited to 1 G for the general public (10 G for short-term exposure) and 5 G (50 G for short-term exposure) for the occupationally exposed.

SUMMARY AND CONCLUSIONS

As mentioned at the beginning of this chapter, radiation is a ubiquitous component of our environment. There is no way to avoid exposure to radiation. Furthermore, as our technology advances, our exposure to various forms of radiation increases in both amount and complexity.

It has clearly been established that ionizing and UV radiation are both carcinogens. They represent the primary physical carcinogens in our environment and most efforts at reducing cancer risks are appropriately focused on these agents. There remain questions as to mechanisms of carcinogenesis for ionizing radiation, and in particular the effects of low-level exposures. Most ongoing studies in this area are attempting to address these questions. The answers may have a profound effect on radiation protection standards.

In contrast to ionizing and UV radiation, the evidence for an association between cancer and microwave and RF radiation, ultrasound or EMF exposure is weak and inconsistent. The levels of exposure to these types of radiation are increasing, as are the numbers of individuals exposed. There is also some uncertainty as to potential interactions between different types of radiation. Hence there are likely to be continuing investigations into the effects of these different types of radiation exposures.

ACKNOWLEDGEMENTS

This work was supported by the National Cancer Institute (NCI) and the National Aeronautics and Space Administration (NASA) through NIH grant CA-73931 and by a grant from the Department of Energy (DE-FC03-00ER62908).

REFERENCES

de Gruijl, F. R. (1999). Skin cancer and solar UV radiation. *European Journal of Cancer*, **35**, 2003–2009.

Green, A., *et al*. (1999). Sun exposure, skin cancers and related skin conditions. *Journal of Epidemiology*, **9**, S7–S13.

Miller, D. L. (1987). A review of the ultrasonic bioeffects of microsonation, gas-body activation, and related cavitation-like phenomena. *Ultrasound in Medicine and Biology*, **13**, 443–470.

Moulder, J. E., *et al*. (1999). Cell phones and cancer: what is the evidence for a connection? *Radiation Research*, **151**, 513–531.

Murnane, J. P. and Kapp, L. N. (1993). A critical look at the association of human genetic syndromes with sensitivity to ionizing radiation. *Seminars in Cancer and Biology*, **4**, 93–104.

National Institute of Environmental Health Sciences (1999). *Health Effects from Exposure to Power-line Frequency Electric and Magnetic Fields*. NIH Publication No. 99-4493. (NIEHS, Washington, DC).

National Research Council (1990). Committee on the Biological Effects of Ionizing Radiation (BEIR V). *Possible Health*

Effects of Exposure to Low Levels of Ionizing Radiation: BEIR V. (National Academy Press, Washington, DC).

National Research Council (1997). Committee on the Possible Effects of Electromagnetic Fields on Biologic Systems. *Effects of Exposure to Residential Electric and Magnetic Fields.* (National Academy Press, Washington, DC).

National Research Council (1999). Committee on Health Risks of Exposure to Radon (BEIR VI). *Health Effects of Exposure to Radon: BEIR VI*. (National Academy Press, Washington, DC).

Pollycove, M. (1998). Nonlinearity of radiation health effects. *Environmental Health Perspectives*, **106**, Supplement 1, 363–368.

Schwartz, J. L. (1995). Radiation carcinogenesis and the development of radiation injury. In: Arcos, J. C. and Arcos, M. F. (eds). *Chemical Carcinogenesis: Modulation and Combination Effects*. 473–508 (Bikrkhäuser, Springer International, Boston).

van der Leun, J. C. (1984). UV-carcinogenesis. *Photochemistry Photobiology*, **39**, 861–868.

Wertheimer, N. and Leeper, E. (1979). Electrical wiring configurations and childhood cancer. *American Journal of Epidemiology*, **109**, 273–284.

FURTHER READING

Hall, E. J. (2000). *Radiobiology for the Radiobiologist*, 5th edn. (Lippincott, Williams & Wilkins, Philadelphia).

Websites

www.epa.gov
www.lowdose.org
www.osha.gov

Chapter 23

Non-genotoxic Causes of Cancer

Susan K. Murphy and Randy L. Jirtle
Duke University Medical Center, Durham, NC, USA

CONTENTS

INTRODUCTION

Genotoxic agents are characterized by their ability to induce DNA changes that alter the coding information within the primary nucleotide sequence either directly or through a reactive metabolite. These nucleotide mutations are widely accepted as the molecular events responsible for inactivating both alleles of a tumour-suppressor gene according to the Knudson 'two-hit model' of carcinogenesis (Knudson, 1971). However, agents that induce epigenetic alterations to the genome, or DNA structural changes in the absence of nucleotide sequence alterations, have increasingly been demonstrated to also play a fundamental role in cancer formation. These non-genotoxic agents induce heritable changes in the DNA that can disrupt gene regulatory regions. Epigenetic changes are potentially even more potent than genetic mutations in causing cancer since the resulting chromatin structural alterations can exert regional influences, thereby disrupting the normal transcriptional activity of multiple genes. This chapter will focus on the mechanisms of chromatin structure modulation and the potential role of epigenetics in the formation and treatment of cancer.

EPIGENETIC CHARACTERISTICS OF DNA

The double helical structure of DNA is remarkable in its simplicity, yet correct gene function requires not only that the base sequence is faithfully transcribed, but also that expression is both spatially and temporally regulated in a tightly controlled manner. The process of development in multicellular organisms depends on the differential repression or activation of particular genes in a cell type-specific manner, and this programming information must be maintained throughout the life of the individual. This is referred to as 'cellular memory,' and is controlled by epigenetic mechanisms (Riggs and Porter, 1996). Epigenetic regulation is therefore fundamentally important to the control of gene expression. Two major classes of epigenetic modification are instrumental in determining this complex level of gene regulation: histone acetylation and cytosine methylation.

Chromatin Structure

The human genome consists of approximately three billion base pairs of DNA divided amongst 22 pairs of autosomes and two sex chromosomes. If left in its native form, the genome would be several metres in length. The problem of packaging and organizing the DNA within the nucleus is overcome by the coordinated compaction of the genome by specialized DNA-binding proteins, including the histone proteins. 'Chromatin' refers to the combination of DNA together with these proteins and is organized into two major types of subchromosomal domains. Heterochromatin is tightly compacted and less transcriptionally active while euchromatin is less compacted and more likely to be transcribed. The chromatin structure also varies with the phases of the cell cycle. It is relaxed and transcriptionally active during interphase whereas it is condensed and inactive during mitosis. (See chapter on *Regulation of the Cell Cycle*.)

Histones are a family of proteins that provide the scaffolding for chromatin assembly and, consistent with this function, are among the most highly conserved proteins throughout evolution. The histones share the same basic structure consisting of a globular head and a positively charged, nonglobular tail. Histones H2A, H2B, H3 and H4

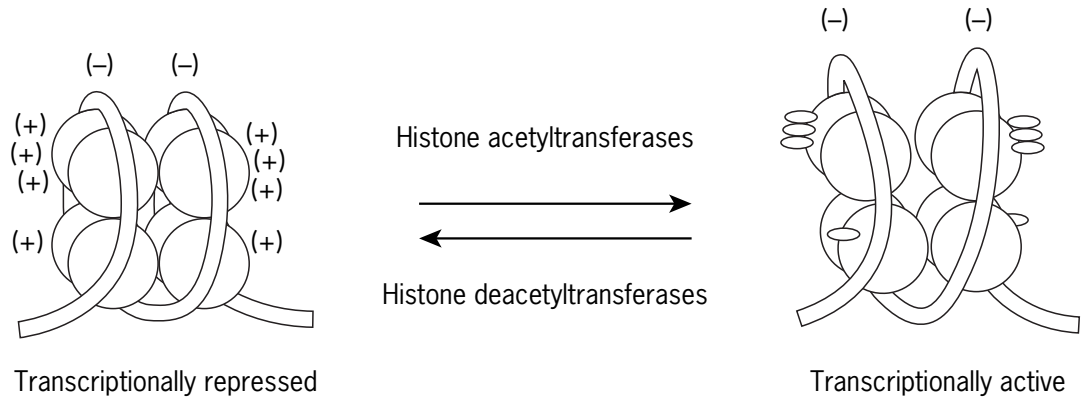


Figure 1 Histone acetylation and gene transcription. The nucleosomal core consists of eight histone proteins: two copies each of histones H2A, H2B, H3 and H4 (spheres). Approximately 150 base pairs of DNA are wrapped around each histone octamer. In transcriptionally inactive chromatin, the histones lack acetyl groups and are tightly compacted with the DNA. In transcriptionally active chromatin, histones H3 and H4 are acetylated (ovals) on their N-terminal tails. This post-translational modification neutralizes the inherent positive charge of the histone tails that in turn is thought to decrease the affinity of the histones for the negatively- charged DNA phosphate backbone. This contributes to an open chromatin structure in which the DNA is more loosely wrapped around the histones, and is receptive to interaction with transcription factors.

together form a heterooctamer core around which approximately 150 base pairs of DNA are wrapped to form discrete units called nucleosomes (**Figure 1**). Internucleosomal segments are bound by histone H1 and link each nucleosome to the next. The nucleosomes are further bundled into higher order structures to form compacted and organized chromatin. Histone–DNA interactions are modulated in part by reversible acetylation of the ε-amino groups of the histones lysine side chains. The lysine residues carry positive charges thought to contribute to the histone's affinity for negatively charged DNA. Acetylation of these lysine side chains neutralizes the positive charge thereby decreasing the ability of histones to interact with DNA, resulting in a more open chromatin configuration and increased transcription (**Figure 1**). Acetylation of histones by histone acetyltransferase (HAT) is reversed by the action of histone deacetylase (HDAC). The latter enzyme reduces transcriptional activity by promoting chromatin condensation and inhibiting access of the transcription machinery to the DNA. Histone modifications at distinct lysine residues may also allow for the recruitment of proteins capable of regulating transcription in a gene-specific manner.

CpG Methylation

Genome structure is also influenced by cytosine methylation, the only known biological base modification of DNA. Indeed, methylated cytosine has been referred to as the 'fifth base' because of its ability to convey heritable information. The extent of cytosine methylation plays a major role in the organisation of the genomic DNA. Densely methylated DNA is located in condensed

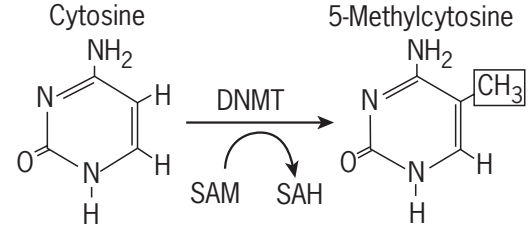


Figure 2 Methylation of cytosine. Cytosine methylation is performed by DNMT (DNA methyltransferase), and occurs predominantly at CpG dinucleotide pairs, although cytosines in the context CpNpG are sometimes also methylated. The DNMT enzyme transfers a methyl group from *S*-adenosylmethionine (SAM) to the 5-carbon position of cytosine (boxed CH_3), forming 5-methylcytosine and leaving *S*-adenosylhomocysteine (SAH).

heterochromatin while sparsely methylated DNA is located in the more relaxed euchromatin. Methylation is the best studied epigenetic modification that occurs in cancer, and will be the main focus of the remainder of this chapter.

The genomic methylation patterns in gametic DNA are erased by a genome-wide demethylation shortly after fertilization. This is followed by *de novo* re-establishment of the methylation patterns after implantation. The mechanisms involved and the proteins guiding the erasure and resetting of the methylation patterns in the genome during embryogenesis are not yet fully understood. Several known DNA methyltransferases (DNMTs) in mammals, including DNMT1, DNMT3a and DNMT3b, catalyse the transfer of a methyl group from *S*-adenosylmethionine (SAM) to the 5-carbon position of cytosine (**Figure 2**).

This reaction occurs most commonly when cytosine is in a CpG dinucleotide sequence. All three enzymes are capable of performing both *de novo* and maintenance methylation (see below), but DNMT3a and DNMT3b appear to act by transferring methyl groups to previously unmodified CpG sequences, a process that occurs predominantly during embryogenesis.

In contrast, DNMT1 is thought to be the major maintenance DNA methyltransferase enzyme. The palindromic nature of the CpG target of DNMT1 is a key feature in the heritability of the DNA methylation profile during cell replication (**Figure 3**). During DNA synthesis, the nascent daughter strand is methylated by the methyltransferase enzyme positioned at the replication fork. DNMT1 preferentially recognizes the hemimethylated state of the two strands, and copies the methylation pattern of the parent strand on to the daughter strand. Thus, somatic cell methylation profiles represent epigenetic information that is faithfully replicated from one generation to the next.

The incidence of CpG dinucleotides in the genome is about 5–10-fold less than the approximate 6% frequency expected from the random distribution of the 16 possible dinucleotide combinations. The mammalian genome is thought to have progressively lost many of the methylated cytosines within CpG dinucleotide pairs during the course of evolution. This most likely occurred by endogenous deamination of methylated cytosines to form thymine (**Figure 4**). Although 60–80% of the CpGs within the mammalian genome are methylated, cytosines that reside in CpG islands, ranging from 200 to 4000 bases in length, are protected from methylation. They are therefore resistant to mutation by methyl group-driven deamination. These CpG islands are located near the promoter regions of approximately 50–60% of the genes within the genome, including all housekeeping genes. It is not unusual to find CpG islands also located within the 5′ coding region of genes or even in downstream introns.

The biological function of CpG methylation is not clearly understood. Methylation of genomic DNA is a modification employed by numerous species, including bacteria, plants and mammals; however, methylation is not detectable in yeast, *Drosophila* or *Caenorhabditis elegans*. Methylation serves as a host defence mechanism in prokaryotes to protect against the introduction of foreign DNA. DNA methylation in eukaryotes is proposed to similarly serve in host defence by protecting cells from transcription and transposition of endogenous retroviral sequences, and/or to reduce transcriptional 'noise' from very large genomes (Baylin *et al*., 1998; Robertson and Wolffe, 2000). The CpG islands associated with many of the genes located on the inactive X chromosome in females are also extensively methylated whereas the same alleles on the active X chromosome are unmethylated. Furthermore, the silenced allele of imprinted genes usually exhibits parent of origin-dependent dense methylation of at least one associated CpG island.

Hypermethylation of promoter region CpG dinucleotides is strongly correlated with the transcriptional silencing of genes. The causal relationship between cytosine methylation and gene silencing in mammals is supported by studies both *in vitro* and *in vivo*. Transfection experiments using reporter constructs with a methylated promoter

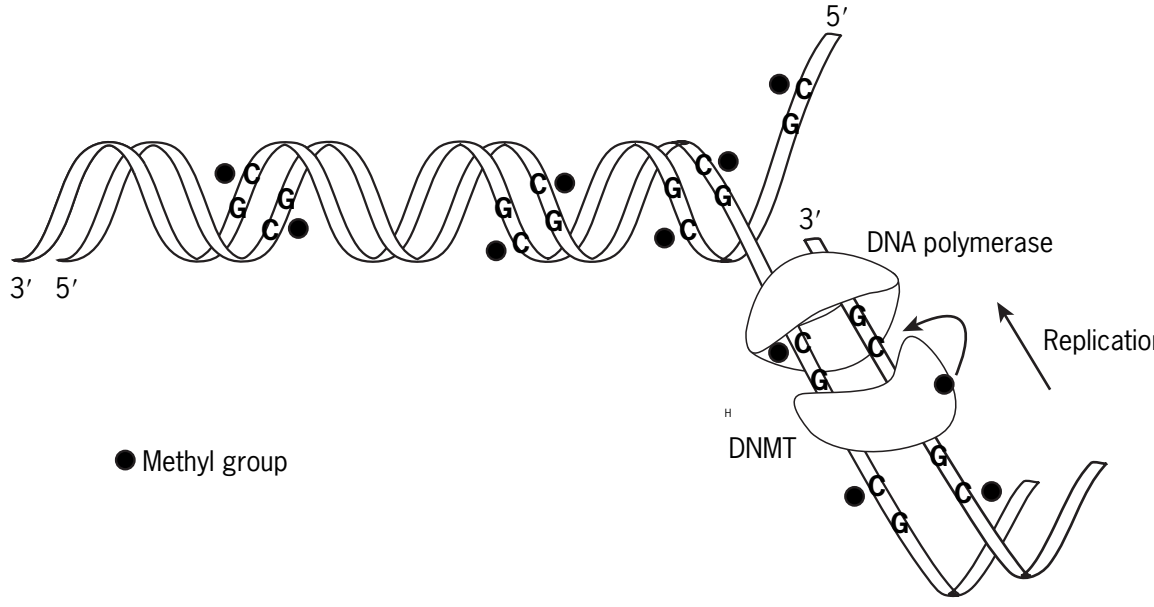


Figure 3 Replication and methylation of hemimethylated DNA. During DNA replication, the newly synthesized daughter strand is methylated by DNMT (DNA methyltransferase) within 1 min of synthesis. DNMT1 recognizes the hemimethylated state of the parent/daughter strand duplex and copies the methylation pattern of the parent strand CpG dinucleotide on to the daughter strand.

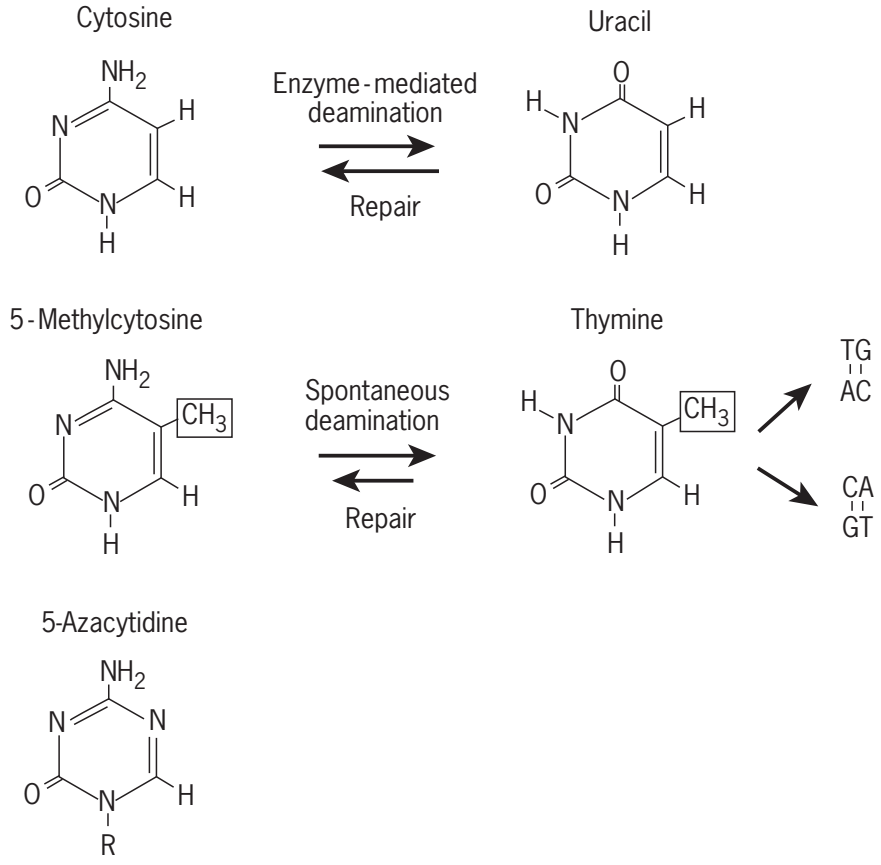


Figure 4 Structures of cytosine, uracil, 5-methylcytosine, thymine and 5-azacytidine (5-azaC). Cytosine is subject to DNMT mediated deamination to form uracil. DNA containing uracil is repaired by uracil DNA glycosylase. Spontaneous deamination of 5-methylcytosine produces thymine and repair of the resultant T–G mismatch is inefficient. Subsequent replication results in the production of either a TpG or CpA dinucleotide pair, depending on the strand affected by the original deamination event. When incorporated into either DNA or RNA, 5-azaC forms a covalent bond to DNMT (DNA methyltransferase) that results in sequestration of the enzyme followed by a progressive depletion of DNA methylation.

region show reduced transcription relative to that for constructs with unmethylated promoters. The DNMT inhibitor 5-azacytidine (5-azaC) (**Figure 4**) causes transcriptional reactivation of endogenous genes with hypermethylated promoters. Furthermore, homozygous disruption of *Dnmt1* in mice results in a three fold reduction in genomic 5-methylcytosine content, embryonic death (Li *et al*., 1992), and biallelic expression of imprinted genes (Li *et al*., 1993). These results emphasize the importance of cytosine methylation in gene regulation and embryogenesis. The deviations from normal methylation patterns frequently observed in cancer cells further suggest that epigenetic perturbations are mechanistically involved in oncogenesis.

MOLECULAR MODULATORS OF EPIGENETIC SIGNALS

Several protein complexes play fundamental roles in transcriptional control by recognizing and binding to methylated DNA. They can block the formation of transcription initiation complexes at methylated promoters by steric hindrance. The methyl CpG proteins (MeCP1 and MeCP2) are also capable of directing transcriptional repression through coordinated chromatin alterations. MeCP1 is a ubiquitously expressed protein that binds densely methylated DNA in a sequence-independent manner. It forms a complex with the methyl binding domain protein (MBD2), which further associates with HDACs. The HDACs are thought to contribute to the MeCP1-mediated transcriptional repression by causing chromatin condensation (Hendrich and Bird, 2000).

MeCP2 recruits several other proteins including a repressor of transcription, Sin3A, and an HDAC complex that induces chromatin condensation by deacetylating histone proteins (**Figure 5**) (Jones *et al*., 1998; Nan *et al*., 1998). Unlike MeCP1, MeCP2 can bind to DNA via a single methylated CpG dinucleotide; however, the efficiency of repression coordinated by the MeCP2 complex is dependent on the density of methylated CpGs (Magdinier *et al*., 2000). Dense promoter methylation may in itself

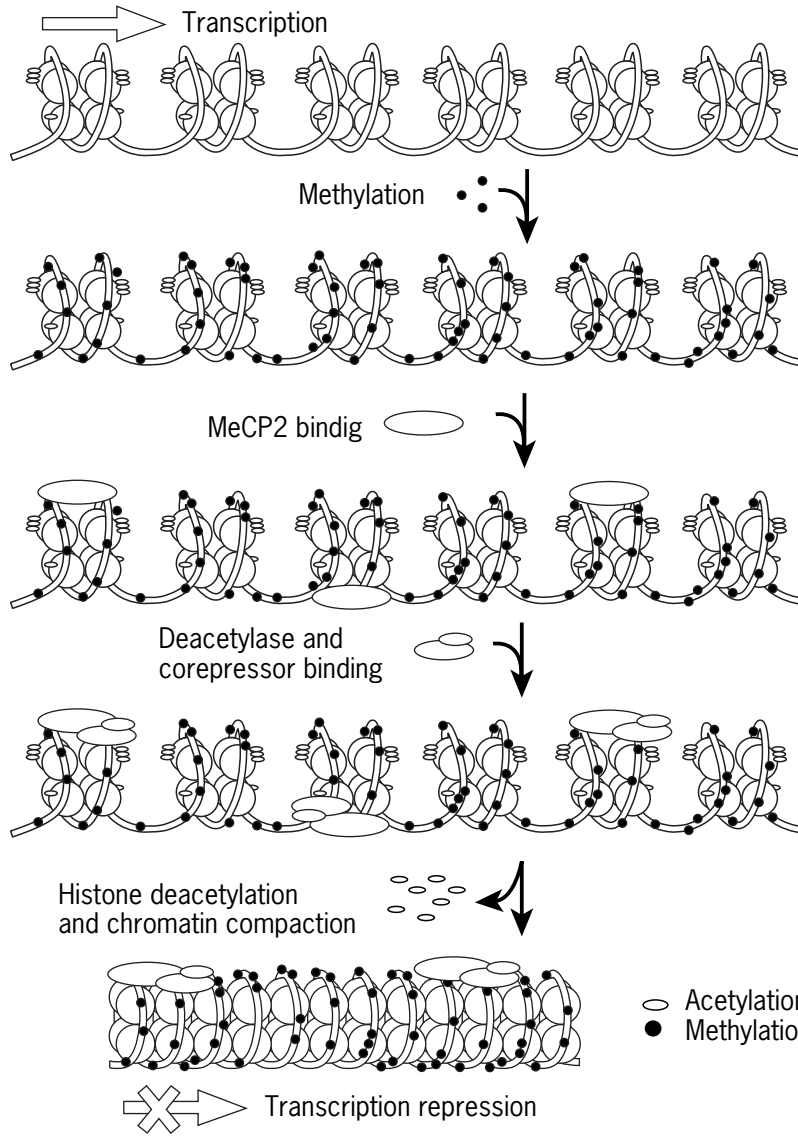


Figure 5 Model for the epigenetic inactivation of DNA transcription. A schematic representation of transcriptionally active open chromatin with acetylated histones is shown at the top. Cytosine methylation of the normally unmethylated promoter region is followed by recognition and binding by MeCP2 (methyl CpG protein). MeCP2 recruits HDAC (histone deacetylase) and other corepressor proteins that function to locally deacetylate histones followed by condensation of the chromatin; the result is promoter silencing. (Adapted from Jones and Laird, 1999.)

further inhibit the access of HAT activity to the hypoacetylated histones, preventing conversion to a more active chromatin configuration.

Recent studies supporting a link between DNA methylation and chromatin structure formation have shown that DNMT1 also interacts with HDAC2 to repress transcription (Fuks *et al*., 2000; Robertson and Wolffe, 2000; Robertson *et al*., 2000; Rountree *et al*., 2000). During DNA replication, unincorporated histones arrive at the replication fork in an acetylated state. DNMT1 and HDAC2 are both positioned at the replication foci when heterochromatin is replicated during late S phase. This association with HDAC2 could provide a mechanism whereby histones are deacetylated during assembly to facilitate the faithful structural reproduction of condensed chromatin from one cell to the next during cell division (Rountree *et al*., 2000).

Cytosine methylation and histone acetylation are proposed to act synergistically in the progressive silencing of genes. One model that accounts for tumour-suppressor gene silencing by epigenetic mechanisms invokes abnormal hypermethylation of the promoter CpG island followed by recruitment of methyl-DNA-binding proteins, including complexes such as MeCP2 that direct HDACs to the area of hypermethylation. Histone deacetylation, probably accompanied by other chromatin remodeling

events, can then cause local chromatin condensation and subsequent transcriptional repression (**Figure 5**) (Jones and Laird, 1999; Rice and Futscher, 2000). Thus, epigenetic events are poised to function mechanistically in carcinogenesis by silencing tumour-suppressor genes.

EPIGENETIC CHARACTERISTICS OF TUMOUR CELLS

There is widespread documentation of significant changes in the epigenome of cancer cells. These changes include an overall level of genomic hypomethylation coupled with gene-specific hypermethylation. Hypomethylation events are more generalized, and could lead to the activation of endogenous retroviral elements and any associated proto-oncogenes. However, gene-specific hypomethylation is unlikely to play a major role in oncogenesis since promoter CpG islands are normally unmethylated, with the notable exceptions of imprinted alleles and genes on the inactive X chromosome. In contrast, promoter-specific hypermethylation can lead to the silencing of tumour-suppressor genes.

Inherent Mutability of Methylated Cytosine

The methylation of CpG dinucleotides creates mutagenic susceptibility targets that can subsequently undergo endogenous deamination to form TpG (CpA on the opposite strand) dinucleotide pairs (**Figure 4**). In contrast to cytosine deamination that results in DNA containing uracil, 5-methylcytosine deamination creates a C to T base substitution that is not as readily recognized by DNA repair proteins as being misplaced in the DNA strand. This contributes to inefficient repair of these lesions, and subsequent accumulation of this type of mutation in the genome. Deamination of 5-methylcytosine in *p53* and *HPRT* is a frequent mutational event associated with human cancers (Pfeifer *et al*., 2000). Although the spectrum of mutations within *p53* varies between different forms of cancer, 50% of all point mutations in colon cancer arise from transition mutations of normally methylated CpG dinucleotides. In contrast, only 10% of liver and lung cancers contain these same mutations. In addition to the potential direct mutational inactivation of a gene by a CpG to TpG transition, these mutated sequences could hinder the interaction between DNA and specific proteins involved in transcriptional regulation.

Alterations in CpG Island Methylation

The genome regions subject to hypermethylation in cancer cells are the CpG islands associated with gene promoters. A study undertaken to identify all differentially methylated CpG islands in cancer estimated that of the 45 000 CpG islands in the human genome, 600 exhibit methylation patterns in tumors different from those in normal tissues (Costello *et al*., 2000). These methylation changes appear to occur early in the neoplastic process, and some are even cancer-type specific, suggesting that CpG island hypermethylation is mechanistically involved in carcinogenesis rather than being a consequence of neoplastic transformation. An increase in DNA methylation also occurs with ageing. Consequently, the increased cancer predisposition observed with ageing may be partially attributable to the age-dependent increases in genome methylation (see below).

A tumour-suppressor gene inactivated by CpG island hypermethylation would have a number of important characteristics. These include (1) dense methylation of the normally unmethylated CpG island present in the promoter region, (2) an absence of coding region mutations in the tumour, (3) a deficiency of gene-specific transcripts in the tumour, (4) gene reactivation in the tumour with DNMT inhibitors (e.g. 5-azaC) and (5) loss of gene function from hypermethylation comparable to that seen for inactivating mutations (Baylin *et al*., 1998). It is important to note that alterations in CpG methylation are mitotically heritable (**Figure 3**), and can potentially result in the clonal expansion of neoplastic cells if the epigenetic changes provide a selective growth advantage.

Hypermethylation of Genes in Cancer

The normally unmethylated promoter for many genes is increasingly methylated during neoplastic progression. **Table 1** lists genes whose promoter regions are hypermethylated in cancer, and a number of them are described in further detail below.

p16 located at human chromosome 9p21 is the best documented example of tumour suppressor gene inactivation by promoter CpG island hypermethylation. *p16* is involved in the process of cell cycle regulation. *p16* encodes for a cyclin-dependent kinase inhibitor that blocks cell cycle progression by inhibiting the action of cyclin D. Cyclin D binds to cyclin-dependent kinases and forms a complex that phosphorylates the Rb protein, allowing progression through the cell cycle. Thus, *p16* inactivation would enable cells to proceed inappropriately through cell cycle check points.

p16 was initially identified as a tumour suppressor in melanoma; however, its function as a tumour suppressor was questioned early on because many tumours with LOH at this gene locus lacked mutations in the remaining allele. This apparent discrepancy was ultimately explained by the finding that the promoter of the nonmutated allele was often hypermethylated. Promoter region hypermethylation is also frequently the only detectable alteration, suggesting that the two 'hits' required to inactivate a tumour suppressor can both be epigenetic in origin. *p16* promoter methylation also appears to be an early event in carcinogenesis since it is

Table 1 Genes subject to hypermethylation cancer

Gene	Locus	Function	Tumour type(s)
Tumour-suppressor genes			
APC	5q21	Regulation of β-catenin; cell adhesion	Colorectal, gastrointestinal
BRCA1	17q21	DNA damage repair	Breast, ovarian
CDH1 (E-cadherin)	16q22.1	Homotypic epithelial cell–cell adhesion	Bladder, breast, colon, liver
LKB1	19p13.3	Serine, threonine kinase	Hamartomatous colon, papillary breast
MLH1	3p21.3	DNA mismatch repair	MSI positive colorectal and endometrial
p15 (CDKN2B)	9p21	Cyclin-dependent kinase inhibitor	Acute leukaemias, Burkitt lymphoma, multiple myelomas
p16 (CDKN2A)	9p21	Cyclin-dependent kinase inhibitor	Lung, gliomas, breast, colon, bladder, nasopharyngeal, melanomas, prostate, thymomas, multiple myelomas, lymphomas
PTEN	10q23.3	Regulation of cell growth and apoptosis	Prostate
RB	13q14.2	Sequesters E2F transcription factor	Retinoblastoma
VHL	3p25	Inhibits angiogenesis, regulates transcription	Renal cell carcinoma
Candidate tumour-suppressor and other genes			
14-3-3-σ (stratifin)	1p	Cell cycle control	Breast
CALCA (calcitonin)	11p15.2-p15.1	Reduces serum calcium	Various carcinomas, leukaemia
CD44	11p13	Metastasis suppressor	Prostate
DAP-kinase 1	9q34.1	γ-IFN-induced promoter of apoptosis	Burkitt lymphoma, other B cell malignancies
EDNRB	13q22	Endothelin receptor	Prostate
ER	6q25.1	Oestrogen-induced transcriptional activation	Breast, colon, lung, leukaemia
FHIT	3p14.2	Dinucleoside oligophosphate hydrolase	Oesophageal
GST-π	11q13	Cellular detoxification	Prostate
H-cadherin	16q24.2-q24.3	Cell adhesion	Lung
HIC1	17p13.3	Zinc finger protien	Brain, breast, colon, renal
IGF2	11p15.5	Growth factor, paternally expressed	Rhabdomyosarcoma, Wilms tumour
MDG1	1p33-p32	Fatty acid-binding protein	Breast
MYO-D1	11p15.4	Myogenesis	Bladder, lung
O^6-MGMT	10q26	DNA repair	Brain, colon, lung, lymphomas
p73	1p36	Apoptotic response	Neuroblastoma, T and B cell malignancies
RASSFIA	3p21.3	Ras effector homolog	Lung
RAR-β2	3p24	Retinoic acid receptor/ transcription factor	Breast, colon, pancreas
TIMP-3	22q12.3	Inhibitor of matrix metalloproteinases	Brain, renal

already present in the preinvasive stages of cancer. In colon cancer, where *p16* is frequently hypermethylated, there is also a paucity of mutations in the *Rb* tumour-suppressor gene, indicating that *p16* inactivation alone is sufficient to circumvent the cell cycle G_1 block (Baylin *et al*., 1998). *p16* hypermethylation has now been implicated in many forms of cancer, including those that originate in the breast, bladder, brain, colon, oesophagus, head and neck, and lung (Liggett and Sidransky, 1998).

p15 and *p16* are positioned 15 kb apart at chromosome location 9p21, and both encode for cyclin-dependent kinase inhibitors. Whereas *p16* inactivation is found principally in solid tumours, loss of *p15* function occurs mainly in haematopoietic cell cancer. The constitutive expression of *p16* differs markedly from that of *p15*, which is more limited and regulated by transforming growth factor-β (Baylin *et al*., 1998). Hypermethylation in this chromosomal region can affect both *p15* and *p16* promoter

CpG islands simultaneously or it can be specific to either gene promoter. For example, in Burkitt's lymphomas both *p15* and *p16* are hypermethylated, whereas only *p15* is commonly hypermethylated in adult acute myelogenous leukaemia (AML), paediatric AML, adult acute lymphocytic leukaemia (ALL) and paediatric B cell ALL (Baylin *et al*., 1998). *p15* promoter hypermethylation is also present in the myelodysplastic state preceding leukaemia, indicating that DNA methylation of the *p15* promoter is an early carcinogenic event.

hMLH1 on chromosome 3p21.3 encodes for a protein essential for DNA mismatch repair. (See chapter on *Genetic Instability and DNA Repair*.) Disruption of *hMLH1* expression is frequently found in patients with hereditary nonpolyposis carcinoma, and also to a lesser extent in sporadic colon cancers (Tycko, 2000). The inactivation of *hMLH1* results in microsatellite instability (MIN + phenotype) which is characterized by errors in replication at one to four base pair repetitive microsatellite DNA sequences. *hMLH1* hypermethylation and the concomitant loss of gene expression is estimated to be present in approximately 70% of sporadic MIN + colorectal carcinomas whereas it is infrequent in MIN− tumours (Herman *et al*., 1998). *DNMT* inhibitors can restore DNA mismatch repair activity by causing the demethylation of the hypermethylated *hMLH1* promoter (Herman *et al*., 1998).

E-CADHERIN is located at chromosomal position 16q22.1. E-CADHERIN is involved in the calcium-dependent regulation of cell growth and differentiation by virtue of its ability to mediate homotypic cell-to-cell adhesion. Disruptions in these cellular adhesions are prominent in cancer cells with reduced *E-CADHERIN* expression, providing support for its normal role in preventing tumour invasion and metastasis. LOH at this chromosome location occurs often in breast cancer, and inactivating mutations in *E-CADHERIN* are also found in a variety of other solid tumours. Furthermore, decreased expression of *E-CADHERIN* in both breast and prostate cancers correlates with hypermethylation of the CpG island that encompasses the transcription start site (Baylin *et al*., 1998).

ER (Oestrogen receptor) at chromosome location 6q25 is expressed in a wide variety of tissues, and it encodes for a transcription factor that is activated only upon ligand binding. The presence of ER protein in breast cancer is prognostic for increased survival and lower risk of relapse. (See chapter on *Signalling by Steroid Receptors*.) Hypermethylation of the *ER* promoter has been found in a variety of tumour types including breast, colon and lung cancer as well as leukaemia (Baylin *et al*., 1998). An increase in *ER* promoter methylation in the normal colon also occurs with ageing, and this epigenetic change may predispose humans to colon cancer (Baylin *et al*., 1998; Issa, 2000).

GELSOLIN encodes for a calcium-dependent actin filament severing and nucleating protein whose expression directly correlates with the induction of cellular differentiation (Hoshikawa *et al*., 1994). *GELSOLIN* expression is always decreased during malignant transformation (Hoshikawa *et al*., 1994). Epigenetic silencing of *GELSOLIN* is a common feature of most human breast malignancies, a condition that is reversed in cultured cells by treatment with either HDAC inhibitors such as trichostatin A or DNMT inhibitors such as 5-azaC (Tycko, 2000). The ability to reverse silencing of *GELSOLIN* by either one of these agents alone is unusual since many other genes silenced by hypermethylation require treatment first with a DNMT inhibitor followed by a HDAC inhibitor to achieve substantial gene reactivation (Cameron *et al.*, 1999).

O^6-MGMT on chromosome 10q26 encodes for the enzyme methylguanine–DNA methyltransferase. This ubiquitous protein is responsible for repairing mutagenic and carcinogenic O^6-alkylguanine adducts. The CpG island associated with the promoter of *O^6-MGMT* is hypermethylated and transcriptionally silenced in multiple solid tumour types, including those originating in the brain, colon and lung (Herman and Baylin, 2000). *O^6-MGMT* loss of function in colon cancer also plays a major role in determining the type and extent of mutations found in the oncogene K-*RAS*. G to A transition mutations result from unrepaired O^6-alkylguanine adducts, and they are the major type of K-*RAS* mutations present in human colon cancer. The finding that 70% of these K-*RAS* mutations are associated with a hypermethylated *O^6-MGMT* promoter suggests that epigenetic inactivation of this gene is involved in the formation of this disease (Herman and Baylin, 2000)

p73 is located at chromosome position 1p36, and encodes for a protein proposed to act as a neuroblastoma tumour suppressor. *p73* maps to an area of 1p36 commonly found to have LOH in neuroblastomas; however, the remaining allele in these tumours frequently lacks mutations. Wild-type p73 is reported to interact with c-Abl in promoting apoptosis, while a truncated p73 functions as an anti-apoptotic protein during neural development in mice. Several studies have concluded that the maternal allele of *p73* is functionally inactivated by the epigenetic process of genomic imprinting (see below). Although it is presently unclear if imprinting at the *p73* locus is mechanistically involved in carcinogenesis, *p73* promoter hypermethylation has clearly been documented in both T cell acute lymphoblastic leukaemias and Burkitt lymphomas (Herman and Baylin, 2000).

14-3-3-σ (***STRATIFIN***) on chromosome 1p is a member of the 14-3-3 family of proteins that mediate signal transduction by binding to other proteins containing phosphoserine motifs. *14-3-3-σ* transcription is induced by DNA damage through a p53 response element in the promoter. It is involved in the maintenance of cell cycle arrest at G_2, and it prevents entry into mitosis by sequestering the cdc2–cyclinB1 mitotic initiation complex in the cytoplasm. The promoter CpG island of *14-3-3-σ* is frequently hypermethylated with transcriptional silencing in 91% of

breast cancers while point mutations in the coding region and LOH occur infrequently (Ferguson *et al*., 2000). This transcriptional silencing is relieved by treatment of 14-3-3-σ-deficient breast cancer cell lines with 5-azaC.

BRCA1 (breast cancer type 1) is located on human chromosome 17q21. It encodes for an ~220-kDa nuclear phosphoprotein component of the RNA polymerase II holoenzyme complex. *BRCA1* is implicated in diverse cellular functions, including cell cycle control, chromatin structure-imposed transcriptional regulation and DNA damage response. Approximately 50% of familial forms of breast and ovarian cancer harbour germ-line mutations in *BRCA1*. Whereas 40–80% of sporadic breast carcinomas and 30–60% of sporadic ovarian carcinomas show LOH at this chromosomal location and have decreased levels of *BRCA1* mRNA, somatic mutations in *BRCA1* are very rare (Bianco *et al*., 2000; Esteller *et al*., 2000a).

The promoter of *BRCA1* is contained within a 2.7-kb CpG island. The 5′ end of this large CpG island is unusual in that it is normally methylated in somatic cells but not in the gametes. A repeat element located within this CpG island may be responsible for facilitating methylation of this region. *BRCA1* expression in normal breast, cervical and kidney cells is enhanced up to fivefold by treatment with DNMT inhibitors, suggesting that methylation of the 5′ portion of the CpG island in normal tissues may function in regulating *BRCA1* expression (Magdinier *et al*., 2000). Interestingly, *BRCA1* promoter methylation is significantly increased in medullary and mucinous subtypes of breast carcinomas relative to ductal subtypes; *BRCA1* methylation is also observed in conjunction with LOH in ovarian cancer (Esteller *et al*., 2000a). Evidence also suggests that *BRCA1* hypermethylation is specific to breast and ovarian cancer since the *BRCA1* promoter is not abnormally methylated in primary colorectal and liver carcinomas or in leukaemias (Bianco *et al*., 2000; Esteller *et al*., 2000a).

APC (adenomatous polyposis coli) tumour-suppressor gene is located at chromosome 5q21. It encodes for a large protein that normally associates with and negatively regulates signalling of the cell adhesion protein, β-catenin. Familial adenomatous polyposis (FAP) is an inherited disorder arising from germ-line mutations in *APC*. (See chapter on *Inherited Predispositions to Cancer*.) The majority of these mutations produce truncated proteins that are incapable of mediating β-catenin degradation, thereby resulting in the activation of genes involved in cell growth (Esteller *et al*., 2000b). *APC* is genetically inactivated early in the genesis of both familial and sporadic forms of colorectal carcinoma. The identification of FAP cases where either one or both alleles of *APC* lacked genetic lesions suggested an alternative form of *APC* repression. Methylation analyses confirmed that the *APC* promoter is commonly hypermethylated in colorectal adenomas and other cancers associated with the gastrointestinal tract, but not in brain, head and neck, lung and ovarian cancer (Esteller *et al*. 2000b).

HIC1 (hypermethylated in cancer) on chromosome 17p13.3 encodes for a zinc finger protein transcription factor that is a candidate tumour suppressor activated by p53 (Baylin *et al*., 1998). The normally unmethylated *HIC1* contains a CpG island that spans the entire gene, and hypermethylation of this CpG island has been documented in many solid tumours including those of the breast, colon, kidney and lung as well as in leukaemia. *HIC1* hypermethylation occurs in both acute lymphoblastic leukaemias (53%) and acute myeloid leukaemias (10%) (Issa *et al*., 1997). Furthermore, *HIC1* promoter hypermethylation is more prominent during blast crisis in chronic myelogenous leukaemia and the progression stage of acute lymphocytic leukaemia. This suggests that *HIC1* methylation is a later event in the genesis of leukaemia.

RAR-β2 (retinoic acid receptor beta) is located on chromosome 3p24. It is one of three known retinoic acid receptors that together belong to the nuclear receptor superfamily. *RAR-β* encodes for a putative tumour-suppressor protein that functions as a transcription factor when bound to retinoic acid (RA). RAR transcription factors bound to RA heterodimerize with retinoid X receptors, and then bind to retinoic acid responsive elements (RAREs) within the regulatory regions of RA-inducible genes. Transfection of *RAR-β* cDNAs into non-expressing tumour cells results in growth suppression, whereas transgenic mice expressing *RAR-β* antisense RNAs demonstrate enhanced tumorigenesis. These results demonstrate that *RAR-β* normally has an antiproliferative function. RAR-β levels are decreased in a number of cancers, including those that develop in the breast, cervix, head and neck, lung and ovary. Analysis of primary breast tumours indicates that approximately one-third have hypermethylation at the *RAR-β2* promoter (Sirchia *et al*., 2000). Loss of both RAR-β and ER function in breast cancers is correlated with resistance to RA treatment. The *RAR-β2* promoter is abnormally hypermethylated in several forms of cancer including breast cancer, colon carcinoma and pancreatic carcinoma, but its methylation status appears to not be related to the presence of functional ER protein (Sirchia *et al*., 2000).

IMPRINTED GENES AS CANCER SUSCEPTIBILITY LOCI

Genomically imprinted genes provide strong evidence that transcriptional silencing results from DNA methylation. This subset of genes is normally monoallelically expressed in a parent of origin-dependent manner. Imprinted gene expression in somatic cells depends upon the sex of the parent from which the allele originated, but not the sex of the individual. Every imprinted gene thus far examined has been associated with at least one differentially methylated CpG island. The epigenetic changes that confer the heritable

imprint mark have not yet been unambiguously defined, but cytosine methylation is the strongest candidate. Histone acetylation has also been proposed to be mechanistically involved in imprinted gene regulation since parental-specific differences in acetylation are associated with imprinted genes (Hu *et al*., 1998). Approximately 40 imprinted genes have been identified in humans to date, and estimates predict the presence of 100–200 imprinted genes in the entire genome (Barlow, 1995). The normal silencing of the imprinted allele is equivalent to a first 'hit' in the 'two-hit hypothesis' for carcinogenesis (**Figure 6**) (Knudson, 1971). A single genetic or epigenetic alteration in the expressed allele could therefore completely abrogate function of an imprinted gene. Since most imprinted genes are involved in cell growth and all are functionally haploid, they represent unique susceptibility loci for cancer development.

Human *IGF2* and *H19* are located in a chromosomal region (11p15.5) harbouring a cluster of imprinted genes. *IGF2* encodes for a potent mitogenic factor involved in cell growth and embryonic development whereas *H19* transcripts are non-coding. The reciprocally imprinted *IGF2* and *H19* genes are expressed from the paternally and maternally inherited alleles, respectively. The epigenetic regulation of this locus has been intensively studied in normal and malignant tissues. IGF2 overexpression occurs commonly in cancer (for a review, see Reik *et al*., 2000), and loss of imprinting is one mechanism responsible for the dysregulation of this influential growth factor (Rainier *et al*., 1993).

Human *IGF2* has four promoters that function in a tissue-specific manner (Vu and Hoffman, 1994). Promoters 2 to 4 are contained within a CpG island, and transcripts derived from these promoters are monoallelically expressed. In contrast, the P1 promoter 20 kb upstream of P2 is not associated with this CpG island, and drives biallelic expression of IGF2 in the liver (Vu and Hoffman, 1994; Baylin *et al*., 1998). During the ageing process, the P2 to P4 promoters of the normally unmethylated paternal allele are subject to incremental increases in methylation, and methylation of these promoters is also enhanced in cancer. Hypermethylation of this region in tumour cells results in decreased expression originating from the P3 promoter with the switching of *IGF2* transcription regulation to the non-imprinted P1 promoter (Vu and Hoffman, 1994). The ability of *IGF2* to undergo such promoter switching can be regarded as an epigenetic mechanism by which tumour cells can gain a selective growth advantage. Interestingly, the mouse *Igf2* does not have the equivalent of the human P1 promoter, and

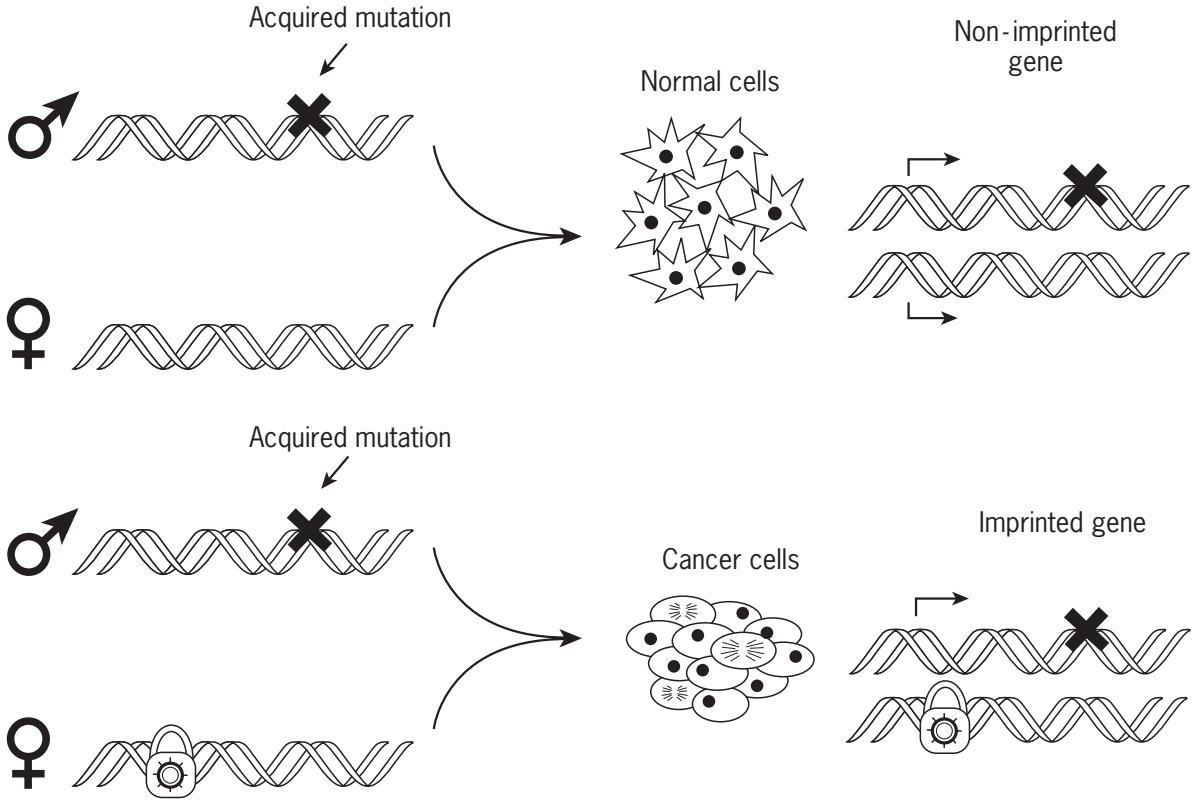


Figure 6 Imprinted genes as susceptibility loci in cancer. For most nonimprinted genes, an acquired mutation (X) does not contribute directly to carcinogenesis because of the presence of a second transcriptionally active wild-type allele. In contrast, acquisition of a genetic or epigenetic mutation on the active allele of an imprinted tumour-suppressor gene (paternal in this example) can directly result in cancer by inactivation of the single functional copy of the gene.

therefore biallelic expression cannot occur via this mechanism in mice. This fundamental difference between species in regulating *IGF2* expression is important to consider when extrapolating carcinogenic risk estimates from mice to humans. (See chapter on *Advantages and Limitations of Models for Cancer and Malignant Cell Progression*.)

The epigenetic control of *IGF2* imprinting has recently been further refined with the demonstration that imprinted expression of both *IGF2* and the adjacent, reciprocally imprinted *H19* depends on the presence of differentially methylated CTCF (vertebrate enhancer blocking protein) DNA-binding sites between these two genes (**Figure 7**) (Bell and Felsenfeld, 2000; Hark *et al*., 2000; Kanduri *et al.*, 2000). The mechanism of CTCF-mediated transcriptional repression also involves CTCF interaction with HDACs (Lutz *et al*., 2000). In the case of *IGF2*, methylation of the paternal allele at the CTCF recognition sites prevents CTCF binding. In contrast, CTCF proteins bind to the unmethylated maternal allele, creating a physical boundary. This prevents the enhancer elements located downstream of *H19* from interacting with the *IGF2* promoter. Consequently, *IGF2* and *H19* are reciprocally imprinted, and transcribed from the paternal and maternal alleles, respectively. Deletion of this CTCF binding region on the maternal allele results in biallelic expression of *IGF2* (Thorvaldson *et al.*, 1998; Srivastava *et al.*, 2000). In colorectal cancers with loss of imprinting for *IGF2*, abnormal methylation of the maternal CTCF binding sites has been observed (Nakagawa *et al.*, 2001). Together with the demonstration that methylation of CpGs within the CTCF binding site block CTCF binding (Kanduri *et al.*, 2000), these results suggest a mechanistic link between the CTCF epigenetic control elements, loss of *IGF2* imprinting, and carcinogenesis.

We have recently provided evidence that another pair of human imprinted genes utilizes a similar mechanism to control their reciprocal imprinting pattern. The *DLK1/GTL2* domain at chromosome 14q32 encodes for the paternally expressed *DLK1* and the maternally expressed *GTL2* genes. *DLK1* is involved in several cellular differentiation processes including adipogenesis, haematopoiesis and

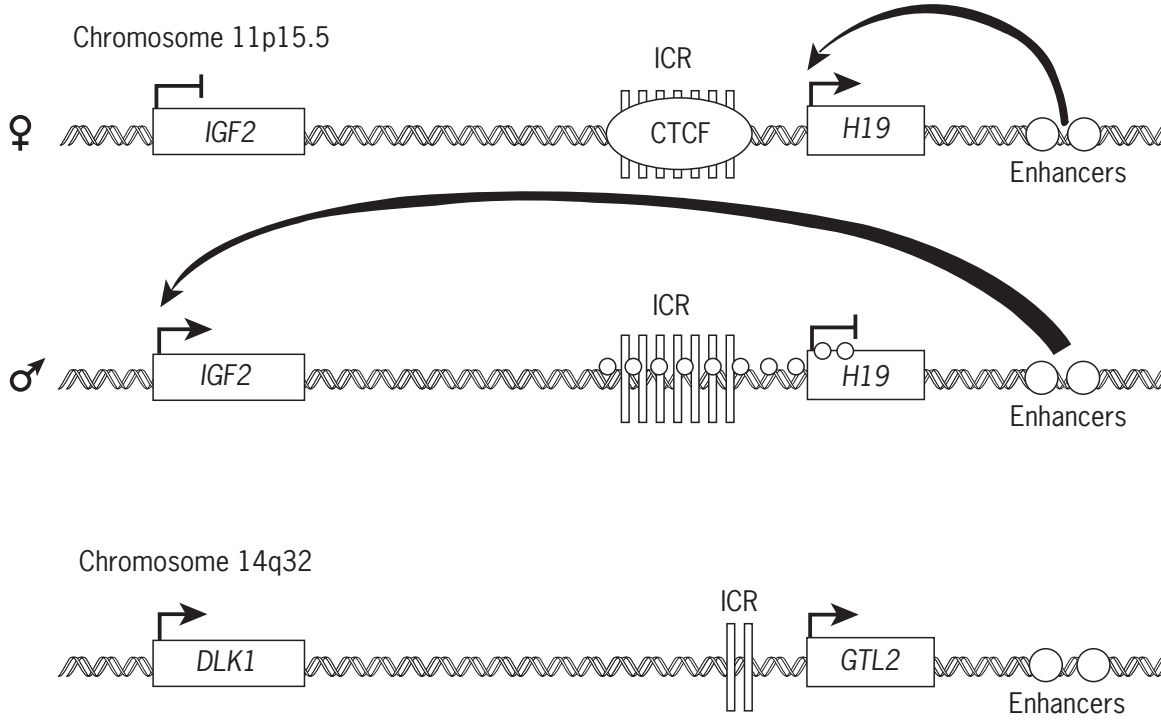


Figure 7 Cancer predisposition and methylation-dependent chromatin boundary elements. The reciprocally imprinted *IGF2* and *H19* genes at 11p15.5 are separated by approximately 100 kb of DNA that contains a differentially methylated ICR (imprint control region). The ICR contains seven methylation-sensitive CTCF (vertebrate enhancer blocking protein) binding sites (vertical bars). CTCF proteins are blocked from binding the methylated (small circles) paternal allele but the unmethylated maternal allele binds CTCF resulting in the formation of a chromatin boundary element. This boundary is thought to divert the enhancer elements, located downstream of *H19*, away from the maternal *IGF2* promoter and toward the available *H19* promoter, resulting in maternal *H19* expression. In contrast, the paternal *IGF2* is expressed because its promoter is subject to enhancer influence since methylation of the paternal ICR prevents CTCF binding and boundary formation. A second set of reciprocally imprinted genes was identified at 14q32 (Wylie *et al.*, 2000). *DLK1* and *GTL2* share spatial, structural, and expression characteristics with the *IGF2/H19* locus and there are two differentially methylated consensus CTCF binding sites between these two genes.

neuroendocrine differentiation, and may play an important role in neuroendocrine tumorigenesis (Laborda, 2000). Like *IGF2* and *H19*, *DLK1* and *GTL2* are also separated by differentially methylated consensus CTCF binding sites. Furthermore, enhancer element sequences are located downstream from *GTL2* that are identical with those found downstream of *H19* (**Figure 7**) (Wylie *et al*., 2000). These findings suggest that this mechanism of epigenetic regulation may be commonly employed to coordinate the expression of juxtapositioned reciprocally imprinted genes. Thus, specific mutation of these CTCF binding sites and/or the CTCF binding proteins could potentially alter the expression of a number of imprinted genes in the genome.

NON-GENOTOXIC MEDIATORS OF CARCINOGENESIS

Non-genotoxic carcinogens induce cancer by causing epigenetic changes in the genome rather than by directly mutating the DNA base composition. Furthermore, both endogenous and exogenous factors can cause these epigenetic changes. Oxidation of guanine by endogenous chemicals often results in the formation of 8-hydroxyguanine that can contribute directly to genetic alteration because it is a potentially miscoding base, and also impair normal epigenetic DNA modification by impeding the methylation of adjacent cytosines. This is also true for photodimers, abasic sites, base alkylations and other oxidative DNA lesions that not only directly damage DNA, but also lead to heritable alterations in normal genomic methylation patterns (Wachsman, 1997; MacPhee, 1998). Additionally, a number of exogenous chemical agents are now known that induce both genome wide and gene specific chromatin structure changes by altering DNA methylation directly or by modifying DNMT activity (**Table 2**) (Baylin *et al*., 1998). Below is a brief description of specific compounds and factors that cause cancer through non-genotoxic mechanisms.

Table 2 Agents that influence DNA methylation

Induce hypomethylation	Induce hypermethylation
1,3-Bis(2-chloroethyl)-1-nitroso-urea	3′-Azidodideoxythymidine (AZT)
4-Nitroquinoline 1-oxide	3-Deazaadenosine
4-Nitrosodiethylamine	5-Fluorouracil
4-Nitroso-*N*-ethylurea	5-Fluorodeoxyuridine
5-Azadeoxycytidine	Aphidicolin
5-Fluorodeoxycytidine	Butyrate
6-Thioguanine	Cisplatinum
7,12-Dimethylbenz[*a*]-anthracene	Colchicine
Aflatoxin B_1	Etoposide
Benzo[*a*]pyrene	Doxorubicin
Butyrate	Nalidixic acid
Butyryl-cAMP	Propionate
Bromobenzene	Trapoxin
Cyclophosphamide	Trichostatin
Ethionine	Vinblastine
Hydralazine	Vincristine
Lead nitrate	
MNNG, other alkylating agents	
N-Methyl-*N*-nitrosourea	
Novobiocin	
Oestradiol	
Procainamide	
Pseudoisocytidine	
Retinoic acid	
Teniposide	
Toposide	
Topoisomerase II inhibitors	

(Adapted from Holliday, 1991 and Wachsman, 1997.)

Nickel

Studies with carcinogenic nickel have led to development of a model for the potential epigenetic mechanisms by which non-genotoxic carcinogens inactivate tumor suppressor genes (Costa, 1995). Water-insoluble Ni^{2+} is a non-genotoxic carcinogen that localizes to the nucleus following its cell entry by phagocytosis. There it is thought to act through its affinity for heterochromatic regions of the genome. Chromosomal damage in these regions presumably results from oxidation that occurs when Ni^{2+} binds to chromatin proteins. These genotoxic events are normally not detrimental since the affected heterochromatin is usually genetically inactive; however, nickel also appears to cause harmful epigenetic modifications. This is postulated to result from nickel binding to oxygen atoms in the DNA phosphate backbone within heterochromatin and inducing a localized increase in DNA methylation. If this nickel-induced DNA methylation spreads outward to encompass adjacent euchromatic regions, tumour-suppressor genes within the proximity of the epimutated heterochromatin can be aberrantly silenced (Costa, 1995).

Peroxisome Proliferators

Peroxisomes are cytoplasmic organelles that are present in all eukaryotic cells except red blood cells. They are bounded by a single membrane and function in the metabolism of many substrates including long-chain fatty acids, sterols, dicarboxylic acids, prostaglandins, xenobiotics and oxygen free radicals. They are also involved in the synthesis of cholesterol, ether lipids, carbohydrates and bile acids (Masters and Crane, 1998). While peroxisomes are present in cells throughout the body, their density is cell-type dependent, with the largest number present in mammalian liver and kidney (Masters and Crane, 1998).

Peroxisomes are induced by a group of compounds collectively called peroxisome proliferators. These agents markedly increase peroxisome number, and they can also stimulate replicative DNA synthesis while suppressing apoptosis (Masters and Crane, 1998). Although this class of agents induces peroxisome proliferation in mice and rats through its interaction with the peroxisome proliferator activated receptor alpha (Peters *et al*., 1997), guinea pigs do not exhibit this effect, and neither do dogs, marmosets or humans. There is strong evidence that peroxisome proliferators are hepatocellular carcinogens in rodents; however, there is a substantial species-specific difference in the carcinogenic response to these agents (Roberts, 1999). They are considered to be non-genotoxic carcinogens since they act independently of covalent DNA binding and without evidence of genetic mutation. The mechanism by which peroxisome proliferators cause cancer is still under active investigation, but it is thought to involve increased oxidative stress caused by prolonged agent exposure (Masters and Crane, 1998). Importantly, accumulated data indicate that peroxisome proliferators do not constitute a serious carcinogenic risk to humans and other primates.

Diet

S-Adenosylmethionine (SAM) is the methyl donor for various methylation reactions in mammalian cells, including DNA methylation (**Figure 2**). A dietary deficiency in SAM precursors or cofactors involved in SAM biosynthesis (such as folate) leads to genomic hypomethylation, and diets deficient in these components have been widely used to induce liver tumours in rodents (Laird, 2000). In humans, dietary deficiency of methyl donors correlates with an increased risk for liver and colon tumours (Giovannucci *et al*., 1993). This hypomethylation is thought to be caused by either a deficiency in methyl group donors or by DNMT-enhanced cytosine deamination, but the mechanism has not yet been established (Laird and Jaenisch, 1996; Zingg and Jones, 1997). With the increasing awareness of the importance of DNA methylation in cancer formation, there is an expanding commercial interest in human dietary supplements fortified with methylation precursors such as SAM and folate. Studies to determine the specific contribution of diet to changes in DNA methylation patterns are certainly called for, and will help clarify the need for dietary intervention as a preventative or therapeutic measure against cancer.

Ageing

DNA hypomethylation was originally suspected to be responsible for the gene expression changes often observed with the ageing process. Interestingly, age-related decreases in DNA methylation occur primarily in the coding and intronic regions of genes, and they correlate poorly with observed reductions in gene expression (Issa, 2000). This disparity was clarified by studies showing that reduced gene expression that occurs with age results from a progressive increase in gene specific promoter methylation rather than generalized genomic hypomethylation.

These age-related increases in promoter CpG island methylation occur in a number of genes involved in cancer, including *IGF2*, *Versican*, *PAX6*, and *N33* in colon cancer and *HIC1* in prostate cancer (Issa, 2000). It is likely that many other genes will also fall into this category, because several studies designed to isolate differentially methylated CpG islands in cancer have identified a number of CpG islands that exhibit increased methylation with both ageing and neoplastic transformation (Issa, 2000). Not all age-dependent hypermethylation events result in cancer. The *ER* gene is hypermethylated in nearly all primary colon cancers, yet the normal colon of patients both with and without colon cancer has about the same yearly increase in *ER* promoter CpG island methylation (Issa, 2000). Since age-related hypermethylation varies among individuals of the same age, it is likely that genetic predisposition to epimutations, as well as exposure to environmental factors are involved in cancer formation. Thus, there is now compelling evidence of a mechanistic link between the ageing process and tumorigenesis in that age-related promoter hypermethylation frequently occurs in genes known to be involved in cancer formation (Issa, 2000).

DIAGNOSTIC AND THERAPEUTIC POTENTIAL OF EPIGENETIC ABNORMALITIES

Cancer Diagnosis

Sequencing data obtained from the human genome project are currently undergoing analysis to construct a human epigenetic map based on CpG content. This knowledge coupled with cross-species comparisons of the epigenome will be invaluable in deciphering the epigenetic elements involved in gene regulation (e.g., Hardison, 2000; Killian *et al*., 2000; Vu and Hoffman, 2000; Wylie *et al*., 2000). Epigenetic alterations in genes are early oncogenic events in some cancers, and detection of these early abnormalities may aid in protecting people from cancer through dietary alterations or pharmacological intervention (Laird, 1997). With increasing awareness of the importance of epigenetics in cancer formation, and the advent of laboratory techniques such as bisulfite DNA sequencing, methylation-sensitive PCR (**Figure 8**) and gene expression profiling by DNA microarrays, it is likely that methylation profiles will ultimately be used to predict an individual's predisposition to cancer, assist in cancer diagnosis and determine optimal therapeutic approaches (Esteller *et al.*, 2001).

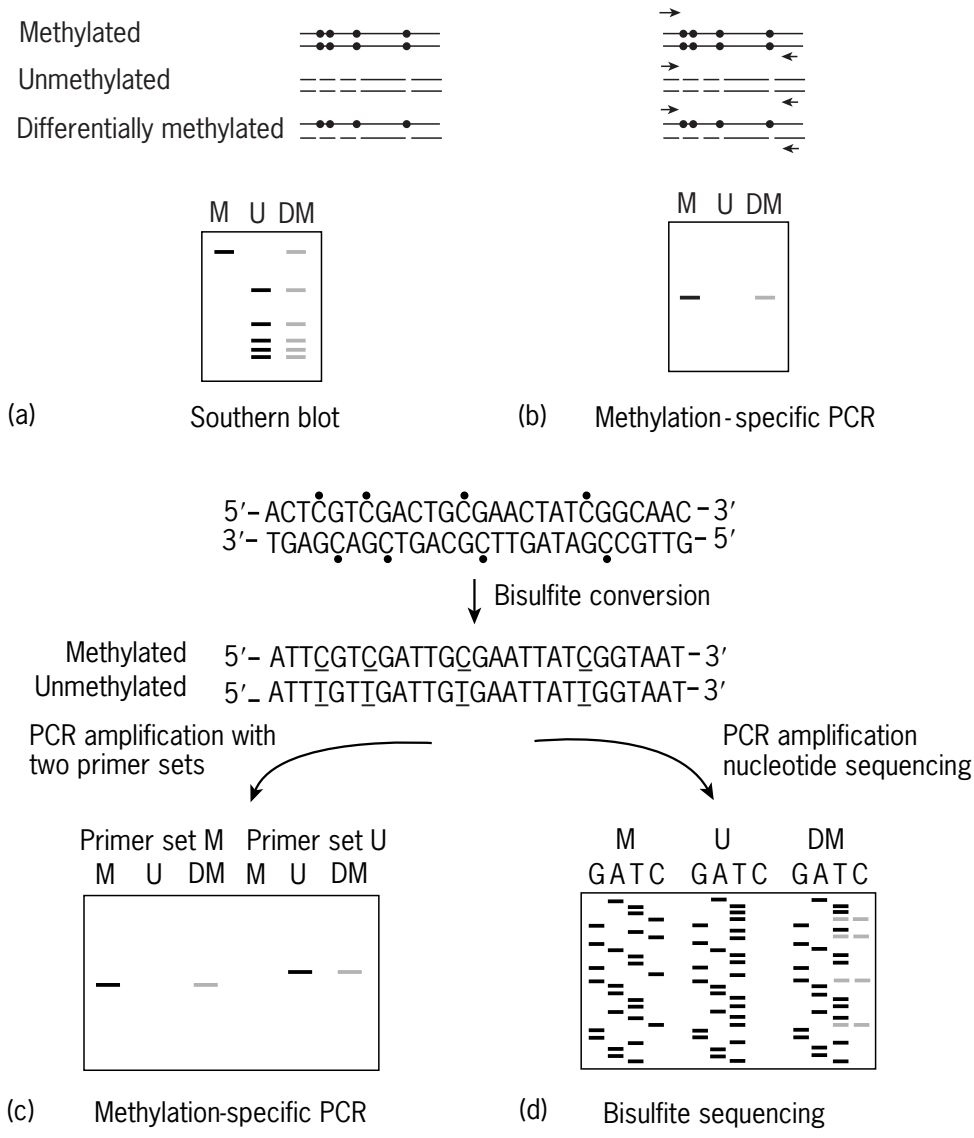


Figure 8 Methods used to analyse CpG methylation. Cytosine methylation is indicated by the circles. (a) Southern blotting depends on methylation-sensitive restriction enzymes to discriminate between methylated and unmethylated alleles. DNA is digested and fractionated on an agarose gel followed by blotting with a probe specific to the region of interest. Methylated recognition sites are resistant to digestion and will yield larger DNA fragments on the blot than unmethylated DNA. (b) Methylation-specific PCR requires digestion with methylation sensitive restriction enzymes, followed by PCR amplification. Methylated (uncut) DNA will yield an amplification product whereas unmethylated DNA will not be amplified. Bisulfite conversion of unmethylated cytosines precedes analysis by either a modification of methylation-specific PCR (c) or bisulfite sequencing (d). Sodium bisulfite treatment of DNA leads to the conversion of unmethylated cytosines to uracils while methylated cytosines are protected from conversion. Subsequent PCR using two independent primer sets designed to bisulfite protected and bisulfite converted sequence amplify methylated (M) and unmethylated (U) alleles, respectively. Bisulfite sequencing is the most direct means of analysing the methylation status of individual cytosines. Fully methylated cytosines are evident in the C lane of a sequencing gel using this method, while unmethylated cytosines are converted to thymines in the PCR amplification step prior to sequencing. Alleles having differential cytosine methylation (DM) are evident by the presence of bands in both the C and T lanes.

Cancer Treatment

A promising feature of alterations in DNA methylation patterns and chromatin structure in cancer cells is their potential for reversibility, because these modifications occur without changing the primary nucleotide sequence. The two major pharmacological targets associated with these epigenetic changes are DNMT

and HDAC. The DNMT inhibitor 5-azaC is structurally similar to cytosine (**Figure 4**), but when incorporated into DNA it forms a stable covalent bond with DNMT that inhibits further methylation by the sequestered enzyme. Consequently, overall genomic hypomethylation develops with subsequent rounds of DNA replication.

5-AzaC is efficacious in treating patients with acute leukaemia. It has also undergone clinical testing for the treatment of solid tumours; however, 5-azaC produces a high level of normal tissue toxicity and mutagenicity. These untoward side effects are not due to the resulting hypomethylation, but are attributed to the presence of the incorporated DNMT–5-azaC complexes in the genomic DNA (Laird, 1997). More specific strategies to inhibit the action of DNMT are being developed, including the use of antisense molecules. In this approach, antisense DNAs complementary to the DNMT mRNA inhibit methyltransferase activity by preventing DNMT translation. HDAC inhibitors, such as trichostatin A and sodium butyrate, have been shown to increase the level of histone acetylation in cultured cells, and to cause growth arrest, differentiation and apoptosis. Consequently, they are currently being tested in clinical trials as therapeutic agents for cancer.

Refinement in our understanding of the specific contributions of methylation and histone deacetylation to tumour-suppressor gene silencing in each type of cancer may make custom-designed treatments for gene reactivation possible. For example, *GELSOLIN* silencing can be reversed by treatment with HDAC inhibitors alone, while other tumour-suppressor genes achieve higher levels of reactivation when DNMT and HDAC inhibitors are used together. *RAR-β2* provides another example of the specificity of epigenetic reactivation. Loss of *RAR-β* expression by promoter hypermethylation can result in tumour resistance to treatment with all-*trans*-retinoic acid. Thus, demethylating agents in combination with all-*trans*-retinoic acid have been proposed for the treatment of cancers lacking *RAR-β* expression because of promoter hypermethylation (Côté and Momparler, 1997). Combining trichostatin A with all-*trans*-retinoic acid not only reactivated *RAR-β* in breast cancer cells, but also significantly reduced cell proliferation (Sirchia *et al*., 2000).

Another novel therapeutic approach proposes to use genetically engineered proteins to reactivate genes with epigenetically silenced promoters. In this approach, a chimeric fusion protein containing DNA-binding zinc finger motifs joined to protein domains having, or capable of recruiting HAT activity for example, might be used to target and alleviate localized areas of chromatin condensation in the promoter regions of silenced tumour suppressor genes. The specificity of the target sequence binding originates from the customized site-specific DNA contacts of the zinc finger domain. Once bound to the target sequence, the chromatin modifier activity would act locally to alleviate chromatin condensation and promote gene reactivation. These proteins could be tailored to virtually any sequence, and would be tethered to the particular protein domains required for the activation of the affected gene promoter (e.g. Liu *et al.*, 2001). Combinations of these chimeric proteins might also expedite a positive clinical outcome by targeting multiple silenced genes simultaneously. Epigenetic cancer therapy also has major potential advantages over conventional therapeutic approaches. First, intact copies of tumour-suppressor genes do not need to be transfected into cells because they are already present in the cancer cell genome; they only need to be reactivated. Second, if gene-specific approaches are used to reactivate epigenetically silenced tumour-suppressor genes there should be little normal tissue toxicity, enabling them to be safely combined with more conventional therapies.

CONCLUSION

Recent years have seen a shift in thinking regarding the molecular basis for gene inactivation in cancer toward accommodating both genetic and epigenetic mechanisms. There has been a dramatic increase in the number of literature reports documenting hypermethylation of specific genes in numerous types of cancer, and it is clear that abnormal promoter hypermethylation is a prominent aetiological event. The specific mechanisms leading to epigenetic inactivation of genes in cancer must be further defined in addition to the roles of repressor complex components that coordinate silencing of these specific genes. With this knowledge, it may be possible to implement strategies in susceptible individuals to prevent deleterious epigenetic alterations that would otherwise lead to cancer. There is also the exciting possibility of developing novel therapeutic approaches for specifically alleviating abnormal promoter hypermethylation in tumours. The sequences involved in establishing the epigenetic profile of chromatin are clearly of fundamental importance to oncogenesis. Cross-species sequence comparisons in the future will greatly facilitate our ability to move from a single-gene to a genome-wide approach to identify conserved regulatory elements and determine their role in the epigenetic control of gene expression.

ACKNOWLEDGEMENTS

Research in the Jirtle laboratory was supported by NIH grants CA25951 and ES08823, DOD grant DAMD17-98-1-8305, Sumitomo Chemical Company, Ltd, and AstraZeneca Pharmaceuticals, Ltd.

REFERENCES

Barlow, D. P. (1995). Gametic imprinting in mammals. *Science*, **270**, 1610–1613.

Baylin, S. B., *et al*. (1998). Alterations in DNA methylation: a fundamental aspect of neoplasia. *Cancer Research*, **72**, 141–196.

Bell, A. C. and Felsenfeld, G. (2000). Methylation of a CTCF-dependent boundary controls imprinted expression of the *Igf2* gene. *Nature*, **405**, 482–485.

Bianco, T., *et al*. (2000). Tumour-specific distribution of *BRCA1* promoter region methylation supports a pathogenic role in breast and ovarian cancer. *Carcinogenesis*, **21**, 147–151.

Cameron, E. E., *et al.* (1999). Synergy of demethylation and histone deacetylase inhibition in the reexpression of genes silenced in cancer. *Nature Genetics*, **21**, 103–107.

Costa, M. (1995). Model for the epigenetic mechanism of action of nongenotoxic carcinogens. *American Journal of Clinical Nutrition*, **61(suppl.)**, 666S-669S.

Costello, J. F., *et al*. (2000). Aberrant CpG-island methylation has non-random and tumour-type-specific patterns. *Nature Genetics*, **25**, 132–138.

Côté, S. and Momparler, R. L. (1997). Activation of the retinoic acid receptor gene by 5-aza-2′-deoxycytidine in human DLD-1 colon carcinoma cells. *Anticancer Drugs*, **8**, 56–61.

Esteller, M., *et al*. (2000a). Promoter region hypermethylation and *BRCA1* inactivation in sporadic breast and ovarian tumors. *Journal of the National Cancer Institute*, **92**, 564–569.

Esteller, M., *et al*. (2000b). Analysis of *Adenomatous Polyposis coli* promoter hypermethylation in human cancer. *Cancer Research*, **60**, 4366–4371.

Esteller, M., *et al*. (2001). A gene hypermethylation profile of human cancer. *Cancer Research*, **61**, 3225–3229.

Ferguson, A. T., *et al*. (2000). High frequency of hypermethylation at the 14-3-3 locus leads to gene silencing in breast cancer. *Proceedings of the National Academy of Sciences of the USA*, **97**, 6049–6054.

Fuks, F., *et al*. (2000). DNA methyltransferase Dnmt1 associates with histone deacetylase activity. *Nature Genetics*, **24**, 88–91.

Giovannucci, E., *et al*. (1993). Folate, methionine, and alcohol intake and risk of colorectal adenoma. *Journal of National Cancer Institute*, **85**, 875–884.

Hardison, R. C. (2000). Conserved noncoding sequences are reliable guides to regulatory motifs. *Trends in Genetics*, **16**, 369–372.

Hark, A. T., *et al*. (2000). CTCF mediates methylation-sensitive enhancer-blocking activity at the *H19/Igf2* locus. *Nature*, **405**, 486–489.

Hendrich, B. and Bird, A. (2000). Mammalian methyltransferases and methyl-CpG-binding domains: proteins involved in DNA methylation. *Current Topics in Microbiology and Immunology*, **249**, 35–54.

Herman, J. G. and Baylin, S. B. (2000). Promoter-region hypermethylation and gene silencing in human cancer. *Current Topics in Microbiology and Immunology*, **249**, 35–54.

Herman, J. G., *et al*. (1998). Incidence and functional consequences of *hMLH1* promoter hypermethylation in colorectal carcinoma. *Proceedings of the National Academy of Sciences of the USA*, **95**, 6870–6875.

Holliday, R. (1991). Mutations and epimutations in mammalian cells. *Mutation Research*, **250**, 351–363.

Hoshikawa, Y., *et al*. (1994). Trichostatin A induces morphological changes and gelsolin expression by inhibiting histone deacetylase in human carcinoma cell lines. *Experimental Cell Research*, **214**, 189–197.

Hu, J. F., *et al*. (1998). The role of histone acetylation in the allelic expression of the imprinted human insulin-like growth factor II gene. *Biochemical Biophysical Research Communications*, **251**, 403–408.

Issa, J. P. (2000). CpG-island methylation in aging and cancer. *Current Topics in Microbiology and Immunology*, **249**, 101–118.

Issa, J. P., *et al*. (1997). *HIC1* hypermethylation is a late event in hematopoietic neoplasms. *Cancer Research*, **57**, 1678–1681.

Jones, P. A. and Laird, P. W. (1999). Cancer epigenetics comes of age. *Nature Genetics*, **21**, 163–167.

Jones, P. L., *et al*. (1998). Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nature Genetics*, **19**, 187–191.

Kanduri, C., *et al*. (2000). Functional association of CTCF with the insulator upstream of the *H19* gene is parent of origin-specific and methylation sensitive. *Current Biology*, **10**, 853–856.

Killian, J. K., *et al*. (2000). *M6P/IGF2R* imprinting evolution in mammals. *Molecular Cell*, **5**, 707–716.

Knudson, A. G., Jr. (1971). Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences of the USA*, **68**, 820–823.

Laborda, J. (2000). The role of the epidermal growth factor-like protein dlk in cell differentiation. *Histology and Histopathology*, **15**, 119–129.

Laird, P. W. (1997). Oncogenic mechanisms mediated by DNA methylation. *Molecular Medicine Today*, **3**, 223–229.

Laird, P. W. (2000). Mouse models in DNA-methylation research. *Current Topics in Microbiology and Immunology*, **249**, 119–134.

Laird, P. W. and Jaenisch, R. (1996). The role of DNA methylation in cancer genetics and epigenetics. *Annual Review of Genetics*, **30**, 441–464.

Li, E., *et al*. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*, **69**, 915–926.

Li, E., *et al*. (1993). Role for DNA methylation in genomic imprinting. *Nature*, **366**, 362–365.

Liggett, W. H. and Sidransky, D. (1998). Role of the *p16* tumor suppressor gene in cancer. *Journal of Clinical Oncology*, **16**, 1197–1206.

Liu, P., *et al*. (2001). Regulation of an endogenous locus using a panel of designed zinc finger proteins targeted to accessible chromatin regions. *Journal of Biological Chemistry*, **276**, 11323–11334.

Lutz, M., *et al*. (2000). Transcriptional repression by the insulator protein CTCF involves histone deacetylases. *Nucleic Acids Research*, **28**, 1707–1713.

MacPhee, D. G. (1998). Epigenetics and epimutagens: some new perspectives on cancer, germ line effects, and endocrine disrupters. *Mutation Research*, **400**, 369–379.

Magdinier, F., *et al*. (2000). Regional methylation of the 5′ end CpG island of *BRCA1* is associated with reduced gene expression in human somatic cells. *FASEB Journal*, **14**, 1585–1594.

Masters, C. and Crane, D. (1998). On the role of the peroxisome in cell differentiation and carcinogenesis. *Molecular and Cellular Biochemistry*, **187**, 85–97.

Nakagawa, H., *et al*. (2001). Loss of imprinting of the insulin-like growth factor II gene occurs by biallelic methylation in a core region of *H19*-associated CTCF-binding sites in colorectal cancer. *Proceedings of the National Academy of Sciences of the USA*, **98**, 591–596.

Nan, X., *et al*. (1998). Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature*, **393**, 386–389.

Peters, J. M., *et al*. (1997). Role of PPAR alpha in the mechanism of action of the nongenotoxic carcinogen and peroxisome proliferator, Wy-14,643. *Carcinogenesis*, **18**, 2029–2033.

Pfeifer, G. P., *et al*. (2000). Mutation hotspots and DNA methylation. *Current Topics in Microbiology and Immunology*, **249**, 1–20.

Rainier, S., *et al*. (1993). Relaxation of imprinted genes in human cancer. *Nature*, **362**, 747–749.

Reik, W., *et al*. (2000). Igf2 imprinting in development and disease. *International Journal of Developmental Biology*, **44**, 145–150.

Rice, J. C. and Futscher, B. W. (2000). Transcriptional repression of *BRCA1* by aberrant cytosine methylation, histone hypoacetylation and chromatin condensation of the *BRCA1* promoter. *Nucleic Acids Research*, **28**, 3233–3239.

Riggs, A. D. and Porter, T. N. (1996). Overview of epigenetic mechanisms. In: Russo, V.E.A., *et al*. (eds), *Epigenetic Mechanisms of Gene Regulation*. 29–45 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY).

Roberts, R. (1999). Peroxisome proliferators: mechanism of adverse effects in rodents and molecular basis for species differences. *Archives of Toxicology*, **73**, 413–418.

Robertson, K. D. and Wolffe, A. P. (2000). DNA methylation in health and disease. *Nature Reviews in Genetics*, **1**, 11–19.

Robertson, K. D., *et al*. (2000). DNMT1 forms a complex with Rb, E2F1 and HDAC1 and represses transcription from E2F-responsive promoters. *Nature Genetics*, **25**, 338–342.

Rountree, M. R., *et al*. (2000). DNMT1 binds HDAC2 and a new co-repressor, DMAP1, to form a complex at replication foci. *Nature Genetics*, **25**, 269–277.

Sirchia, S. M., *et al*. (2000). Evidence of epigenetic changes affecting the chromatin state of the retinoic acid receptor $\beta 2$ promoter in breast cancer cells. *Oncogene*, **19**, 1556–1563.

Srivastava, M., *et al*. (2000). H19 and Igf2 monoallelic expression is regulated in two distinct ways by a shared *cis* acting regulatory region upstream of H19. *Genes and Development*, **14**, 1186–1195.

Thorvaldsen, J. L., *et al*. (1998). Deletion of the H19 differentially methylated domain results in loss of imprinted expression of H19 and Igf2. *Genes and Development*, **12**, 3693–3702.

Tycko, B. (2000). Epigenetic gene silencing in cancer. *Journal Clinical Investigation*, **105**, 401–407.

Vu, T. H. and Hoffman, A. R. (1994). Promoter-specific imprinting of the human *insulin-like growth factor-II* gene. *Nature*, **371**, 714–717.

Vu, T. H. and Hoffman, A. R. (2000). Comparative genomics sheds light on mechanisms of genomic imprinting. *Genome Research*, **10**, 1660–1663.

Wachsman, J. T. (1997). DNA methylation and the association between genetic and epigenetic changes: relation to carcinogenesis. *Mutation Research*, **375**, 1–8.

Wylie, A. A., *et al*. (2000). Regulatory motifs of the novel imprinted domain, *DLK1/GTL2*, mimic those of *IGF2/H19*. *Genome Research*, **10**, 1711–1718.

Zingg, J. and Jones, P. A., (1997). Genetic and epigenetic aspects of DNA methylation on genome expression, evolution, mutation, and carcinogenesis. *Carcinogenesis*, **18**, 869–882.

FURTHER READING

Baylin, S. B., *et al*. (2001). Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer. *Human Molecular Genetics*, **10**, 687–692.

Cheung, P., *et al*. (2000). Signaling to chromatin through histone modifications. *Cell*, **103**, 263–271.

Murphy, S. K. and Jirtle, R. L. (2000). Imprinted genes as potential genetic and epigenetic toxicological targets. *Environmental Health Perspectives*, **108(Suppl. 1)**, 5–11.

Reik, W. and Walter, J. (2001). Genomic imprinting: parental influence on the genome. *Nature Reviews Genetics*, **2**, 21–32.

Wade, P. A. (2001). Transcriptional control at regulatory checkpoints by histone deacetylases: molecular connections between cancer and chromatin. *Human Molecular Genetics*, **10**, 693–698.

Websites

http://www.geneimprint.com (Genomic Imprinting website).

http://www3.mdanderson.org/leukemia/methylation (MD Anderson resource on CpG island methylation in ageing and cancer).

http://dnamethsoc.server101.com/ (DNA Methylation Society homepage).

http://www.cancer.med.umich.edu/learn/1learn.htm (University of Michigan Comprehensive Cancer Center homepage).

http://condor.bcm.tmc.edu/ermb/tgdb (The Tumor Gene Database).

http://nhgri.nih.gov/histones/ (The Histone Sequence Database).

Chapter 24

Infectious Agents and Cancer

Robert Newton
ICRF Radcliffe Infirmary, Oxford, UK

CONTENTS

INTRODUCTION

Infections play an important role in cancer. With the exception of canine venereal transmissible sarcoma, cancer is not itself contagious, but the underlying cause can be. Not only is the process whereby infection leads to cancer important for gaining insights into oncogenesis, but also control or elimination of infection holds promise for cancer prevention.

Historical Perspective

Transmissible agents have a venerable part in the history of cancer research. In 1911, Peyton Rous, often considered to be the father of tumour virology, was the first to demonstrate the acellular transmission of a sarcoma between chickens (the term 'virus' had not yet been coined). The research community was not receptive to the notion that a chronic disease may have an infectious cause and it was to be 55 years before Rous received the Nobel Prize for his seminal discovery. In the 1930s, Shope discovered oncogenic pox viruses and papillomaviruses in rabbits. In 1936, Bittner demonstrated that predisposition to breast cancer in C3H mice was transmitted in breast milk. In 1951, Gross discovered the first murine leukaemia virus and, in 1960, Hilleman identified SV40 virus as a contaminant of polio vaccine grown in monkey kidney cultures. However, the concept that infections might cause chronic diseases, such as cancer, can be traced back even further. For example, in the nineteenth century, the simple epidemiological observation that cancer of the uterine cervix was relatively common in prostitutes, but unknown in celibate nuns, led to the suggestion that the cause might be linked to sexual behaviour and perhaps even be sexually transmitted. In 1905, several years before Rous published his work on chicken sarcomas, Goebel drew attention to 'the occurrence of bladder tumours due to bilharziasis' (shistosomiasis).

In 1964, the first human tumour virus (the Epstein–Barr virus (EBV)) was discovered using electron microscopy, in Burkitt lymphoma cells, by Epstein, Achong and Barr. Later, EBV was also detected in undifferentiated nasopharyngeal carcinoma and subsequently in several other tumours. By the 1970s, cancer viruses were in fashion. President Nixon 'declared war' on cancer (National Cancer Act, 1971) and funding was increased for the National Cancer Institute's 'special virus cancer programme.' Although the 1970s saw many important developments, including the discovery of oncogenes and tumour-suppressor genes (*TP53*), no new cancer viruses were identified and interest began to wane. This was to change in the early 1980s, with several major discoveries. In 1980, Poiesz, Gallo and colleagues discovered the human T cell leukaemia virus, which is associated with endemic leukaemia/lymphoma, particularly in Southern Japan and the Caribbean. In 1981, the large-scale prospective epidemiological studies of Beasley *et al*., in Tiawan, confirmed the long-suspected causal association between the hepatitis B virus and liver cancer. In 1983, zur Hausen and colleagues isolated HPV 16 from a human cervical cancer specimen, Marshall and Warren identified *Helicobacter pylori* (later associated with gastric cancer), and the HIV (discovered by Barré-Sinoussi) emerged as an important cause of several cancers. Hepatitis C virus, a cause of liver cancer, was discovered in 1989 and, in 1994, Chang and Moore identified the Kaposi sarcoma-associated herpesvirus (HHV-8).

The Global Burden of Cancers Caused by Infections

It is estimated that approximately 15% of cancers (between about 1.2 and 1.5 million cases per year, worldwide) are attributable to viral (11%), bacterial (4%) and helminth (0.1%) infections (**Table 1**). Collectively, infectious agents are the most important known cause of cancer after tobacco.

Table 1 Cancers attributable to infections (these are conservative estimates, adapted from Parkin *et al.*, 1999)

Infection	Cancer(s)	No. of cases worldwide
Human papillomaviruses	Cervical cancer	360 000
	Other female genital cancers	
Hepatitis B virus	Liver cancer	230 000
Epstein–Barr virus	Burkitt lymphoma	100 000
	Hodgkin disease	
	Nasopharyngeal cancer	
Epstein–Barr virus/HIV	Non-Hodgkin lymphoma	9 000
Human herpesvirus-8/HIV	Kaposi sarcoma	45 000
Human T cell leukaemia virus	Leukaemia	3 000
Hepatitis C virus	Liver cancer	110 000
Helicobacter pylori	Gastric carcinoma	350 000
	Gastric lymphoma	
Schistosomes	Bladder cancer	10 000
Liver flukes	Cholangiocarcinoma	1 000

A better understanding of the role of infectious agents in the aetiology of cancer is a public health imperative, because such cancers are theoretically preventable by vaccination or early treatment of infection. Furthermore, cancer-causing infections often cause substantial morbidity and mortality from non-malignant conditions. Therefore, an additional benefit of any scheme to reduce the burden of cancers caused by infections would also involve a reduction in the incidence of other diseases.

The majority of infection attributable cancers (perhaps 1 million cases per year) occur in the developing world, reflecting the higher prevalence of the major causative agents, particularly hepatitis B, human papillomaviruses (HPV), *H. pylori* and human immunodeficiency virus (HIV). It is conservatively estimated that if these infectious diseases were controlled, up to one in four cancers in developing countries and one in 10 cancers in developed countries might be prevented. This chapter briefly reviews the association between certain infections and cancer, outlines the mechanism by which disease might be caused (if known) and presents comments on the potential for prevention of such tumours.

VIRUSES

Human Papillomaviruses (HPV)

The papillomaviruses are double-stranded DNA viruses. About 100 subtypes have been distinguished to date, many of which can infect humans, and at least as many again await formal characterization. Several specific subtypes have been associated with cancer in humans. Indeed, more cancers are attributable to HPV infection than to any other transmissible agent (**Table 1**). Most of these are cervix cancers, but other anogenital tumours and cancers at distant sites may also be caused by HPV.

After cancer of the breast, cervix cancer is the most common female malignancy worldwide. The risk of disease is increased among women reporting multiple sexual partners, early age of first sexual intercourse and among those whose male partners have multiple partners, all features that implicate a sexually transmitted aetiological agent. Of the known HPV types, about 30 can infect the female genital tract. Some of these are associated with benign lesions, such as warts (e.g. HPV 6 and 11) while others, so-called 'high-risk' types, are associated with invasive cancer and advanced precancerous lesions (e.g. HPV 16, 18, 31, 33, 45, 51, 52, 58, 59).

HPV infection is one of the most common sexually transmitted infections of women and probably also of men. Viral DNA is detectable in a large proportion of women shortly after becoming sexually active and the main determinant of infection is the number of sexual partners. The prevalence of infection varies between populations, but is of the order of 20–30% in women aged 20–24 years, declining to 5–10% in women over the age of 40 years. Follow-up of young, sexually active cohorts of women suggest that the incidence of HPV infection is about 15% per year, with more than 50% of women becoming infected at some stage in their lives. Most infections are cleared spontaneously, but a small proportion become persistent and it is these that carry the risk of neoplastic change. Much less is known about HPV infection in men and their reservoir of infection has not been clearly identified, although the glans penis and internal meatus are the most likely locations.

HPV infection of the genital tract can be latent, or associated with cellular alterations known as cervical intraepithelial neoplasia (CIN) (sometimes called squamous intraepithelial lesions (SIL)), graded according to severity from 1 to 3, depending on the degree of nuclear and cytoplasmic change. All grades are considered to be manifestations of HPV infection; grade 1 lesions are generally

benign and often resolve, whereas high-grade lesions are more likely to persist and progress. The time taken to progress from one grade to another and then to invasive disease is unknown, but has been estimated to be several years, possibly more than a decade. The features associated with HPV-mediated progression of CIN are not clear, but persistent infection is more likely in older women infected with 'high-risk' HPV subtypes. The quality of the immune response to infection is likely to be important and some HLA associations have been reported. Other possible factors include high parity, concomitant venereal infections and tobacco use (see IARC (1995) and Herrero and Muñoz (1999) for a more thorough review).

The evidence for a causal role of HPV infection in the aetiology of CIN and invasive cancer of the cervix is overwhelming. Virtually all squamous cell cancers and more than 90% of adenocarcinomas of the cervix contain HPV-DNA. A recent study by Walboomers *et al*. (1999) identified HPV-DNA in 99.7% of almost 1000 invasive cancers in a worldwide study. HPV 16 accounts for about 55% of tumours and is particularly dominant in Western countries. HPV 18 and 45 are relatively more important in tropical areas and account for about 15 and 10% of cervix cancers, respectively. Data from case-control and cohort studies consistently suggest that the risk of invasive cervix cancer or CIN is very high in association with HPV infection (relative risks of greater than 50 in HPV infected women compared with uninfected women). Furthermore, the risk increases with increasing viral load, as measured by the amount of HPV-DNA.

Molecular analyses support the epidemiological evidence. The high-risk HPV subtypes exhibit transforming potential and can immortalize cells, processes fundamental to the development of malignancy. The HPV genome contains two oncogenes, E6 and E7, and despite the frequent loss of much of the viral genome in cervical cancer cells, these regions are consistently maintained and expressed. The protein products of the E6 and E7 oncogenes affect the normal function of cellular proteins, essential for regulating cell growth. TP53 and Rb are important tumour-suppressor proteins and loss of their function is a common theme in most human cancers, regardless of origin. E6 interferes with p53 and E7 with the Rb protein and this is likely to be central to the oncogenic activity of certain HPVs (Phillips and Vousden, 1999).

There is some evidence that high-risk HPV types also play a role in the aetiology of cancers at other anogenital sites (IARC, 1995). These include cancers of the anus, penis, vagina and certain histological subtypes of vulval cancer. However, these tumours are sufficiently rare (particularly in comparison with cancer of the cervix), as to have only a limited impact on public health. A proportion of tumours of the head and neck (including conjunctiva), oesophagus, lung, bladder and prostate may also be associated with HPV infection, although the evidence remains scant. Of more public health importance is the suggestion that HPV subtypes 5 and 8 may cause a proportion of squamous cell carcinomas of the skin. Epidermodysplasia verruciformis is a rare hereditary disease characterized by impaired cell mediated immunity. Patients with this condition suffer from multiple skin warts, which often progress to squamous carcinoma. HPV 5 and 8 have been consistently identified in these tumours (McGregor and Proby, 1996); similar subtypes have been found in warts and malignant skin lesions of renal transplant and other immunosuppressed patients. The role of HPV 5 and 8 in the aetiology of squamous cell skin cancers occurring in the general population is uncertain.

Prevention of HPV-associated cancers, particularly cervix cancer, is a public health priority. In the long-term, vaccination holds the most promise for the eradication of such tumours, but of more immediate value is screening. Cervical cancer screening programmes rely on the detection of treatable precancerous lesions by exfoliative cervical cytology, and are known to be effective at reducing the incidence of and mortality from invasive disease. Several studies suggest that detection of high-risk HPV types is more sensitive for detecting CIN than conventional cytology, although the rates of false positives are also higher (Cuzick, 1999). The inclusion of HPV detection as part of the normal cytological screening programme may, therefore, lead to better disease detection, increased intervals between screens and early cessation of screening in women without HPV infection (IARC, 1992). Large trials are needed to determine the value for prevention and cost effectiveness of such an approach.

Prophylactic HPV vaccines are based on the induction of neutralizing antibodies able to prevent infection and associated lesions. Their development has been slowed by the inability to propagate HPV in tissue culture and the consequent lack of a source of antigens, in particular, the structural proteins L1 and L2. These problems have largely been solved with the generation of virus-like particles (VLPs), obtained by expression of the major capsid proteins, which because they are indistinguishable from authentic virions except that they lack a viral genome, can induce neutralizing antibodies. Such vaccines have proven efficacy in protecting against infection in many animal models. Several VLP-based vaccines have been produced to protect against human HPV types (6, 11, 16, 18, 31 and 33) and are being tested in phase I and II trials (Coursaget and Muñoz, 1999). It will be many years before large-scale phase III clinical trials yield results, but such vaccines hold great promise for the ultimate prevention of HPV-associated cancers.

Hepatitis B Virus

The role of chronic infection with hepatitis B virus (HBV) in the aetiology of primary liver cancer, specifically hepatocellular carcinoma, is well established (IARC, 1994a). Hepatitis B is one of the most common infectious

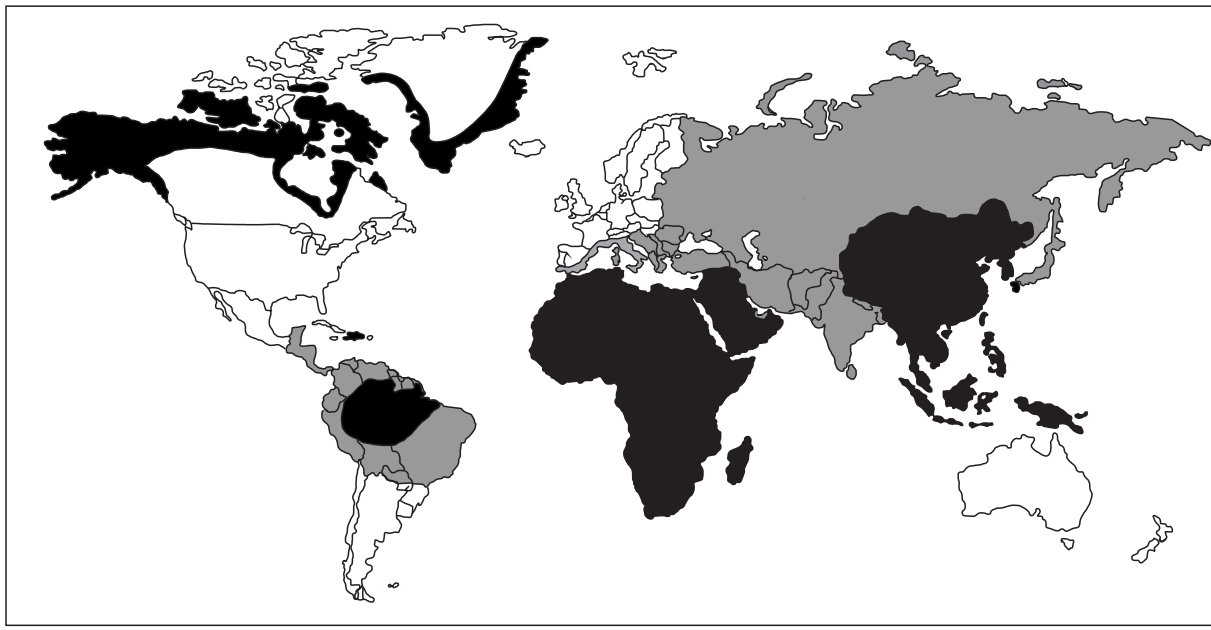


Figure 1 The geographical distribution of hepatitis B infection. Black, ≥8%, high; grey, 2–7.9%, intermediate; white, <2%, low (from WHO).

diseases worldwide, with between 300 and 350 million chronic carriers – up to 10% of the population in high prevalence areas – two-thirds of whom will develop chronic hepatitis. Of those, 20–25% will eventually die from primary liver cancer or cirrhosis. As a result, primary liver cancer is one of the most common cancers, particularly in parts of sub-Saharan Africa, China and South-east Asia, where the virus is prevalent (**Figure 1**). The hepatitis B virus is responsible for causing up to a quarter of a million cancers per year (**Table 1**), about 50–55% of the total number of liver cancers, and is the second most important oncogenic virus (Parkin *et al*., 1999).

Hepatitis B virus belongs to a group of hepatotropic double-stranded DNA viruses (with a single-stranded region of variable length), called hepadnaviruses and is characterized by the fact that it replicates through an RNA intermediate, via reverse transcription. The genome contains four genes, S, C, pol (the largest gene, coding for reverse transcriptase) and X. The virion, or 'Dane particle,' has an outer protein coat encoded by the S gene (hepatitis B surface antigen (HBsAg), the so-called 'Australia antigen,' discovered by Blumberg in 1963, in the serum of an aboriginal man), and an inner protein core (hepatitis B c antigen (HbcAg)) coded by the C gene. The terminal region of the C gene encodes the hepatitis B e antigen (HbeAg) and the X gene encodes a protein that upregulates transcription from all the viral and some cellular promoters.

Transmission of HBV occurs from a person with acute infection or carrier status (i.e. HBsAg positive on at least two occasions, 6 months apart) who has circulating virus. Individuals who are HBeAg positive are particularly infectious since the presence of this antigen correlates with high serum levels of HBV-DNA. The mode of transmission is not completely clear but varies with age, occurring primarily at three stages in life. First, transmission can occur from a mother to her child. Neonates born to HBeAg-positive mothers have about a 90% chance of becoming infected, whereas children of HBeAg-negative mothers have a 30% probability of infection. This is a particularly important means of transmission in China, but is probably less important in Africa where, for reasons that are not known, fewer infected mothers are HBeAg positive. Transmission in childhood is probably the most important age for infection globally and is associated with residence with infected siblings; the exact means of virus spread within households is unknown. In adult life, the major routes of transmission are by needles and sexual intercourse. Since the introduction of screening for HBV in blood products, the predominant mode of needle transmission is by the sharing of needles amongst intravenous drug users. Sexual transmission has been most clearly documented in countries of Europe and North America, where the prevalence of HBV is low, but may also have a limited role in endemic countries (IARC, 1994a).

The principle risk factor for HBV-induced liver cancer (and cirrhosis) is persistence of replicative infection (as it is with human papillomavirus infection and cancer of the cervix) and a key determinant of persistence is the age of infection. About 90% of children infected in the perinatal period become chronic carriers; up to 5 years old, around 20–30% of those infected become chronic carriers and above this age, the proportion is probably less than 5%. Presumably, the relationship between persistence of

infection and age is due to changes in the ability of the immune system to contain the virus, but the determinants of this are unclear (Wild and Hall, 1999). At all ages, males are more likely than females to become chronic carriers. Other factors are known to contribute to the risk of liver cancer in those chronically infected with HBV, including coinfection with the hepatitis C virus. Similarly, exposure to aflatoxins, common in parts of Africa and Asia, leads to a roughly 10-fold increase in the risk of malignancy above the excess associated with HBV infection.

The mechanisms whereby HBV may induce primary liver cancer are not fully understood. In most instances, the tumour is preceded by several decades of chronic hepatitis or cirrhosis and the associated rapid cell turnover may render host DNA more susceptible to malignant change. Mutagenesis may be induced by the associated inflammatory response or exogenous carcinogens, or may result from genetic instability following integration of the viral genome. Another possible mechanism involves the activation of cellular genes by the hepatitis B virus X gene product: it encodes a protein that may interact with the TP53 and deregulate the cell cycle.

Prospects for prevention of HBV-induced liver cancer are good. Screening of blood and organ donors has reduced the spread of infection among adults in developed countries. Similarly, screening for infection among high-risk pregnant women, such as those from endemic countries, is of value. There is ample evidence that the use of hepatitis B immunoglobulin within hours of birth, followed by hepatitis B vaccine on multiple occasions, is effective in reducing the infection rate in infants born to infected mothers by more than 80% and the carrier rate by more than 90% (reviewed by Cuzick *et al*., 1999). The best hope for prevention, however, lies with mass vaccination against HBV, particularly in populations where infection is most prevalent. Furthermore, there is also some interest in the development of therapeutic vaccines; these might eventually be used to treat liver cancer patients, by inducing an effective immune response to virus-infected cells.

The safety and effectiveness of prophylactic HBV vaccines have been clearly demonstrated in numerous studies, since they became widely available in 1982. Early vaccines were based on purified HbsAg from patients with HBV-induced liver cancer, but these were superseded in 1987 by vaccines produced using recombinant technology. Such vaccines are highly immunogenic and induce protective antibodies in 90–100% of healthy recipients (reviewed by Coursaget and Muñoz, 1999). Some studies have demonstrated protection from infection for up to 12 years. In a study in the Gambia with 10 years of follow-up, hepatitis B vaccine prevented 83% of infections and 94% of chronic carriage (Gambia Hepatitis Study Group, 1987). Although it will be many years before an effect on the incidence of liver cancer is noted in clinical trials, the introduction of mass vaccination in Taiwan has been associated with a sharp decline in the incidence of liver cancer in children (Chang *et al*., 1997). Currently, 80 countries, mostly in high-risk endemic areas, have routine hepatitis B vaccination programmes. Despite this, it is estimated that only a third of infants who might benefit from vaccine worldwide actually receive it. (See also chapter on *Human DNA Tumour Viruses*.)

Epstein–Barr Virus (EBV)

The Epstein–Barr virus is a ubiquitous human herpesvirus of the family Gammaherpesviridae that infects more than 90% of the world's population. The virion consists of a core wrapped with DNA and is contained within a capsid, which is surrounded by a membranous envelope with glycoprotein spikes. The genome is a relatively large 172-kb double-stranded DNA molecule encoding about 100 genes. There are two major subtypes of EBV which differ with respect to some of the genes coding for nuclear proteins in latently infected cells, but there is no evidence that one subtype is more strongly associated with disease than the other.

The EBV is transmitted orally, either by exchange of virus particles or infected cells in buccal fluid and, like other herpesviruses, it establishes a latent infection, with life-long persistence in the infected host. Asymptomatic primary infection usually occurs in childhood, although in more developed countries infection may be delayed. If primary infection occurs in adolescence, about 50% of cases may develop the clinical syndrome of infectious mononucleosis and most of what is known of the events occurring at primary infection are extrapolations from the study of such individuals.

It is not known whether the initial target cell for orally transmitted virus is a B lymphocyte, made accessible by damage to the oral mucosa, or an epithelial cell of the mucosa itself. However, following primary infection, foci of productive (lytic) infection are established in the oropharyngeal mucosa and a large pool of latently infected B cells can be found both in the blood and in lymphoid tissues. The overgrowth of virally transformed B cells is controlled by specific cytotoxic T cell responses, the absence of which (in allograft recipients and others with impaired T cell function) can result in EBV-driven lympho-proliferation and even lymphoma.

A recent review of the role of the EBV in the aetiology of cancer concluded that there is sufficient evidence for the carcinogenicity of EBV in the causation of several types of lymphoma (including Burkitt lymphoma, Hodgkin disease and immunosuppression-related lymphomas) and undifferentiated nasopharyngeal carcinoma (IARC, 1997). Other cancer types have also been linked to infection with EBV, although the evidence for causality is less clear. In total, it is estimated that the EBV may be responsible for up to about 100 000 cancers per year worldwide (**Table 1**; Parkin *et al*., 1999).

Burkitt lymphoma occurs throughout the world with varying frequency, but everywhere it represents a

significant proportion of malignant lymphomas in children. There are two broad types of Burkitt lymphoma, which although histologically indistinguishable, differ in several other important ways. The 'endemic' form of the disease was first described in 1958 in parts of sub-Saharan Africa, South-east Asia and Papua New Guinea. It is particularly frequent where malarial infection is heavy and widespread. In these areas it is the most common childhood malignancy, with a peak incidence between 5 and 8 years old. It often presents with facial lesions and is associated with infection with EBV. Viral DNA is almost invariably identified in tumour tissue in monoclonal form, indicating that the original malignant clone must have arisen from a single virus infected cell. Furthermore, high antibody titres to EBV infection in young children are predictive of the subsequent development of the tumour (de-Thé *et al*., 1978). This is in contrast to the sporadic form of Burkitt lymphoma seen throughout the rest of the world, which usually presents with abdominal lesions, occurs at older ages (particularly in teenagers) and is only associated with EBV infection in about a third of cases. In Western populations, sporadic Burkitt lymphoma is about 1000 times more common in HIV-infected individuals than in those uninfected with HIV and about half of these tumours are also associated with EBV infection. Both the endemic and sporadic Burkitt lymphoma involve c-*MYC* gene translocations, but the breakpoints in each type are different: in the endemic disease, the breakpoints tend to be far upstream of the oncogene, whereas in the sporadic form, the breakpoints tend to be adjacent to or within the oncogene. It is therefore possible that the endemic and sporadic forms of Burkitt lymphoma have different aetiologies, with only a proportion (mostly the endemic cases) being caused by EBV infection.

Nasopharyngeal carcinoma of undifferentiated or poorly differentiated type is an epithelial EBV-associated tumour, which, like Burkitt lymphoma, is characterized by marked geographic and population differences in incidence. It is common in southern China and South-east Asia, where it may represent up to 20% of all cancer cases. It also occurs relatively frequently in Eskimo populations and in Mediterranean Africa. Despite this geographical restriction, more cases of nasopharyngeal cancer are attributed to infection with the EBV, than any other cancer (Parkin *et al*., 1999).

The identification of an association between nasopharyngeal cancer and EBV occurred by chance, when patients with the tumour were chosen as controls in a case-control study of Burkitt lymphoma and were found to have high antibody titres to the virus. Specifically, elevated immunoglobulin A (IgA) antibodies to the EBV viral capsid antigens (VCA) predate the development of nasopharyngeal cancer by several years and are also correlated with tumour burden and recurrence. Measurement of IgA levels has formed the basis of a screening programme in southern China, where about 5% of individuals aged 30 years or more are positive for IgA/VCA and 5–12% of these have nasopharyngeal cancer (IARC, 1997). In addition, clonal EBV-DNA is consistently detected in tumour tissue (but not in normal nasopharyngeal epithelium), suggesting that the tumour develops from a single EBV-infected cell.

Because the EBV is ubiquitous, the geographical variation in the incidence of nasopharyngeal cancer may be explained by the combined influences of genetic predisposition in particular racial groups and of local environmental or dietary factors, although the exact nature of these effects is not clear. Salted and preserved food in the diet and a lack of fresh fruits and vegetables are currently the most clearly established cofactors. Another possibility is that particular strains of EBV, which carry a higher risk for nasopharyngeal carcinoma, are common where the disease occurs, although evidence for this remains scant.

Hodgkin's disease is a malignant lymphoma characterized by the loss of lymph node architecture, with the majority of infiltrating cells being benign. Indeed, the malignant cells – the Reed–Sternberg cells – constitute only about 2% of tumour mass. The disease occurs worldwide and in Western populations it has a distinctive age distribution with two peaks of incidence, at ages 25–30 and over 45 years. Sero-epidemiological studies indicate that high antibody titres to EBV precede the development of Hodgkin's disease and a history of infectious mononucleosis is a strong risk factor. Furthermore, about half of cases have evidence of clonal EBV-DNA in tumour tissue. Parkin *et al*. (1999) estimated that about half of the 60 000 new cases annually are related to infection with the EBV.

Some other lymphomas may be related to EBV, but the overall number of cases is likely to be relatively small. Non-Hodgkin lymphomas associated with HIV, or immunosuppressive therapy, are characterized by a number of features, including an aggressive clinical course. High-grade disease is common and extra-nodal sites are often involved, with lesions in the central nervous system being virtually unknown, except in the immunosuppressed. The most common subtype of non-Hodgkin lymphoma both in those with HIV infection and in immunosuppressed transplant recipients, is B cell immunoblastic lymphoma. Post-transplant immunoblastic lymphomas are nearly always associated with EBV and probably represent the end result of an EBV-driven lymphoproliferation in the absence of effective T cell immunity. EBV sequences are detectable in about 50% of HIV-associated immunoblastic lymphomas (although 100% of primary cerebral lymphomas), suggesting that other factors may also be important (Newton *et al*., 1999a). Several nonspecific host factors have been suggested to play a role in lymphomagenesis in immunosuppressed individuals, such as disrupted immunosurveillance, chronic antigenic stimulation and cytokine dysregulation, all of which might be responsible for expanding the B cell population from which a lymphoma subsequently develops.

Infection with the EBV has also been associated with a number of very rare T cell lymphoma subtypes and with

lymphoepithelial carcinomas of the stomach, lung and salivary gland. In addition, smooth muscle tumours (leiomyosarcomas) in immunosuppressed individuals uniformly contain EBV. The number of such cancers, however, is small.

The development of vaccines to control the diseases associated with EBV infection are at an early stage of development, but may be used in a number of ways. First, prophylactic vaccines may modify or prevent primary infection. Normally, infection occurs during the first few years of life, but such vaccines would be particularly useful for the prevention of infectious mononucleosis, which results in 50% of individuals in whom infection is delayed until adolescence. Second, postinfection vaccination may be used to modify the existing immune status of an infected individual and could have some value in preventing the development of nasopharyngeal carcinoma in those who have high IgA antibodies against EBV. Finally, therapeutic vaccination might be selectively targeted against viral antigens expressed in tumour cells. However, for the time being, the effect of any form of vaccination against the EBV remains speculative.

Human Herpesvirus-8 (HHV-8/KSHV)

Before the HIV epidemic, Kaposi sarcoma showed a greater geographical variation in incidence than almost any other cancer. It was as common in parts of sub-Saharan Africa, such as Uganda and eastern Zaire, as colon cancer is in Europe and the USA, representing up to 9% of all cancers in men. Kaposi sarcoma was also endemic, although much rarer, in countries around the Mediterranean, particularly Italy, Greece and the Middle East, but was almost nonexistent elsewhere in the world, except in immigrants from those endemic countries. In all of these areas, Kaposi sarcoma was more common in men than in women (Newton *et al*., 1999a).

It was the appearance of aggressive forms of Kaposi sarcoma in the USA in the early 1980s that heralded the onset of the HIV epidemic in Western countries. Although the incidence of Kaposi sarcoma has increased in populations at high risk of HIV in northern Europe and the USA, it existed at such a low level before the onset of the epidemic that it remains a relatively rare tumour. However, parts of Africa with a high prevalence of HIV and where Kaposi sarcoma was relatively common even before the era of acquired immunodeficiency syndrome (AIDS), have seen an explosion in the incidence of the disease. In the last 10–15 years the incidence of Kaposi sarcoma has increased about 20-fold in Uganda and Zimbabwe, such that it is now the most common cancer in men and the second most common in women (IARC, 1997).

In 1994, Chang, Moore and colleagues identified sequences of a new herpesvirus in a biopsy specimen of Kaposi sarcoma from an HIV-infected homosexual man, using representational difference analysis. The virus – human herpesvirus-8 (HHV-8) or Kaposi sarcoma-associated herpesvirus (KSHV) – has been consistently associated with Kaposi sarcoma and is now considered to be the principal cause of the disease. Genomic sequences of HHV-8 are present in tumour cells of Kaposi sarcoma lesions (specifically in the spindle cells, which constitute the bulk of the tumour) in virtually all subjects, but are not found in other tissues (with the exception of blood). The presence of HHV-8, detected by polymerase chain reaction (PCR) or serology, in peripheral blood, predicts the subsequent development of Kaposi sarcoma, particularly in individuals with high anti-HHV-8 antibody titres.

HHV-8 is not a ubiquitous virus, but is most prevalent in groups or populations at highest risk of developing Kaposi sarcoma, such as HIV-infected homosexual men in the USA and in African populations where the tumour has long been endemic (Boshoff, 1999). The proportion of adults in the general population with antibodies against HHV-8 ranges from fewer than 5% in northern America and northern Europe to around 10% in southern Europe and more than 30% in black Africans. In the United States and Europe, more than 30% of HIV-infected homosexual men have been found to have antibodies against HHV-8.

The modes of transmission of HHV-8 are yet to be fully elucidated. In the USA, sex between men may be an important route of transmission since this is the main behavioural risk factor for Kaposi sarcoma and indeed there is now some evidence that this is so. In some African countries, where HHV-8 is relatively common, the seroprevalence does not vary by sex and has been found to increase with age, from birth, through childhood and into adult life. This suggests that some transmission from a mother to her child and from child to child is also likely.

HHV-8 is a gamma-herpesvirus, closely related to the EBV, and infects CD19+ B cells as well as the endothelial derived spindle cells of a Kaposi sarcoma lesion. In addition to its role in Kaposi sarcoma, the virus also causes a rare type of lymphoma (primary effusion lymphoma) and a lymphoproliferative B cell disorder (a subtype of Castleman disease). The mechanism by which HHV-8 causes disease is controversial, although the genome encodes several putatively transforming genes in addition to genes encoding a number of regulatory cytokines and angio-proliferative factors that may facilitate tumour growth. HHV-8 infection alone, however, may not be enough to induce disease. Cofactors such as immunosuppression, as a result of advancing age, HIV/AIDS or therapy following tissue transplantation are presumably required.

Human T Cell Leukaemia Virus Type 1 (HTLV-1)

HTLV-1 is the main causal agent of adult T cell leukaemia/lymphoma, a disease characterized by malignant proliferation of CD4-positive T lymphocytes. Clinical features

include hypercalcaemia, lymphadenopathy, skin lesions due to leukaemic cell infiltration, involvement of the spleen and liver and immunodeficiency. The prognosis of patients with acute adult T cell leukaemia/lymphoma is poor, and few survive more than a few months following diagnosis. HTLV-1 also causes slowly progressive myelopathy (tropical spastic paraparesis) and uveitis.

HTLV-1 is an enveloped retrovirus, of the family Oncornavirinae, containing two covalently bound genomic RNA strands, which are combined with several viral enzymes, including reverse transcriptase. The prevalence of infection with HTLV-1 varies widely worldwide, with high levels in diverse geographic areas. Antibodies to HTLV-1 are found in 5–15% of indigenous adult populations in southern Japan, the Caribbean, South America, central Africa, Papua New Guinea and the Solomon Islands. Within endemic areas, clusters of especially high prevalence can occur. Carriers can be found elsewhere in the world, but are mostly individuals who moved from endemic areas. It is estimated that there are between 15 and 20 million infected individuals in the world (IARC, 1996).

Three modes of transmission of HTLV-1 have been identified: mother-to-child transmission, mainly due to breast-feeding beyond 6 months, sexual transmission, predominantly from men to women, and transmission by transfusion of cellular blood products and through intravenous drug use. Control and prevention of infection depends on reducing transmission by these three major routes. Perinatal transmission has been greatly reduced in Japan by avoidance of prolonged breast feeding and several countries have introduced universal screening of blood donors. Passive and active immunization is effective in animal models but no preventive vaccine is yet available for humans.

Adult T cell leukaemia/lymphoma (ATLL) occurs almost exclusively in areas where HTLV-1 is endemic, such as Japan, the Caribbean and West Africa, and cases described elsewhere have generally been in immigrants from those endemic regions, or their offspring. Early studies showed that infection with HTLV-1 is so closely associated with adult T cell leukaemia that it is now part of the diagnostic criteria used for defining the disease. All antibody-positive cases of adult T cell leukaemia have monoclonally integrated HTLV provirus in the malignant cells, suggesting that the tumour is an outgrowth of an individual T cell clone. The virus is able to immortalize human T lymphocytes, a property that has been related to a specific viral gene *tax*, which has been identified as a transforming factor. ATLL develops in about 2–5% of HTLV-1-infected individuals and is especially frequent among those infected early in life. No other environmental cofactors for disease have so far been identified. (See chapter on *RNA Viruses*.)

Hepatitis C Virus (HCV)

The identification of HCV in 1989 by Choo and colleagues arose from an investigation of the causes of post-transfusion non-A, non-B hepatitis. It is a single-stranded RNA virus assigned to a separate genus within the family Flaviviruses (which includes yellow fever virus and dengue virus) and is completely unrelated to the HBV. To date, six major subtypes of HCV have been identified, which have different geographical distributions. Acute infection often causes only mild illness, but it is becoming increasingly clear that HCV is responsible for substantial morbidity and mortality, particularly from chronic liver disease and hepatocellular cancer (IARC, 1994a). It may also play a role in the aetiology of other malignancies, such as non-Hodgkin lymphoma and cancers of the oral cavity (Tanaka and Tsukuma, 1999).

The prevalence of infection with HCV varies around the world and is estimated to be about 1–1.5% in Europe and the USA, about 3% in Japan and Oceania (excluding Australia and New Zealand) and up to 3.6% in Africa (Parkin *et al*., 1999). In most countries, the prevalence of infection is the same in men as it is in women and increases steadily with age. Transmission is primarily by the parenteral route and, before the introduction of screening for hepatitis C, blood transfusions were a major source of infection. Intravenous drug users comprise a substantial proportion of identified cases in western populations and health-care workers, renal dialysis patients and those with clotting disorders are also at an increased risk. Both sexual and perinatal transmission occur and household contact with an infected family member may also account for a proportion of cases. However, almost half of all hepatitis C-infected individuals have no identifiable risk factors.

In contrast to hepatitis B, as many as 85% of hepatitis C virus infections become persistent and at least two-thirds of those individuals go on to develop chronic liver disease, including hepatocellular carcinoma. The risk of liver cancer in chronically infected individuals is around 20-fold higher than in the general population (more in people coinfected with HBV) and it has been estimated that HCV causes around 110 000 cancers per year (Parkin *et al*., 1999; **Table 1**). This represents about 25% of all liver cancers, with particularly high proportions in Africa (41%), Japan (36%) and Oceania (33%). Although a similar proportion of men and women become chronic carriers of HCV, the development of cancer is more frequent in men. Alcohol and tobacco have been implicated as cofactors and may account for this discrepancy.

The mechanism by which HCV causes cancer is not clear. The virus can replicate in hepatocellular carcinoma cells, but there is no evidence that DNA sequences are integrated into the host genome. Nearly all cases of liver cancer associated with hepatitis C occur in the presence of cirrhosis or severe chronic hepatitis. Indeed, progression from chronic active hepatitis to cirrhosis to hepatocellular carcinoma has been documented in prospective studies (Tanaka and Tsukuma, 1999). In the absence of genomic integration, it is possible that the emergence of a malignant clone reflects a multifactorial process of inflammation,

necrosis and cellular regeneration. Whether HCV contributes more directly to carcinogenesis is not known.

There are currently no practical strategies for the prevention of HCV infections, with the exception of blood-screening programmes (which have greatly reduced post-transfusion hepatitis). Although evidence remains scant, safe sexual practices and distribution of clean needles to intravenous drug users might reduce a small proportion of cases and perinatal transmission might be reduced with carefully managed deliveries and limitation of breast feeding. The efficacy of postexposure prophylaxis with immunoglobulin has yet to be confirmed. Efforts to develop a vaccine have been hampered by the inability to produce large amounts of immunogen *in vitro*, the fact that correlates of immunity are ill-defined and by the lack of an animal model. Although offering the greatest hope of prevention, vaccination against hepatitis C remains only a theoretical possibility.

Human Immunodeficiency Virus (HIV)

There is little evidence that HIV has a direct oncogenic effect in relation to the development of a specific cancer. Instead, it appears to facilitate the development, via its effects on the immune system, of a number of cancers, all of which are known (or thought) to be caused by other infectious agents. The abbreviation HIV is used throughout this chapter and refers specifically to HIV-1; reports on the association of cancers with HIV-2 are infrequent.

The human immunodeficiency virus was discovered in 1983 (Barré-Sinoussi *et al*., 1983) and firmly associated with AIDS in 1984 (Gallo *et al*., 1984). It is a human retrovirus, belonging to the lentivirus sub-family and is distinguished by its single-stranded RNA genome, which replicates via a DNA intermediate through the action of the enzyme 'reverse transcriptase' and integrates into the host chromosomal DNA. Although the first cases of HIV disease were reported as recently as 1981, the epidemic continues to escalate and the World Health Organisation (WHO) estimates that over 38 million young adults have been infected, the majority in sub-Saharan Africa. The three primary routes of transmission – sexual intercourse, blood contact and from mother to infant – were proposed on the basis of the epidemiology of AIDS, even before the identification of HIV. Of those, heterosexual transmission accounts for over 80% of new infections worldwide, while the importance of contact with infected blood products is declining since the introduction of routine screening procedures in blood banks.

The immunosuppression resulting from infection with HIV is causally associated with Kaposi sarcoma and non-Hodgkin lymphoma and, in the light of data emerging from sub-Saharan Africa, with squamous cell carcinoma of the conjunctiva. Recent evidence for two other cancers, Hodgkin disease and leiomyosarcoma in children, also suggests a definite increase in risk associated with HIV infection. The scale of the excess risk of these cancers in HIV-infected compared with uninfected individuals tends to be very large, 10-fold or more. However, people with HIV infection do not experience large excess risks of most cancers, including cancer of the uterine cervix and hepatocellular carcinoma, neither of which appear to be increased markedly in people with AIDS (IARC, 1996; Newton *et al*., 1999a).

The cancers identified as being HIV-associated have been linked (with varying degrees of certainty) to other infectious agents. Kaposi sarcoma is caused by the newly discovered human herpesvirus 8. Certain types of non-Hodgkin lymphoma have been linked to infection with the EBV and HHV-8, and conjunctival carcinoma has been linked to human papillomavirus (HPV) infection in some studies, but not others (IARC, 1995; Newton, 1999a). Thus, infection with the HIV appears to facilitate the development of certain cancers with an infectious aetiology. It is not clear why other cancers which are caused by infections, such as hepatocellular carcinoma, are not also AIDS associated.

Many of the increases in cancer risk found in people with HIV disease are similar to the findings in immunodeficient children and in transplant recipients, suggesting that it is the impairment of immune function that is the major factor leading to the appearance of these tumours (Beral and Newton, 1998). In addition, the risk of Kaposi sarcoma and non-Hodgkin lymphoma increases with increasing severity of immunosuppression, suggesting that this is the principle mechanism favouring their development. Also, both tumour types have been shown to regress following treatment with highly active antiretroviral therapy, which leads to improvements in immune function (IARC, 1996).

It has been estimated that, in 1990, there were about 52 000 additional cases of cancer that were a consequence of infection with the HIV (**Table 1**, Parkin *et al*., 1999). This conservative estimate is based on the known impact of infection with the virus on just two cancers with which it has been most clearly linked: Kaposi sarcoma and non-Hodgkin lymphoma. In populations with a high prevalence of HIV infection, the impact of the epidemic is clearly reflected in cancer registry statistics, although if Kaposi sarcoma and non-Hodgkin lymphoma are excluded, there is little evidence of an increase in the incidence of all other cancers combined. In Uganda and Zimbabwe, for example, the incidence of Kaposi sarcoma has increased between 10- and 20-fold in the era of AIDS, such that it is now the most common cancer in males in both countries and amongst the most common in females. There is evidence of a reduction in risk of both Kaposi sarcoma and non-Hodgkin lymphoma in those on antiretroviral therapy. However, in the developing world, where such treatment is prohibitively expensive and therefore not widely available, the incidence of HIV-associated cancers is likely to increase with the spread of the HIV epidemic.

In the absence of an effective vaccine, behavioural change is still the most important method of controlling the spread of HIV. Transmission of the virus in blood and blood products has been largely halted in developed countries, with the introduction of screening and education in combination with needle exchange programmes, which have been shown to be effective in reducing the spread of HIV among intravenous drug users. The bulk of transmission of HIV is sexual, however, and preventive activities include reducing the number of sexual partners, modifying the types of sexual contact and the use of condoms. Several behavioural interventions in high-risk populations have been tried, with variable results, but continued education remains a high priority.

BACTERIA

Helicobacter pylori

Helicobacter pylori is a spiral, flagellated, Gram-negative bacteria that colonizes the human gastrointestinal tract and lives beneath the mucus overlaying gastric epithelium. It causes gastritis in all infected people and although many cases remain asymptomatic, some result in gastric or duodenal ulceration. In a very small proportion of infected individuals, *H. pylori* may be involved in the aetiology of gastric adenocarcinomas and the much rarer primary gastric non-Hodgkin lymphoma – about a third of all gastric cancer cases have been attributed to *H. pylori* infection (IARC, 1994b; Danesh, 1999; Parkin *et al*., 1999).

It is estimated that about 50% of the world's population are chronically infected with *H. pylori*. The prevalence of infection is highest in developing countries and increases rapidly during the first two decades of life, such that 80–90% of the population may be infected by early adulthood. In most developed countries, the prevalence of infection is substantially lower at all ages, particularly in childhood. Everywhere, the prevalence of *H. pylori* is strongly correlated with markers of poverty and, indeed, has been decreasing in developing countries for decades, presumably because of improvements in living conditions. Transmission occurs from person to person, probably from mouth to mouth, faecal–orally or both.

In 1990, there were about 750 000 deaths attributable to gastric cancer, making it the fourteenth leading cause of death in the world and the second leading cause of cancer death (Murray and Lopez, 1997). Despite rapidly declining incidence rates in developed countries, gastric cancer is set to remain a major cause of death for many years, a result of population ageing, population growth in developing countries and a poor prognosis. Data from prospective sero-epidemiological studies suggest that infection with *H. pylori* results in a 2–4-fold increase in the risk of gastric cancer. With average seroprevalence rates of about 80% in developing countries and 50% in the developed world, it is estimated that about 340 000 new cases of gastric cancer each year are attributable to *H. pylori* infection – about 40% of the world total of gastric cancers (Parkin *et al*., 1999; **Table 1**). Similarly, about 4000 cases per year of gastric non-Hodgkin lymphoma (which represent about 3% of all gastric cancers) have been attributed to *H. pylori* infection.

The mechanisms by which *H. pylori* might increase the risk of gastric cancer are unclear. The bacteria cause lifelong inflammation, possibly leading to the production of mutagenic compounds, as well as loss of gastric acidity and epithelial cell proliferation. Any or all of the above might contribute to carcinogenesis. There is evidence that the development of cancer is preceded by progressive changes to the stomach mucosa, from inflammation, to atrophy and cellular proliferation, a process that is thought to be related to infection with *H. pylori*. This may be particularly true of the proinflammatory strains of *H. pylori* – those that possess the cytotoxin-associated gene A – because they markedly affect gastric cytokine levels and promote cell turnover, without a corresponding increase in apoptosis. Therefore, *cagA* strains of *H. pylori* might be expected to show a stronger association with gastric cancer, but the evidence for this remains scant.

Drug therapy consisting of two antibiotics in combination with either a bismuth preparation or an acid inhibitor for 14 days is effective in eradicating *H. pylori* in about 80% of cases. Given the high incidence of stomach cancer, the availability of screening tests and eradication regimens, but the relatively low progression rates to cancer in people with *H. pylori* infection, very large randomized trials are needed to establish the value of eradication for the prevention of gastric cancer. In order to achieve statistically reliable results, up to 100 000 people aged 60 years may have to be randomized to eradication therapy or placebo and followed for at least two decades (Danesh, 1999). However, the large-scale use of antimicrobial treatment is problematic – eradication has proved to be difficult in some developing countries and reinfection is common. Furthermore, extensive use of antibiotics may lead to the development of resistant strains (Coursaget and Muñoz, 1999).

In the future, immunization may be a better strategy for the prevention of *H. pylori*-associated diseases, particularly in developing countries. It has been demonstrated in mouse models that *H. pylori* vaccines not only can protect against infection, but may also induce regression of associated lesions. However, in models more relevant to humans, such as monkeys, the results have been disappointing. In phase I trials of a recombinant vaccine in humans, no adverse events were observed, but neither were there any changes in gastric bacterial density. Further work is required to identify appropriate target antigens and delivery systems and to understand the mechanism of protective immunity.

HELMINTHS

Schistosomes

Schistosomiasis (or 'bilharzia') is the generic term given to disease caused by the parasitic blood flukes of the genus *Schistosoma*, class Trematoda of the phylum Platyhelminthes (or flatworms). The genus contains 19 species of which three (*S. mansoni, S. haematobium* and *S. japonicum*) are of major importance to humans. Most infections are subclinical, but in those who develop severe disease the clinical features vary, depending on the species of schistosome. Cancer is an important but relatively rare outcome of infection; the bulk of morbidity and mortality is caused by nonmalignant conditions, such as renal or hepatic failure.

Infection occurs via exposure to water containing the larvae (cercariae). The worms mature in the veins that drain the bladder (*S. haematobium*) or intestine (other species). The adults can survive in the body for several years producing eggs, some of which leave the body in urine or faeces and hatch in water, freeing the miracidium larva. This stage infects certain types of fresh water snail, within which the parasites multiply asexually to produce free-swimming cercariae. These infect humans via skin penetration. Retained eggs elicit hypersensitivity reactions and cause disease of the urogenital system (*S. haematobium*) or of the liver and intestines (other species).

It is estimated that 200 million people in 74 countries are infected with schistosomes and over 600 million are at risk of infection. The geographical distribution of schistosomiasis corresponds to the distribution of susceptible snail hosts, which are present in many tropical and subtropical regions. *S. mansoni* is the most widespread species and is found in 54 countries in Africa, the eastern Mediterranean, South America and the Caribbean. *S. haematobium* has a similar distribution to *S. mansoni* in Africa and the eastern Mediterranean, where coinfection is relatively common, but does not occur in the Americas. *S. japonicum* is endemic in China, the Philippines and Indonesia (WHO, 1993; **Figure 2**). Within endemic areas, however, transmission tends to be highly focal (depending as it does on exposure to contaminated fresh water) and the prevalence and intensity of infection may vary between different communities, or even between households.

Contact with contaminated freshwater is the most important risk factor for infection with schistosomes. The level of contamination can depend on the size and distribution of the intermediate snail population, human population density and behaviour in relation to bodies of fresh water and, climatic and hydrological features (WHO, 1993). Infection is acquired cumulatively over a period of years and the severity is strongly related to the worm burden, or intensity of infection. The age distribution of all schistosome infections is similar, with a characteristic peak in both prevalence and intensity (as measured by active egg excretion) in the second decade of life and a gradual decline thereafter, probably resulting from changes in behaviour in relation to water exposure (WHO, 1993; IARC, 1994b).

Once inside a human host, worm pairs of the species *S. haematobium* reside primarily in the small venules that drain the bladder and ureters and are associated with disease of the urinary system. A causal association between infection with *S. haematobium* and bladder cancer was first postulated near the beginning of the twentieth century and has since been supported by clinical and experimental observations. Bladder cancers associated with schistosomiasis are primarily of the squamous cell type, rather than the transitional cell carcinomas that predominate in nonendemic parts of the world.

The link between infection with *S. haematobium* and bladder cancer is established first from the observation of elevated incidence rates in places where infection is endemic. Furthermore, the sex ratio of squamous cell carcinoma of the bladder in different countries correlates with the relative involvement of men and women in agricultural work (a risk factor for infection). Numerous case series and case-control studies confirm that *S. haematobium* is an important cause of bladder cancer in endemic countries. Infection may play a role in the aetiology of other cancers, in particular cancer of the uterine cervix, although this has yet to be established.

Worm pairs of the species *S. japonicum* reside in the venules that drain the gastrointestinal tract or in the liver, and chronic infection is associated with diseases of these organs. Data from ecological studies, case series and case-control and cohort studies indicate that in endemic areas, *S. japonicum* is probably an important cause of primary liver cancer and colorectal cancer (Chen *et al*., 1990; IARC, 1994b; Newton *et al*., 1999b). There are case reports of *S. japonicum* occurring in conjunction with a range of other cancers, but whether it plays an aetiological role is not clear.

The precise mechanism(s) by which schistosomiasis induces cancer is not known, although several possible explanations have been proposed (IARC, 1994b). These can be broadly categorized as involving (1) exogenous or endogenous agents which either induce DNA damage or have a tumour-promoting activity, (2) altered host metabolism, (3) pathological changes leading to increased cell proliferation and (4) altered immune responses. In relation to *S. haematobium* and bladder cancer, the first suggestions involved the effects of chronic inflammation and urinary retention. More generally, alteration of liver function by hepatic schistosomiasis leads to the production and excretion of potentially carcinogenic tryptophan metabolites, although the importance of these *in vivo* is not clear. In addition, chronic bacterial infection of the bladder can complicate schistosomal infestation, leading to the production of carcinogenic nitrosamines from precursors in urine.

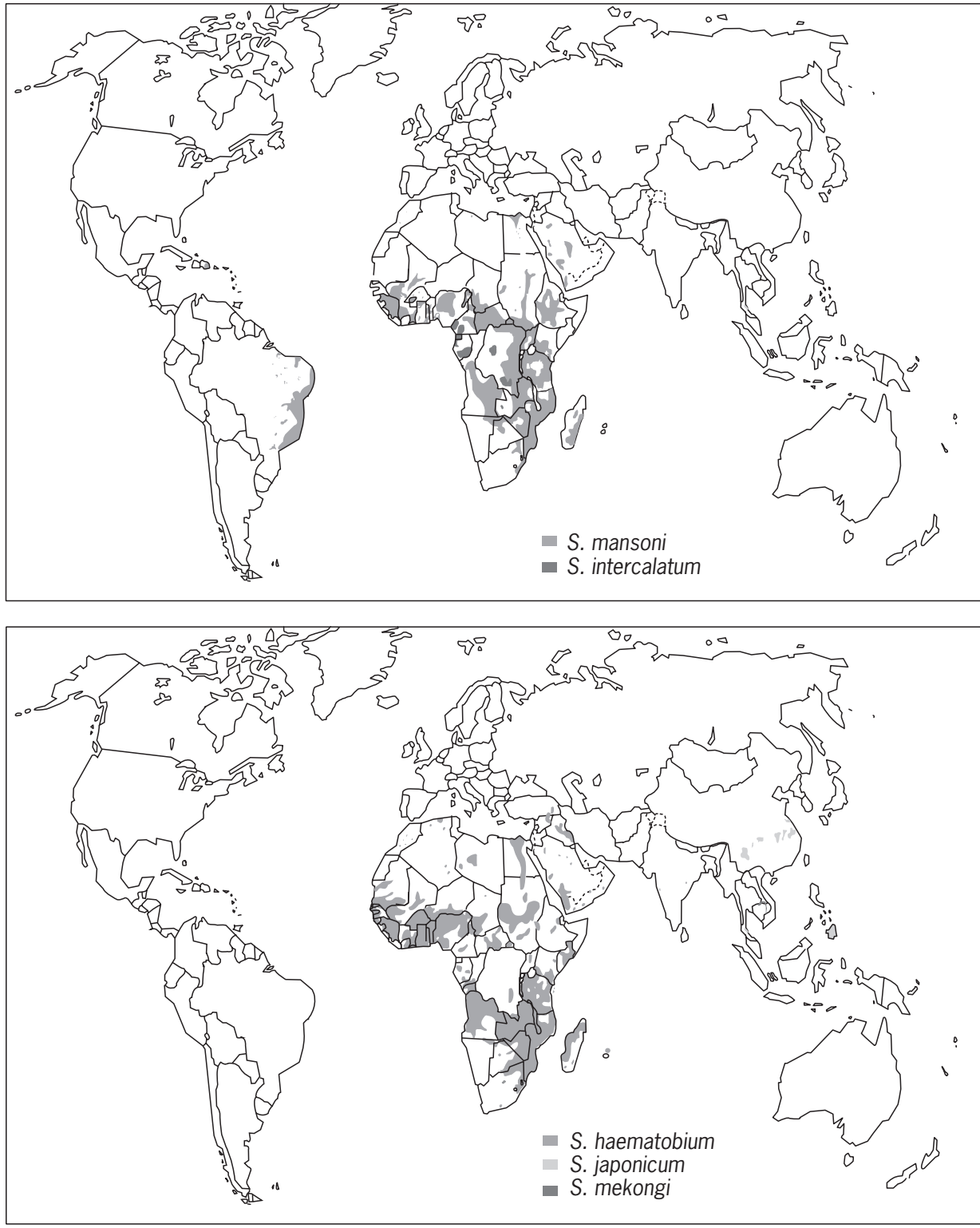


Figure 2 The geographical distribution of schistosomiasis (from WHO).

Safe, effective chemotherapy against all the schistosomes that infect humans has been available for more than two decades (WHO, 1993). The most versatile drug currently available is praziquantel, which is effective in a single oral dose, although it is relatively expensive if used extensively (approximately US $0.35 per treatment). However, treatment can result in resolution of infection, prevention or arrest of disease in heavily infected people and reversal of some manifestations of infection, such as haematuria.

Avoidance of contaminated water would prevent infection with schistosomes and is a relatively simple thing for occasional visitors to endemic areas to achieve. Control and prevention of infection in the community, however, where residents do not always have the luxury of avoiding contact with contaminated water, are complex. Many countries have initiated control programs involving a many pronged approach, including (1) the use of chemotherapy to remove adult worms, (2) elimination of the snail intermediate hosts by habitat modification or chemical attack, (3) changing human behaviour through health education and (4) providing safe water supplies and sanitation.

Use of these integrated control measures over many decades has led to the recent eradication of schistosomiasis in Japan, Tunisia and Monserrat (WHO, 1993). In China, 40 years of unremitting control measures have reduced the prevalence of infection by about 90%. Elsewhere in the World, including Brazil, Egypt, Iran, the Philippines and Venezuela, significant reductions in disease prevalence have been achieved. Even in places where the prevalence of infection has remained high, serious manifestations of disease are becoming less common with the use of effective chemotherapy, although declines in cancer incidence are not yet apparent (WHO, 1993). Despite this, the number of cases of schistosomiasis worldwide was estimated to be the same in 1993 as it was in 1984 (WHO, 1993). In endemic areas, populations (and hence the number of susceptible hosts) continue to grow. In addition, developments in water resource management, land use and irrigation have led to a spread of schistosomiasis to new areas. There is currently no vaccine available, although intensive efforts are being made to develop one, and so the use of complex, integrated control measures remains paramount.

Liver Flukes

Three species of food-borne liver flukes, *Opisthorchis viverrini*, *O. felineus* and *Clonorchis sinensis*, of the class Trematoda, are pathologically important to humans. They establish chronic infection within the intrahepatic bile ducts and occasionally in the pancreas and gall-bladder and have been associated with diseases of these organs. Infection is acquired by eating raw or undercooked freshwater fish, which contain the infective stage (metacercaria). The flukes migrate to the biliary tree via the ampulla of Vater and mature in the intrahepatic bile ducts, producing eggs, which are excreted in faeces. If the eggs reach fresh water and are consumed by an appropriate species of snail, they hatch, undergo asexual reproduction to produce free-living larvae, which can infect freshwater fish and become encysted metacercariae (IARC, 1994b; Vatanasapt *et al*., 1999).

O. viverrini is common in north-east Thailand and Laos, where it is estimated that up to 9 million people are infected (about one-third of the population). About 1.5 million people are infected with *O. felineus*, mainly in central Russia, and 7 million are infected with *C. sinensis*, in Korea, China, Macao and Vietnam (**Figure 3**). The distribution of human infection depends not just on that of the flukes, but also on the habit of eating raw freshwater fish. Infection usually occurs in the first decade of life and men are often more affected than women.

Cholangiocarcinoma has long been recognized as a serious complication of liver fluke infection. This tumour is very rare throughout most of the world, but in endemic areas the number of cases of cholangiocarcinoma is usually higher than that of hepatocellular carcinoma. Evidence from ecological studies relating the prevalence of antibodies to liver fluke infection (or of eggs in faeces) to the incidence of cholangiocarcinoma suggest a causal link. Case-control studies indicate that infected individuals have about a fivefold increase in the risk of cancer compared with the uninfected, a finding that is supported by data from animal studies.

The pathological changes associated with acute liver fluke infection include oedema, desquamation and acute inflammation in the bile ducts. Chronic infection is characterized by goblet cell metaplasia, adenomatous hyperplasia and thickening of the walls. It is thought that cholangiocarcinoma arises from this pre-existing damage, as part of a progressive process. The mechanisms by which this might occur are poorly understood.

The antihelminthic drug praziquantel is currently the drug of choice for the treatment of liver fluke infection, both for the individual patient and for community-based treatment programmes. The best way to reduce the incidence of cholangiocarcinoma is to control liver flukes. Eradication programmes need continuous and intensive health education, together with drug treatment for existing infections. Mass treatment without health education is unlikely to be effective. Health education should focus on the need to cook fish and fish products properly. As yet, eradication programmes have had little effect on the incidence of cholangiocarcinoma and a vaccine is not available.

THE FUTURE

Given that the discovery of an infectious cause of cancer (or any other chronic disease) has such important implications for prevention, how does one identify and confirm a causal association? First, this requires an insight into which diseases may be linked to infection. Certain epidemiological features of a cancer may offer clues as to an infectious aetiology. These include a high incidence of disease in people who are prone to infections (such as the immunosuppressed), large variations in incidence by geographical region or other patterns of clustering, apparent improvements with antimicrobial treatments and a strong correlation with markers of poverty. So what

Figure 3 The geographical distribution of liver fluke infections (from WHO).

cancers might next be linked to infections? It is dangerous to prophesy, but there is considerable speculation about leukaemias and lymphomas, squamous cell skin cancers and even breast cancer. Second, the relevant infection must be identified and isolated from tumour tissue or diseased individuals. Finally, causality must be established. The relative risk of a specific cancer in an infected individual is usually very high, but it is harder to identify large excess risks in infected individuals when the cancer is a rare outcome of a common, or even ubiquitous, infection. In such instances, establishing causality may depend in part, for example, on identifying abnormal immunological responses to infection in diseased individuals, compared with those who are without disease (e.g. differences in antibody titre to infection).

As for prevention of cancers caused by infections, it is clear that vaccination programmes, although ultimately cost effective, require a long-term commitment, involving substantial investment of time, money and effort. In the meantime, an understanding of the biology and epidemiology of the relevant infections is essential to reduce the associated cancer burden in other ways. In certain circumstances, many infections can be avoided by behaviour modification or screening of blood products, for example. In addition, the risk of transmission of an infection from a pregnant mother to her child may be reduced with changes in breast-feeding behaviour or the use of anti-viral therapies. Perhaps one of the most exciting possibilities involves the introduction of HPV testing within the framework of the cervical screening programme, as a means of identifying women at high risk of disease, for intensive follow-up.

All of the above makes a strong case for a coordinated effort by clinicians, molecular biologists and epidemiologists to search for new – as well as known – infections in human tissues and to study their possible associations with disease. Research into a problem, however, is not the whole story. Understanding the causes of cancer is often difficult – prevention requires something more: a commitment to invest in public health.

REFERENCES

Barré-Sinoussi, F., *et al*. (1983). Isolation of a T-lymphotropic retrovirus from a patient at risk for aquired immunodeficiency syndrome (AIDS). *Science*, **220**, 868–871.

Beral, V. and Newton, R. (1998). Overview of the epidemiology of immunodeficiency-associated cancers. *Monographs of the National Cancer Institute*, **23**, 1–6.

Boshoff, C. (1999). Kaposi's sarcoma associated herpesvirus. In: Newton, R., *et al*. (eds), *Cancer Surveys, Vol. 33, Infections and Human Cancer* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY), pp. 157–190.

Chang, M. H., *et al*. (1997). Universal hepatitis B vaccination in Taiwan and the incidence of hepatocellular carcinoma in children. *New England Journal of Medicine*, **336**, 1855–1859.

Chen, J., *et al*. (1990). *Diet, Life-style and Mortality in China: A Study of the Characteristics of 65 Chinese Counties* (Oxford University Press, Oxford, Cornell University Press, Ithaca, NY and People's Medical Publishing House, Beijing).

Coursaget, P. and Muñoz, N. (1999). Vaccination against infectious agents associated with human cancer. In: Newton, R., *et al*. (eds), *Cancer Surveys, Vol. 33, Infections and Human Cancer* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY), pp. 355–382.

Cuzick, J. (1999). Screening. In: Newton, R., *et al*. (eds), *Cancer Surveys, Vol. 33, Infections and Human Cancer* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY), pp. 345–354.

Danesh, J. (1999). Is *Helicobacter pylori* infection a cause of gastric neoplasia? In: Newton, R., *et al*. (eds), *Cancer Surveys, Vol. 33, Infections and Human Cancer* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY), pp. 263–272.

de Thé, G., *et al*. (1978). Epidemiological evidence for a causal relationship between the Epstein–Barr virus and Burkitt's lymphoma from a Ugandan prospective study. *Nature*, **274**, 756–761.

Gallo, R. C., *et al*. (1984). Frequent detection and isolation of cytopathic retrovirus (HTLV-III) from patients with AIDS and at risk for AIDS. *Science*, **224**, 500–503.

Gambia Hepatitis Study Group (1987). The Gambia Hepatitis Intervention Study. *Cancer Research*, **47**, 5782–5787.

Herrero, R. and Muñoz, N. (1999). Human papillomavirus and cancer. In: Newton, R., *et al*. (eds), *Cancer Surveys, Vol. 33, Infections and Human Cancer* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY), pp. 75–98.

IARC (1992). *The Epidemiology of Cervical Cancer and Human Papillomavirus*. IARC Scientific Publications, Vol. 119 (International Agency for Research on Cancer, Lyon).

IARC (1994a). *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, Vol. 59, Hepatitis Viruses* (International Agency for Research on Cancer, Lyon).

IARC (1994b). *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, Vol. 61, Schistosomes, Liver Flukes and Helicobacter pylori* (International Agency for Research on Cancer, Lyon).

IARC (1995). *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, Vol. 64, Human Papillomaviruses* (International Agency for Research on Cancer, Lyon).

IARC (1996). *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, Vol. 67, Human Immunodeficiency Viruses and Human T-cell Lymphotropic Viruses* (International Agency for Research on Cancer, Lyon).

IARC (1997). *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, Vol. 67, Epstein–Barr Virus and Kaposi's Sarcoma Herpesvirus/Human Herpesvirus 8* (International Agency for Research on Cancer, Lyon).

McGregor, J. M. and Proby, C. M. (1996). The role of papillomaviruses in human non-melanoma skin cancer. In: Leigh, I. M.,

et al. (eds), *Cancer Surveys, Vol. 26, Skin Cancer* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY), pp. 219–236.

Murray, C. J. L. and Lopez, A. D. (1997). Mortality by cause for eight regions of the world: global burden of disease study. *Lancet*, **349**, 1269–1276.

Newton, R., *et al*. (1999a). Human immunodeficiency virus infection and cancer. In: Newton, R., *et al*. (eds), *Cancer Surveys, Vol. 33, Infections and Human Cancer* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY), pp. 237–262.

Newton, R., *et al*. (1999b). Schistosomes and human cancer. In: Newton, R., *et al*. (eds), *Cancer Surveys, Vol. 33, Infections and Human Cancer* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY), pp. 291–312.

Parkin, D. M., *et al*. (1999). The global health burden of infection associated cancers. In: Newton, R., *et al*. (eds), *Cancer Surveys, Volume 33, Infections and Human Cancer* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY), pp. 5–34.

Phillips, A. C. and Vousden, K. H. (1999). Human papillomavirus and cancer: the viral transforming genes. In: Newton, R., *et al*. (eds), *Cancer Surveys, Vol. 33, Infections and Human Cancer* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY), pp. 55–74.

Tanaka, H. and Tsukuma, H. (1999). Hepatitis C virus. In: Newton, R., *et al*. (eds), Cancer Surveys, Vol. 33, Infections and Human Cancer (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY), pp. 55–74.

Vatanasapt, V., *et al*. (1999). Liver flukes and liver cancer. In: Newton, R., *et al*. (eds), *Cancer Surveys, Vol. 33, Infections and Human Cancer* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY), pp. 313–344.

Walbloomers, J. M., *et al*. (1999). Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *Journal of Pathology*, **189**, 12–19.

WHO (1993). *The Control of Schistosomiasis*. Second Report of the WHO Expert Committee. WHO Technical Report Series, No. 830 (World Health Organization, Geneva).

Wild, C. P. and Hall, A. J. (1999). Hepatitis B virus and liver cancer: unanswered questions. In: Newton, R., *et al*. (eds), *Cancer Surveys, Vol. 33, Infections and Human Cancer* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY), pp. 35–54.

FURTHER READING

Newton, R., *et al*. (eds) (1999). *Cancer Surveys, Vol. 33, Infections and Human Cancer* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY).

Parsonnet, J. (ed.) (1999). *Microbes and Malignancy: Infection as a Cause of Human Cancer* (Oxford University Press, Oxford).

Chapter 25

Short-term Testing for Genotoxicity

David Gatehouse
GlaxoWellcome Research and Development Limited, Ware, UK (Present address: Concave Laboratories Ltd, Harrogate, UK)

CONTENTS

INTRODUCTION

It has been known for several hundred years that exposure to particular chemicals or complex mixtures can lead to cancer in later life. There has been accumulating evidence that many cancers can arise from damage to DNA and the resulting mutations. This has been discussed in detail in the preceding chapters. As a consequence of this, it has become necessary to determine whether widely used chemicals or potentially useful new chemicals possess the ability to damage DNA. Data concerning the genotoxicity of a new chemical have become part of the basic toxicological information package. They are needed for decision making and to reduce risks that might otherwise be unforeseen.

The field of genetic toxicology began in the 1960s when several seminal conferences were held focusing on chemical mutagens and in particular their effects on germ cells and the risk to future generations. Although germ cell risk was the initial concern, this was broadened in the 1970s when evidence relating genotoxicity and carcinogenicity began to accumulate. This was further supported by the use of *in vitro* metabolic activation systems capable of producing electrophilic metabolites, and the fact that early analysis of rodent carcinogens and noncarcinogens suggested that almost all carcinogens were also genotoxic. This view has now been modified, since it is clear that nongenotoxic carcinogens also exist, as discussed in an earlier chapter (see the chapter *Non-Genotoxic Causes of Cancer*). From this time onwards, various national expert committees were formed to advise governments on the type of approach that should be taken to screen new chemicals for carcinogenic risk (and any potential heritable effects). Consequently, numerous guidelines have been prepared over the past 20 years describing the tests which should be used to investigate the genotoxicity of chemicals. It is not the intention of this chapter to give an exhaustive list of these guidelines, except to mention that a harmonized approach to the genotoxicity testing of drugs has recently been introduced as a result of the International Conference on Harmonisation (ICH) programme. In this process, areas of disharmony were identified and differences in regulation with respect to genotoxicity were discussed, resulting in the creation of two guidelines (referenced at the end of this chapter). At the same time as the ICH process was occurring, the Organisation for Economic Cooperation and Development (OECD) also updated a number of its genotoxicity guidelines. Both processes influenced each other resulting in similar recommendations. A summary of the testing strategy recommended by the ICH is given in **Figure 1**.

In this chapter, the individual test systems that are required for genotoxicity screening will be described together with a discussion on how the results obtained should be interpreted. As well as the primary test systems, there will also be a description of the supplementary assays that may be required when investigating positive effects. It is not intended to give detailed guidance on the performance of these tests. For this information, the reader is directed to the Further Reading list at the end of the chapter. The chapter concludes with a brief overview of some of the new developments in the field.

1. A test for gene mutation in bacteria
2. An *in vitro* test with cytogenetic evaluation of chromosomal damage with mammalian cells or an *in vitro* mouse lymphoma tk assay
3. An *in vivo* test for chromosomal damage using rodent haematopoietic cells

Figure 1 Recommended ICH genotoxicity test battery.

PRIMARY TEST SYSTEMS

In Vitro Metabolic Activation

Before describing the individual *in vitro* test systems, it is necessary to mention briefly a factor of critical importance in genotoxicity screening, namely the need to include some form of *in vitro* metabolizing system. This is because most of the indicator cells (bacteria and mammalian cells) possess a very limited capacity for endogenous metabolism of xenobiotics. Many carcinogens and mutagens are unable to interact with DNA unless they have undergone some degree of metabolism (see also the chapter *Mechanisms of Chemical Carcinogenesis*). To improve the ability of the test systems to detect as many authentic *in vivo* mutagens and carcinogens as possible, extracts of mammalian liver (usually rat) are incorporated. The liver is a rich source of mixed-function oxygenases capable of converting carcinogens to reactive electrophiles. Crude homogenate such as the 9000 *g* supernatant (S9 fraction) is used, which is composed of free endoplasmic reticulum, microsomes, soluble enzymes and some cofactors. The oxygenases require the reduced form of nicotinamide adenine dinucleotide phosphate (NADPH), which is normally generated *in situ* by the action of glucose-6-phosphate dehydrogenase on glucose-6-phosphate and reducing NADP, both of which are normally supplied as cofactors. Normal uninduced S9 preparations are of limited value for screening as they are deficient in particular enzyme activities. In addition, species and tissue differences are most divergent in such preparations. These problems are reduced when enzyme inducers are used, and most commonly preparations are made from rat livers after enzyme induction with Aroclor 1254, which is a mixture of polychlorinated biphenyls. Concern about the toxicity, carcinogenicity and persistence of this material in the environment has led to the introduction of alternatives, such as a combination of phenobarbitone and β-naphthoflavone. This combination induces a similar range of mono-oxygenases and has been recommended as a safer alternative to Aroclor (Elliott *et al*., 1992).

It should be noted that this system is only a first approximation to the complex metabolic processes that occur *in vivo*, and in particular there is little account taken of the phase II detoxification reactions. Such factors should be considered when interpreting positive *in vitro* results which are only seen in the presence of S9 mix.

In Vitro Tests for Gene Mutation in Bacteria

The most widely used assays for detecting chemically induced gene mutations are those employing bacteria. These assays feature in all test batteries for genotoxicity as it is relatively straight forward to use them as a sensitive indirect indicator of DNA damage. Bacteria can be grown in large numbers overnight, permitting the detection of rare mutational events. The extensive knowledge of bacterial genetics that was obtained during the twentieth century allowed the construction of special strains of bacteria with exquisite sensitivity to a variety of genotoxins. An offshoot of the studies on genes concerned with amino acid biosynthesis led to the development of *Escherichia coli* and *Salmonella typhimurium* strains with relatively well defined mutations in known genes. The most commonly used bacteria are the *S. typhimurium* strains which contain defined mutations in the histidine operon. These were developed by Bruce Ames, and form the basis of the 'reverse' mutation assays (Ames *et al*., 1971). In these assays, bacteria which are already mutant at the histidine locus are treated with a range of concentrations of test chemical to determine whether the compound can induce a second mutation that directly reverses or suppresses the original mutations. Thus, for the *S. typhimurium* strains which are histidine auxotrophs, the original mutation resulted in the loss of ability to grow in the absence of histidine. The second mutation (induced by the chemical) restores prototrophy, i.e. the affected cell is now able to grow in the absence of histidine, if provided with inorganic salts and a carbon source. This simple concept underlines the great strength of these assays for it provides enormous selective power which can identify a small number of the chosen mutants from a population of millions of unmutated cells and cells mutated in other genes. Each of the *S. typhimurium* strains contains one of a number of possible mutations in the histidine operon, and each can be reverted by either base-change or frameshift mutations. The genotype of the most commonly used strains is shown in **Table 1**, together with the types of reversion events that each strain detects.

In order to make the bacteria more sensitive to mutation by chemical agents, several additional traits have been introduced. Ames and colleagues realized that many carcinogens (or their metabolites) are large molecules that are often unable to cross the protective cell wall of the bacteria. Wild-type cells produce a lipopolysaccharide that acts as a barrier to bulky hydrophobic molecules. Consequently, an *rfa* mutation was introduced into the *Salmonella* strains, which resulted in defective lipopolysaccharide and increased permeability.

Bacteria possess several major DNA repair pathways that appear to be error-free. The test strains were constructed, therefore, with a deletion removing the *uvrB* gene. This codes for the first enzyme in the error-free excision repair pathway, and so gene deletion renders the strains excision repair deficient, thus increasing their sensitivity to many genotoxins by several orders of magnitude. Lastly, some of the bacterial strains do not appear to possess classical error-prone repair as found in other members of the Enterobacteria such as *E. coli*. This results from a deficiency in *umuD* activity. This deficiency is overcome by insertion of a plasmid containing *umuDC* genes. Plasmid