1 Basic genetics

This chapter provides a review of basic genetics. It concentrates on the general principles that apply to normal, healthy animals. The exceptions to these principles are often the basis of genetic diseases, which are discussed in subsequent chapters.

Chromosomes

When a culture of rapidly dividing white blood cells is treated with the alkaloid colchicine (which halts cell division), and the cells are then stained and viewed under a light microscope, structures called **chromosomes** become clearly visible. They are scattered randomly within clusters, and each cluster contains all the chromosomes from just one cell. The area of genetics concerned with chromosomes is called **cytogenetics**.

In order to study chromosomes more closely, a suitable cluster is chosen, as shown in Fig. 1.1a. Each item in the cluster consists of two rod-like structures joined together at a constricted point. Each rod-like structure is a **chromatid** and the constriction is a **centromere**. The two chromatids that are joined at the centromere have just been formed from one original chromosome. If the cell division had been allowed to proceed, the centromere would have split and each separate chromatid would then be called a new chromosome. For convenience, we talk of each pair of chromatids joined at the centromere as being just one chromosome, referring in fact to the chromosome that has just given rise to them.

All the chromosomes in the cluster are then rearranged in order of size. An arrangement such as this provides a picture of the complete set of chromosomes or **karyotype** of a cell (Fig. 1.1b). If many such arrangements are examined from normal, healthy individuals of both sexes of any species of mammal or bird, two facts become evident: each species has a characteristic karyotype and, within any species, each sex has a characteristic karyotype.

Karyotypes of different species differ in the shape, size, and number of their chromosomes. Within any species, all the chromosomes occur in pairs. In individuals of one sex, both members of each chromosome pair have the same size and shape. In the other sex, all but two chromosomes occur in such pairs, with the remaining

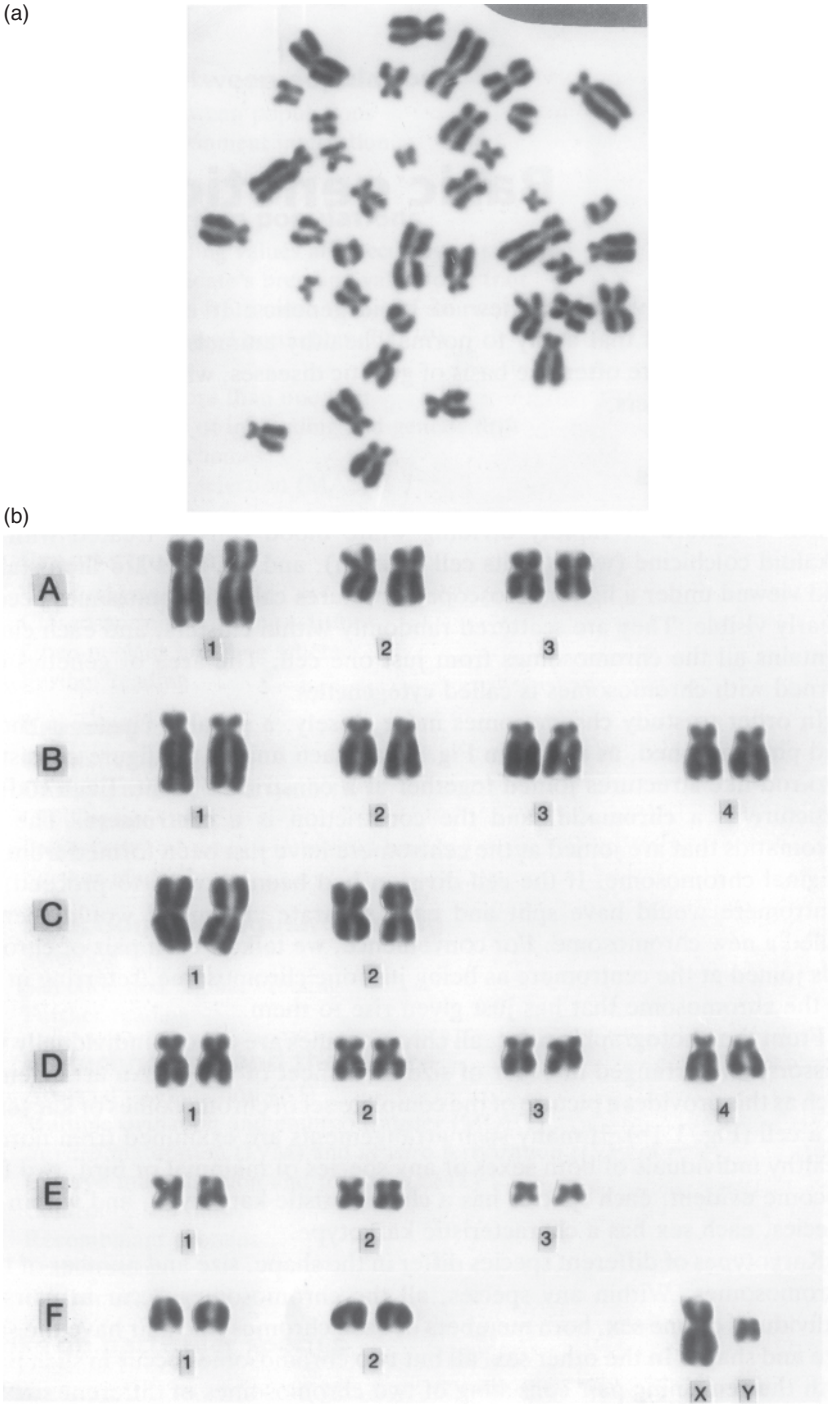


Fig. 1.1 (a) The chromosomes of a male cat, as seen through a light microscope. (b) The karyotype of a male cat, as obtained by rearranging individual chromosomes from (a). (Reproduced courtesy of P. Muir.)

pair consisting of two chromosomes of different size and shape. In this unequal pair, one chromosome has the same shape and size as members of one of the pairs in the opposite sex.

The difference in karyotype between the two sexes is the key to sex determination. In mammals, the two chromosomes that form the unequal pair occur in males, and are called the X and Y chromosomes. In female mammals, one of the pairs of chromosomes consists of two X chromosomes. Thus in mammals, males are XY and females are XX. The X and the Y chromosomes are known as **sex chromosomes**. In birds, the sex chromosomes are given different names, and their relationship to sex is the opposite of that in mammals: male birds are ZZ and female birds are ZW. For convenience, we shall refer only to mammals in the following discussion, although all statements apply equally to birds if the names of the sexes are reversed.

Chromosomes other than the sex chromosomes are called **autosomes**. Within any species, males and females have the same set of autosomes, occurring in pairs. The sex chromosomes plus the autosomes constitute a **genome**, which is the total set of chromosomes in a cell. Genomes in which chromosomes occur in pairs are said to be **diploid**, and the two members of a pair are called **homologues**. In order to emphasize that chromosomes occur in pairs, the total number of chromosomes is called the $2n$ number, where n is the number of pairs. For example, the number of chromosomes in the karyotype illustrated in Fig. 1.1 is $2n = 38$. To enable identification of each pair of chromosomes in a karyotype, the autosome pairs are labelled according to an internationally agreed convention, as shown in Fig. 1.1b. The two sex chromosomes are placed at the end.

In order to describe karyotypes more fully, chromosomes are often classified according to whether the centromere is at one end (**acrocentric**), closer to one end than the other (**sub-metacentric**) or in the middle (**metacentric**). In this book, we shall follow common practice in using metacentric to cover both metacentric and sub-metacentric. The short arm of each chromosome is designated p (think of petite = small), and the long arm is designated q. (If the centromere is in the centre of the chromosome, the designation of which arm is called p is arbitrary, but is agreed by international convention; for acrocentric chromosomes, e.g. cattle autosomes, there is only one arm which, by convention, is designated the q arm). A summary description of the karyotypes of common domestic species is given in Table 1.1. Avian karyotypes are somewhat different to mammalian karyotypes, in that their smallest chromosomes are far smaller than the smallest mammalian chromosome. Until recently, chicken chromosomes were categorized as either macrochromosomes or microchromosomes, but now, in recognition of the continuum of sizes, they are classified into four groups, from group A (the largest chromosomes) to group D (the smallest).

Banding

When karyotypes were first investigated, individual pairs of chromosomes could be identified only according to their shape and size. Since then, various methods of staining chromosomes have been developed, giving rise to alternating light and dark

Table 1.1 A summary description of the karyotypes of some domestic species

Species	Total diploid number (2*n*=)	Autosomal pairs	
		Metacentrics	Acrocentrics
Cat, *Felis catus*	38	16	2
Dog, *Canis familiaris*	78	0	38
Pig, *Sus scrofa domesticus*	38	12	6
Goat, *Capra hircus*	60	0	29
Sheep, *Ovis aries*	54	3	23
Cattle, *Bos taurus*	60	0	29
Horse, *Equus caballus*	64	13	18
Donkey, *Equus asinus*	62	24	6
Alpaca, *Lama pacos* Llama, *Lama glama*	74	16	20
Rabbit, *Oryctolagus cuniculus*	44	19	2
Chicken, *Gallus gallus*	78	7	31

regions called **bands**. The main types of bands are broadly classified as G (stained with Giemsa), Q (stained with quinacrine), R (reverse of G-banding), C (centromeres are stained), and T (telomeres are stained). Among specific banding strategies, we have G-banding by trypsin treatment and Giemsa staining (GTG), G-banding by early bromodeoxyuridine (BrdU) incorporation and Giemsa staining (GBG), R-banding by late BrdU incorporation and acridine-orange staining (RBA), and R-banding by late BrdU incorporation and Giemsa staining (RBG).

As an example of banding, the GBG bands of cattle are illustrated in Fig. 1.2. Since the position, width, and number of bands are different for each pair of chromosomes, each chromosome pair can be identified by its banding pattern. By studying many cells treated in the same way, it is possible to draw up an **idiogram**, which is a representation of the characteristic banding pattern for each pair of chromosomes. The bands are uniquely identified according to a convention known as the International System for Cytogenetic Nomenclature of Domestic Bovids (ISCNDB). Each arm is divided into a small number of regions which are numbered sequentially starting from the centromere. Then, in each region, the bands are numbered sequentially starting nearest the centromere. For example, the second band in the third region of chromosome 1 in cattle is designated q132, while the second band in the fourth region of the long arm of the X chromosome is Xq42. The ISCNDB idiogram for cattle is illustrated in Fig. 1.3. Banded karyotypes of other domestic species are illustrated in Appendix 1.1.

Meiosis and mitosis

For many thousands of years, humans have observed two phenomena in relation to sex determination in animals: first, that there is considerable variation in the numbers of each sex among the offspring of pairs of parents; and second, that despite this

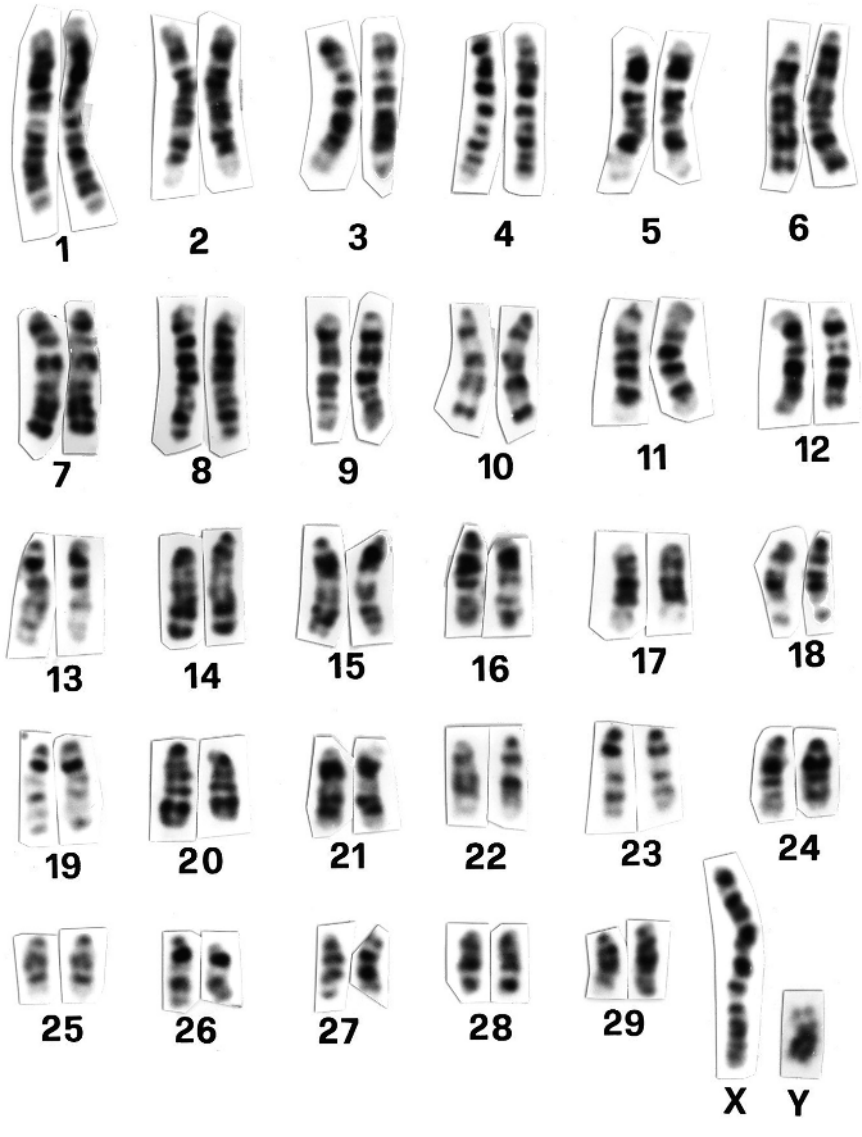


Fig. 1.2 The standard GBG-banded cattle karyotype. (Reprinted by permission from S. Karger AG, Basel: ISCNDB (2000), Di Berardino, Di Meo, Gallagher, Hayes and Iannuzzi, *Cytogenetics and Cell Genetics*, **92**, 283–99.)

variation, the overall numbers of males and females across families are approximately equal.

As noted above, the difference in sex chromosomes between the two sexes is the key to sex determination. The reason why XX individuals are females and XY individuals are males is explained in Chapter 4. For the present, we shall ask simply: why is there so much variation in the numbers of XX and XY individuals in the

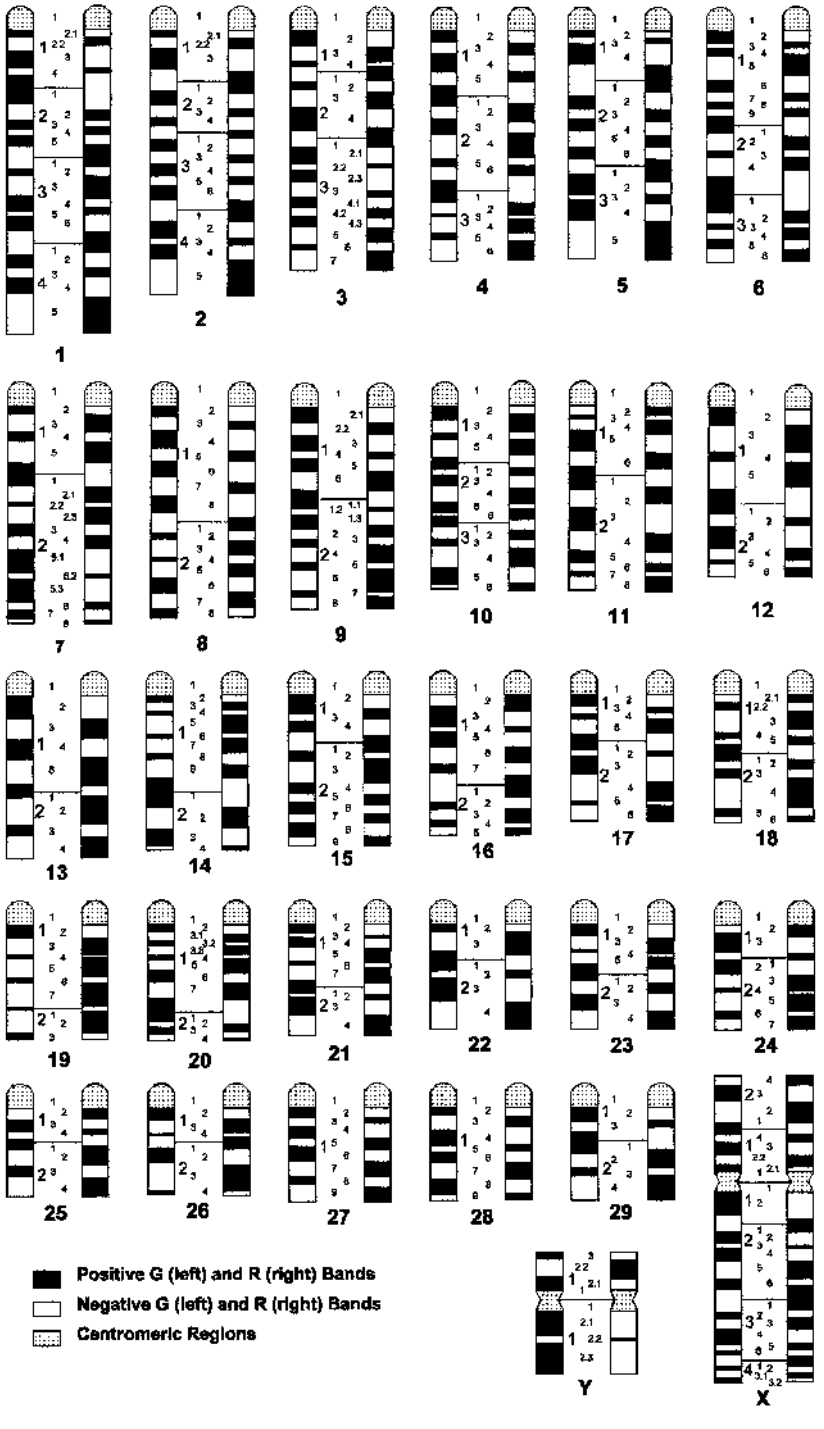


Fig. 1.3 The standard cattle idiogram, showing both G-bands (*left*) and R-bands (*right*). (Reprinted by permission from S. Karger AG, Basel: ISCNDB (2000), Di Berardino, Di Meo, Gallagher, Hayes and Iannuzzi, *Cytogenetics and Cell Genetics*, **92**, 283–99.)

offspring of pairs of parents, and yet at the same time approximately equal numbers of each sex overall? The answer lies in an understanding of gamete formation.

Meiosis

Meiosis is the process of gamete formation in which sperms are formed in testes of males and ova are formed in ovaries of females. The main result of meiosis is that each sperm and each ovum contains one member of each pair of chromosomes. Containing exactly one half of the usual diploid number of chromosomes, gametes are said to be **haploid**. The union of a sperm with an ovum at fertilization produces a **zygote** with the usual diploid number of chromosomes.

The process of meiosis commences with a normal cell containing the usual diploid set of chromosomes. To make the explanation easier, we shall consider what happens to just one pair of chromosomes (the sex chromosomes) in one sex (females), as illustrated in Fig. 1.4. In order to distinguish the two X chromosomes in females, we shall refer to them as X_p (paternal: originating from the father) and X_m (maternal: originating from the mother).

Meiosis occurs in two stages. Meiosis I begins with each chromosome duplicating itself, giving rise to two identical chromatids joined at the centromere. Then homologous chromosomes, in our case X_p and X_m, line up next to each other in the centre of the cell, in a process known as **pairing** or **synapsis**. This is facilitated by a protein structure called the **synaptonemal complex**, which 'zips' the two homologues together. The pair of homologues is called a **bivalent**. Because each chromosome has already duplicated itself into two chromatids, there are now four chromatids side by side in the cell; two X_p chromatids and two X_m chromatids. The two X_p chromatids are still joined at their centromere, as are the two X_m chromatids. At this stage, a process called **recombination** or **crossing-over** occurs, in which homologous chromatids each break at the same site and, in the process of re-uniting, exchange segments. This produces a cross-like structure called a **chiasma** (plural: **chiasmata**). In order to simplify the present discussion, we shall continue to refer to the chromatids as X_p or X_m, realizing that, as a result of crossing-over, any one chromatid may in fact consist of parts of both X_p and X_m. (A full discussion of the genetic implications of crossing-over is presented later in this chapter.) In the next stage of meiosis I, the two centromeres are pulled to opposite ends or **poles** of the cell, with the result that the two X_p chromatids move to one pole of the cell and the two X_m chromatids move to the other pole. Since this process involves the two pairs of chromatids disjoining from their previous paired arrangement, it is known as **disjunction**. In the final stage of meiosis I, the original cell divides into two cells; one contains the two X_p chromatids still joined at their centromere, and the other contains the two X_m chromatids, still joined at their centromere.

Following disjunction in females, only one cell continues to function normally; the other degenerates into a dark-staining structure known as the **first polar body**. It is entirely a matter of chance which of the cells remains functional. Consequently, there is an equal chance of either the two X_p chromatids or the two X_m chromatids ending up in the functional cell. (In Fig. 1.4, it happens to be the X_p chromatids that

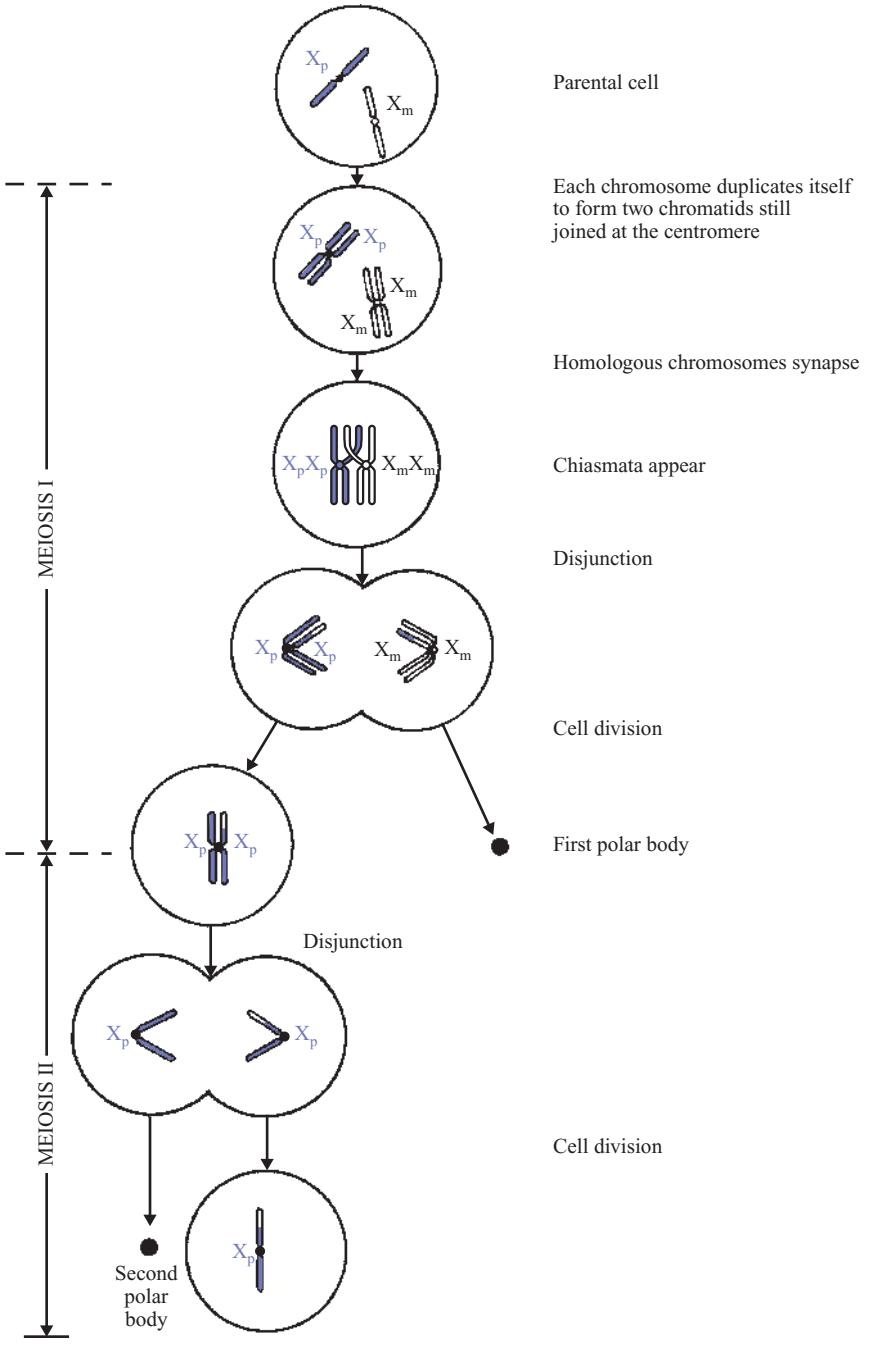


Fig. 1.4 Meiosis in a female, illustrated in terms of the sex chromosomes. A single cross-over has created two recombinant and two non-recombinant chromatids. The functional gamete in this example contains a recombinant chromosome with a paternal centromere. Exactly the same processes occur for all pairs of autosomes.

have survived.) In meiosis II in females, the two chromatids in the functional cell move apart (disjoin) and the cell divides into two cells, each containing one chromatid which is now called a chromosome. Once again, only one of the two cells remains functional; the other degenerates into the **second polar body** and, once again, it is entirely a matter of chance as to which of these two cells becomes the second polar body.

It is evident that in females, only one functional gamete results from each cell that originally underwent meiosis. It is also obvious that, irrespective of which cell ultimately remains functional, all gametes produced by females are the same in the sense that each contains one X chromosome. For this reason, females are known as the **homogametic** sex.

In males, meiosis is basically the same as described above: a disjunction followed by a cell division in meiosis I, and the same in meiosis II (Fig. 1.5). There are, however, two important differences. The first is that the X and Y chromosomes have only a small homologous region at the end of one arm (called the **pseudo-autosomal region**) where synapsis occurs; for the remainder of their length, the arms are not joined together. Despite this unusual arrangement, their subsequent disjunction is normal, and gives rise to two functional cells at the end of meiosis I: one contains two X chromatids still joined at their centromere, and the other contains two Y chromatids still joined at their centromere. The second difference between meiosis in females and in males is that polar bodies are not formed in males. Instead, both of the cells formed at the end of meiosis I undergo a cell division in meiosis II, giving rise to four functional gametes (sperms), two of which contain an X chromosome and two of which contain a Y chromosome. Since males produce two different types of gametes, they are known as the **heterogametic** sex.

Having now produced the gametes, the next stage is fertilization which, genetically speaking, is largely a matter of chance.

Chance and variation

Since all female gametes contain an X chromosome, the chance of a female gamete containing an X chromosome is one. In contrast, males produce equal numbers of X-bearing gametes and Y-bearing gametes. There is, therefore, a chance of $\frac{1}{2}$ that a particular sperm contains an X and the same chance that it contains a Y. It follows that the chance of obtaining an XY zygote is $1 \times \frac{1}{2}$ which equals $\frac{1}{2}$. Similarly, the chance of obtaining an XX zygote is $1 \times \frac{1}{2}$, or $\frac{1}{2}$. We can represent this situation by using a common genetic device called a checkerboard or **Punnett square**, in which the proportion at the head of each column is multiplied by the proportion at the head of each row, to give the expected proportions of offspring in the body of the checkerboard:

	Male gametes	
	½ X	½ Y
Female gametes all X	½ XX	½ XY

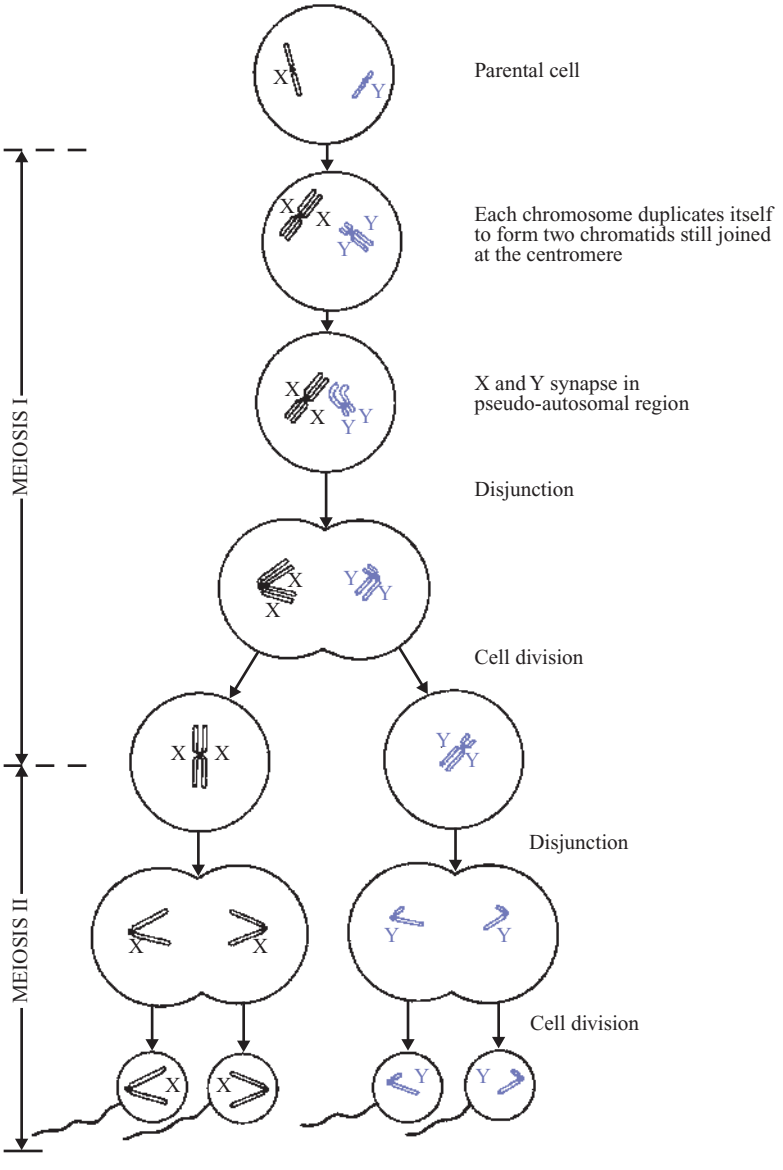


Fig. 1.5 Meiosis in a male, illustrated in terms of the sex chromosomes. With the exception of the unusual pairing in meiosis I, exactly the same processes occur for all pairs of autosomes.

We have now seen how meiosis enables the production of an expected equal proportion of each sex, which accounts for one of our original observations. How can we account for the second observation, concerning the considerable variation in numbers of each sex among the offspring of different pairs of parents? There is just one fact that enables us to explain this variation: each fertilization is an independent event. By this we mean that irrespective of whether an X-bearing or Y-bearing sperm

is successful with a particular ovum, the result of that fertilization has no bearing on subsequent fertilizations, even if they occur at the same time. For example, in a female that ovulates four ova, the chance that the last ovum is fertilized by a Y-bearing sperm is exactly $\frac{1}{2}$ irrespective of which type of sperm fertilized the other ova. In fact, any particular sequence of sexes, e.g. MMFM, is just as likely as any other sequence, e.g. FFFF.

We have now provided adequate explanations for each of the observations described earlier. In so doing, we have discussed chromosomes, simple inheritance, and chance, each of which is basic to an understanding of genetics. In order to complete the cycle of reproduction on which we embarked when discussing meiosis, we need to pass, by a process known as mitosis, from the zygote to an adult capable of producing its own gametes.

Mitosis

The growth of a single-celled zygote into a multicellular adult involves a mechanism whereby the number of cells can be expanded rapidly, while at the same time ensuring that each cell has exactly the same set of chromosomes as the original single-celled zygote. Mitosis is such a mechanism. For convenience we shall consider just two chromosomes (the sex chromosomes) in a male; but the process is exactly the same for all chromosomes in both sexes. As shown in Fig. 1.6, mitosis begins when each chromosome duplicates itself to form two chromatids still joined at their centromere. Each duplicated chromosome moves to the centre of the cell but does not, as in meiosis, synapse with its homologue. This stage, which is known as **metaphase**, is the one at which chromosomes are most readily visible. Karyotypes, therefore, consist of metaphase chromosomes. After metaphase, the centromere splits and the chromatids separate (disjoin), one going to each pole of the cell. A constriction forms in the centre of the cell and two cells are formed, each containing both an X and a Y. In this way, the two cells have exactly the same set of chromosomes as did the original cell.

In both meiosis and mitosis, chromosomes are duplicated. How does this happen?

The biochemistry of inheritance

Chemically, chromosomes consist of mostly deoxyribonucleic acid (DNA) with a small amount of a protein called histone. The latter has a binding and structural function, while the former constitutes the genetic information that is passed from 'parent' cell to 'offspring' cell during mitosis, and from one generation to the next, via meiosis.

DNA

DNA consists of two strands, each of which is a linear arrangement of **nucleotides**. All nucleotides of DNA contain an identical pentose sugar molecule (deoxyribose)

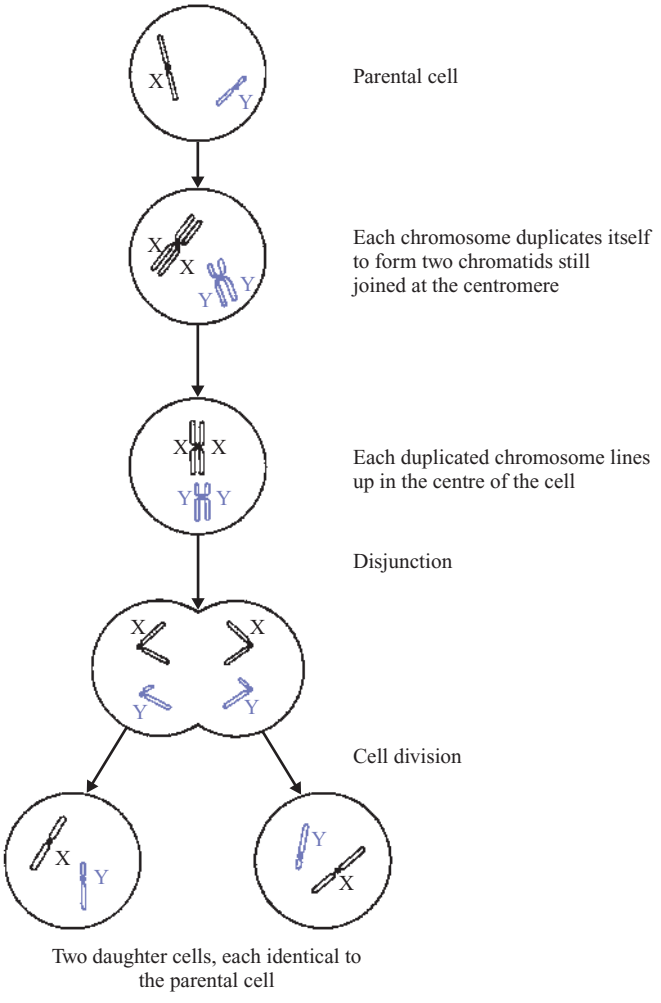


Fig. 1.6 Mitosis in a male, illustrated in terms of the sex chromosomes. The process is exactly the same for all chromosomes, and in all cells of each sex.

and an identical phosphate group. Their third component, a nitrogenous base, exists in four different forms (adenine: A; guanine: G; thymine: T; cytosine: C), giving rise to four different nucleotides, as illustrated in Fig. 1.7a. The bases A and G have a similar structure and are called **purines**; T and C have a similar structure and are called **pyrimidines**. A strand of nucleotides is held together by covalent bonding between the phosphate attached to the 5′ (pronounced 'five prime') carbon of one nucleotide and the OH attached to the 3′ carbon of the adjacent nucleotide, as shown in Fig. 1.7b. It follows that a strand of DNA has a 5′ phosphate at one end (called the **5′ end**) and a 3′ OH at the other end (called the **3′ end**).

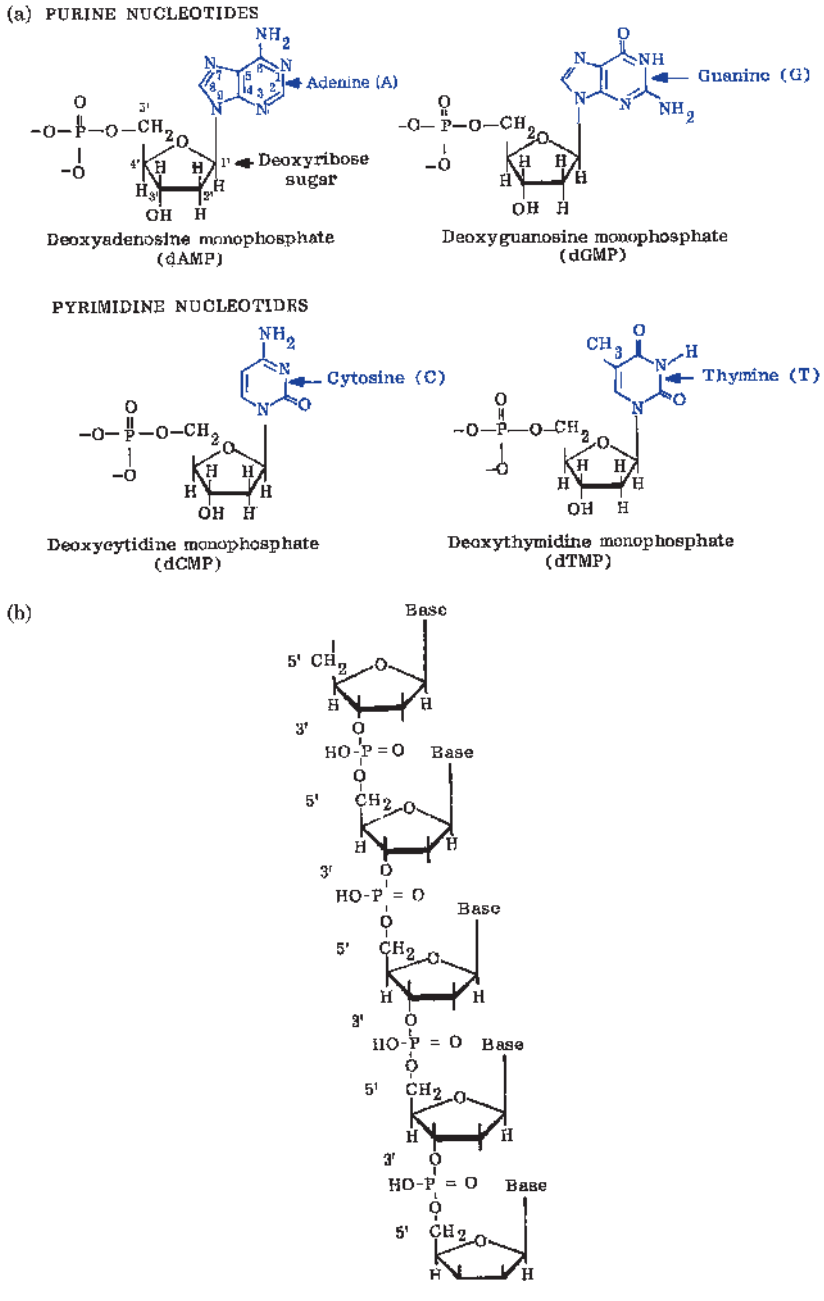


Fig. 1.7 (a) The chemical structure of the four nucleotides that are the building blocks of DNA. (b) The basic structure of a strand of DNA. ((a) Reprinted by permission from Freeman and Co.: Suzuki, Griffiths and Lewontin (1981) *An Introduction to Genetic Analysis* (2nd ed.). (b) Adapted courtesy of the estate of R.H. Symons from Symons (1981) in H. Messel (ed.) *The Biological Manipulation of Life*, Pergamon.)

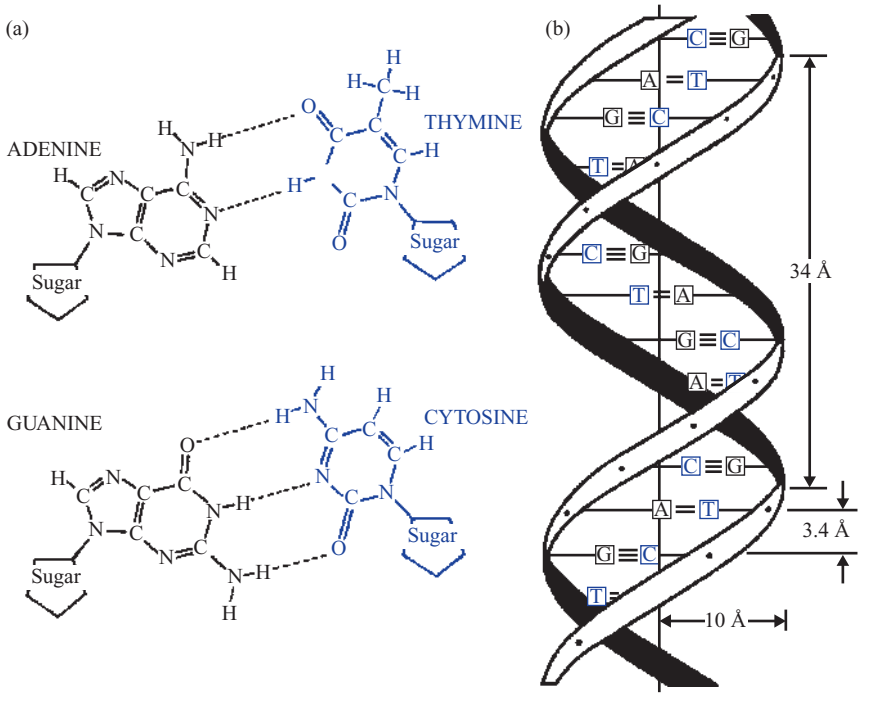


Fig. 1.8 (a) The two types of pyrimidine : purine base pairs formed by hydrogen bonding between the two strands of DNA. (b) A double helix of DNA. The two ribbons represent the sugar–phosphate 'backbones'. The structure repeats every 10 base pairs. (Adapted courtesy of the estate of R.H. Symons from Symons (1981) in H. Messel (ed.) *The Biological Manipulation of Life*, Pergamon.)

The two strands that constitute DNA are held together by very specific hydrogen bonding between purines and pyrimidines (A with T and G with C; Fig. 1.8a), giving rise to the **base pairs** A : T and G : C. Since A binds only with T and G binds only with C, one strand of DNA is complementary to the other; the sequence of bases in one strand can be predicted from the sequence in the other strand. A further consequence of the pairing arrangements is that the two strands occur together in a helix. Because two strands are involved, it is known as a **double helix** (Fig. 1.8b). The length of a short segment of DNA is usually measured in terms of the number of base pairs (bp). Longer segments are measured in terms of kilobases (1 kb = 1000 bases) or even megabases (1 Mb = 1000 kb).

The most important aspect of DNA structure is that it immediately suggests a mechanism for replication. If the double helix begins to unwind and the two strands separate, free nucleotides present in the cell are able to pair with the bases of each strand, forming a new and complementary strand for each of the original strands. As the unwinding proceeds (Fig. 1.9), two double helixes are produced from one original double helix; DNA has been replicated; a chromosome has been duplicated. The formation of each new strand by the addition of nucleotides is accomplished with the aid of the enzyme **DNA polymerase**. However, this enzyme can add nucleotides only

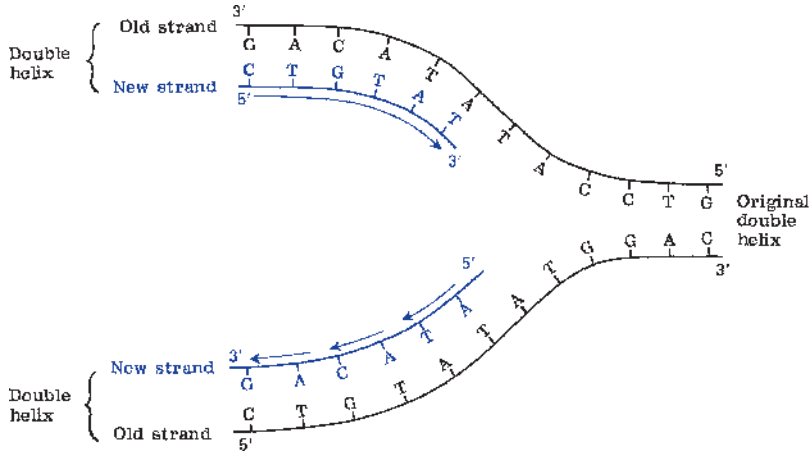


Fig. 1.9 Replication of DNA.

at the 3′ end of a growing strand, which means that replication can occur only in the 5′ to 3′ direction. Consequently, one new strand (top of Fig. 1.9) is synthesized continually, while the other new strand (bottom of Fig. 1.9) is assembled in small segments (called **Okazaki fragments**) which are each synthesized in the 5′ to 3′ direction. The Okazaki fragments are subsequently joined together by another enzyme called **DNA ligase**. The ability of these two enzymes to perform these functions has been put to good use in molecular biology, as described in Chapter 2.

It remains now to relate our knowledge of the structure of DNA to the structure of a chromosome as seen through a microscope. The total length of DNA in a mammalian cell is approximately 2 metres, which is more than 8000 times the total length of metaphase chromosomes viewed through a microscope! Obviously, therefore, a chromosome is composed of DNA that is very tightly folded or coiled. This raises the question as to how such a tight coil is unwound each time a chromosome replicates itself. The histone proteins are certainly involved in chromosome replication, but the actual mechanism has not yet been fully revealed.

The structure of DNA is the key to understanding the way in which genetic information is stored in the chromosomes, and is transmitted to the cell in such a way as to produce a particular effect. In fact, the sequence of bases in DNA has a very specific meaning recorded in the form of a code.

The genetic code

Proteins are chemical compounds with a wide range of specific roles in living organisms. Some are involved in transport (e.g. haemoglobin), support (e.g. collagen), or immunity (e.g. antibodies); some are enzymes that catalyse the innumerable biochemical reactions that occur in living cells (e.g. alcohol dehydrogenase). Some are hormones (e.g. growth hormone); some are receptors for hormones (e.g. oestrogen receptor). Some control the flow of molecules or ions in and out of cells (e.g. calcium release channel). In addition, the commercial products commonly obtained from

animals either consist almost solely of protein, e.g. meat and wool, or have protein as an important component, e.g. milk and eggs.

Proteins consist of one or more polypeptides, each of which is a chain of amino acids. There are 20 different amino acids. Each polypeptide has a specific sequence of amino acids that confers upon it a specific set of physical and chemical properties.

The information necessary for producing a specific sequence of amino acids is contained in code form within the sequence of bases in a segment of DNA. This code, which is called the **genetic code**, exists as triplets of bases (Table 1.2). With $4 \times 4 \times 4 = 64$ different possible triplets and only 20 amino acids, there is obviously some **redundancy**; in fact, the first two bases of a triplet are often sufficient to specify a particular amino acid, e.g. the triplets GTT, GTC, GTA, and GTG all specify valine. Three triplets (TAA, TAG, and TGA) do not code for any amino acid, and are known as **stop** triplets; they bring about the termination of a polypeptide chain. Another triplet (ATG) acts as a **start** signal for polypeptide synthesis. (It also codes for methionine.) The DNA between and including the start and stop triplets is called an **open reading frame** (ORF) or **coding sequence** (CDS) or **coding region**, within which the base sequence is 'read' in triplets, each of which encodes an amino acid.

Equipped with the genetic code, we can now follow the processes involved in the synthesis of proteins.

Protein synthesis

As shown in Fig. 1.10, the synthesis of polypeptides begins with the relevant segment of DNA unwinding, and the two strands separating. The sequence of DNA bases in one of the strands (called the **template** strand) acts as a template for the synthesis of a different nucleic acid (ribonucleic acid, RNA, so-called because its nucleotides contain ribose rather than deoxyribose). The synthesis is catalysed by the enzyme **RNA polymerase**, which, like DNA polymerase, adds nucleotides at the 3′ end of the growing strand, i.e. RNA is also synthesized in the 5′ to 3′ direction. Three of the bases in RNA are the same as in DNA, and the fourth, uracil (U), occurs instead of thymine. The formation of a complementary strand of RNA on the DNA template is called **transcription** (because the base sequence in DNA has been transcribed to RNA).

Before the next stage can commence, the RNA has to move from the nucleus, where the chromosomes are, to structures called **ribosomes** in the cytoplasm, where polypeptides are synthesized. (Obviously, this step is necessary only in organisms whose cells have a nucleus, i.e. eukaryotes. In prokaryotes, which have no nucleus, ribosomes become attached directly to RNA even before transcription has finished.) Because the RNA carries the code between DNA and protein, it is called **messenger RNA** or **mRNA**. Its triplets are called **codons** (shown in Table 1.2).

As also shown in Fig. 1.10, the second stage of protein synthesis involves a second type of RNA known as **transfer RNA** or **tRNA**. For each of the 20 amino acids, there is one or more specific tRNA molecules which bind to the relevant amino acid

Table 1.2 The genetic code

Triplet in DNA coding strand[1]	mRNA codon	Amino acid[2]	Triplet in DNA coding strand[1]	mRNA codon	Amino acid[2]
TTT	UUU	Phenylalanine	TAT	UAU	Tyrosine
TTC	UUC	(Phe; F)	TAC	UAC	(Tyr; Y)
TTA	UUA		TAA	UAA	STOP
TTG	UUG		TAG	UAG	
CTT	CUU	Leucine	CAT	CAU	Histidine
CTC	CUC	(Leu; L)	CAC	CAC	(His; H)
CTA	CUA		CAA	CAA	Glutamine
CTG	CUG		CAG	CAG	(Gln; Q)
ATT	AUU	Isoleucine	AAT	AAU	Asparagine
ATC	AUC	(Ile; I)	AAC	AAC	(Asn; N)
ATA	AUA		AAA	AAA	Lysine
ATG	AUG	START/ Methionine (Met; M)	AAG	AAG	(Lys; K)
GTT	GUU		GAT	GAU	Aspartic acid
GTC	GUC	Valine	GAC	GAC	(Asp; D)
GTA	GUA	(Val; V)	GAA	GAA	Glutamic acid
GTG	GUG		GAG	GAG	(Glu; E)
TCT	UCU		TGT	UGU	Cysteine
TCC	UCC	Serine	TGC	UGC	(Cys; C)
TCA	UCA	(Ser; S)	TGA	UGA	STOP
TCG	UCG		TGG	UGG	Tryptophan (Trp; W)
CCT	CCU		CGT	CGU	
CCC	CCC	Proline	CGC	CGC	Arginine
CCA	CCA	(Pro; P)	CGA	CGA	(Arg; R)
CCG	CCG		CGG	CGG	
ACT	ACU		AGT	AGU	Serine
ACC	ACC	Threonine	AGC	AGC	(Ser; S)
ACA	ACA	(Thr; T)	AGA	AGA	Arginine
ACG	ACG		AGG	AGG	(Arg; R)
GCT	GCU		GGT	GGU	
GCC	GCC	Alanine	GGC	GGC	Glycine
GCA	GCA	(Ala; A)	GGA	GGA	(Gly; G)
GCG	GCG		GGG	GGG	

[1] Following the usual convention, DNA sequences are written in the DNA-equivalent of RNA, which is the sequence in the non-template (coding) strand (see Fig. 1.10).
[2] The symbols in brackets after each amino-acid name are the standard three-letter and single-letter abbreviations.

and which have a nucleotide triplet (called an **anticodon**) that is complementary to an mRNA codon. Being complementary, a tRNA anticodon pairs with the relevant mRNA codon, bringing the correct amino acid into position in the polypeptide chain. This second and final stage of protein synthesis is called **translation**, because the

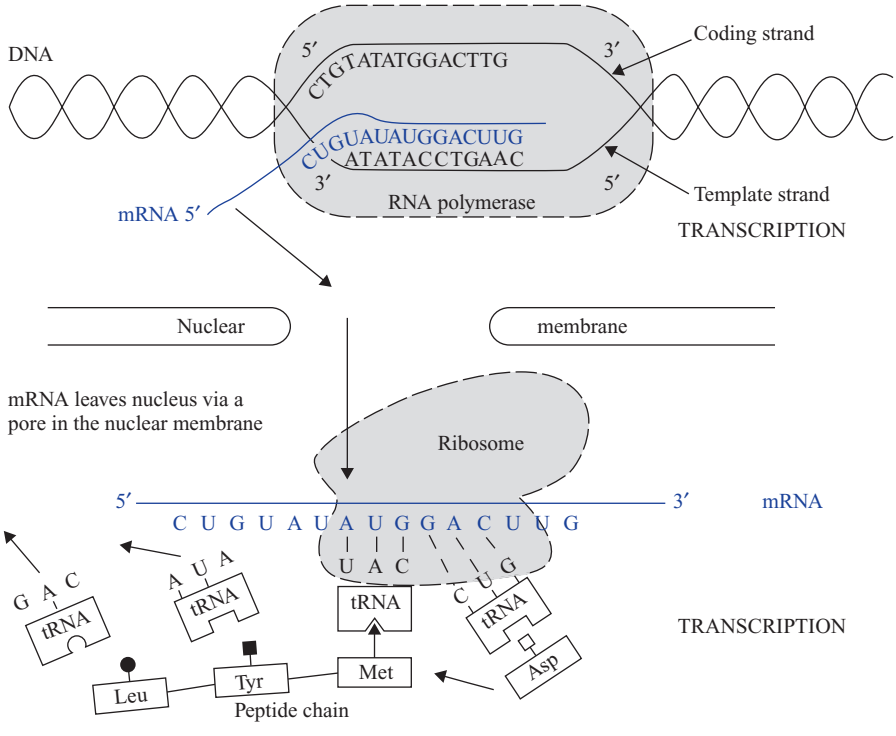


Fig. 1.10 Synthesis of polypeptide in eukaryotes, by means of transcription and translation.

base sequence has been translated (via the genetic code) to a sequence of amino acids.

What is a gene?

From the above account, it is evident that particular segments of DNA code for particular polypeptides, which are produced via mRNA. A segment of DNA including all the nucleotides that are transcribed into mRNA is called a **structural gene** (Fig. 1.11).

Since it is the mRNA sequence that is actually translated into polypeptide, and since translation starts at the 5′ end of mRNA, there is a convention that the base sequence of a gene is written as the DNA equivalent of the mRNA sequence, in the 5′ to 3′ direction. Recalling that the base sequence of mRNA is complementary to the sequence in the DNA strand from which it was transcribed, it follows that the DNA strand whose sequence is equivalent to the mRNA sequence is not the template strand, but the other strand. For this reason, the non-template strand is called the

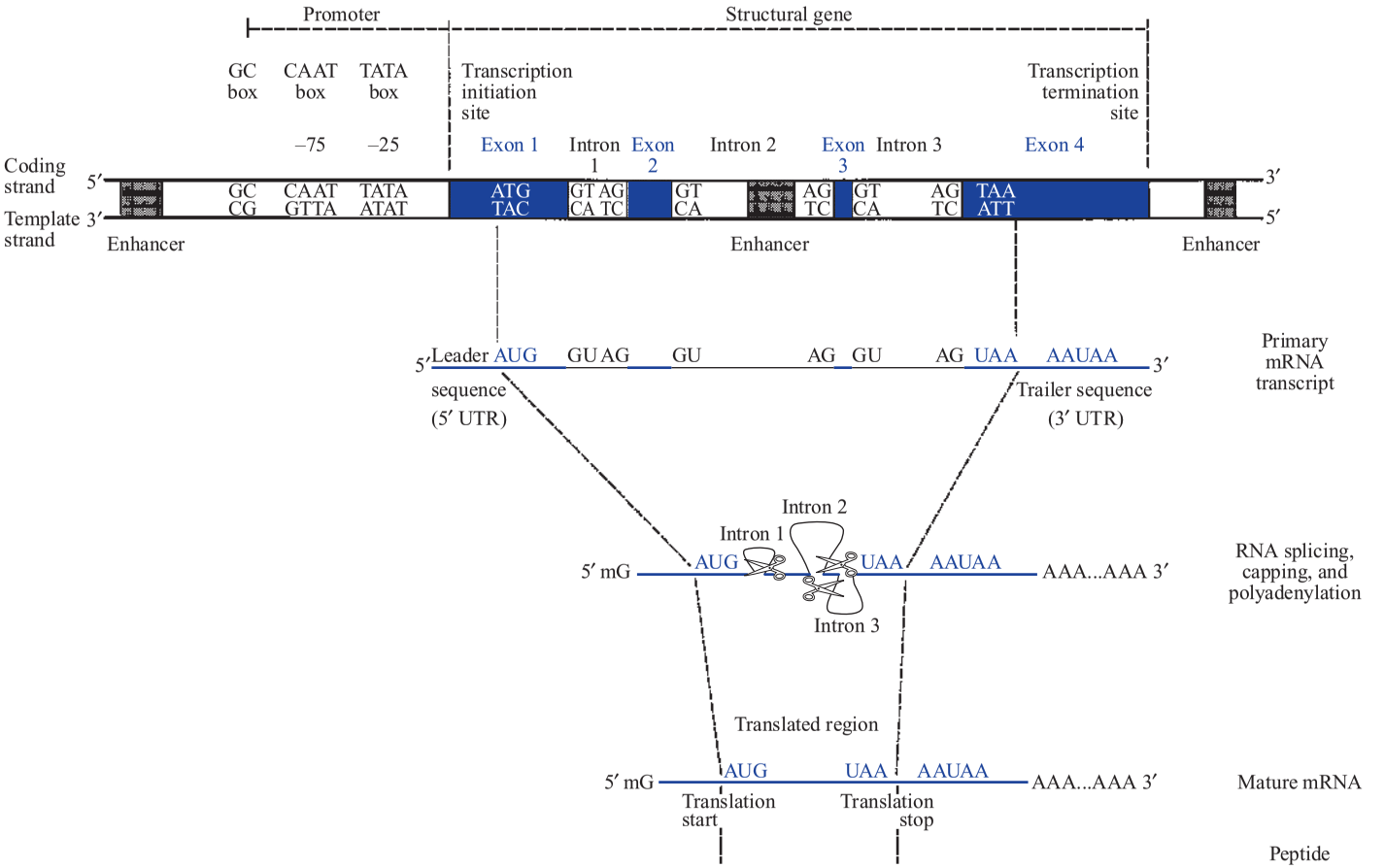


Fig. 1.11 The essential features of a gene, its messenger RNA and its peptide product.

sense or **coding** strand, while the template strand is called the **antisense** or **anticoding** strand.

Obviously a structural gene includes all of the nucleotides that are eventually translated into polypeptide, i.e. everything from the first nucleotide of the start triplet to the third nucleotide of the triplet immediately prior to the stop triplet. But it includes more than this. In fact, transcription starts before the start triplet and finishes after the stop triplet. This means that mRNA has **untranslated regions** (UTRs) at each end. The region that occurs before the start triplet (the **leader sequence** or 5′ UTR) is where ribosomes become attached. The untranslated region at the other end (the **trailer sequence** or 3′ UTR) is required for the processing of mRNA. The definition of a structural gene includes the sections of DNA that correspond to these regions. Thus, the first nucleotide of a structural gene is the nucleotide at the point where transcription actually commences, i.e. at the **transcription initiation site**. The end of a structural gene is called the **transcription termination site**. By convention, nucleotides of a structural gene are numbered from the start of the transcription initiation site, and bases preceding the site are numbered negatively, i.e. −1, −2, etc.

The region immediately preceding, i.e. **upstream** from, the transcription initiation site is particularly important, because it is the site to which RNA polymerase becomes attached prior to the initiation of transcription. This region is called the **promoter**. It contains specific sequences that are highly conserved, by which we mean that the same or very similar base sequence occurs in most genes. By studying the sequence at these conserved regions in many genes in many species, a **consensus sequence** can be deduced, consisting of the most common base at each position. The first such sequence in the promoter of eukaryotic genes is TATAAAA (called the **TATA box**), which is located approximately 25 bases upstream, i.e. at around position −25. Further upstream is the sequence GGCCAATCT (the **CAAT box**) at around position −75, and GGGCGG (the **GC box**) at around −90. These boxes are the sites for recognition and binding of regulatory proteins called **transcription factors**, which enable RNA polymerase to be positioned correctly for the initiation of transcription. In this way, they exert control over transcription. Prokaryotes have similarly conserved sequences in their promoters, namely TATAAT and TTGACA.

Termination of transcription is less well understood than initiation. It is known, however, that transcription actually extends beyond what we have called the termination site. Then the transcript is cleaved at the termination site. There is no conserved sequence corresponding to this site, but there is a highly conserved region with consensus sequence AATAAA (AAUAAA in the mRNA) located 10–30 bases before the termination site, which appears to be a recognition site in mRNA for a factor that controls the cleavage.

Split genes

Until 1977, it was thought that structural genes were just long enough to encode the mRNA that is involved in translation. In that year, it was shown that in eukaryotes, most structural genes are longer than this – they contain sections that are not represented in the mRNA by the time it undergoes translation. These sections are called

introns (because they were originally called *intr*agenic regi*ons*). Sections that are represented in the final (**mature**) version of mRNA are called **exons** (*ex*pressed regi*ons*). Genes that contain introns are said to be **split**. The original (**primary**) mRNA is a copy of the whole of a structural gene, i.e. exons and introns. Before the mRNA moves from the nucleus to the cytoplasm, the introns are removed and the exons are spliced together (in a process called **RNA splicing**), so that the mature mRNA consists solely of exons. The recognition and removal of introns is aided by the fact that in each intron, the first two bases at the 5′ end and the last two bases at the 3′ end are highly conserved – the same bases (GU and AG respectively in the mRNA) occur at these sites in most, if not all, introns. Recalling that sequences are usually expressed in terms of the coding strand of DNA, this is called the **GT–AG rule**. The base sequence immediately on either side of the highly conserved sequences, i.e. across the boundaries between introns and exons, is also conserved, but to a lesser degree. These boundaries are called **splice sites**, with the 5′ site being called the **donor site** and the 3′ site the **acceptor site**.

In the domesticated animal species for which good data exist, an average structural gene is around 43 000 bases (43 kb), comprising 10 exons and nine introns; with an average mature mRNA molecule (mature transcript) of around 2000 bases (2 kb). However, there is substantial variation around these averages. Exons range in size from just a few bases to around 17 000 bases (17 kb), with an average of around 200 bases. In contrast, introns range in size from just a few bases to 1 000 000 bases (1 Mb), with an average of around 4500 bases (4.5 kb). It is obvious from these figures that exons constitute only a small proportion of structural genes (a reasonable estimate is between 4 and 5 per cent, on average); most of the DNA in structural genes is in introns. The existence of so much 'non-functional' DNA within structural genes remains one of the great unsolved mysteries of biology. We should not be surprised, however, if the solution to the mystery is that introns really do have important functions, which are just waiting to be discovered. We already have some hints, in the form of evidence that some intronic RNA actually has a functional role in gene regulation (see next section). This is just one example of **non-protein-coding RNA** (ncRNA). In addition, introns tend to act as spacers between functional units, i.e. many exons correspond to functional units which can be assembled in different combinations to produce different polypeptides.

Apart from the excision of introns, the primary mRNA transcript is also modified in two other ways. Its 5′ (front) end is protected by the addition of a **5′ cap**, which consists of a methylated guanine nucleotide. At the other end, a **poly-A tail** is added, consisting of a variable number (typically 100–200) of adenine nucleotides. Since the cap and the tail are such important features of mature mRNA, the translation initiation and termination sites of a gene are sometimes called the **cap site** and the **poly-A site**. The AAUAAA sequence that occurs 10–30 bases upstream from the poly-A site is called the **polyadenylation signal**.

The type of gene described above is by far the most common. However, it is important to realize that there are segments of DNA whose sole function is the production, via transcription, of either tRNA or **ribosomal RNA** (**rRNA**, which is the major constituent of ribosomes). These are also called genes. Since a large

quantity of rRNA is required for the construction of sufficient ribosomes to satisfy each cell's requirements for translation, there are hundreds of genes for rRNA in the genome. They occur in several clusters of tandemly repeated rRNA genes. Each cluster produces a **nucleolus**, which is a discrete structure found within the nucleus, and which consists mainly of ribosomal RNA plus the enzymes necessary for the assembly of ribosomes. A cluster of rRNA genes is called a **nucleolar organizer region** (NOR).

In order to accommodate all the possibilities outlined above, we can redefine a gene as a stretch of DNA that produces a functional RNA molecule.

Gene regulation

It is obvious that there would be complete chaos if all genes were transcribed in all cells all at the same time. In fact, only a small proportion of genes are transcribed at any one time in any one cell. From the moment of fertilization until death, the development of each living organism is determined by genes being switched on and off at the appropriate time(s) in the appropriate cell(s). This switching is achieved by various proteins (sometimes in association with steroid hormones) becoming attached to, or being released from, specific sequences of DNA that are often highly conserved. We have already encountered three such sequences in promoters. Sequences having a similar function other than within promoters are called **enhancers**. These are located upstream, downstream, and sometimes even within structural genes (i.e. within introns). Some are located in the near vicinity of the structural gene, but others are quite some distance (20 kb or even further) away.

The regulatory proteins that exert control over transcription by binding to promoters or enhancers have in common one or more amino-acid sequences (called amino-acid **motifs**) that have been given some rather exotic names. For example, **zinc fingers** are structures which, typically, involve the repeated occurrence of a pair of cysteine molecules separated by two or three other amino acids, followed 10 or so amino acids later by a pair of histidine molecules, also separated by two or three other amino acids. In combination with a zinc atom, the two cysteines link with the two histidines, and the intervening amino acids form a finger-like loop which binds with DNA. Another example is the **leucine zipper**, which contains a periodic repeat of leucine every seventh amino acid, giving rise to a helix with the leucines aligned along one face. Two such molecules readily join together (in the manner of a zipper), creating a dimer. One end of this dimer binds to DNA.

In general, it is the binding of regulatory proteins (sometimes in conjunction with steroid hormones) to DNA that determines whether or not RNA polymerase can attach to the promoter of a gene, thereby exerting control over the first step in transcription. Regulatory proteins that increase the rate of transcription are called **activators**; those that decrease the rate of transcription are called **repressors**. In essence, these are the means by which genes are switched on and off.

For example, early embryonic development is heavily influenced by a set of **homeotic** genes (each of which determines the developmental fate of one segment of the embryo, e.g. hindbrain or spinal cord). Proteins encoded by homeotic genes have DNA-binding motifs such as zinc fingers or leucine zippers. Many of these genes also have a highly conserved 180-bp region called the **homeobox**, which encodes a DNA-binding motif called the **homeo domain** that is highly conserved throughout eukaryotes. Thus the genes that control early embryonic development encode regulatory proteins that switch genes on and off in a controlled manner.

A specific example of gene regulation later in life is provided by the way in which oestrogen controls the transcription of the ovalbumin gene in chickens. Molecules of the hormone enter a cell and bind with a protein called the oestrogen receptor. The complex of oestrogen plus receptor then binds to a region approximately 250 bases upstream from the TATA box of the gene for ovalbumin, switching it on.

Not surprisingly, groups of genes that all need to be controlled in a similar manner have a region of their promoters in common. These regions are called **response elements**. For example, all genes that need to be activated by glucocorticoids have a glucocorticoid response element, to which the glucocorticoid receptor binds, following activation by glucocorticoid. Its consensus sequence is TGGTACAAATGTTCT.

From the above discussion, it is clear that the sequences on either side of a gene are just as important as the gene itself. In fact, when the word ‘gene’ is used on its own, it is often taken to include the promoter and enhancer regions as well as the region in between.

Mutation

We have seen how DNA replicates, and how it gives rise to protein. Although the processes involved are remarkably elegant and usually operate faultlessly, mistakes do occur from time to time. Many mistakes have no effect at all, as they are corrected by the cell’s own repair mechanisms. Uncorrected mistakes in DNA replication, however, result in an alteration in the DNA in at least one of the new cells. And because DNA replication is usually so faultless, the altered DNA is passed on unchanged to all descendent cells; until the next mistake occurs. Uncorrected mistakes in DNA replication are called **mutations**.

We shall start by considering **point mutations** (also called **gene mutations**), which involve the substitution of one nucleotide for another, or the addition or deletion of one or a few nucleotides. Other types of mutation will be considered in subsequent chapters.

There are several different possible consequences of point mutations. At one extreme, a base substitution can change a functional triplet into a stop triplet (called a **nonsense** mutation). For example, TAT codes for tyrosine; but if T in the third position is replaced with A, the resulting triplet (TAA) means stop (check this in Table 1.2). If the new stop triplet occurs before the usual stop triplet, the resultant

polypeptide is shorter than usual, and is therefore probably not functional. If a base substitution changes a triplet so as to cause an amino-acid substitution, it is called a **mis-sense** mutation. For example, substituting A for T at the third position of CAT (histidine) results in CAA (glutamine).

At the other extreme, many base substitutions have no effect on the amino acid sequence of the gene product, because the mutant triplet happens to specify the same amino acid as the original triplet. These so-called **silent** mutations are a direct consequence of the redundancy in the genetic code. For example, substituting C for T at the third position of CAT (histidine) results in CAC, which still codes for histidine (check this in Table 1.2).

Another type of point mutation involves the insertion or deletion of one or two bases, known collectively as **indels**. These are called **frameshift** mutations, because each of the triplets that occurs downstream from the site of such a mutation is shifted out of phase from the original open reading frame. At the very least, a frameshift mutation results in a completely different sequence of amino acids downstream from the site of mutation. For example, consider the following case (check it against Table 1.2):

Original coding strand TCCGAGTATCAGTCCCAG …
Amino acid sequence Ser Glu Tyr Gln Ser Gln …

If the second base is deleted, we have:

Mutant coding strand TCGAGTATCAGTCCCAG …
Amino acid sequence Ser Ser Ile Ser Pro …

The mutant amino acid sequence is obviously very different from the original sequence. In many cases, one of the 'new' triplets is a stop triplet, causing premature termination of translation. Whether or not a new stop triplet is created, it is most unlikely that the mutant polypeptide will be functional.

If a mutation occurs in cells other than those that give rise to sex cells, it is called a **somatic mutation**. The stage of development of the animal when a somatic mutation occurs determines the total number of cells that contain the altered or mutant DNA; in general, the earlier the mutation occurs, the larger the number of cells affected.

In contrast, a mutation that occurs in cells that give rise to sex cells is known as a **germ-line mutation**, which may lead to the formation of a gamete that contains the altered DNA. If this gamete is successful in fertilization, the mutation is passed on to the resultant offspring, in every cell of which it is faithfully reproduced.

Genes, alleles, and loci

The different forms of a segment of DNA that can exist at a particular site in a chromosome are called **alleles**. The particular site of a gene in a chromosome is

called the **locus** (plural **loci**). The word 'gene' is commonly used in the sense of either allele or locus. When used in this way, the appropriate meaning of the word is usually quite evident from its context.

If an offspring results from the union of a sperm with normal DNA and an ovum with an altered or mutant DNA segment in one of its chromosomes, that offspring has one normal chromosome and one mutant chromosome making up the pair of homologues. More specifically, there is one normal allele and one mutant allele at the relevant locus. In the examples below, we shall give these two alleles the symbols *B* and *b* respectively. Animals with two different alleles at a particular locus are said to be **heterozygous** at that locus. In contrast, if an animal has two copies of the same allele, the animal is **homozygous** at that locus.

Although any one animal can have a maximum of only two different alleles at a locus, the number of different alleles in a population of animals can be much greater than two. If more than two alleles exist in a population at a particular locus, that locus is said to have **multiple alleles**.

Simple or Mendelian inheritance

The passage of genes from one generation to the next is called **inheritance**. One of the major breakthroughs in science was the realization that the results of inheritance can be predicted. The first person to formulate these predictions was the Augustinian monk Gregor Mendel, who conducted his research in a monastery in Brün (now Brno in the Czech Republic) in the middle of the 19th century.

Single locus

Consider, for example, the mating of a heterozygote (*Bb*) with a homozygote (*bb*). This is exactly analogous to the situation with sex chromosomes. Consequently, the results of the mating *Bb* × *bb* can be represented in exactly the same way as that used for the inheritance of sex discussed earlier, with the aid of a checkerboard:

		Gametes from heterozygous parent	
		½ *B*	½ *b*
Gametes from homozygous parent	all *b*	½ *Bb*	½ *bb*

The result of the mating *Bb* × *bb* is expected to be an equal proportion of *Bb* and *bb* offspring. The separation of alleles at a locus during meiosis is called **segregation**, and the ratio of different types of offspring is known as the **segregation ratio**. For the mating *Bb* × *bb*, the segregation ratio is $\frac{1}{2}$ *Bb*: $\frac{1}{2}$ *bb*, which is often written as 1 *Bb*:1 *bb*.

A checkerboard can be used to predict the outcome of any particular mating involving a single locus. The segregation ratios expected from all possible types of mating with respect to a single locus are listed in Table 1.3.

Table 1.3 Segregation ratios expected in the offspring arising from all possible types of mating in relation to a single autosomal locus, as obtained from a checkerboard

Type of mating	Segregation ratio				
	BB		*Bb*		*bb*
BB × *BB*	1	:	0	:	0
BB × *Bb*	1	:	1	:	0
BB × *bb*	0	:	1	:	0
Bb × *Bb*	1	:	2	:	1
Bb × *bb*	0	:	1	:	1
bb × *bb*	0	:	0	:	1

More than one locus

In the absence of any evidence to the contrary, it is assumed that segregation at one locus is independent of segregation at other loci. This was the assumption made by Mendel, and it is true for many situations observed in animals today. If segregation at each locus is independent of segregation at other loci, the chance of obtaining a gamete with a particular allele (say *B*) at the first locus and a particular allele (say *d*) at the second locus is simply the product of the probabilities associated with each allele independently. For example, if an individual is heterozygous at two loci (*Bb Dd*), there are four possible types of gametes, *BD*, *Bd*, *bD*, and *bd*, each of which is produced with a frequency of $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$.

The results of independent segregation at two loci can also be shown in a checkerboard:

		Gametes from one parent			
		¼*BD*	¼*Bd*	¼*bD*	¼*bd*
Gametes from other parent	¼*BD*	*BBDD*	*BBDd*	*BbDD*	*BbDd*
	¼*Bd*	*BBDd*	*BBdd*	*BbDd*	*Bbdd*
	¼*bD*	*BbDD*	*BbDd*	*bbDD*	*bbDd*
	¼*bd*	*BbDd*	*Bbdd*	*bbDd*	*bbdd*

Combining all cells of the checkerboard having identical offspring, and realizing that the offspring in each cell occur with a frequency of $\frac{1}{4} \times \frac{1}{4} = \frac{1}{16}$, the segregation ratio is:

1*BBDD* : 2*BBDd* : 1*BBdd* : 2*BbDD* : 4*BbDd* : 2*Bbdd* : 1*bbDD* : 2*bbDd* : 1*bbdd*

Although checkerboards quickly become rather large, in principle they can be used to derive expected segregation ratios for any type of mating involving any number of independently segregating loci.

Sex-linkage

The above patterns of inheritance illustrate simple autosomal inheritance because they describe what happens in relation to loci on autosomes. Some loci, however, are on the sex chromosomes and consequently have different patterns of inheritance. Such loci are said to be **sex-linked**. To consider the most common form of sex linkage, the inheritance patterns of **X-linked** loci can be illustrated in a checkerboard:

		Male gametes	
		½X^H	½Y
Female gametes	½X^H	¼X^HX^H (female offspring)	¼X^HY (male offspring)
	½X^h	¼X^HX^h (female offspring)	¼X^hY (male offspring)

and are summarized in Table 1.4. Very few loci have been identified on Y chromosomes (**Y-linked**).

Table 1.4 Segregation ratios expected from all possible types of matings in relation to an X-linked locus, as obtained from a checkerboard

Type of mating	Segregation ratio								
	Among females					Among males			
	X^HX^H		X^HX^h		X^hX^h	X^HY		X^hY	
$X^HX^H \times X^HY$	1	:	0	:	0	1	:	0	
$X^HX^h \times X^HY$	1	:	1	:	0	1	:	1	
$X^hX^h \times X^HY$	0	:	1	:	0	0	:	1	
$X^HX^H \times X^hY$	0	:	1	:	0	1	:	0	
$X^HX^h \times X^hY$	0	:	1	:	1	1	:	1	
$X^hX^h \times X^hY$	0	:	0	:	1	0	:	1	

At the beginning of this section it was implied that sometimes segregation at two or more loci is not completely independent. We shall now examine why this is so.

Linkage

There are probably around 22 000 genes in mammals. In contrast (as we saw earlier), there are only relatively few chromosomes. Inevitably, therefore, each

chromosome consists of many different genes, each of which has a specific position (locus) on that chromosome. If chromosomes were inherited as integral units, then for all the loci on a particular chromosome, the alleles present in that chromosome would always segregate together. Consider, for example, one chromosome containing allele *B* at one locus and allele *D* at another locus, and its homologue containing alleles *b* and *d*, respectively. If chromosomes segregated as integral units, only two types of gametes would result, namely *BD* and *bd*, with equal frequency.

In practice, chromosomes are not inherited as integral units. Instead, as described earlier, recombination or crossing-over occurs when homologous chromosomes are synapsed during the first stage of meiosis. During synapsis, breakage and rejoining of chromatids occur. If the two segments of a broken chromatid rejoin, that chromatid is still inherited as an integral unit. If, however, a break occurs in the same position in two adjacent chromatids, sometimes the segments change partners, forming **recombinant** chromatids. If the two chromatids originated from just one homologue, i.e. are joined at the centromere (called **sister chromatids**), the cross-over has no effect, since sister chromatids are exact copies of each other. If, however, the two chromatids are **non-sister chromatids** (one from one homologue and one from the other), the cross-over results in the reciprocal exchange of genes between homologous chromosomes, as shown in Fig. 1.12. To the extent that breakages occur more or less randomly along the length of each chromosome, there is a direct relationship between the physical distance separating two loci on a chromosome and the chance (or probability) of a cross-over occurring between them. Unfortunately, this probability cannot be directly estimated. However, by observing the progeny of certain matings, we can calculate the **recombination fraction**, which is the proportion of gametes from one parent that can only have resulted from crossing-over during meiosis in that parent. Fig. 1.13 illustrates the concept of recombination fraction, for the two extreme cases of complete linkage and independence, and for an intermediate case.

If two loci are so close together on a chromosome that a cross-over never occurs between them (Fig. 1.13a), only two types of gametes are possible from the double-heterozygote parent, namely $\frac{1}{2}$ *EF* and $\frac{1}{2}$ *ef*. Both of these types of gametes are non-recombinants, because no recombination (crossing-over) has occurred between the two loci. Notice that non-recombinant gametes correspond to the two chromosomes of the double-heterozygote: one contains allele *E* and allele *F*, while the other contains allele *e* and allele *f*. When this double-heterozygote parent is mated with a double-homozygote parent, the checkerboard in Fig. 1.13a shows that two types of offspring (both non-recombinants) result with equal frequency. We can conclude that when loci are so close that recombination never occurs between them, the result is 100 per cent non-recombinant offspring and 0 per cent recombinant offspring.

If the two loci are some distance apart on the same chromosome (Fig. 1.13b), then in some germ cells undergoing meiosis, a cross-over will occur between the two loci. As we saw in Fig. 1.12, whenever a cross-over occurs between two loci on the same chromosome, the result is two types of recombinant gametes and two types of non-recombinant gametes. These are shown in Fig. 1.13b. For those meioses in

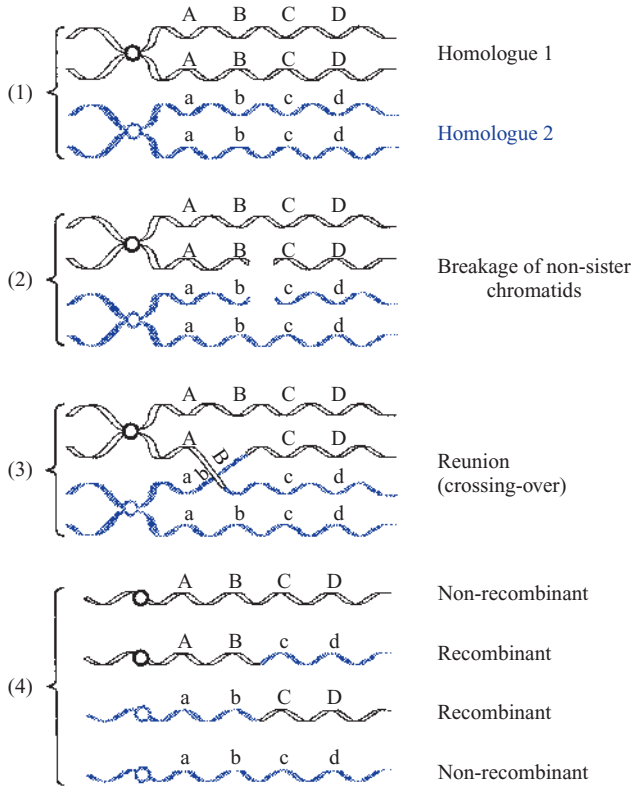


Fig. 1.12 The four stages involved in crossing-over between a pair of homologous chromosomes.

which a cross-over does not occur, only two types of gametes are produced, namely the two types of non-recombinant gametes. If the distance between the two loci is such that the chance of a cross-over between the two loci is, for example, 20 per cent, then 20 per cent of meioses in this parent will result in two recombinant gametes and two non-recombinant gametes, and the remainder (80 per cent) of meioses will result in four non-recombinant gametes. In Fig. 1.13b, we consider the results of 100 meioses (producing 400 gametes), with a 20 per cent chance of crossing-over between the two loci. Among the 400 gametes produced from this parent, the total number of non-recombinant gametes is 160 + 160 = 320 (from the meioses in which recombination did not occur) plus 20 + 20 = 40 from the meioses in which recombination did occur, making a total of 180 + 180 = 360 non-recombinant gametes. The other 40 gametes are recombinant gametes. The resultant offspring are shown in Fig. 1.13b: two types of recombinant offspring and two types of non-recombinant offspring. We can now estimate the proportion of offspring that can only have resulted from recombination between the two loci. This is the recombination fraction. In this particular example it is 40/400, i.e. 10 per cent. If the two loci had been closer together on the chromosome, the recombination fraction would

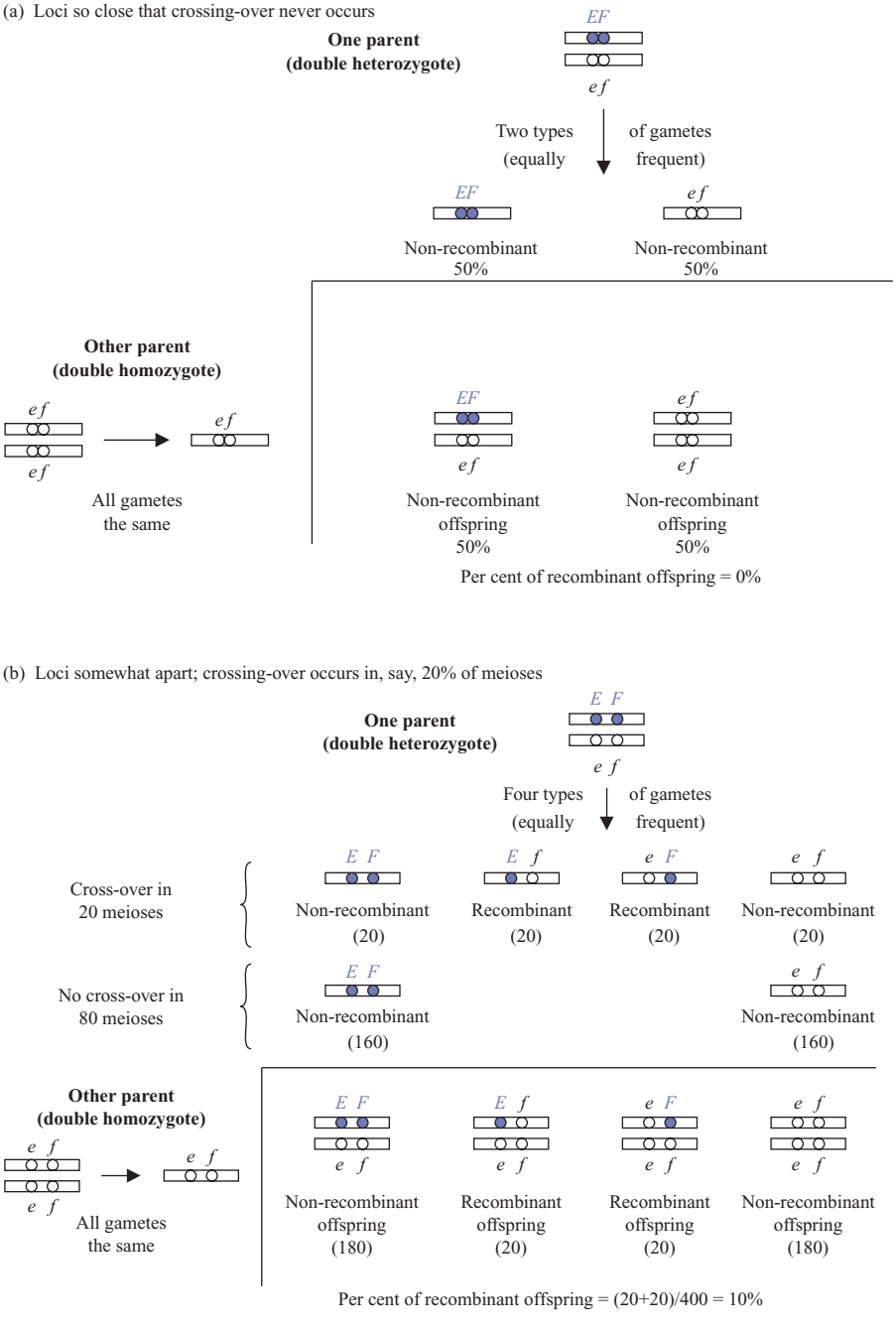


Fig. 1.13 The concept of recombination fraction, illustrated with a checkerboard. (a) Loci so close that crossing-over never occurs. (b) Loci somewhat apart; crossing-over occurs in, for example, 20 per cent of meioses. (c; next page) Loci so far apart that a cross-over occurs in every meiosis.

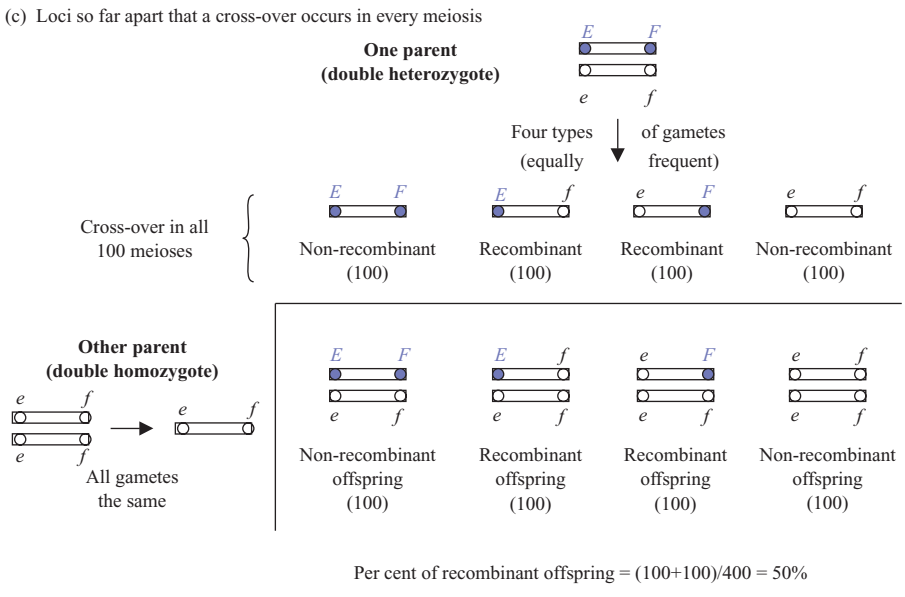


Fig. 1.13 *Continued*

have been less than 10 per cent; if the two loci were further apart on the same chromosome, the recombination fraction would have been greater than 10 per cent. From this, it can be seen that the recombination fraction is an indirect indicator of the distance between the two loci. This distance is called the **map distance**, and is measured in units of centimorgans (cM), named after Thomas Morgan, who discovered the phenomenon of linkage (for which he was awarded the Nobel prize – the first geneticist so honoured). For low recombination fractions (say, less than 10 per cent), 1 map unit (1 cM) equals 1 per cent recombination fraction. For distances greater than this, there is an appreciable chance that more than one cross-over will occur between the loci. Consequently, as the distance between a pair of loci increases, recombination fraction increasingly underestimates map distance. To cater for this, a mapping function such as that shown in Fig. 1.14 is used for estimating map distance from recombination fraction.

The final situation to consider is that shown in Fig. 1.13c, in which the two loci are so far apart that a cross-over always occurs between them. Since every meiosis in this situation involves a recombination between the two loci, each meiosis produces equal numbers of recombinant and non-recombinant gametes: all four types of gametes occur with an equal frequency (of 25 per cent), and, as shown in Fig. 1.13c, the recombination fraction is 50 per cent. You may recognize that this result is exactly the same as that seen for the segregation of two loci that are located on different chromosomes, as shown in the two-locus checkerboard on page 26. In other words, loci that are far apart on the same chromosome segregate independently, just as if they were located on different chromosomes. It should also now be clear that

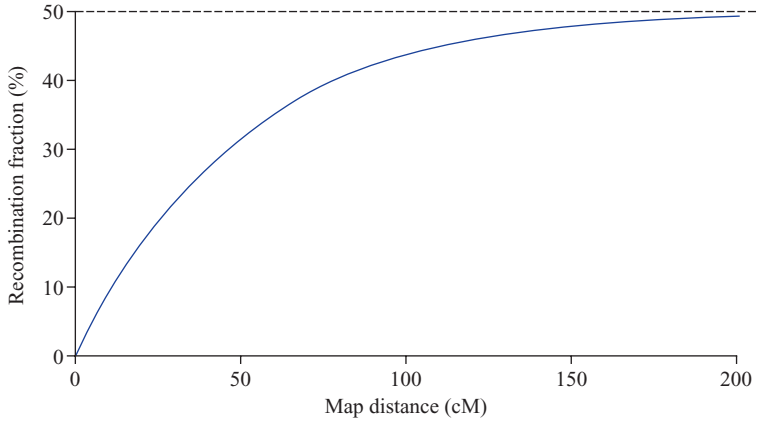


Fig. 1.14 A mapping function, showing the relationship between recombination fraction and map distance.

the maximum recombination fraction is 50 per cent, as shown in the mapping function in Fig. 1.14.

The relationship between the recombination fraction and the distance between loci enables the construction of linkage maps, in which loci are positioned according to the map distance between them. The construction of linkage maps, which have very important practical applications, is discussed in Chapter 2.

Inactivation

X-inactivation and dosage compensation

Among the many coat colours seen in cats, the mosaic of orange and non-orange, which is known as tortoiseshell (Fig. 1.15a) is one of the most attractive. The orange hairs result from the action of an X-linked allele *O*, which prevents the production of dark pigment (black and brown), but enables the production of yellow pigment. The non-orange hairs are due to the normal (**wild-type**) allele at the same locus, *o*, which enables the production of dark pigment, in whatever manner is determined by alleles at other coat-colour loci. (See Chapter 11 for more information on the genetics of coat colour.) Since obviously both alleles must be present to produce the mosaic of orange and non-orange, tortoiseshell cats must be heterozygous, $X^{O}X^{o}$, at this X-linked locus. But why do some parts of the body express the effect of the orange allele, while other parts express that of the non-orange allele? And why are the patches of non-orange and orange approximately equal in total area, and why are they scattered more or less randomly throughout the coat?

The answers to these questions lie partly in another observation, first made in cats, by Barr and Bertram, who in 1949 reported that the nucleus of non-dividing nerve cells in females usually contains a small dark-staining body, whereas that in males

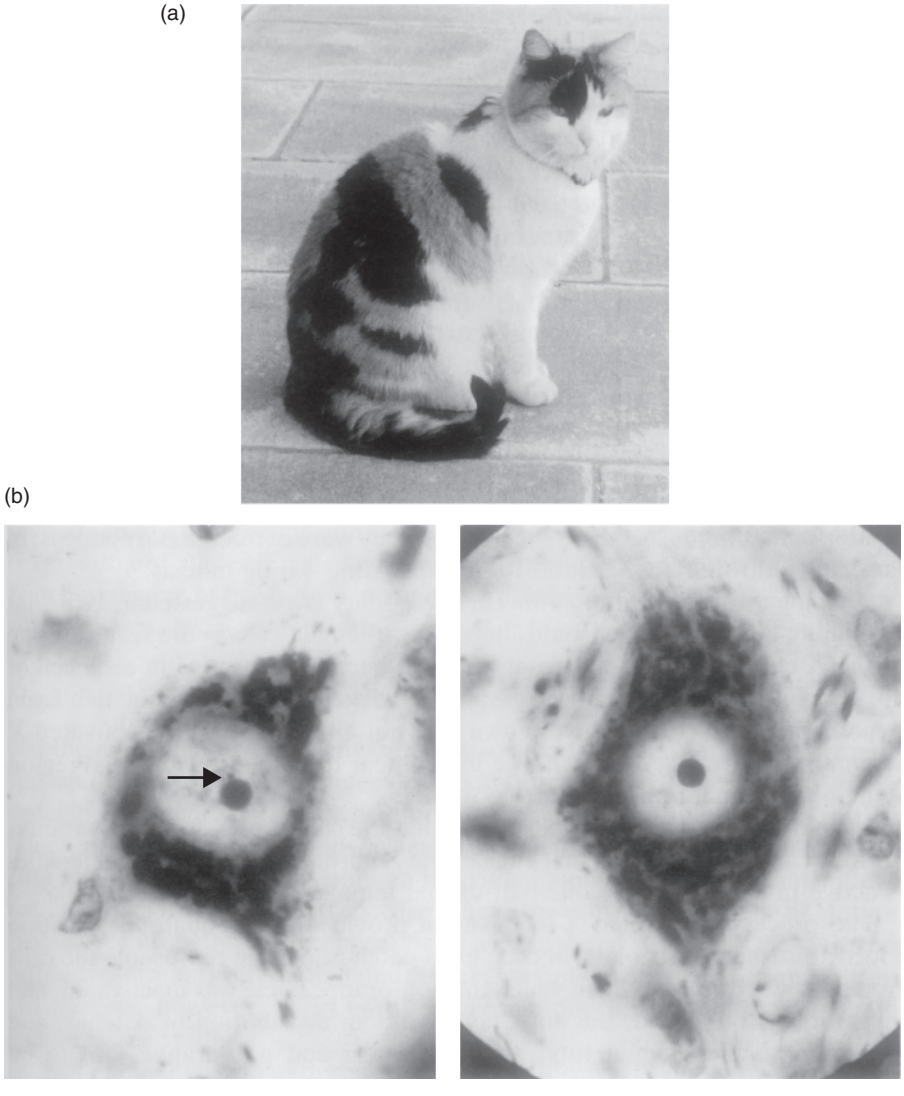


Fig. 1.15 (a) A tortoiseshell cat with white spotting. The white spotting is due to an allele at an autosomal locus (see Chapter 11).
(b) Nucleus of a hypoglossal nerve cell from a mature female cat (left) and a mature male cat (right). The large dark-staining body in each nucleus is a nucleolus. The small dark-staining body (arrowed) in the female cell is a Barr body. ((b) Reprinted by permission from Macmillan Publishers Ltd: *Nature*, (1949) **163**, pp 676–7.)

does not (Fig. 1.15b). This dark-staining body is now called a **Barr body** or **sex chromatin**. Although it had been observed by many previous researchers, Barr and Bertram were the first to note that the dark-staining body occurs in only female cells. In an attempt to explain their observations, they speculated that it might be an X chromosome which had become very highly condensed. Other researchers showed

that they were correct; a Barr body is, in fact, an X chromosome that is late in replicating during mitosis.

Drawing on similar observations in mice, Mary Lyon suggested in 1961 that the highly condensed X chromosome seen in female cells is the result of one of the X chromosomes (chosen at random) becoming inactive in each cell of all female embryos at an early stage of development. This is the **Lyon hypothesis**. (In fact, it is now known that not all genes on the inactivated X chromosome are inactivated; those in and near the pseudo-autosomal region remain functional in both X chromosomes.)

Since the Lyon hypothesis postulates that the choice of X for inactivation is entirely random, it follows that each of the X chromosomes in normal females is active in approximately one-half of that female's cells.

The process of **random X-inactivation** provides an adequate explanation for tortoiseshell coat colour in cats; each patch of orange represents the cells that descended from a cell in which the non-orange allele was inactivated, and *vice versa*. In addition, the apparently random distribution of patches and the approximately equal total area of orange and non-orange are to be expected if the inactivated X is chosen at random.

In passing, it should be noted that since tortoiseshell cats are heterozygous at an X-linked locus, they must have two X chromosomes, in which case they should be female. Normal male cats, having only one X chromosome, can be either orange ($X^{O}Y$) or non-orange ($X^{o}Y$), but not tortoiseshell. It is therefore a fairly safe bet that any tortoiseshell cat is a female. Occasionally male tortoiseshells are reported, but they often turn out to be abnormal males having an extra X chromosome, as described in Chapter 4.

The result of random X-inactivation is that each female is a **mosaic**, consisting of two distinct populations of cells derived from a common source; in one population of cells the maternal X chromosome is inactive, and in the other cell population, the paternal X is inactive.

The only well-documented exception to random X-inactivation occurs in marsupials, where it is the paternal X chromosome that is inactivated, and often only incompletely so. The reason for this is not known.

Obviously the result of X-inactivation in females is that each female cell has the same quantity of gene product from X-linked genes as do males. Thus, X-inactivation is a mechanism that compensates for the difference in gene 'dosage' between males and females in relation to X-linked genes. This effect of X-inactivation is called **dosage compensation**.

Finally, an important difference between mammals and birds must be noted: while X-inactivation appears to occur in all mammals, Z-inactivation does not occur in birds. The reasons for this are not known.

Imprinting

Inactivation is not confined to the X chromosome. At certain loci on other chromosomes, the extent to which an allele is expressed (or even whether it is expressed at

all) depends on the parent from which it came. This differential expression of alleles is called **genomic imprinting**. As might be imagined, this can be a source of frustration in attempts to determine the mode of inheritance of disorders, because imprinting can result in atypical inheritance patterns.

Inactivation results from methylation

At the molecular level, inactivation is associated with the addition of a methyl group (CH_3) to cytosine molecules that occur immediately on the 5′ side of guanine molecules, i.e. inactivation is associated with methylation of cytosines in so-called **CpG islands**, where p stands for the phosphate link between the two adjacent bases. Within an individual animal, all descendants of each cell in which inactivation first occurred have the same inactive gene or chromosome, because after each replication of a methylated strand of DNA, the new strand is automatically methylated at the same CpG sites as in the original strand. In meiosis, however, or in early embryonic development, the methylation patterns are reset.

Not all methylation patterns are set for the lifespan of the animal. In fact, in regions that are not subjected to X-inactivation or imprinting, methylation of CpG islands in promoters is a common attribute of inactive genes, and demethylation is a prerequisite for transcription of many genes. Thus, methylation is another means by which genes are regulated.

Inactivation and imprinting are just two examples of **epigenetics** – the inheritance between cell generations and sometimes between animal generations of changes in 'phenotype' that are not the result of changes in DNA sequence. As we shall see in later chapters, there are some fascinating biological stories emerging from epigenetics research. However, as the name implies, these are additions to, rather than (as some people tend to think) replacements of, the principles of Mendelian inheritance.

Types of DNA

How can we categorize total DNA, and where do structural genes fit into the picture?

The most common category of DNA consists of **unique** or **single-copy** sequences, which account for 60–70 per cent of the genome of mammals. These single-copy sequences are dispersed throughout the genome. A small proportion of this DNA accounts for most structural genes. Evidence emerging from genome sequence assemblies (see Chapter 2) suggests that the total number of structural genes in animals is around 22 000.

Some structural genes occur in **multigene families**, which consist of sets of identical or very similar genes, whose individual members are usually scattered around the genome, or, in some cases, occur as sets of adjacent genes. Not surprisingly, the structural genes that occur in multigene families are those whose products are required in relatively large quantities, e.g. histone, keratin, collagen, ribosomal RNA, and transfer RNA.

The third major category of DNA is **repetitive** DNA, which consists of multiple copies of particular sequences called **repeat units**, which range in size from a single base to several thousand bases. Repetitive DNA has assumed increasing importance in recent years, with the realization that it holds the key to some important inherited diseases, and provides a major tool for the practical application of molecular biology to measurement of genetic diversity within and among animal populations and, to a lesser extent in recent years, animal health and animal improvement. We shall discuss these aspects of repetitive DNA in following chapters.

There is one other category of chromosomal DNA that should be mentioned. Scattered throughout the genome are small DNA fragments called **transposable genetic elements** (**TGEs**) or **jumping genes**. A notable property of a TGE is that the nucleotide sequence at one end is an **inverted repeat** (or occasionally a **direct repeat**) of the sequence at the other end. In cattle, for example, there is a TGE which is 611 bases long. Its terminal sequences are:

5′ GCCGGGGA … TCCCCGGC 3′
3′ CGGCCCCT … AGGGGCCG 5′

Notice that the sequence TCCCCGGC at the 3′ end of the top strand is really a repeat of the GCCGGGGA sequence at the other end of the same strand, only in an inverted (i.e. reverse) and complementary form. Another way to look at this is to note that reading the sequence in the top strand from 5′ to 3′ is exactly the same as reading the bottom sequence from 5′ to 3′. Because they contain exactly the same message when read in either direction, inverted repeats are said to be **palindromes** (by analogy with palindromic sentences such as ABLE WAS I ERE I SAW ELBA).

The repeats are homologous, and are therefore able to pair with each other, just as homologous chromatids pair during meiosis I. In the case of TGEs, the pairing of the repeats results in the TGE itself forming a loop. Often when this occurs, the whole of the TGE excises itself from wherever it happens to be, and moves to another section of the same or a different chromosome, into which it then inserts itself by the reverse process. Sometimes, the TGE is replicated, and the resultant copy moves elsewhere, leaving the original one in its original position. As we shall see in Chapter 10, TGEs are extremely important in relation to the rapid spread of multiple antibiotic resistance in bacteria. In eukaryotes, TGEs are surprisingly ubiquitous: there are hundreds of thousands of them in the mammalian genome, making up from 1 to 5 per cent of total DNA. In cattle, for example, the TGE described above occurs 35 000 times. Many of the copies of this and other TGEs have lost the ability to move from one site to another (to transpose). However, those that are able to transpose themselves are extremely important, because if a TGE inserts itself into a structural gene, it will most likely inactivate that gene. Alternatively, if a TGE inserts itself in the control region of a gene, it may interfere with normal control, resulting in the gene being expressed at inappropriate times and places, or not being expressed when it should be. Because of these effects, TGEs are important sources of mutation. The mutations they create are called **insertion mutations**. TGEs are also important causes of cancer, as we shall see in Chapter 11.

Summary

- A chromosome is, in essence, a double helix of DNA
- Each species has a characteristic set of chromosomes (karyotype)
- Typically, an animal karyotype comprises two subsets of chromosomes (diploid): one inherited from the mother and one from the father, with each maternal chromosome having a matching paternal chromosome (pairs of homologues), except for the sex chromosomes
- The first stage of meiosis and mitosis involves the duplication of each chromosome (i.e. the replication of a double helix), resulting in two chromatids (two double helixes) joined at the centromere
- In mitosis, this is the stage at which karyotypes become visible
- Meiosis results in gametes having one set of chromosomes (haploid), comprising one member (chosen at random – maternal or paternal) of each homologous pair
- Recombination (crossing-over) during meiosis results in some chromosomes having a portion of maternal and paternal double helix
- The sequence of bases in some segments of DNA acts as a template for the production of a specific sequence of amino acids which form a particular polypeptide
- This process involves transcription of the DNA sequence into mRNA, followed by translation of the mRNA into peptide
- A segment of DNA including all the nucleotides that are transcribed into mRNA is called a structural gene; more generally, a gene is a stretch of DNA that produces a functional RNA molecule
- From the moment of fertilization until death, the development of each living organism is determined by genes being switched on and off at the appropriate time(s) in the appropriate cell(s), by means of regulatory proteins (activators and repressors)
- Uncorrected mistakes in DNA replication are called mutations
- Each mutation creates an altered DNA sequence (allele) at a particular location (locus) on a chromosome
- An inevitable consequence of meiosis is that the alleles at any locus segregate into gametes in a predictable manner (Mendelian inheritance)
- Alleles at different loci typically segregate independently
- If two loci are close together on the same chromosome, their alleles do not segregate independently – they are said to be linked
- In mammals, one of the X chromosomes (chosen at random soon after fertilization) in females is inactivated by methylation of cytosine molecules in its DNA, resulting in dosage compensation for genes on the X chromosome (in relation to males, which have only one X chromosome)
- Some sections of autosomes are also inactivated by methylation (imprinting)
- Inactivation and imprinting are just two examples of epigenetics – the inheritance between cell generations and sometimes between animal generations of changes in 'phenotype' that are not the result of changes in DNA sequence.
- Only a few per cent of DNA accounts for all structural genes; the remainder includes vast quantities of repetitive DNA and transposable genetic elements (jumping genes)

Further reading

Ellegren, H., Hultin-Rosenberg, L., Brunstrom, B., Dencker, L., Kultima, K. and Scholz, B. (2007) Faced with inequality: chickens do not have a general dosage compensation of sex-linked genes. *BMC Biology*, **5**, 40.

Feschotte, C. and Pritham, E.J. (2007) DNA transposons and the evolution of eukaryotic genomes. *Annual Review of Genetics*, **41**, 331–68.

Hartl, D.L. and Jones, E.W. (2009) *Genetics: analysis of genes and genomes*, 7th edn. Jones and Bartlett, Boston.

King, R.C., Stansfield, W.D. and Mulligan, P.K.(2006) *A dictionary of genetics*, 7th edn. Oxford University Press, New York.

Mercer, T.R., Dinger, M.E. and Mattick, J.S. (2009) Long non-coding RNAs: insights into functions. *Nature Reviews Genetics*, **10**, 155–9.

O'Brien, S.J., Menninger, J.C. and Nash, W.G. (eds) (2006) *Atlas of mammalian chromosomes*. Wiley, Hoboken, NJ.

Payer, B. and Lee, J.T. (2008) X chromosome dosage compensation: how mammals keep the balance. *Annual Review of Genetics*, **42**, 733–72.

Presole, G. (2008) What is a gene? An updated operational definition *Gene*, **417**, 1–4.

Robinson, T.R. (2005) *Genetics for dummies*. Wiley, Hoboken, NJ.

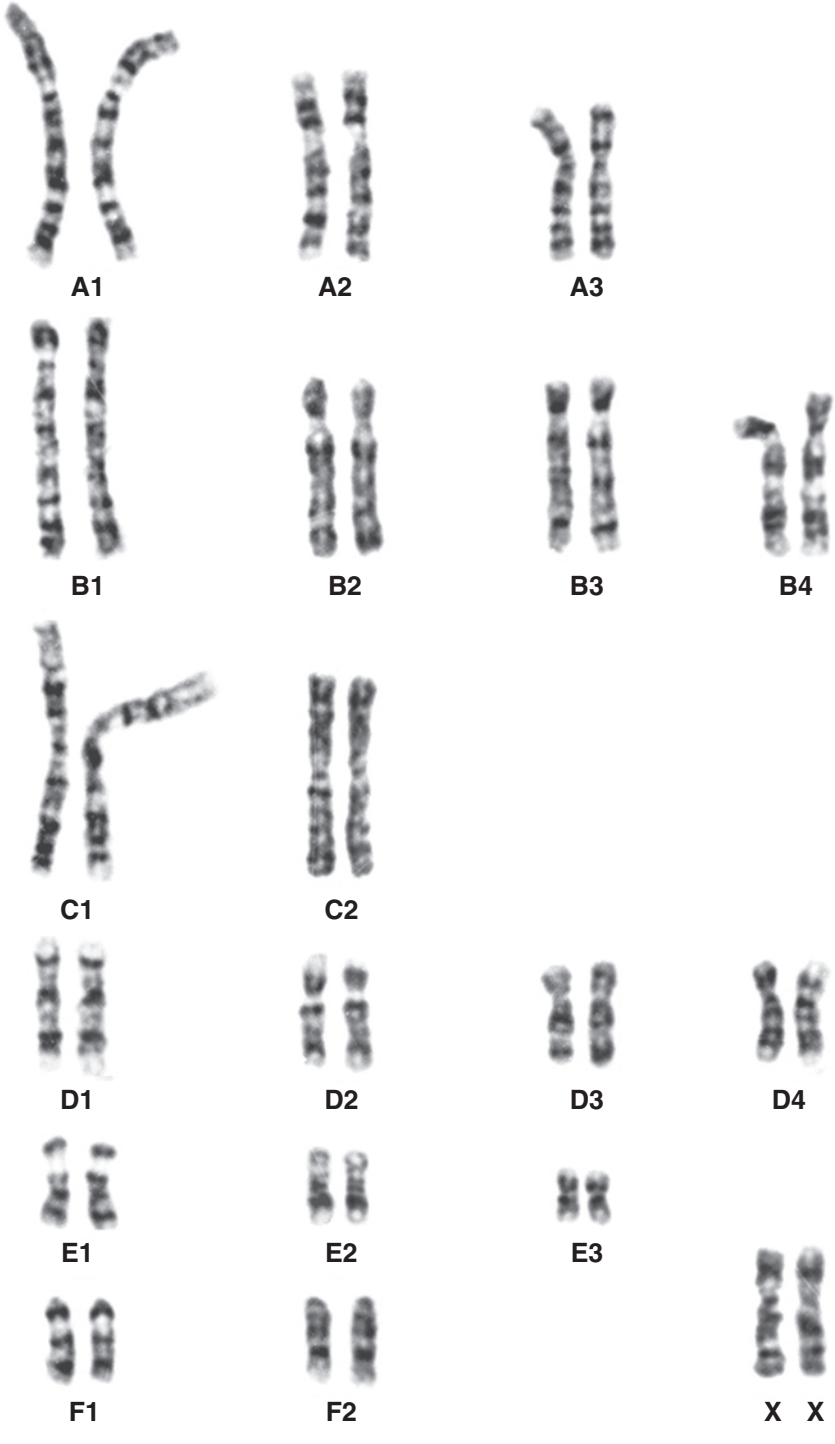
Snustad, D.P. (2009) *Principles of genetics*, 5th edn. Wiley, Hoboken, NJ.

Suzuki, M.M. and Bird, A. (2008) DNA methylation landscapes: provocative insights from epigenomics. *Nature Reviews Genetics*, **9**, 465–76.

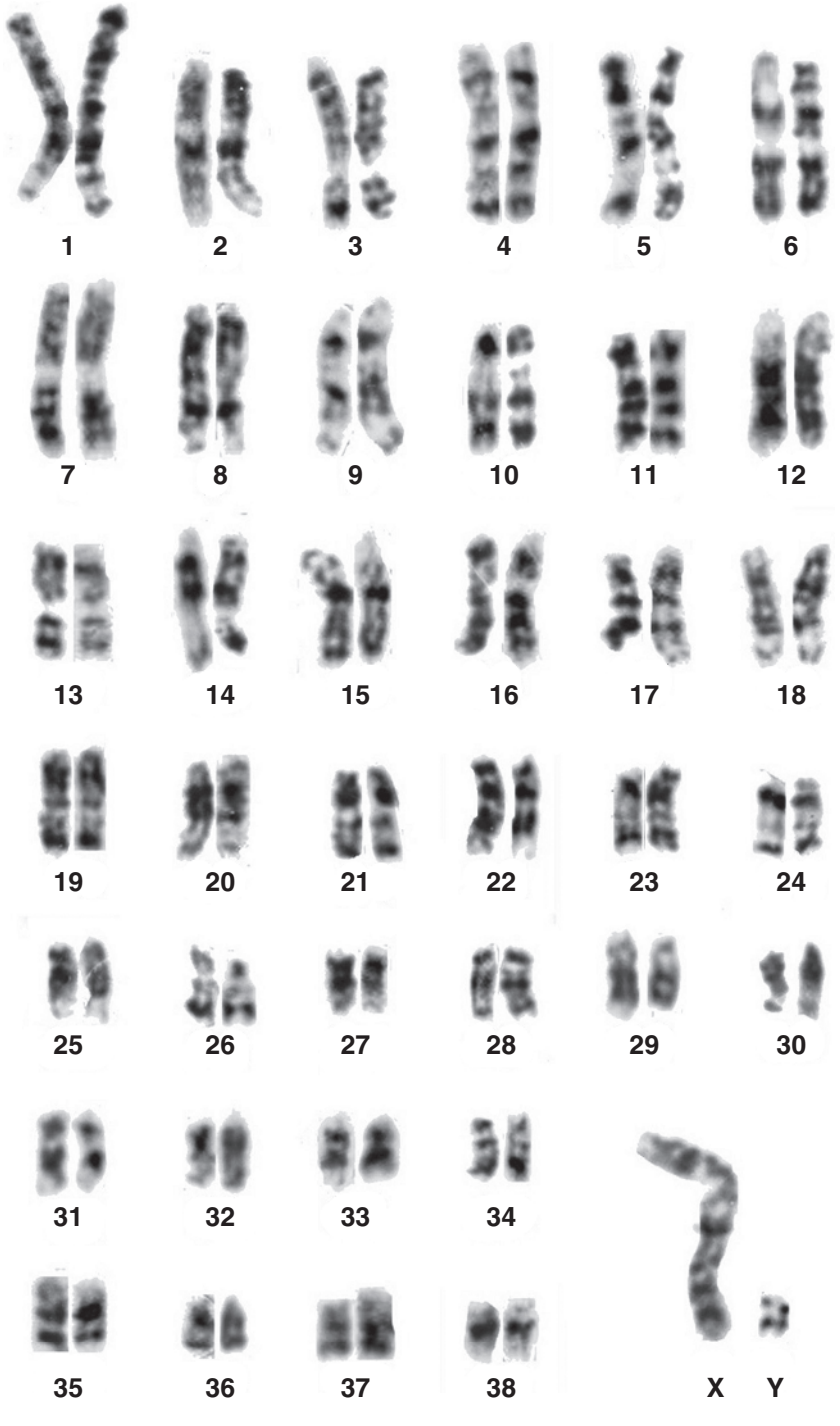
Watson, J.D., Baker, T.A., Bell, S.P., Gann, A., Levine, M. and Losick, R. (2008). *Molecular biology of the gene*, 6th edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor & Benjamin Cummins, San Francisco.

Yang, F. and Wang, P.J. (2009) The mammalian synaptonemal complex: a scaffold and beyond. *Genome Dynamics*. **5**, 69–80.

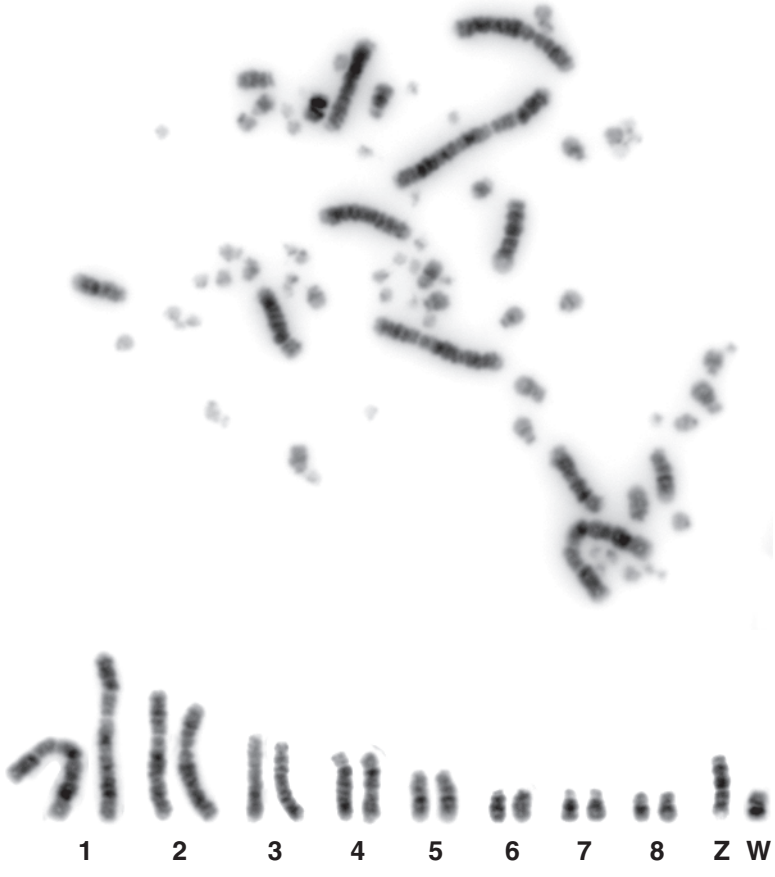
Appendix 1.1 Banded karyotypes of domestic species



Cat (G-banding). (Reproduced by permission of W. Nash from O'Brien, S.J., Menniger, J.C. and Nash, W.G. (eds) (2006) *Atlas of Mammalian Chromosomes*, Wiley-Liss, p. 514.)

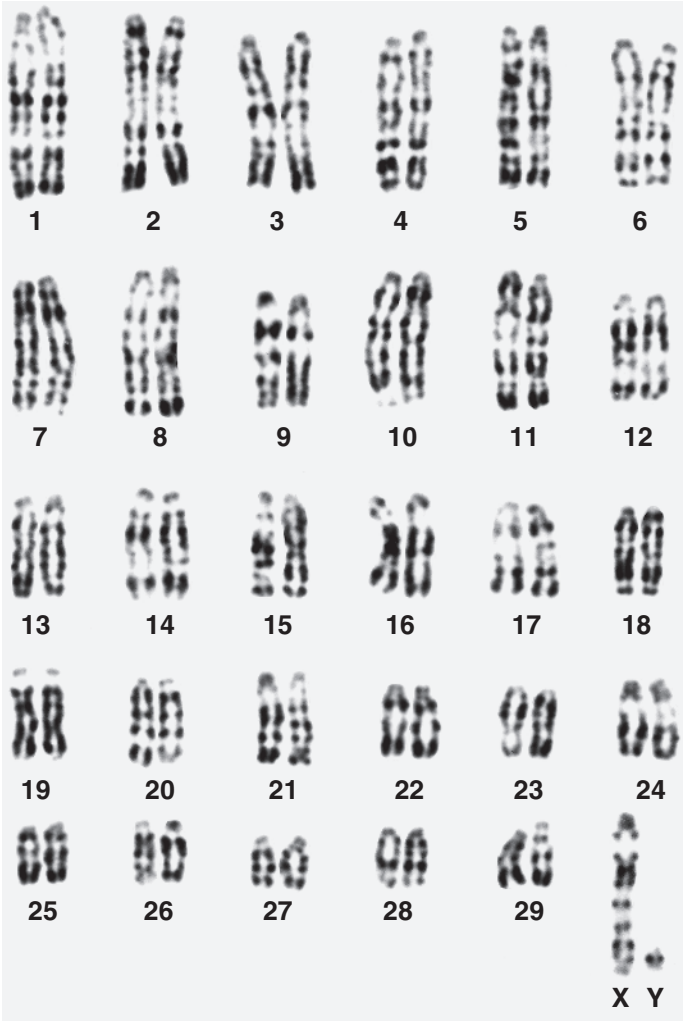


Dog (GTG-banding). (Reproduced by permission of A. Graphodatsky from O'Brien, S.J., Menninger, J.C. and Nash, W.G. (eds) (2006) *Atlas of Mammalian Chromosomes*, Wiley-Liss, p. 467.)

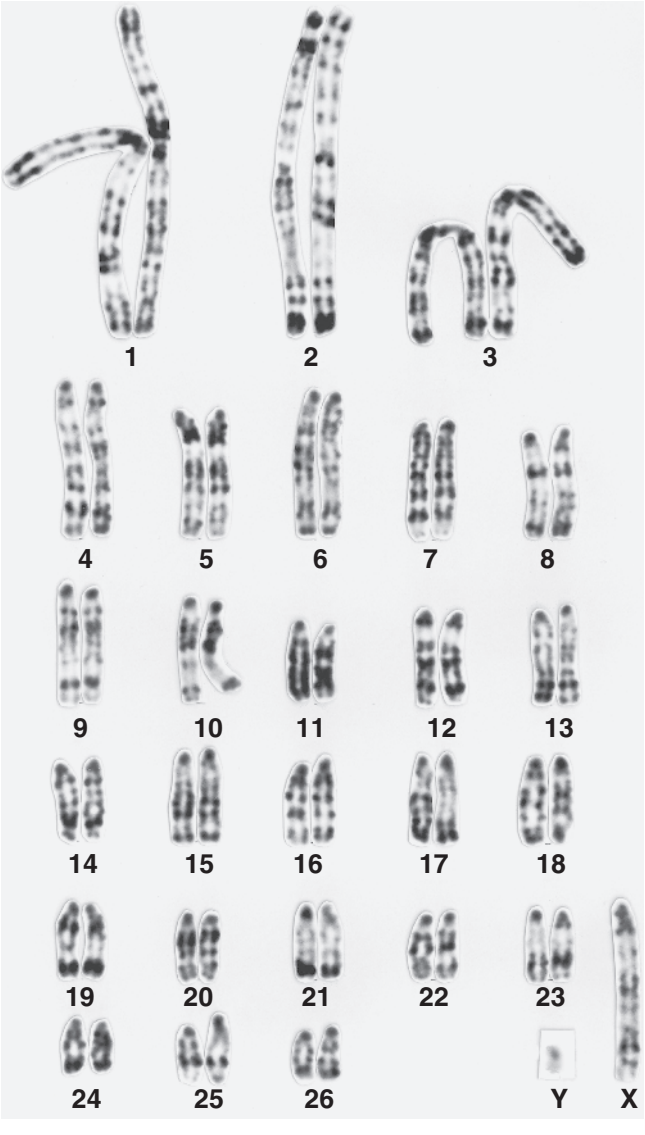


Chicken: (above) the original chromosome spread; (below) the eight largest autosomes and the sex chromosomes. (DAPI/AMD banding; producing the equivalent of G bands). (Reproduced courtesy of M. Völker and D.K. Griffin.)

Pig (GTG-banding). (Reproduced by permission of A. Graphodatsky from O'Brien, S.J., Menninger, J.C. and Nash, W.G. (eds) (2006) *Atlas of Mammalian Chromosomes*, Wiley-Liss, p. 566.)



Goat (RBG-banding). (Reprinted by permission from S. Karger AG, Basel: ISCNDB (2000), Di Berardino, Di Meo, Gallagher, Hayes and Iannuzzi, *Cytogenetics and Cell Genetics*, **92**, 283–99.)



Sheep (RBG-banding). (Reprinted by permission from S. Karger AG, Basel: ISCNDB (2000), Di Berardino, Di Meo, Gallagher, Hayes and Iannuzzi, *Cytogenetics and Cell Genetics*, **92**, 283–99.)

Horse (GTG banding). (Reproduced by permission of R. Stanyon from O'Brien, S.J., Menninger, J.C. and Nash, W.G. (eds) (2006) *Atlas of Mammalian Chromosomes*, Wiley-Liss, p. 669.)