Chapter 1

Basic genetics and cytogenetics: a brief reminder

Karen Sermon

Introduction

This brief reminder chapter aims to freshen up what professionals in reproduction may have learned a while ago at university, and will also serve the reader as a source of information to comprehend the following, more complex chapters. At the end of this chapter, basic study books recommended for further reading are given to help the reader in the further understanding of this textbook [1, 2].

Human reproduction and genetics are intimately intertwined and indeed often confused and rolled into one. Understanding reproduction is impossible without a firm basis in genetics, and the readiness to acquire more knowledge when needed. However, human genetics is much broader than just reproduction – think of for instance oncogenetics – so in this chapter I will summarize what basic genetics is indispensable for the specialists in reproduction.

In this chapter, I will introduce the general organization of our genome, how this genome behaves when it goes through a reproductive cycle (meiosis), how our genome is used as a template for making proteins, and how this is broadly regulated, major genetic and hereditary abnormalities both at the chromosome and at the monogenic level, and a brief overview of current genetic diagnostic techniques.

The organization of the human genome

The basic building material: DNA

Deoxyribonucleic acid (DNA) consists of four different nucleotides [1, 2]. Each nucleotide consists of sugar, which is deoxyribose in DNA, a phosphate group, and a base. Four different bases are present in the nucleotides of DNA: adenine (A), cytosine (C), guanine (G), and thymine (T) (Fig. 1.1A [3]). These four nucleotides are strung together in long strands of DNA, alternating a sugar (with a base attached) and a phosphate group, and where the order of the different bases defines the genetic code (Fig. 1.1B). The phosphate group can be either bound to the 5′ carbon of the sugar, or to the 3′ carbon of the sugar, while the base is bound to the 1′ carbon. When strung together, the first sugar in the DNA strand has a free 5′ carbon, while its 3′ carbon is covalently bound to a phosphate group. This phosphate group is bound further down the strand to the 5′ of the next sugar. This is why in a DNA strand the 5′ carbon of the first sugar is free and at the end of the strand the 3′ carbon of the last sugar is free. This is why DNA base pairs are always read from 5′ to 3′.

Moreover, cellular DNA is usually found in a double helix form. The bases A and T on one hand and C and G on the other hand are complementary to each other and will pair up by forming hydrogen bonds, thus stabilizing the double helix (Fig. 1.1C). According to the Ensembl database [4], our nuclear DNA contains 3 287 209 763 base pairs (bp). An estimated half of the human genome consists of transposons, also known as mobile DNA elements or "jumping genes," and will be extensively discussed in Chapter 6.

DNA is organized in chromosomes

The DNA of our whole genome is not ordered in one long strand, nor does it lie naked and unprotected in the nuclei of our cells. Nuclear DNA is organized in 46 separate strands, so each chromosome contains one long DNA strand. Twenty-two of these chromosomes are paired, one inherited from the mother and one from the father. These are called the autosomes.

Textbook of Human Reproductive Genetics, eds Karen Sermon and Stéphane Viville. Published by Cambridge University Press.

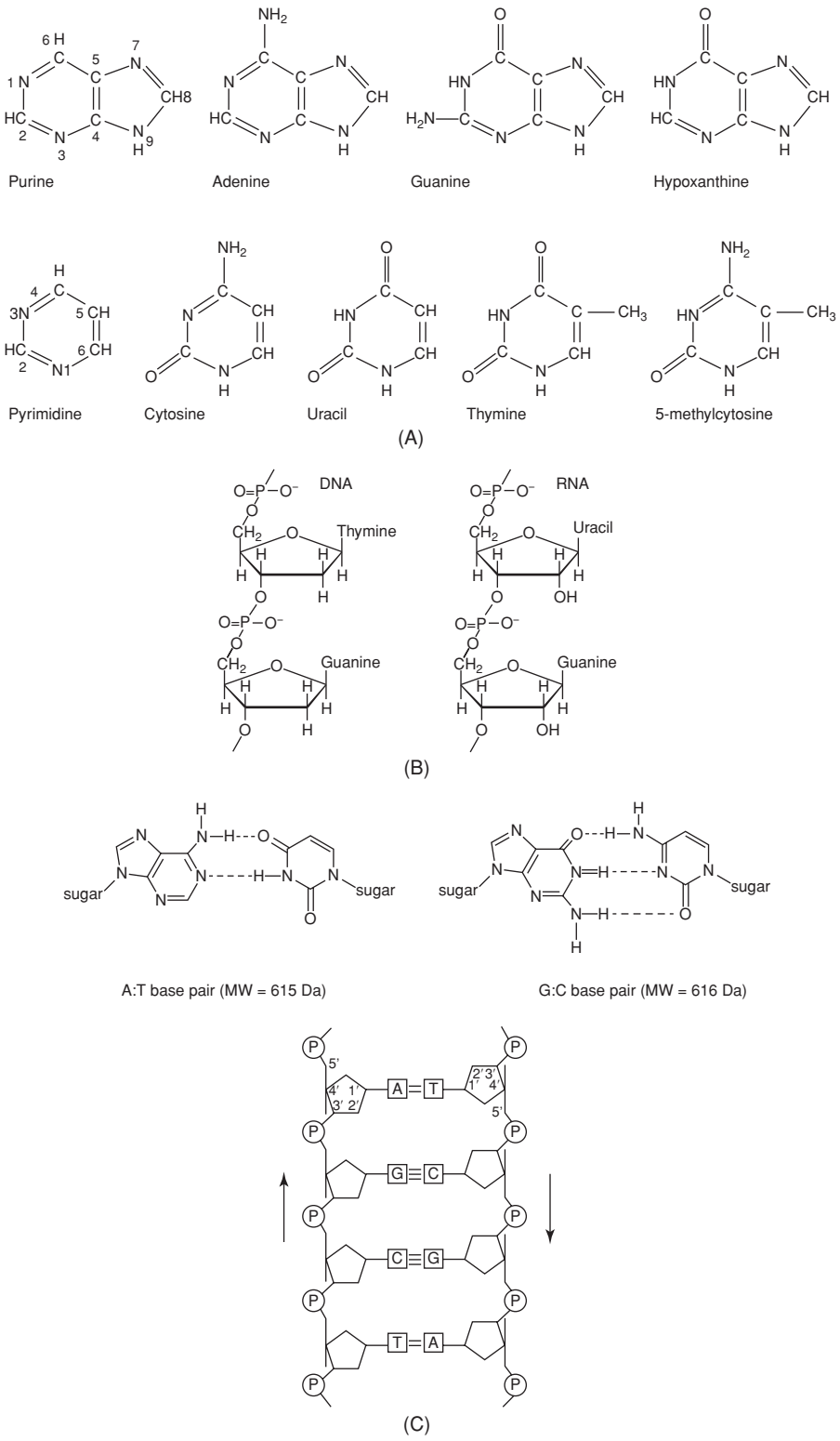

Figure 1.1 (**A**) The four bases in DNA and RNA. (**B**) Sugar–phosphate backbone of nucleotides. DNA has 2′-deoxyribose as sugar, RNA has ribose as sugar. (**C**) A and T form two hydrogen bonds while G and C form three hydrogen bonds. Together with the sugar–phosphate backbone this forms the double helix of DNA. Source: Figs 2.4 and 2.7– 2.9 from Ringo [3].

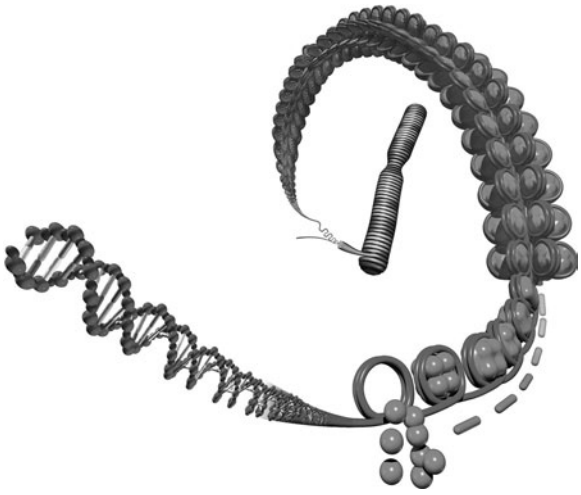

Figure 1.2 DNA is organized in chromatin. See plate section for color version. © Science Photo Library.

The two remaining chromosomes are the sex chromosomes: females have two X chromosomes, while males have one X and one smaller Y chromosome.

When the cell is not dividing, the DNA is organized in chromatin. The double helix is wound around protein structures called histone octamers, which in their turn are further coiled into structures called solenoids. These solenoids are attached to a protein scaffold within the nucleus to form loops (Fig. 1.2). The way the chromatin is organized in a specific cell, and the way for instance some histones are chemically modified, defines which genes can or cannot be transcribed and expressed. For instance, some genes will be so tightly packed in the chromatin that the transcription machinery cannot reach them and thus these genes are silenced in this particular cell. The chromatin structure is thus one of the determinants of the cell's expression pattern and therefore its function.

A particular type of DNA can be found in mitochondria. Mitochondria are cell organelles that are essential for the respiration and energy production of the cell. They carry their own circular DNA of about 16 000 bp which replicates independently of the DNA in the nucleus. The genes on the mitochondrial DNA encode for their own ribosomal RNA (rRNA), transfer RNA (tRNA), and ribosomal proteins, as well as a handful of aerobic metabolism enzymes. Many mitochondrial proteins are however encoded by the nuclear DNA and are later imported into the mitochondrion to contribute to the mitochondrial function.

The functional genetic entity: the gene

A gene can be defined as "a sequence of DNA in the genome that is required for the production of a functional product which can be a polypeptide or a functional RNA molecule" [2]. The number of genes in our genome is estimated to be around 25 000 ([4]). These are not equally scattered along the chromosomes: some parts of chromosomes are very gene-rich while other stretches of more than a million base pairs contain no genes at all and are called gene deserts. Some genes are quite small and count only a few kilobase pairs, while others span a million or more base pairs. One such large gene is the dystrophin gene spanning more than two million base pairs. The two gene copies on the autosomes are usually both transcribed (or both silenced), except for a number of developmentally important genes where only one copy is transcribed. In these so-called imprinted genes, either the paternal or the maternal copy is exclusively expressed. Some of these genes have been implicated in congenital defects occurring more frequently after assisted reproductive technology (ART): this topic is discussed in Chapter 13.

Genes that typically code for a polypeptide have recurring structural features (Fig. 1.3 [3]). Not all base pairs in a gene will be translated into a protein: genes typically contain exons that are the translated

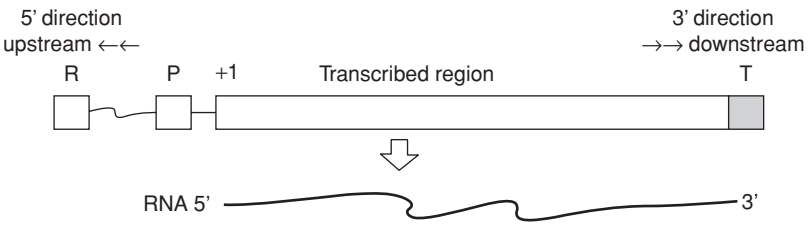

Figure 1.3 Anatomy of a gene: R is a regulatory sequence, P is the promoter sequence, and T is the terminator sequence. Source: Fig. 7.4 from Ringo [3].

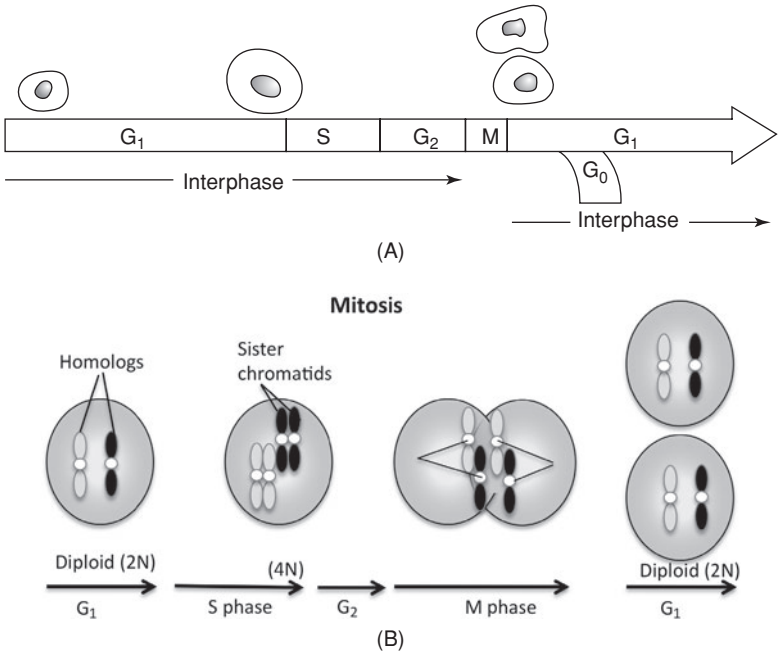

Figure 1.4 (**A**) The cell cycle of a somatic cell and (**B**) the different steps in mitosis. Source: Figure 19.1 from Ringo [3] and Chapter 3.

parts interspersed with introns that are not translated. In many genes, the introns represent a significantly larger proportion of the gene than the exons. Other recurring structural features are sequences conserved among many different genes that give the transcription machinery of the cell appropriate signals of when and where to transcribe specific genes. Every gene has a start and a stop signal, as well as a promoter sequence in the 5′ end. This promoter specifies the pattern as well as the level of expression of a gene. Other regulatory elements include enhancers, silencers, and locus control regions and can be either in the 5′ or 3′ untranslated region, or in intronic sequences of the gene. Some can even lie far away from the coding sequence of a gene.

Alleles are alternative forms of the same gene: changes in the DNA sequence will lead to a different protein with different function. A well-known example of multiple alleles is the ABO blood group system.

Making more copies: the cell cycle

Different parts of the cell cycle

In order to get from the fertilized zygote to the estimated 100 trillion cells in the human body, cells need to undergo continuous divisions. Cell division and mitosis are also crucial for differentiation.

When a cell is not in mitosis, it is said to be in interphase (Fig. 1.4A,B [3]). The first part of the interphase, immediately after mitosis, is called the G_1 phase. In this phase, the chromosomes each contain only one copy of the DNA strand. This G_1 phase typically lasts for several hours until the S phase is reached, although some terminally differentiated cells (neurons or white blood cells) may withdraw from the cell cycle altogether and are said to be in G_0.

When the cell starts to replicate its DNA, it is said to enter the S phase. The two DNA strands are separated, and a complementary strand is made of each strand. The chromosomes at the end of the S phase consist of two sister chromatids, each containing an identical DNA strand. The chromatids are held together at the centromere that, associated with specialized proteins, forms the kinetochore by which the chromatids will be attached to the mitotic spindle. Before the cell enters mitosis, it goes through a brief control G_2 phase in which the cell visibly grows after the accumulation of synthesized proteins during the whole cell cycle. The whole interphase of a typical cell lasts between 16 and 24 hours but may extend to

months, whereas mitosis is completed in a couple of hours.

Passing on the information: mitosis

At the end of G_2, every chromosome consists of two chromatids and one of each of these chromatids needs to end up in one of the daughter cells in orderly chromosome segregation. The first step in mitosis is the prophase, in which chromosomes start to condense and a mitotic spindle starts to form (see Figs 1.4B, and 3.1A in Chapter 3). The formation of the mitotic spindle is organized through two centrosomes from which microtubules will radiate to form the spindle. During prometaphase, the nuclear membrane breaks up and the chromosomes are attached to the mitotic spindle by their kinetochore. Led by the microtubules of the spindle, the chromosomes move to the metaphase plate in a process called congression. During the metaphase, the chromosomes have reached their maximal condensation at the equatorial plane and are easiest to visualize. The anaphase starts when the chromosomes separate into two chromatids and each chromosome moves to the different poles. The mitosis is ended by the telophase during which the chromosomes decondense and the nuclear membrane is built up again, this process is called karyokinesis. Concomitantly to telophase, the cytoplasm is divided over the two daughter cells and karyokinesis is completed.

Making more humans: meiosis

During the process of mitosis, two daughter cells are produced that carry exactly the same genetic information both in content as in volume. Mitosis could thus not be used to form gametes, since with every generation the amount of DNA would double. Meiosis as a specialized form of cell division solves this problem by resulting in cells (gametes) with only half of the DNA content, i.e. one of each chromosome, of the somatic cells. An added bonus is that during meiosis, the DNA from the two parental chromosomes is exchanged to form new chromosomes built with genetic material from the two parents. This process of recombination is important to generate genetic diversity in a species and thus to secure its evolution. Meiosis is the most important step for the survival and evolution of sexually reproducing species.

The different steps in male and female meiosis are discussed in depth in Chapter 3.

From DNA to protein: the transcription and translation machinery

RNA comes in many forms and functions

If DNA is important as the keeper of genetic information, the role of RNA is at least as important – some say more important – because of the versatility in form and function of RNA. In three aspects RNA differs from DNA: (1) the sugar moiety is different and is ribose instead of deoxyribose; (2) the thymine base is replaced by a uracil base; and (3) RNA can be found in many different three-dimensional structures, usually as a single strand [1, 2].

RNA is transcribed from the DNA template by RNA polymerases that first unwind the DNA and then synthesize an RNA strand that forms a temporary double helix with the DNA. The RNA is synthesized on the 3′–5′ DNA strand in the 5′-3′ direction. This 3′–5′ template DNA strand is often called the antisense strand because it is in the opposite sense from the RNA, while the 5′–3′ DNA strand, which does not serve as a template, has the same nucleotide sequence as the synthesized RNA strand (except that thymine is replaced by uracil) and is often called the sense strand.

Different examples of RNA forms and functions are messenger RNA (mRNA) that is transcribed from polypeptide-encoding genes, ribosomal RNA (rRNA) that will make up the ribosomes and transfer RNA (tRNA) that will ultimately translate the information in the mRNA sequence to an amino-acid sequence. When RNA is transcribed from the DNA strand, it will need to undergo a lot of processing before it can fulfill its function. Introns will have to be spliced out of the mRNA so that only the exons are translated into protein and a polyA tail will be added to the mRNA to ensure its stability. The rRNA and tRNA also undergo extensive changes before they are functional.

A more recently discovered type of RNA is transcribed from non-coding RNA genes. These RNAs have an important function in the regulation of other genes. Well-known examples of this are the microRNAs that can regulate the amount of mRNA available for translation. Other examples are other small non-coding RNAs such as piwi-protein interacting RNA (piRNA) and short interfering RNA (siRNA) as well as long non-coding RNAs. After the chromatin structure, this is the second example of how gene expression is regulated and explains the increasing importance that scientists are giving to RNA.

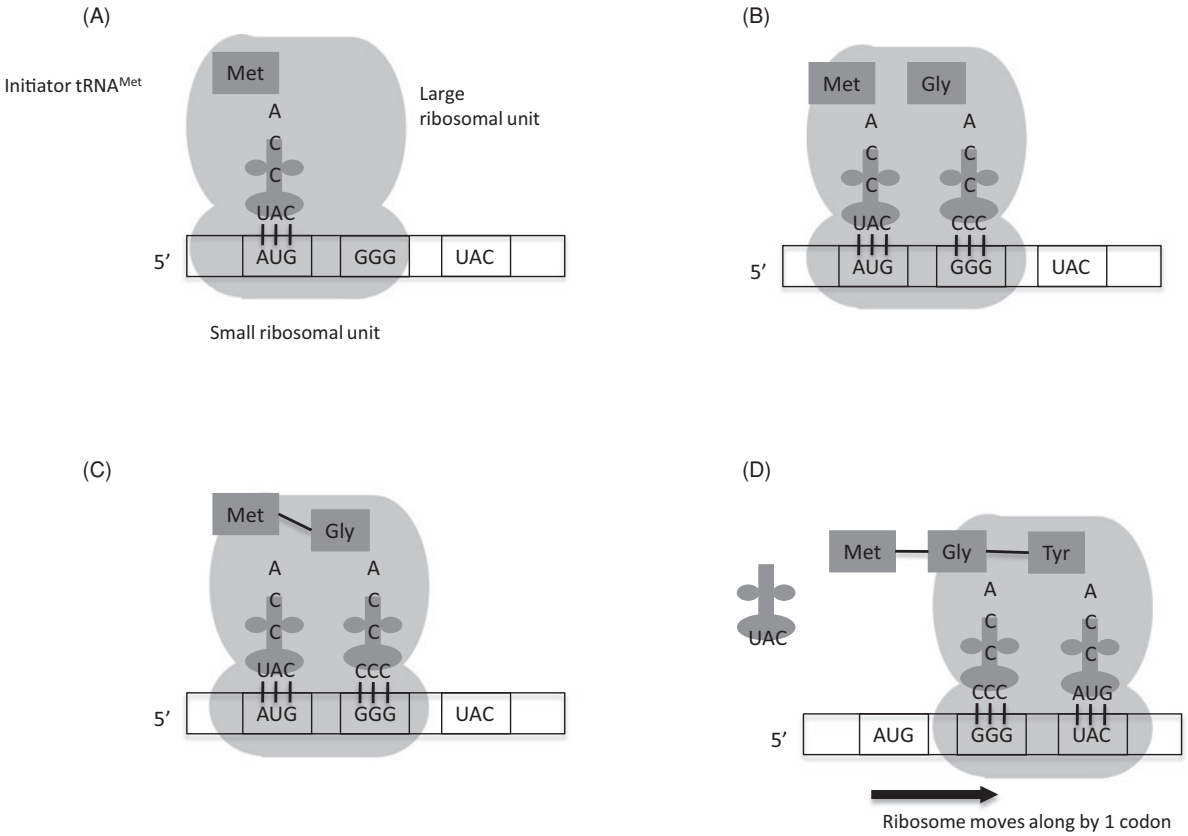

Figure 1.5 Translation of mRNA into protein (A–D). For explanation, see text.

Different RNAs work together to translate DNA into protein

Genetic code

In the DNA sequence of a polypeptide gene, a set of three base pairs constitutes a codon. One codon codes for one amino acid. Looking at all the possible codons using combinations of three of the four base pairs, we get to 4^3 or 64 possibilities. These 64 codons are the genetic code. Because there are only 20 amino acids, one amino acid can be represented by more than one codon. There is only one codon for the amino acid methionine, and this codon also represents the start of the polypeptide. Finally, three codons represent a stop signal that tells the translation machinery that the end of the polypeptide is reached [1, 2].

Translation machinery

Once the mRNA, rRNA, and tRNA have been transformed into their mature and functional forms, they are transported to the cytoplasm of the cell where the translation into proteins takes place. The large ribosomes, built of several rRNA molecules, will start to read the mRNA at the start codon, which is always AUG coding for methionine (Fig. 1.5A–D). Within the ribosome, a tRNA that carries an anticodon that is complementary to the methionine codon on the mRNA will then bind to the methionine codon. The other end of the tRNA carries a methionine amino acid. The ribosome then moves one codon along the mRNA to the next codon. In Figure 1.5, this codon is GGG coding for glycine. A tRNA that has CCC as anticodon binds to the mRNA codon and carries a glycine amino acid. This glycine amino acid is then bound to the first methionine amino acid, and the two first amino acids of the polypeptide are formed. The whole mRNA is read in this fashion until the ribosome reaches one of the three stop codons. The synthesis of the polypeptide is then terminated, and the polypeptide is released from the ribosome to be further processed into a functional protein by other

cytoplasmic organelles such as the endoplasmic reticulum before reaching its cellular localization.

Behold the genome: tools in human genetics

We will discuss here only a handful of tools as they appear in the next chapters. This should help the reader to understand the contents, but is by no means a complete story. We refer the interested reader to the study books in the reference list for additional information [1, 2].

The big view: cytogenetics

Know your classics: G-banding and karyotypes

Cytogenetics is the study of chromosomes, their structure, and their inheritance. At the end of the 1950s, the exact number of chromosomes was known and, at the first Conference on Standardization in Human Cytogenetics in Denver, the chromosomes were classified according to their size: the largest chromosome is chromosome 1, the smallest is chromosome 22, and, in addition, chromosome X and chromosome Y were also included. Later on, banding techniques became available and the current classification was established in Paris in 1971 (Fig. 1.6) [5]. The complete picture with every chromosome of an individual arranged from the largest to the smallest is called a karyotype.

Most commonly, chromosomes are visualized from lymphocytes from peripheral blood. These are put in culture and forced to divide and then arrested when they go into metaphase. Once in metaphase, the microtubules in the spindle are stopped, the cells lysed and spread on a glass slide. The spread metaphases can be dyed with Giemsa staining, which is the classical G-banding found on most karyotype protocols. According to the position of the centromere, three types of chromosomes can be distinguished: metacentric chromosomes, with a centromere approximately in the middle, submetacentric chromosomes with two arms of clearly unequal length, and acrocentric chromosomes with centromeres at or near an end. Each chromosome has a long arm, called the q-arm and a short arm called the p-arm. The banding patterns visible after G-banding characterize each individual chromosome, and allow cytogeneticists to classify and number the chromosomes and to identify possible rearrangements (Fig. 1.6).

Fluorescence *in situ* hybridization: count the dots

Although fluorescence in situ hybridization (FISH) was the first step from classic cytogenetics to molecular cytogenetics, the method gives only limited information in contrast to currently developed molecular cytogenetics methods. Fluorescence *in situ* hybridization is a relatively quick and easy method that allows fast screening of, for example, prenatal samples for the most common aneuploidies such as trisomy 13, 18, and 21. The sampled cells are spread on a glass slide, fixed, and the DNA is denatured so that the single-stranded DNA is accessible for fluorescently labeled DNA probes. These probes are chosen so that they are complementary to the DNA region of interest, are allowed to form double-stranded DNA with the DNA in the sample ("to hybridize"), after which the place in the nucleus where the probe has bound can be seen as a fluorescent spot. The FISH probes are carefully chosen to serve a particular purpose. For a quick chromosome count such as for aneuploidy detection in prenatal samples, centromere probes are usually chosen, because they give large, easy-to-read signals. For the detection of more specific chromosome regions, such as in translocations involving small fragments or microdeletion syndromes such as DiGeorge syndrome, more care has to be taken in the design of the FISH probes. If the probes are mixtures covering a whole chromosome, they are called chromosome paints and are very useful for visualizing a chromosomal translocation.

More on FISH and its application in preimplantation genetic diagnosis and screening can be found in Chapters 4 and 11.

The detailed view: DNA and RNA analysis

Polymerase chain reaction

Before the advent of polymerase chain reaction (PCR), the only way to amplify a DNA fragment of interest was to clone it into a vector (say a plasmid), introduce the vector in a host (say an *Escherichia coli* bacteria), culture the bacteria, and then isolate the expanded plasmid with the DNA of interest. This was very time consuming, necessitating a large amount of DNA starting material and special equipment for bacteria culture. Polymerase chain reaction very quickly took its place in many applications in molecular biology: PCR products are used as templates for restriction enzymes or sequencing reactions; probes used

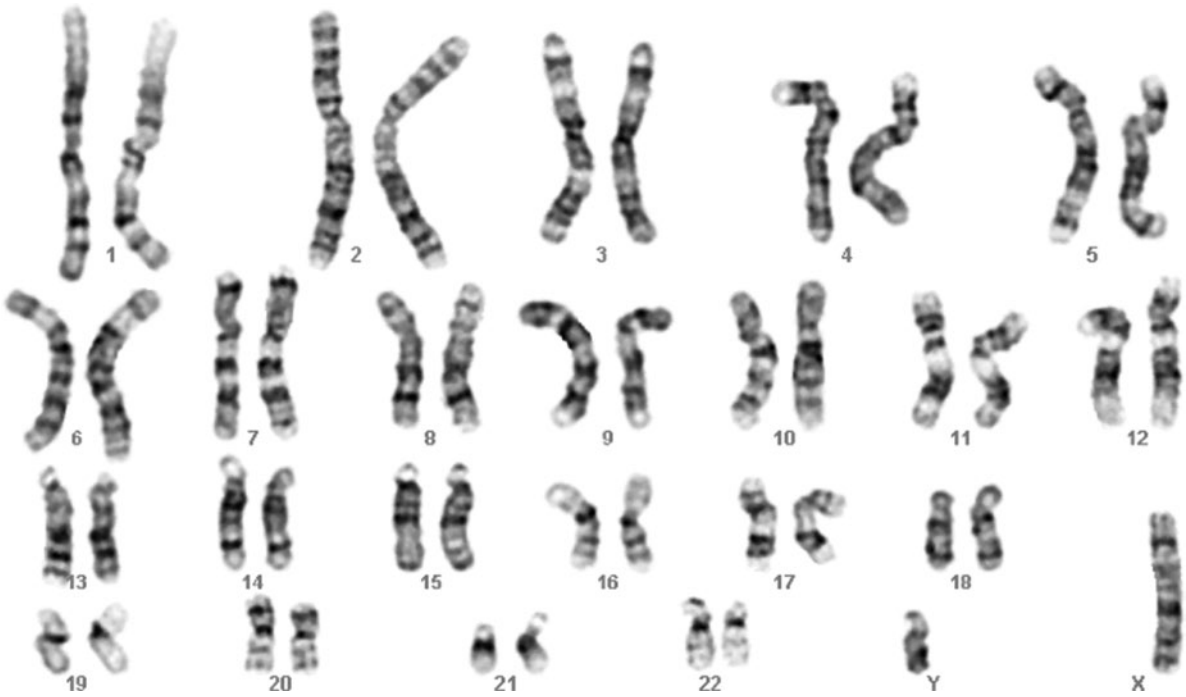

Figure 1.6 Normal human male karyotype (46,XY). © Science Photo Library.

in other applications (FISH, hybridizations of all kinds) are synthesized using PCR; even making transgenic animals has become much simpler thanks to PCR.

Polymerase chain reaction is in essence a DNA copier machine (Fig 1.7 [3]). In a first step, the native DNA strands are denatured, i.e. the two strands are separated by heating them up, typically to 95°C. In the second step, at a lower temperature, two single-strand short DNA fragments bind to the complementary DNA. These short DNA fragments are called the primers, and are chosen so that they delineate a fragment of the genomic DNA that has to be amplified; for instance, because a disease-causing mutation resides in that fragment. The primers then serve as an anchor for the polymerase, which synthesizes the complementary DNA strand resulting in double-stranded DNA at a higher temperature, typically 72°C, during the third part of the reaction. Ordinary DNA polymerases are not heat resistant and will thus be destroyed at each PCR cycle. Therefore, the discovery of thermophilic DNA polymerases was a major breakthrough in PCR technology. Taq DNA polymerase was the first and most widely used thermophilic DNA polymerase and was first described in *Thermus aquaticus*, a thermophilic bacterium isolated from the hot springs of Yellowstone National Park in the USA.

Like any ordinary chemical reaction, PCR reactions have typical kinetics. First, during the exponential phase, the number of PCR fragments generated increases in an exponential fashion. At a certain point, the reaction components are depleted and PCR reaches a plateau phase, meaning that the number of PCR molecules does not increase anymore. When PCR fragments are analyzed on agarose gels, or other means of fragment analyses, the reaction usually had reached this plateau phase.

Sequencing

The still most widely used chemistry for DNA sequencing is called Sanger sequencing. The chemical principle has remained the same, whether using radioactively labeled nucleotides for detection or the currently widely used fluorescence sequencing. The starting material is nowadays usually a PCR product, but before the advent of PCR, DNA fragments cloned into vectors and flanked by known primers were used. The principle is shown in Fig. 1.8: the DNA of interest is denatured and primers are allowed to bind to the complementary sequences. Then, a polymerase

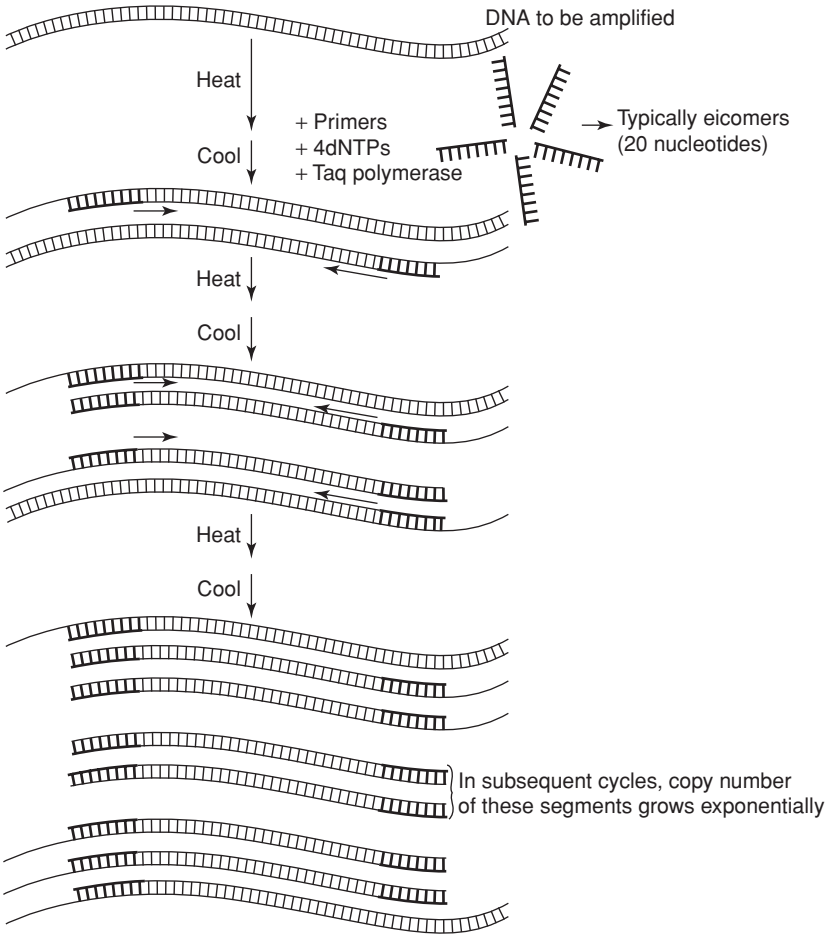

Figure 1.7 Polymerase chain reaction scheme. Source: Fig. 27.7 from Ringo [3].

reaction is initiated. In the reaction mix, four nucleotides with a different chemistry are added next to the usual deoxynucleotides: the dideoxynucleotides. There are four dideoxynucleotides: dideoxy A, dideoxy C, dideoxy G, and dideoxy T. Each dideoxynucleotide is labeled with a different fluorescent dye, represented as shades of gray in Fig. 1.8. In the example, dideoxy G is labeled black, dideoxy T is labeled white, dideoxy A is light gray and dideoxy C is labeled dark gray.

When the polymerase synthesizes the complementary strand, each time a normal nucleotide is incorporated, the synthesis continues. If a dideoxy is incorporated, the synthesis is stopped. This reaction creates a mixture of fragments of different lengths, each differing only one nucleotide in size. The last nucleotide of the fragment is always a fluorescent dideoxynucleotide, and is complementary to the last nucleotide in the analyzed fragment. By separating the different fragments according to their length, and analyzing which fluorescence they emit, the sequence can be read.

Quantitative-real time-reverse transcription PCR

To know if a specific gene is transcribed in a particular cell type, and to what level, quantitative-real time PCR is often used. After RNA extraction from the cells, the mRNA is reverse transcribed into complementary DNA (cDNA). This cDNA can then be used as a template for real-time PCR. The difference between regular PCR and real time PCR is that the number of fragments generated is read during the PCR reaction, through the exponential phase and to the plateau phase, and not at the end after the PCR has reached the plateau phase. For this, special PCR machines are

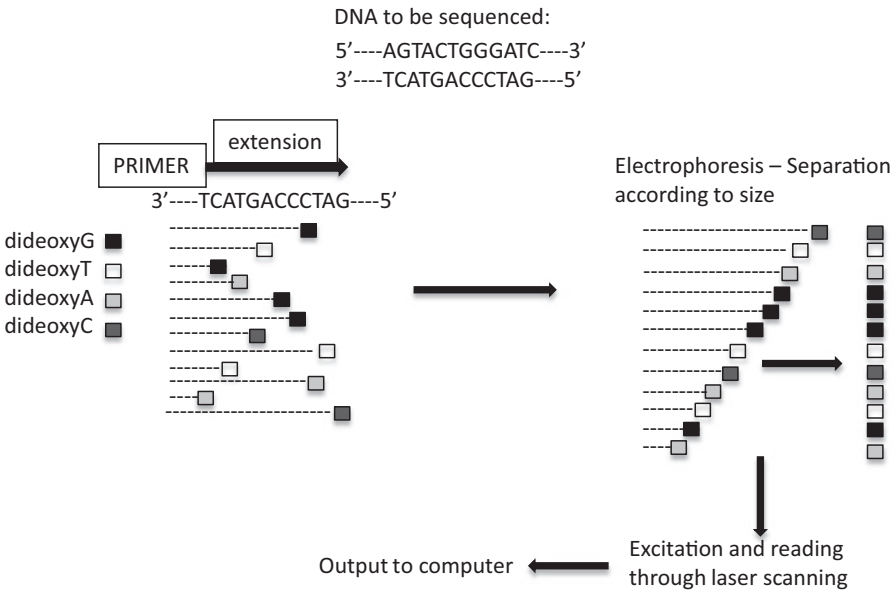

Figure 1.8 Sanger sequencing – dideoxy termination method. For explanation, see the text.

necessary that can read fluorescence incorporated in the PCR fragments after each PCR cycle. The more cDNA of the gene of interest is present in the original sample, the earlier, or the lower number of PCR cycles, the exponential phase of the PCR will start. An often-used measure of the abundance of the cDNA in the original sample is the C_t or threshold cycle. The C_t is defined as the cycle where the number of detected PCR fragments reaches a previously set threshold, usually when the exponential phase has started without doubt. It is a measure of the number of fragments in the original sample. The abundance of the mRNA of interest is usually normalized to the abundance of mRNA from household genes supposed to be expressed in all cells such as HPRT or GAPDH. This allows for the comparison of the expression of a gene of interest between different samples (e.g. *OCT4* in human embryonic stem cells) or for changes that can be measured in time.

The holistic view: modern tools in genetic analysis

Microarrays

Microarrays were developed in the 1990s when the need was first felt to analyze the whole genome or transcriptome of cells. They are always built following the same principles: DNA fragments are spotted upon a suitable support, mostly glass, according to a very precise location, hence the name array. The sample to be analyzed can be DNA or cDNA in solution and fluorescently labeled. After hybridization to the support, the amount of fluorescence for every spot in the array can be read in a high-resolution laser scanner. The abundance of a given DNA fragment or cDNA in a sample will determine the amount of fluorescence. Hence, the amount of fluorescence is a measure for the abundance of the DNA fragment or cDNA.

Originally, DNA fragments from large Bacterial Artificial Chromosome (BAC) libraries were spotted onto the glass slides. The more versatile and amenable oligonucleotide arrays (i.e. very short DNA fragments of only about 20 bp) have now largely taken over, except for very specific applications such as array comparative genomic hybridization (aCGH). Single nucleotide polymorphism (SNP) arrays interrogate SNPs dispersed over the human genome and can be used for haplotyping (i.e. identification of the maternal or paternal origin of the chromosome) and copy number analysis.

Despite the development of higher resolution technologies, aCGH is still a method of choice in prenatal diagnosis because of its robustness and low cost. In reproductive medicine, it is now of course widely used in chromosome analysis in embryos. Chapters 2, 4, and 11 will go more deeply into this topic.