1

Molecular Genetics

Working on genetic epidemiological questions, we have to deal with a variety of information. On the one hand, as in other epidemiological approaches, we use clinical and environmental information. For instance, we might be interested in the relationship between fairer or darker skin, the extent of sun exposure, and the development of malignant melanoma. On the other hand, our specialty is to incorporate genetic information. We might therefore want to look at whether these relationships are mediated by genetic factors. The result might suggest that because of their genetic background, some people are more susceptible to melanoma even though they have darker skin. But what exactly is this genetic background? The aim of this chapter is to familiarize ourselves with this kind of genetic data.

Specifically, we first need to understand what the biological substance of the genetic information is, where it is located in the human body, what it means, and how it is translated for further use. These questions will be answered in Section 1.1. What distinguishes genetic information from other information is that parts of it are transmitted from generation to generation. Understanding the biological mechanisms that underlie this concept forms the basis for later test statistics, and this will be the focus of Section 1.2. Finally, to be useful for statistical purposes, genetic information has to be subject to variation. Hence, the last section of this chapter is devoted to the questions of how individuals differ with regard to their genetic information, how variations can occur, how they can be detected, and how likely the detection is.

This chapter will not give a comprehensive overview on human molecular genetics. Instead, we will focus on issues that are important to understand the statistical methods introduced later. For more in-depth descriptions, the reader is referred to standard textbooks on this topic (e.g., Refs. [254, 636]). Readers who are already familiar with molecular genetics could certainly skip this chapter, especially those who feel comfortable with the problems at the end of this chapter.

1.1 WHAT IS THE NATURE OF GENETIC INFORMATION?

1.1.1 Where is the genetic information located?

Every human cell except the red blood cells has a nucleus that carries an individual's genetic information in *chromosomes*. At the same time, chromosomes are found almost exclusively in the nucleus of the cell. Hence, almost every cell of the body carries the information that is required for the entire organism. The chromosomes are composed of *deoxyribonucleic acid (DNA)* and proteins. The DNA is the carrier of the genetic information, whereas the protein components provide different functions.

The DNA is a large molecule consisting of two strands. Each strand has a linear backbone of alternating sugar (deoxyribose) and phosphate residues. To facilitate the description of the structure, the five carbon atoms of the deoxyribose are consecutively numbered from $1'$ to $5'$ (see Figure 1.1, left-hand side). Covalently attached to the backbone is a sequence of bases. Here, four bases are found, with adenine (A) and guanine (G) being purines, and cytosine (C) and thymine (T) being pyrimidines. The structural unit of one sugar with an attached base is called a nucleoside, and one nucleoside with a phosphate group tied to the carbon atom $5'$ or $3'$ makes one *nucleotide*. In addition to this structure of a single strand, the two strands of the DNA molecule are connected by a hydrogen bond between two opposing bases of the two strands. Specifically, thymine always bonds with adenine via two hydrogen bonds, and cytosine with guanine via three hydrogen bonds. The resulting DNA resembles a ladder whose sides are connected by the bases. However, because of the chemical nature of its components, the ladder does not go straight up but is slightly twisted, which is why it has been described as a *double helix* (Figure 1.1, right-hand side).

Any two DNA fragments differ only with respect to the order of their bases. Therefore, the genetic information we are looking for is coded exactly in the linear sequence of the bases of the DNA fragment. This linear sequence of the bases is called the *primary structure* of the DNA. Because the bonding of the bases between the two strands is specific, the two strands can be said to be complementary, meaning that the sequence of one strand can be exactly inferred from the other. As a consequence, if one wants to describe the sequence, it suffices to write the sequence of one strand. Here, it has become customary to write the sequence in the $5'$ to $3'$ direction. The basic length unit of the DNA is one nucleotide, or one *basepair (bp)*, which refers to the two bases that connect the two strands. In total, the human DNA contains about 3.3 billion bp.

As a second carrier of genetic information in addition to the DNA in chromosomes, copies of parts of the DNA are found in smaller molecules called *ribonucleic acid* (*RNA*) in the nucleus and the surrounding plasma of the cell. The RNA is constructed in a way very similar to the DNA but shows four main differences. In addition to being much shorter, RNA consists of only a single strand instead of two. Also, this single strand is slightly different from a single DNA strand in that the sugar component is made up of ribose instead of deoxyribose. Finally, although the other bases are the same, uracil (U) is found instead of thymine.

The total set of genetic information is distributed in a series of 23 chromosomes. Of these, 22 are autosomes that are consecutively numbered from 1, the longest chromosome, to 21 and 22, the shortest chromosomes. Chromosomes 1 and 2 encompass more than 240 million bp, whereas chromosomes 21 and 22 have no more than 50 million bp. The remaining chromosome is one of the sex chromosomes X and Y, with lengths of about 152 million bp and 50 million bp, respectively. A cell containing a single set of chromosomes with all 22 autosomes and one of the two sex chromosomes is termed to be *haploid*. A regular human cell, however, is *diploid*,

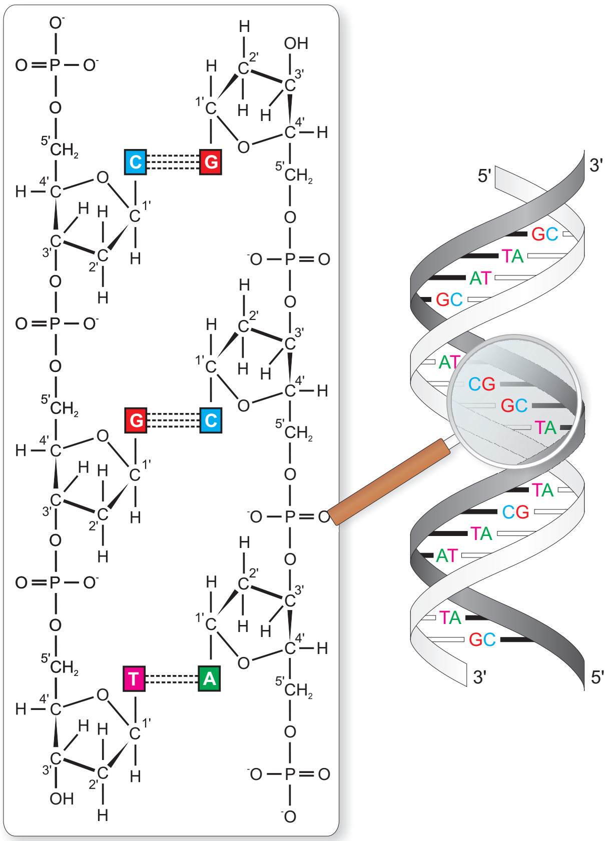

Fig. 1.1 Schematic structure of the DNA. The left-hand side shows the sugar phosphate backbone of the DNA with attached carbon atoms and bases. Two backbones are connected via the bases with two or three hydrogen bonds. The right-hand side is zoomed out to depict that the two strands are slightly twisted, forming the double helix.

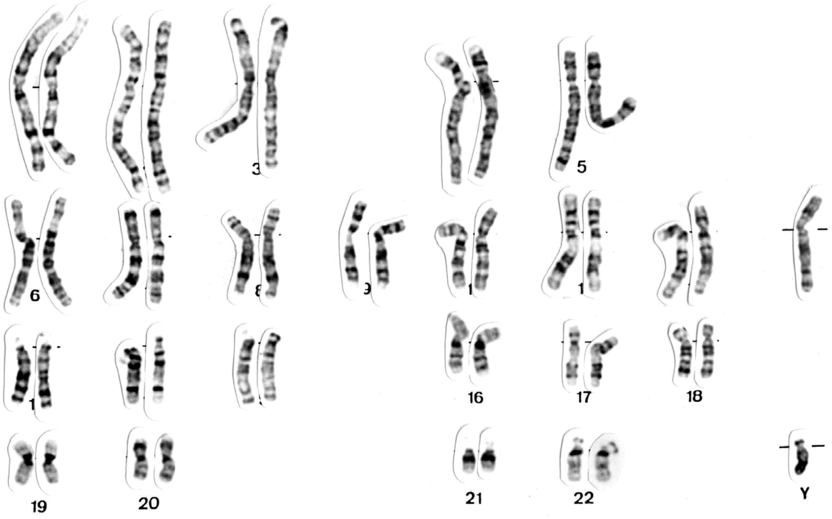

Fig. 1.2 Karyogram of a healthy male. This figure was kindly provided by the Institut für Humangenetik, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Germany.

meaning that it contains a double set, with one coming from the father and the other coming from the mother. Hence, a regular cell has $2 \cdot 22 = 44$ autosomes and two sex chromosomes. Specifically, cells in a female contain two X chromosomes, whereas males carry one X chromosome and one Y chromosome in their cells. There are two regions on the X and Y chromosomes deserving special attention. They are the *pseudoautosomal regions* PAR1 and PAR2 which are homologous sequences of nucleotides on the X and Y chromosomes. The PARs behave like an autosome. Thus genes in this region are inherited in an autosomal rather than a strictly sex-linked fashion. PAR1 comprises 2.6 million bp of the short-arm tips of both X and Y chromosomes in humans. PAR2 is located at the tips of the long arms, spanning 320 kilo bp; the PARs are comprehensively reviewed in Ref. [437].

At a specific point in the division of a cell, the chromosomes can be made visible under the light microscope. When they are stained with certain dyes, they reveal a specific pattern of light and dark bands reflecting regional variations in the amounts of the bases. Differences with respect to length and the banding pattern allow the chromosomes to be easily distinguished from each other, as is visible in Figure 1.2. This

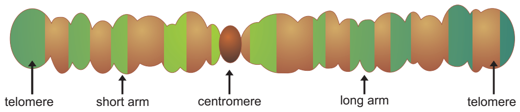

Fig. 1.3 Schematic structure of a chromosome with its characteristic elements.

figure shows an example of the *karyotype* of a healthy male, which is the constitution of the chromosomes of an individual according to standard classification systems. On a single chromosome, different structural elements are distinguished (see Figure 1.3). At the ends of the chromosome, the *telomeres* have special functions involved in the duplication of the chromosomal ends during cell division. A corresponding structure nearclose to the middle of the chromosome is the *centromere*. The short arm of the chromosome is usually termed p for *petit* (small) and the long arm, q for *queue* (tail). Accordingly, the telomeres are referred to as *pter* and *qter*, respectively.

1.1.2 What does the genetic information mean?

After explaining that the genetic information is stored in the linear sequence of the bases of DNA or RNA, we now need to read this information. For this, it is important to note that most functions in human organisms are carried out by proteins. Proteins consist of polypeptides, which are nothing but a linear sequence of repeating units that are called *amino acids*. In humans, 20 different amino acids occur. Hence, there needs to be a translation between the linear sequence of four different bases in the DNA or RNA into a linear sequence of 20 different amino acids for a protein.

How does a base sequence translate into protein structure? It has been found that three bases, a triplet, code for one amino acid. Accordingly, the sequence of three bases is called a *codon*. Using such a three-letter code, it is possible to form $4^3 = 64$ different codons from four possible bases. Because there are only 20 amino acids that need to be coded, the genetic code can be said to be degenerate, with the third position often being redundant. Depending on the starting point of reading, there are three possible variants to translate a given base sequence into an amino acid sequence. These variants are called *reading frames*. The beginning of the translation process from bases into amino acids is signaled by special functional start codons, mostly A(denine)-U(racil)-G(uanine). The opposite stop codons, for instance, U(racil)-A(denine)-G(uanine), will terminate the translation. It should be noted that in practice, the translation of bases into amino acids does not use DNA but RNA; the base uracil appears in the code that is displayed in Figure 1.4.

1.1.3 How is the genetic information translated?

As we have seen, genetic information is basically a construction plan for proteins. Hence, we now need to understand how, beginning with the DNA, proteins are actually synthesized. Because of the importance of this process, protein synthesis has been termed the *central dogma of molecular biology*. To anticipate the major pathway, this dogma has been expressed as:

DNA makes RNA, RNA makes proteins, proteins make us.

The overall process of protein synthesis can be partitioned into two steps. First, we have to remember that DNA is stationary and located in the nucleus of the cell. In contrast, protein synthesis takes place in the surrounding plasma of the cell. Therefore, the first step involves the *transcription* of DNA into messenger RNA

Amino acid	Notation
Alanine	Ala
Arginine	Arg
Aspartate	Asp
Asparagine	Asn
Cysteine	Cys
Glutamine	Gln
Glutamate	Glu
Glycine	Gly
Histidine	His
Isoleucine	Ile
Leucine	Leu
Lysine	Lys
Methionine	Met
Phenylalanine	Phe
Proline	Pro
Serine	Ser
Threonine	Thr
Tryptophan	Trp
Tyrosine	Tyr
Valine	Val

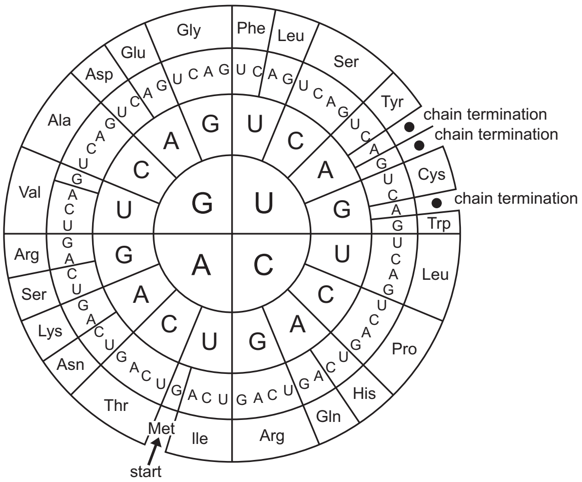

Fig. 1.4 Codon table to translate RNA into amino acid. The codon table is read from center to outside, so that, for example, the sequence A (center) U (middle) G (outside) codes for the amino acid Met (methionine) and at the same time serves as an initiation site for translation (see Section 1.1.3).

(mRNA) that then carries the information from the nucleus into the plasma. More specifically, the DNA double helix is unwound and unzipped into single strands. Catalyzed by the RNA polymerase enzyme, the strand in the $5'$ to $3'$ direction is used as a template. Led by the polymerase that migrates from one nucleotide to the next, free RNA nucleotides anneal to the DNA and are tied together to a strand. After coming across a stop signal, the RNA strand uncouples from the DNA that is then again zipped together with its complementary strand. The resulting RNA subsequently leaves the nucleus of the cell. Because the RNA now has the same direction and base sequence as the strand of the DNA that was not used as a template, this non-template strand is called the *sense strand*. In contrast, the template DNA strand is termed the *antisense strand*.

After the transcription, the RNA molecule is further edited. Specifically, the so-called *introns* are cut out, whereas *exons* are spliced together. For a large number of segments, multiple alternative splicing variants exist. Finally, unique features are added to each end of the transcript to make a mature mRNA.

The transcription process is regulated by a number of factors. For example, several kinds of the enzyme RNA polymerase and associated protein transcription factors regulate the specificity and rate of transcription. Specific regions of the DNA called *promoters* that include binding sites for the RNA polymerases control the initiation of the transcription. In addition, transcription can also be regulated by variations in the DNA structure or by chemical changes in the bases. One such important chemical modification is methylation. The degree of methylation tends to be negatively correlated with the functional activity of DNA and plays an important role, for example, in genomic imprinting (see Section 2.3.3).

The second step of protein synthesis involves the *translation* of the genetic information that is now contained in the mRNA from genetic code into proteins. The process of translation takes place in the cell plasma at cell organelles called *ribosomes*. After the actual synthesis of amino acid sequences, the proteins are further modified. The advantages of this procedure are that only small segments of DNA are used at a time and that many transcriptions and translations can be rapidly made.

After describing the nature of our genetic information in some depth, we should briefly turn to the question of what a gene actually is. Although definitions vary from source to source, a *gene* is often defined to be a functional unit of the DNA that encodes a product, usually a protein, and includes exons, introns, and small regions of DNA immediately preceding or following the transcribed region. The overall genetic makeup of one individual is termed the *genotype*. In contrast, the observable characteristics of the individual are called the *phenotype*. With *coding* DNA the section that will finally be translated into proteins is described, consisting mostly of exons, whereas *non-coding* DNA includes both the introns and the sequences between genes.

It is important to realize that only a small proportion of the DNA is ever transcribed, and only a proportion of the transcribed mRNA is finally translated into protein. With the length of a gene going up to $10{,}000$ bp and the human genome containing about $20{,}000$ to $25{,}000$ genes according to the National Human Genome Research Institute, genes make up only about 1% of the total DNA. The non-coding DNA contains repetitive DNA, pseudo-genes, or regulatory sequences. These provide critical functions such as indicating the beginning and end of a coding sequence or providing binding sites for proteins that turn the transcription on and off. In addition, in different cells, different units of DNA are transcribed.

1.2 HOW IS GENETIC INFORMATION TRANSMITTED FROM GENERATION TO GENERATION?

By now, we know what the genetic information is, where it is located, and what it means. The specific peculiarity, however, is that it is partly transmitted from generation to generation and that information from two individuals, the mother and the father, is combined in the offspring. This sequence of distribution and combination forms the basis for a number of test statistics that will be introduced later.

Two steps can be distinguished in the process of transmission. First, in *meiosis*, the genetic information of each parent is transferred into germ cells, the *gametes*. Second, gametes of the father and the mother fuse to form the zygote that carries information from both parents.

Meiosis is a special form of cell division in which a single cell is divided into four daughter cells. Because the resulting gametes each have only half the number of chromosomes of the progenitor cell, these are haploid. To be specific, meiosis includes one round of chromosome duplication and two rounds of cell division (Figure 1.5):

1. The process begins with a regular diploid cell $(2n)$, which means that there is one paternal and one maternal double strand or chromosome. Matching maternal and paternal chromosomes are termed to be *homologous*.
2. DNA replication: The DNA is duplicated, and as a result, each chromosome now contains two identical DNA double strands. These are termed *sister chromatids*. Hence, there is a total of four double strands $(4n)$ in the cell.
3. Forming of bivalents: Homologous chromosomes are connected to form *bivalents*.
4. Possible crossing over: In this stage, exchange of genetic material between maternal and paternal strands is possible.
5. Meiotic division I: Non-sister chromatids are separated, while sister chromatids remain paired. This results in two diploid cells containing the sister chromatids.
6. Meiotic division II: The sister chromatids are separated, resulting in four haploid cells (gametes).
7. During fertilization, the fusion of the genetic material from two gametes leads to a restitution of the diploid status.

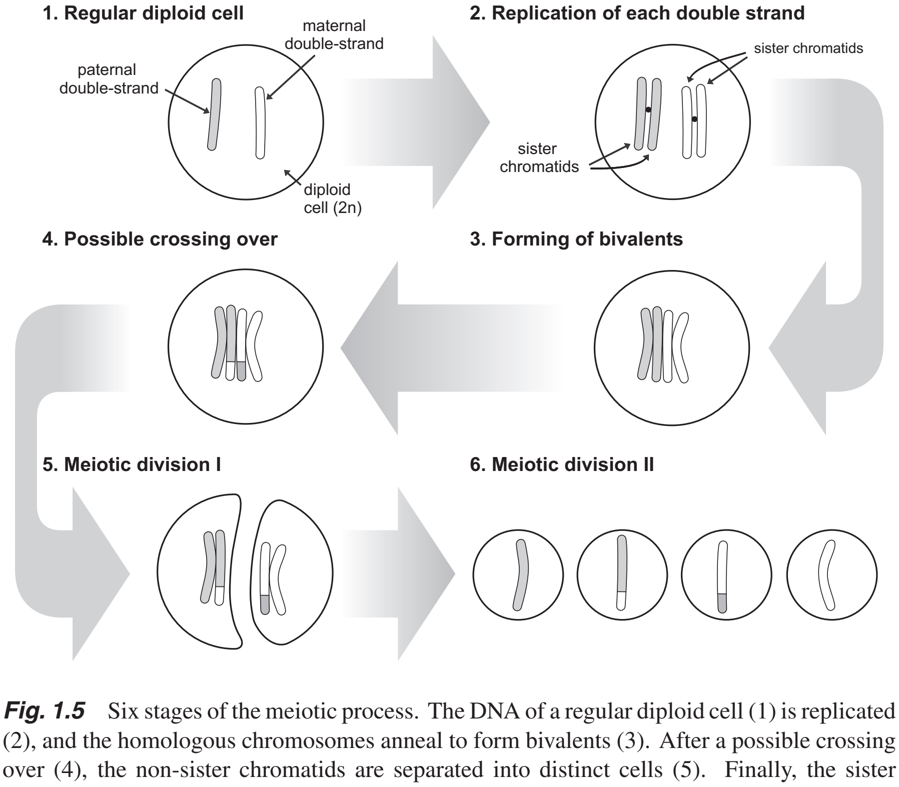

Fig. 1.5 Six stages of the meiotic process. The DNA of a regular diploid cell (1) is replicated (2), and the homologous chromosomes anneal to form bivalents (3). After a possible crossing over (4), the non-sister chromatids are separated into distinct cells (5). Finally, the sister chromatids are distributed to separate haploid gametes (6).

An important feature of meiotic division is that homologous chromosomes are distributed randomly and independent of each other to the gametes. Thus, the resulting gamete most likely contains some chromosomes that the individual has inherited from the mother and others that have come from the father, but the specific combination of chromosomes is random. Hence, there is a large number of possible combinations of the chromosomes in a single cell. With a total of 23 chromosome pairs, the number of possible combinations in one gamete from one parent is 2^{23}. When the fusion of one gamete with another is considered, the number of possible chromosomal combinations is doubled.

In addition, further shuffling of genetic material is possible. In step 4 of the meiotic process, an exchange of genetic material between the maternal and paternal DNA strand, might occur. During meiotic division I, non-sister chromatids are crossed over and form visible chiasmata. At their points of crossing over, the chromatids may break at homologous points and reunite with the non-sister chromatid. This can happen not only once, at one point on the chromosome, but several times, and the results are quite different. If we consider the two chromosomal segments A and B on the two chromatids 1 and 2, one possibility would be that no crossing over occurs between these segments (see Figure 1.6a). Hence, $A1$ will remain on the same chromatid with $B1$, and so will $A2$ with $B2$. In contrast, one crossing over might have taken place (see Figure 1.6b). The result will be a shuffling of the chromosomal segments with $A1$ and $B2$ on one chromatid and $A2$ and $B1$ on the other. Consequently, a *recombination* between segments A and B has happened. A third possibility is that two crossing overs occur (see Figure 1.6c). Although the intermediate segment is exchanged, this leads to the same result regarding the upper and lower segments as with no crossing over, namely, $A1$ and $B1$ being on the first chromatid and $A2$ and $B2$ on the second. Hence, no recombination has taken place between the segments. To generalize from this, we can deduce that a recombination between two chromosomal segments can be observed whenever there is an odd number of crossing overs between them. Even numbers of crossing overs, on the other hand, will even out any interchanging.

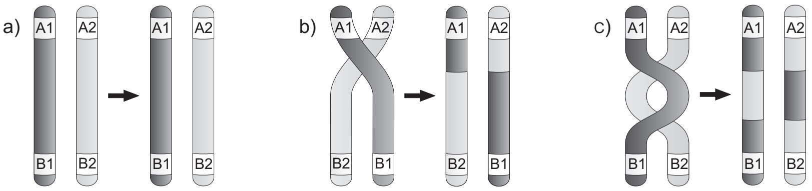

Fig. 1.6 Recombination between chromosomal segments A and B on chromatids 1 and 2. a) No crossing over might occur. b) One crossing over might occur, leading to a recombination between A and B. c) Two crossing overs might occur, leading again to no visible recombination between A and B.

Crossing overs are not rare events. Specifically, the mean number per cell is about 55 in males and is approximately 50% higher in females. The further apart two segments are from each other on the chromosome, the greater the probability is that a crossing over will occur between them. Hence, the greater the distance, the higher the probability for a recombination to be observed between the segments. Turned around, the probability for a recombination, termed the *recombination fraction* θ, can be used as a measure of the distance of two chromosomal segments. If the segments were located very close to each other, they would almost never be separated by a crossing over, hence θ would approximate 0. If, at the other extreme, they were situated on different chromosomes, the first meiotic division would distribute them to different cells in about half the cases, thus leading to a θ of 0.5. Similarly, two segments on the same chromosome but very far apart will almost certainly be subject to a recombination because of the high number of mean crossing overs. Hence, they will again be separated by the first division with equal probability, and θ again approximates 0.5. This forms the basis for different measures of distance between segments and will be discussed in detail in Chapter 5.

1.3 WHAT IS INDIVIDUAL VARIATION IN GENETIC INFORMATION?

1.3.1 How do individuals differ with regard to their genetic information?

In the previous sections, we have assumed that the DNA strands inherited from father and mother are not identical. Generally, we can state that different individuals usually have the same genes, but not exactly the same DNA sequence. An exception is identical twins, who share the same DNA sequence. To use a more general term, we focus on the variability of a DNA *locus*, which is a specific DNA segment defined by its sequence of bases. Variants at one DNA locus are termed *alleles*. Concerning diploid cells, the two alleles at one locus define the genotype at this locus. If an individual carries the same alleles on both DNA strands, he is said to be *homozygous* at the locus. If the alleles are different, the individual is *heterozygous* for the respective DNA locus. But how do differences at a DNA locus come about in the first place?

There are two main sources for any variation in the DNA sequence:

1. During cell division, an abnormal pairing of the chromatids leads to a rearrangement of chromosomal segments. This in turn can result in insertions, deletions, inversions, translocations, or duplications of whole chromosomal segments.
2. During DNA replication, point mutations with substitutions of one nucleotide for another occur. This may result in transitions (exchanging one purine for another or exchanging one pyrimidine for another) or transversions (exchanging a purine for a pyrimidine or vice versa), leading to a synonymous or a non-synonymous base sequence. If instead one or more (except for multiples of three) nucleotides are inserted or deleted, a frameshift mutation emerges.

Mutation, the second source of variation, is defined as a state that permanently changes the function of the gene. Mutations happen because the DNA replication is not absolutely flawless. Most of the errors are corrected immediately, but if an error is not detected or repaired, a mutation has occurred. In fact, the mutation frequency for a single gene ranges from around 10^{-4} to 10^{-6}. As a result, the base sequences of any two human individuals differ with a frequency of about one base out of 1000.

The consequences of a mutation depend on the kind and location of the variation. If one nucleotide is substituted for another, the redundancy of the genetic code may yield the same protein being synthesized, which is called a synonymous mutation. If the resulting codon codes for a different amino acid, the synthesized protein will be different and hence the mutation is called non-synonymous. If nucleotides are inserted or deleted from the DNA, we have to distinguish between two cases. First, the insertion or deletion of a multiple of three nucleotides leads to additional or fewer amino acids being coded. Second, any other number of insertions or deletions will lead to a shift of the reading frame; hence, a different protein will be synthesized.

A term that is sometimes used interchangeably with mutation is *polymorphism*. Although there are definitions that include the consequences of the variation, we will use the term polymorphism to signify a variation in DNA sequence with a frequency of at least 1% in at least one human population.

With novel genetic variants being detected and described, it is important to have a standardized nomenclature to denote them [29]. As a general recommendation, variations should be described on the level of DNA. In this case, the naming itself contains the following elements:

- *g.* for genomic DNA,
- Position of the variation. This is given by the number of the nucleotide in which the variation begins. For this, the base adenine from the start codon is denoted nucleotide $+1$ and the nucleotide in $5'$ to $+1$ is denoted -1. If you want to specify a range, the symbol "_" should be used.
- Specific change. The most common changes are symbolized by $>$ for a substitution, *del* for a deletion, *ins* for an insertion, *dup* for a duplication, and *inv* for an inversion.

Let us consider the following examples: g.23G>T is a substitution of guanine by thymine at position 23, g.3_5delTCC means that the bases thymine, cytosine, and cytosine are deleted from nucleotides 3 to 5, and g.3_4insT symbolizes the inversion of thymine between nucleotides 3 and 4.

If the change has taken place in an intron, the nomenclature suggests to denote the change by *IVS* followed by the number of the intron and the position within the intron. For instance, IVS4 + 2A>C means that adenine was substituted by cytosine in intron 4 at position 2.

It is out of scope of this book to describe the nomenclature of variations on different levels such as genes, RNA, or proteins. These and more complex cases as well as current updates can be found on the web site of the Human Genome Variation Society and/or in Refs. [159, 309, 682].

1.3.2 How can individual differences be detected?

We now know how variations in the DNA sequence may occur and which kinds of mutations are possible. A series of molecular biological laboratory techniques have been developed to detect a novel mutation in a given DNA sequence. In this section, we will describe the basic procedure for sequencing a DNA segment. Because larger amounts of DNA are required for this and other techniques, the amplification of DNA segments plays an important role. Hence, we will explain the basics of the polymerase chain reaction as an amplification technique. Common technologies for single nucleotide polymorphism (SNP) genotyping are described in Section 3.3.

1.3.2.1 Sequencing of DNA segments Different sequencing techniques have been developed over the past few decades. Because our aim is to explain the principles underlying these laboratory techniques, we will focus on one specific technique: the *dideoxy sequencing* method developed in the late 1970s, also termed the *chain-termination method* or, after its developer, the *Sanger method*.

The requirements for the original sequencing reaction are four test tubes with the following reagents:

- A template DNA segment that is to be sequenced. This is prepared as a single-stranded DNA (denatured).
- A set of DNA deoxynucleotides (dATP, dTTP, dCTP, dGTP) to build new segments of DNA.
- Single-stranded labeled DNA segments complementary to a short region on either side of the template sequence. These are called *primers*.
- A specific polymerase enzyme to catalyze the synthesis of the new DNA.
- In each tube, a small amount of one dideoxynucleotide triphosphate that terminates the chain growth wherever it is incorporated (a specific base stop).

The actual sequencing reaction works as follows (see Figure 1.7):

1. Annealing of the primer to the template.
2. Elongation of the primer to complementary DNA using the available deoxynucleotides.
3. Termination of the elongation as soon as a base-specific stop nucleotide is incorporated that cannot form a phosphodiester bond with the next incoming deoxynucleotide.

In each test tube, the result will be a set of new DNA strands of different lengths. To be specific, each length signals the incorporation of a stop nucleotide, hence the occurrence of the respective base in the template DNA. The fragments of different sizes from the four test tubes are then separated by a procedure called gel electrophoresis.

The purpose of gel electrophoresis is to sort a number of DNA segments according to their length. Basically, an agarose gel is poured with small molds and the fragments are poured into the molds, meaning that the gel is loaded with DNA fragments. Then, an electrical field is charged. Because of the current, the fragments start to migrate through the gel. The important point is that their speed of migration depends on their length: the longer they are, the more weight they have and the slower they travel.

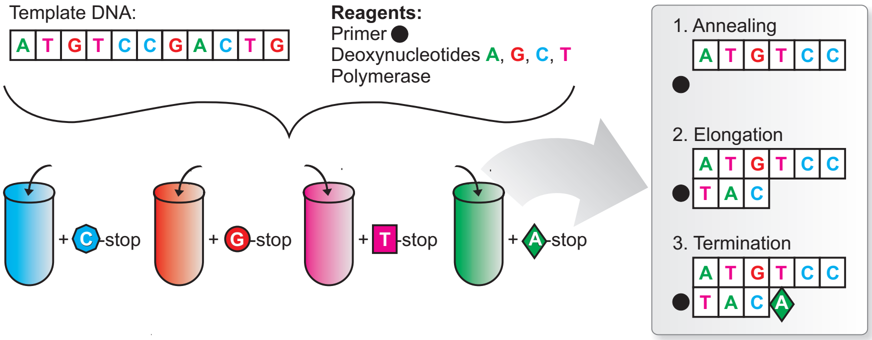

Fig. 1.7 Sequencing of DNA segments. The template DNA and reagents (top left) are mixed in four test tubes that differ only with regard to the added dideoxynucleotide triphosphate serving as a base-specific stop sign (bottom left). In each test tube, the primer will anneal to the template (top right) and be elongated to form a complementary strand (middle right). If, for example, dideoxyadenosine triphosphate (di-dATP) has been added to the mixture, as in the right test tube, instead of the usual deoxyadenosine triphosphate (dATP), di-dATP will be eventually incorporated (bottom right). This will terminate the elongation.

Nowadays, seminal variations of the original procedure have been developed for automated sequencing machines (for reviews, see Refs. [458, 459, 555, 724]). For example, in cycle sequencing the dideoxynucleotides are labeled with fluorescent dyes instead of the primers. Hence, all four reactions can be performed in the same tube, and the products can be loaded into the same lane of the gel. To differentiate the reaction products during electrophoresis, a laser is directed at a fixed position of the gel. As the products migrate past the laser, the laser causes the dyes to fluoresce, and the fluorescence is electronically recorded and interpreted with regard to the wavelength. The result is an electropherogram showing the fluorescence units over time and fragment position (see Figure 1.8).

An important prerequisite for sequencing a DNA segment is that the flanking sequence must be known for primers to be constructed. The two utilized primers have to be specific sequences that are unique in the genome. From a number of approximately 11 bases upward, a locus usually is sufficiently unambiguous in the human genome. Most error prone in this sequencing technique is the reading of the results of the electrophoresis. Hence, it is normally recommended that the output is read twice and additionally once from both directions. The technique is rather expensive and can be performed more efficiently if several loci are simultaneously sequenced. Here, in order to correctly read the result, it is important that they are unambiguously distinguishable with regard to their length. Several factors influence the success of the sequencing. For example, DNA with high *guanine-cytosine content (GC-content)* which are termed *GC-rich* regions can be difficult to sequence. GC-content is the percentage of guanine or cytosine bases on a DNA molecule. GC pairs are bound by three hydrogen bonds, while AT pairs are bound by two hydrogen bonds. DNA with high GC-content is more stable than DNA with low GC-content.

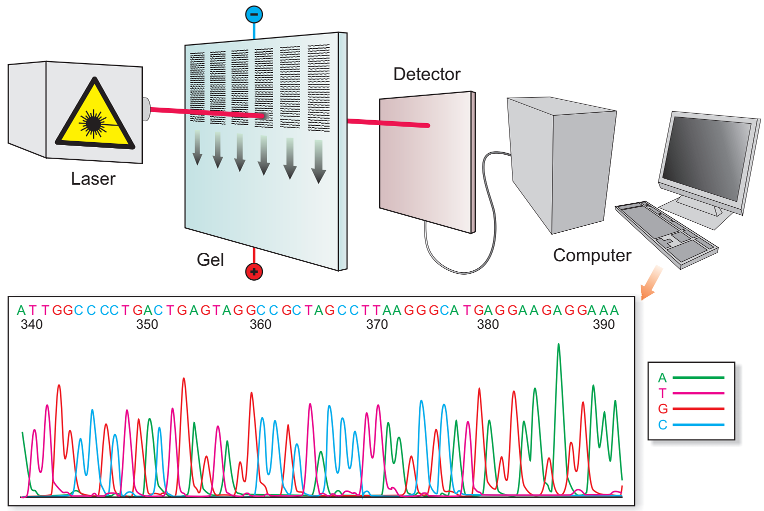

Fig. 1.8 An electropherogram depicting the results of an automated DNA sequencing using fluorescent primers. During electrophoresis, DNA fragments migrate in an electrically charged field, with their speed depending on their length: longer fragments are slower. A laser beam directed at a fixed position of the gel causes the dyes to fluoresce. The recorded fluorescence units over time, thus the position on the fragment, are given in the electropherogram. The example represents sequencing of clone *RP*11-386*I*14 from chromosome 1. This figure was kindly provided by Jeanette Erdmann.

Therefore, the sequence will start out strong but the signal strength will rapidly decrease until there is no sequence data. As a consequence, read lengths are typically shorter for these templates. Problems can also arise for homopolymeric regions and in the presence of high repetitions.

1.3.2.2 Amplification of DNA segments Laboratory techniques such as sequencing require a certain amount of DNA that may not always be available. One possibility for amplifying DNA segments in vitro is the *polymerase chain reaction* (*PCR*).

For the PCR, the following reagents are required:

- A template DNA segment containing the target sequence that is to be amplified.
- A set of DNA nucleotides to build the new DNA.
- Primers complementary to a short region on either side of the target sequence.
- A specific polymerase to catalyze the synthesis of the new DNA.

The PCR procedure is run in a number of cycles (see Figure 1.9, upper half). In each cycle, the following steps are performed:

1. Denaturation: At a high temperature, the template DNA is melted and thus denatured, i.e., the two strands are unzipped.

2. Hybridization: At a low temperature, the primers anneal to the complementary sequences at the sides of the target sequence.
3. Elongation: At a medium temperature, the new DNA is synthesized by the polymerase enzyme. The elongation of the two primers is performed in the direction of the amplified sequence.

As a result, in the first cycle both primers are elongated along the original template DNA until the end of the template. In the second cycle (see Figure 1.9, lower half), the primers additionally initiate synthesis on the products of the first cycle. Hence, half the strands can be elongated as before without a specified stop. The others terminate where the original primers end. In further replications, amplification is more and more restricted to the sequence between the sites defined by the primers, which is the target sequence.

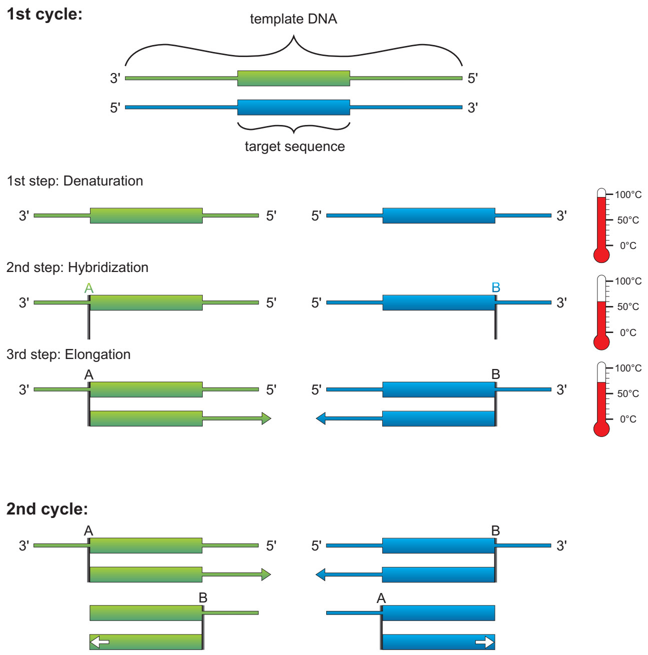

Fig. 1.9 Polymerase chain reaction (PCR) for DNA amplification. A DNA segment containing the target sequence (first from top) is denatured at high temperature so that the two strands become separated (second). Then, the primers hybridize to the ends of the target sequence at low temperature (third). At medium temperature, the primers are elongated unidirectionally (fourth). This cycle of high, low, and medium temperature is repeated (bottom). Importantly, the DNA segment of the previous cycle now serves as an additional template, the only difference being that it is restricted on one side by the primer. Hence, the following synthesized DNA segments will increasingly be restricted to the region between the two primers.

The number of DNA fragments that are thus created increases exponentially with the number of cycles. After 30 to 50 cycles, about 2^{30} to 2^{40} copies are produced.

As in the sequencing reaction, the two primers have to be unique sequences. However, their unambiguity being granted, the length is no longer important for the success of the PCR. What is more important is that their chemical features are similar; specifically, they should have comparable melting temperatures.

As polymerase, different enzymes are used. The most common is the Taq polymerase from the bacterium *Thermus aquaticus*. Because the bacterium lives in hot springs, its polymerase is adapted to high temperatures and is not destroyed during the melting phase of the PCR. Hence, it is not necessary to add new polymerase in each cycle, allowing the entire procedure to be automatized. However, because the Taq polymerase is often connected with a higher error rate, polymerase from different bacteria, for example, *Pyrococcus furiosus*, which has a better correction of errors, is used.

The advantages of the PCR are that the reaction is fast, highly specific, easily automated, and capable of amplifying small amounts of the DNA segment. As before, the sequence to be amplified has to be known in advance in order to construct primers.

1.3.3 How likely is it that individual differences are detected?

Although we now know how to amplify a variation, the last question still has to be answered: How probable is it to detect a variation that has a specific frequency in a given sample? Alternatively, one might be interested in calculating the sample size required to detect a variation with a pre-specified confidence. For simplicity, we first assume that all subjects are homozygous with one of the variants. Furthermore, we assume that the variation is the so-called single nucleotide polymorphism (SNP), i.e., that the variation has only two states (for more details on this, see Section 3.2). If the frequency of the less frequent variant is p and $1-p$ for the more frequent variant, then the probability of observing only the rare variant is p^n and the probability of observing only the frequent variant is $1-p$. Subsequently, the probability of observing only one variant $(1v)$ in n subjects is

$$P(1v) = p^n + (1-p)^n \,. \tag{1.1}$$

So far, we have assumed that all subjects are homozygous. If subjects can be heterozygous and if the frequency for homozygous and heterozygous subjects is in Hardy–Weinberg equilibrium (HWE; see Chapter 2), we do not need to consider subjects but can base our calculations on alleles. In this case, the probability of observing only one variant $(1v_{\text{HWE}})$ in n subjects with $2n$ alleles is

$$P(1v_{\text{HWE}}) = p^{2n} + (1-p)^{2n} \,. \tag{1.2}$$

Equations (1.1) and (1.2) can be used to calculate the number of samples that need to be investigated until both variants are detected. To this end, $1 - P(1v)$ needs to be considered and solved for n given p. This is best achieved by an algorithm, and several solutions including MINSAGE and the package GENETICS in R are available. These packages are able to do the necessary calculations in case of multiple variants as discussed in detail by Gregorius [263].

Table 1.1 Probability for observing both variants of a single nucleotide polymorphism (SNP) as a function of the frequency of the rarer variant (p) and the sample size (n). The third column gives the probability $P(1v)$ when only homozygous subjects are available, and the probabilities $P(1v_{\text{HWE}})$ reported in the fourth column are based on alleles, i.e., under the assumption of Hardy–Weinberg equilibrium.

p	n	$P(1v)$	$P(1v_{\text{HWE}})$	p	n	$P(1v)$	$P(1v_{\text{HWE}})$
0.5	10	0.9980	1.0000	0.01	100	0.6340	0.8660
0.4	10	0.9938	1.0000	0.01	200	0.8660	0.9820
0.3	10	0.9717	0.9992	0.01	500	0.9934	1.0000
0.3	20	0.9992	1.0000	0.005	10	0.0489	0.0954
0.2	10	0.8926	0.9885	0.005	20	0.0954	0.1817
0.2	20	0.9885	0.9999	0.005	30	0.1396	0.2597
0.2	30	0.9988	1.0000	0.005	50	0.2217	0.3942
0.1	10	0.6513	0.8784	0.005	100	0.3942	0.6330
0.1	20	0.8784	0.9852	0.005	200	0.6330	0.8653
0.1	30	0.9576	0.9982	0.005	500	0.9184	0.9933
0.1	50	0.9948	1.0000	0.005	1000	0.9933	1.0000
0.05	10	0.4013	0.6415	0.001	10	0.0100	0.0198
0.05	20	0.6415	0.8715	0.001	20	0.0198	0.0392
0.05	30	0.7854	0.9539	0.001	30	0.0296	0.0583
0.05	50	0.9231	0.9941	0.001	50	0.0488	0.0952
0.05	100	0.9941	1.0000	0.001	100	0.0952	0.1814
0.01	10	0.0956	0.1821	0.001	200	0.1814	0.3298
0.01	20	0.1821	0.3310	0.001	500	0.3936	0.6323
0.01	30	0.2603	0.4528	0.001	1000	0.6323	0.8648
0.01	50	0.3950	0.6340	0.001	2000	0.8648	0.9817

Table 1.1 displays the number of subjects, the frequency of the rarer variant p, the probability $P(1v)$ of not observing both variants when only homozygous subjects are available, and the probability $P(1v_{\text{HWE}})$ of not observing both variants under the assumption of HWE. If the frequency of both variants is high, only few subjects like 10 or 20 need to be screened. However, if the frequency of the rarer variant is as low as 1%, several hundreds need to be screened for observing both variants with a probability of at least 99%. If the frequency of the rare variant is as low as 0.001,

the probability for observing both variants is 0.8648 for a sample size as large as 1000 if HWE holds. This would represent the best case because it allows to consider the single alleles. When only homozygous subjects can be observed, the probability of not observing both variants is 0.6323, thus substantially higher. Although this situation is unrealistic in practice, the two scenarios considered here are the two extremes. While the assumption of HWE yields the lowest number of subjects that need to be screened for identifying both variants, the assumption of homozygosity of all subjects represents the worst case scenario and gives the highest numbers.

1.4 PROBLEMS

Problem 1.1

Problem 1.1.1. Answer the following questions:

1. What are the main constituents of a DNA molecule?
2. How are the two strands of one DNA molecule connected?
3. What are the main differences between DNA and RNA?

Problem 1.1.2. Define the following terms:

1. Nucleotide and nucleoside
2. Diploid and haploid
3. Gene, genotype, and phenotype
4. Introns, exons, and promoters

Problem 1.1.3. Consider the following base sequence of a given DNA segment:

$$\boxed{\text{T A C A A T G A T C T G A C G A T T}}$$

1. What is the sequence of the complementary DNA strand?
2. Given the original template strand, what would be the sequence of an RNA strand after transcription?
3. What would be the result of translation?

Problem 1.2

Problem 1.2.1. Describe the process of meiosis. Specifically, recall how many chromosomes each cell contains before meiosis, after each stage, and after meiosis.

Problem 1.2.2. Consider the following five chromosomal segments A, B, C, D, and E on two chromatids (Figure 1.10).

1. Between which segments would you observe recombinations if there is one crossing over between segments B and C?
2. Between which segments would you observe recombinations if there is one crossing over between segments A and B and a second between segments C and D?

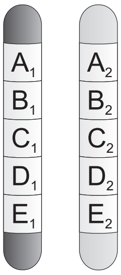

Fig. 1.10 Illustration of five chromosomal segments on two chromatids for Problem 1.2.2.

Problem 1.3

Problem 1.3.1. Define the following terms:

1. Locus and allele
2. Homozygous and heterozygous
3. Mutation and polymorphism

Problem 1.3.2. What would the resulting peptide look like if the following modifications of the original template strand from Problem 1.1.3 had occurred?

1. Insertion of the single base adenine at position 8, between adenine and thymine
2. Substitution of the single base thymine at position 9 with the base adenine
3. Deletion of bases adenine, thymine, and cytosine at positions 8, 9, and 10

Problem 1.3.3. Describe the following techniques:

1. Dideoxy sequencing
2. Polymerase chain reaction

Problem 1.4 What is the probability of not observing both alleles when a variant is expected to exist in 1 out of 10,000 subjects? Assume that 1000, 10,000, 20,000, and 500,000 subjects are available for screening.

URLs

Deciphering the genetic code
http://history.nih.gov/exhibits/nirenberg/
DNA from the beginning
http://www.dnaftb.org/dnaftb/
genetics
http://cran.r-project.org/web/packages/genetics/index.html
minsage
http://www.imbs-luebeck.de/software/minsage.html
Human Genome Variation Society
http://www.dmd.nl/mutnomen.html