Vogel and Motulsky's Human Genetics

Problems and Approaches

Bearbeitet von
Michael Speicher, Stylianos E. Antonarakis, Arno G. Motulsky

ISBN 978 3 540 37653 8
Format (B x L): 20,3 x 27,6 cm
Gewicht: 2075 g

Weitere Fachgebiete > Chemie, Biowissenschaften, Agrarwissenschaften > Tierkunde, Zoologie > Tiergenetik, Reproduktion

schnell und portofrei erhältlich bei

beck-shop.de

Die Online-Fachbuchhandlung beck-shop.de ist spezialisiert auf Fachbücher, insbesondere Recht, Steuern und Wirtschaft. Im Sortiment finden Sie alle Medien (Bücher, Zeitschriften, CDs, eBooks, etc.) aller Verlage. Ergänzt wird das Programm durch Services wie Neuerscheinungsdienst oder Zusammenstellungen von Büchern zu Sonderpreisen. Der Shop führt mehr als 8 Millionen Produkte.
Abstract The knowledge of the content of the individual human genomes has become a sine qua non for the understanding of the relationship between genotypic and phenotypic variability. The genome sequence and the ongoing functional annotation require both comparative genome analysis among different species and experimental validation. Extensive common and rare genomic variability exists that strongly influences genome function among individuals, partially determining disease susceptibility.

2.1 The Human Genome

In order to be able to understand the biological importance of the genetic information in health and disease (assign a particular phenotype to a genome variant) we first needed to know the entire nucleotide sequence of the human genome. Thus an international collaborative project has been undertaken named “The Human Genome Project” to determine the nucleotide sequence of the human genome. The project was initiated on 1 October 1990 and was essentially completed in 2004.

The potential medical benefits from the knowledge of the human genome sequence were the major rationale behind the funding of this international project. In addition, the involvement and contributions of the biotechnology company Celera may have provided the necessary competition for the timely completion of the project. The last (third) edition of this book was published in 1997 before the knowledge of the human genome sequence; thus, this fourth (“postgenome”) edition of the book proudly begins with the discussion of “genome anatomy,” as the genomic sequence was named by Victor McKusick.

The goals of the different phases of the Human Genome Project were to: (1) determine the linkage map of the human genome [1, 60]; (2) construct a physical map of the genome by means of cloning all fragments and arrange them in the correct order [32, 69]; (3) determine the nucleotide sequence of the genome; and (4) provide an initial exploration of the variation among human genomes.

As of October 2004 about 93% of the human genome (which corresponds to 99% of the euchromatic portion of the genome) had been sequenced to an accuracy of better than one error in 100,000 nucleotides.

S.E. Antonarakis
Department of Genetic Medicine and Development, University of Geneva Medical School, and University Hospitals of Geneva, 1 rue Michel-Servet, 1211 Geneva, Switzerland
email: Stylianos.Antonarakis@unige.ch
The DNA that was utilized for sequencing from the public effort came from a number of anonymous donors [84], while that from the industrial effort came from five subjects of which one is eponymous, Dr. J.C. Venter [85, 137]. The methodology used was also different between the two participants: the public effort sequenced cloned DNA fragments that had been previously mapped, while that of Celera sequenced both ends of unmapped cloned fragments and subsequently assembled them in continuous genomic sequences. Detailed descriptions of the genome content per chromosome have been published; the first “completed” chromosome published was chromosome 22 in 1999, chromosome 21 was published in 2000, and all other chromosomes followed in the next 6 years [38, 39, 45, 46, 55, 57, 62, 63, 66, 67, 70, 91, 97, 98, 101, 102, 111, 120, 121, 125, 131, 152, 153]. Figure 2.1 shows the parts of the genome (mainly the heterochromatic fraction) that have not yet been sequenced: the pericentromeric regions, the secondary constrictions of 1q, 9q, 16q, the short arms of acrocentric chromosomes (13p, 14p, 15p, 21p, 22p), and the distal Yq chromosome.

The total number of nucleotides of the finished sequence is 2,858,018,193 while the total estimated length that includes the current gaps is ~3,080,419,480 nucleotides (see Table 2.1, taken from the last hg18 assembly of the human genome http://genome.ucsc.edu/goldenPath/stats.html#hg18). The length of the human chromosomes ranges from ~46 Mb to ~247 Mb. The average GC content of the human genome is 41%. This varies considerably among the different chromosomes and within the different bands of each chromosome. Chromosomal bands positive for Giemsa staining have lower average GC content of 37%, while in Giemsa-negative bands the average GC content is 45%. Interestingly, Giemsa-negative bands are gene-rich regions of DNA (see Chap. 3, Sect. 3.2.4).

Figure 2.2 shows the current status of the “completion” of the human genome sequence [3]. Red bars above the chromosomes represent the sequence gaps. The DNA content of the red blocks (heterochromatin) is still unknown. Heterochromatic regions of chromo-
somes are those that remain highly condensed throughout the cell cycle (see Chap. 3, Sect. 3.2.1); it is thought that transcription is limited in these regions that contain a considerable number of repetitive elements that renders the assembly of their sequence almost impossible.

The sequence of the human genome is freely and publicly available on the following genome browsers, which also contain many additional annotations (see also Chap. 29):

(a) http://genome.ucsc.edu/
(b) http://www.ensembl.org/
(c) http://www.ncbi.nlm.nih.gov/genome/guide/human/

Representative pages of two of these browsers are shown in Fig. 2.3.

There is now a considerable effort internationally to identify all the functional elements of the human genome. A collaborative project called ENCODE (ENcyclopedia Of DNA Elements) is currently in progress with the ambitious objective to identify all functional elements of the human genome [2, 19].

The genome of modern humans, as a result of the evolutionary process, has similarities with the genomes of other species. The order of genomic elements has been conserved in patches within different species such that we could recognize today regions of synteny in different species, i.e., regions that contain orthologous genes and other conserved functional elements. Figure 2.4 shows a synteny map of conserved genomic segments in human and mouse.

The current classification of the functional elements of the genome contains:
1. Protein-coding genes
2. Noncoding, RNA-only genes
3. Regions of transcription regulation
4. Conserved elements not included in the above categories

2.1.1 Functional Elements

2.1.1.1 Protein-Coding Genes

The total number of protein-coding genes is a moving target, since this number depends on the functional annotation of the genome, the comparative analysis with the genomes of other species, and the experimental validation. The so-called CCDS set (consensus coding sequence) is built by consensus among the European Bioinformatics Institute (http://www.ebi.ac.uk/), the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/), the Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/), and the University of California, Santa Cruz (UCSC; http://www.cbse.ucsc.edu/). At the last update (5 July 2009; genome build 36.3) CCDS contains 17,052 genes. This is the minimum set of protein-coding genes included in all genomic databases. The reference sequence (RefSeq) collection of genes of the NCBI contains 20,366 protein-coding gene entries (http://www.ncbi.nlm.nih.gov/
RefSeq/); the UCSC collection of genes contains 23,008 entries (http://genome.ucsc.edu/); the Ensembl browser contains 21,416 entries (23 June 2009; build 36; http://www.ensembl.org/Homo_sapiens/Info/StatsTable). The total number of annotated exons listed in the Ensembl database is 297,252 (23 June 2009; build 36). The discrepancy among the databases reflects the ongoing and unfinished annotation of the genome.

Table 2.2 lists the number of protein-coding and other genes in humans taken from different databases. The human genes are not equally distributed in the chromosomes. In general, Giemsa pale bands are gene rich, and this results in unequal numbers of genes per size unit for the different chromosomes. Figure 2.5 from [84] displays the gene density per megabase for each chromosome and the correlation with CpG-rich islands.

Chromosomes 22, 17, and 19 are unusually gene-rich, while chromosomes 13, 18, and X are relatively gene-poor (interestingly, trisomies for chromosomes 13 and 18 are among the few human trisomies at birth). The average number of exons per gene is nine, and the average exon size is 122 nucleotides. Thus, the total number of annotated exons range from 210,000 to 300,000 (depending on the database), and the total exonic genome size is up to 78 Mb.

The mapping position of the genes can be seen in the genome browsers, and their names can be found in the gene nomenclature Web site, which contains 28,182 entries (http://www.genenames.org/; 30 June 2009).
A single gene may have different isoforms due to alternative splicing of exons, alternative utilization of the first exon, and alternative 5' and 3' untranslated regions. There are on average 1.4–2.3 transcripts per gene according to the different databases (Table 2.2); this is likely an underestimate since, in the pilot ENCODE 1% of the genome that has been extensively studied, there are 5.7 transcripts per gene [19, 61]. The average number of exons per gene, depending on the database, ranges from 7.7 to 10.9.

The size of genes and number of exons vary enormously. The average genomic size of genes (according to the current annotation) is 27 kb. There are, however, small genes that occupy less than 1 kb, and large genes that extend to more than 2,400 kb of genomic space. There are intronless genes (e.g., histones) and others with more than 360 introns (e.g., titin).

The initial results of the ENCODE and other similar projects provided evidence for additional exons to the annotated genes; these exons could be hundreds of kilobases away (usually 5') to the annotated gene elements [19, 40, 44]. In addition, there is evidence for chimeric transcripts that join two “independent” genes [103]. The investigation of these complicated transcripts is ongoing, and the functional significance of them is unknown.

Protein-coding genes can be grouped in families according to their similarity with other genes. These families of genes are the result of the evolutionary processes that shaped up the genomes of the human and other species. The members of the gene families could be organized in a single cluster or multiple clusters, or could be dispersed in the genome. Examples of gene families include the globin, immunoglobulin, histones, and olfactory receptors gene families. Furthermore, genes encode proteins with diverse but recognizable domains. The database Pfam (http://pfam.sanger.ac.uk/, http://www.uniprot.org/) is a comprehensive collection of protein domains and families [48]; the current release
S.E. Antonarakis

Pfam (23.0) contains 10,340 protein families. For example, the WD40 domain family (PF00400) includes 609 human genes, while the homeobox domain family (PF00046) has 430 genes. The identification of domains helps in the prediction of the function and structure of a protein.

Pseudogenes are “dead” nonfunctional genes. These sequences that could be transcribed and spliced contain mutations that render them inactive. Pseudogenes could be generated by several mechanisms that include:

1. Gene duplication events in which one of the duplicated copies accumulates inactivating mutations; alternatively, the duplicated genes may be truncated. These pseudogenes are also called nonprocessed pseudogenes.
2. Transposition events in which a copy of cDNA is reinserted into the genome. These pseudogenes, also called “processed,” are not functional, usually because they lack regulatory elements that promote transcription. In addition, inactivating mutations also occur in processed pseudogenes.

The current estimated number of human pseudogenes (according to one of the databases http://www.pseudogene.org/human/index.php) [151] is 12,534 (~8,000 are processed and ~4,000 duplicated pseudogenes; build 36); while according to the Ensembl browser the number is 9,899 (build 36; 23 June 2009). These pseudogenes belong to 1,790 families; e.g., the immunoglobulin gene family has 1,151 genes and 335 pseudogenes, while the protein kinase gene family has 1,159 genes and 159 pseudogenes (http://pseudofam.pseudogene.org/pages/psfam/overview.jsf).

The total number of human genes is not dramatically different from that of other “less” complex organisms. Figure 2.6 depicts the current estimate of the protein-coding gene number for selected species.

### 2.1.1.2 Noncoding, RNA-Only Genes

Besides the protein-coding genes, there is a growing number of additional genes (transcripts) that produce an

---

**Table 2.2** Human gene, exon, and transcript counts from various databases

<table>
<thead>
<tr>
<th>Database (June 2009)</th>
<th>Protein-coding genes</th>
<th>RNA-only genes</th>
<th>Total genes</th>
<th>Total number of transcripts</th>
<th>Total number of exons</th>
<th>Average exons per gene</th>
<th>Average transcripts per gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCDS</td>
<td>17,052</td>
<td>5,732</td>
<td>22,784</td>
<td>45,428</td>
<td>297,252</td>
<td>10.9</td>
<td>2.3</td>
</tr>
<tr>
<td>Ensembl</td>
<td>21,416</td>
<td>19,155</td>
<td>39,565</td>
<td>62,877</td>
<td>246,775</td>
<td>7.7</td>
<td>2.1</td>
</tr>
<tr>
<td>UCSC</td>
<td>23,008</td>
<td>9,155</td>
<td>32,163</td>
<td>66,802</td>
<td>211,546</td>
<td>9.4</td>
<td>1.4</td>
</tr>
<tr>
<td>RefSeq</td>
<td>20,366</td>
<td>2,044</td>
<td>22,410</td>
<td>31,957</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
RNA that is not translated to protein (see the databases http://biobases.ibch.poznan.pl/ncRNA/, http://www.ncrna.org/frnadb/search.html, http://www.sanger.ac.uk/Software/Rfam/ and [56]). Table 2.2 contains the current number of these genes, which ranges from 2,044 in RefSeq to 9,155 in the UCSC browser.

The different classes of RNA-only genes are briefly discussed below:

Ribosomal RNA (rRNA) Genes [53, 82, 84]: ~650–900. These are genes organized in tandemly arranged clusters in the short arms of the five acrocentric chromosomes (13, 14, 15, 21, and 22). The transcripts for 28 S, 5.8 S, and 18 S rRNAs are included in one transcription unit, repeated 30–50 times per chromosome. These tandemly arranged genes are continuously subjected to concerted evolution, which results in homogeneous sequences due to unequal homologous exchanges. The transcripts for the 5 S rRNAs are also tandemly arranged, and the majority map to chromosome 1qter. There exist also several pseudogenes...
for all classes. The total number of these genes is polymorphic in different individuals. The best estimates of the number of rRNA genes are:

<table>
<thead>
<tr>
<th>Type</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>28 S (components of the large cytoplasmic ribosomal subunit)</td>
<td>~150–200</td>
</tr>
<tr>
<td>5.8 S (components of the large cytoplasmic ribosomal subunit)</td>
<td>~150–200</td>
</tr>
<tr>
<td>5 S (components of the large cytoplasmic ribosomal subunit)</td>
<td>~200–300</td>
</tr>
<tr>
<td>18 S (components of the small cytoplasmic ribosomal subunit)</td>
<td>~150–200</td>
</tr>
</tbody>
</table>

Transfer RNA (tRNA): ~500 (49 Types). At the last count there are 497 transfer RNA genes (usually 74–95 nucleotides long) encoded by the nucleus and transcribed by RNA polymerase III (additional tRNAs are encoded by the mitochondrion genome). There are also 324 tRNA pseudogenes [84]. The tRNA nuclear genes form 49 groups for the 61 different sense codons. Although the tRNA genes are dispersed throughout the genome, more than 50% of these map to either chromosomes 1 or 6; remarkably 25% of tRNAs map to a 4-Mb region of chromosome 6.

Small Nuclear RNA (snRNA) [84, 87, 105]: ~100. These are heterogeneous small RNAs. A notable fraction of these are the spliceosome [139] RNA genes many of which are uridine-rich; the U1 group contains 16 genes, while U2 contains six, U4 4, U6 44, and the other subclasses are represented by one member. Some of these genes are clustered, and there is also a large number of pseudogenes (more than 100 for the U6 class).

Small Nucleolar RNA (snoRNA): ~200. This is a large class of RNA genes that process and modify the tRNAs and snRNAs [135, 147]. There are two main families: C/D box snoRNAs that are involved in specific methylations of other RNAs; and H/ACA snoRNAs, mostly involved in site-specific pseudouridylation. Initially, there were 69 recognized in the first 15 and 15 in the second [84]; however, the total number is probably larger. A cluster of snoRNAs maps to chromosome 15q in the Prader–Willi syndrome region (at least 80 copies); deletions of which are involved in the pathogenesis of this syndrome [26, 117]. Another cluster of snoRNAs maps to chromosome 14q32 (~40 copies). The majority of snoRNAs map to introns of protein-coding genes and can be transcribed by RNA polymerase II or III.

Micro RNAs (miRNA): (706 Entries on 26 June 2009). These are single-stranded RNA molecules of about 21–23 nt in length that regulate the expression of other genes. miRNAs are encoded by RNA genes that are transcribed from DNA but not translated into protein; instead they are processed from primary transcripts known as pri-miRNA to short stem-loop structures called pre-miRNA and finally to functional miRNA. Mature miRNA molecules are complementary to regions in one or more messenger RNA (mRNA) molecules, which they target for degradation. A database of the known and putative miRNAs, and their potential targets, can be found in http://microrna.sanger.ac.uk/. miRNAs have been shown to be involved in human disorders.

Large Intervening Noncoding RNAs (LincRNAs): ~1,600. This new class has been recently identified using trimethylation of Lys4 of histone H3 as a genomic mark to observe RNA PolIII transcripts at their promoter, and trimethylation of Lys36 of histone H3 marks along the length of the transcribed region [95] to identify the spectrum of PolII transcripts. Approximately 1,600 such LincRNA transcripts have been found across four mouse cell types (embryonic stem cells, embryonic fibroblasts, lung fibroblasts, and neural precursor cells) [59]. Among the “exons” of these LincRNAs, approximately half are conserved in mammalian genomes, and are thus present in human. Since this class was described in 2009, further work is needed for its characterization and validation, as well as the potential overlap of its members with the other classes.

Other Noncoding RNAs [7, 75, 113, 126, 136]: ~1,500. The field of noncoding RNA series is constantly expanding. Some of these RNA genes include molecules with known function such as the telomerase RNA, the 7SL signal recognition particle RNA, and the XIST long transcript involved on the X-inactivation [23]. There are also numerous antisense noncoding RNAs, and the current effort to annotate the genome suggests that a substantial fraction of the transcripts are noncoding RNAs.

2.1.1.3 Regions of Transcription Regulation

The genome certainly contains information for the regulation of transcription. The current list of these regulatory elements includes promoters, enhancers, silencers, and locus control regions [92]. These elements are usually found in cis to the transcriptional
unit, but there is growing evidence that there is also trans regulation of transcription. The discovery of the regulatory elements, their functional interrelationship, and their spatiotemporal specificity provides a considerable challenge. A systematic effort during the pilot ENCODE project has provided initial experimental evidence for genomic regions with enriched binding of transcription factors [19, 80, 86, 133]. A total of 1,393 regulatory genomic clusters were, for example, identified in the pilot ENCODE regions; remarkably only ~25% of these map to previously known regulatory regions and only ~60% of these regions overlap with evolutionarily constrained regions. These results suggest that many novel regulatory regions will be recognized in the years to come, and also that there exist regions of transcriptional regulation that are not conserved and thus novel for different clades and species. The use of model organisms facilitates the experimental validation of regulatory elements, and there are systematic efforts underway for the exploration of conserved elements ([106] and http://enhancer.lbl.gov/).

### 2.1.1.4 Conserved Elements Not Included in the Above Categories

Since it is assumed that functional DNA elements are conserved while nonfunctional DNA diverges rapidly, it is expected that all other conserved elements are of interest and should be studied for potential pathogenic variability. How much of the human genome is evolutionarily conserved? The answer to this question depends on the species compared and the time of their common ancestor. Comparative genome analysis between human and mouse, for example, is particularly instructive, since the time of the common ancestor between these two species is estimated to be ~75 million years ago, and thus the conserved elements are likely to be functional. Approximately 5% of the human genome is conserved compared to mouse [145] (and to several other mammalian genomes). Of this, ~1–2% are the coding regions of protein-coding genes, and ~3% are conserved non-coding DNA sequences (CNCs; Fig. 2.7) [41, 42]. The function of the majority of CNCs is unknown. Please note that this 5% conserved fraction between human and mouse is an underestimate of the functional fraction of the human genome, which is likely to be bigger and to contain additional sequences not conserved with the mouse.

The ENCODE pilot project [19, 90], with data from 1% of the human genome and sequences from the orthologous genomic regions from 28 additional species, also estimated that the constrained portion of the human genome is at least ~4.9%; remarkably, 40% of this genomic space is unannotated and thus of unknown function (Fig. 2.8).

---

**Fig. 2.7** The pie-chart depicts the different fractions of the genome. CNCs, conserved noncoding sequences
2.1.2 Repetitive Elements

The function of the majority of the human genome is unknown. Remarkably, ~45% of the genome is composed of repetitive elements, and another ~43% is not conserved and does not belong to the functional categories mentioned above. The different interspersed repeats of the human genome are shown in the Fig. 2.9 (from [84]):

- LINEs (long interspersed nuclear elements [76, 77]) are autonomous transposable elements, mostly truncated nonfunctional insertions (average size of 900 bp). More than 20% of the human genome is polluted by LINEs. Transposable elements are mobile DNA sequences which can migrate to different regions of the genome. Autonomous are those that are capable of transposing by themselves. A small fraction of LINEs (~100) are still capable of transposing. The full LINE element is 6.1 kb long, has an internal PolII promoter, and encodes two open reading frames, an endonuclease, and a reverse transcriptase. Upon insertion a target site duplication of 7–20 bp is formed. There are a few subclasses of LINEs according to their consensus sequence. The subfamily LINE1 is the only one capable of autonomous retrotransposition (copy itself and pasting copies back into the genome in multiple places). These LINEs enable transposition of SINEs (defined below), processed pseudogenes, and retrogenes [76, 77]. LINE retrotransposition has been implicated in human disorders [78]. LINEs are more abundant in G-dark bands of human chromosomes.

- SINEs (short interspersed nuclear elements [18]) mainly include the Alu repeats, which are the most abundant repeats in the human genome, occurring on average in every 3 kb. Thus, 13% of the genome is polluted by Alu sequences and other SINEs. They are inactive elements originated from copies of tRNA or from signal recognition particle (SRP; 7SL) RNA. The full-length element is about 280 nt long and consists of two tandem repeats each ~120 nt followed by polyA.

![Fig. 2.8](image)

**Fig. 2.8** The fractions of different genomic annotations among the 4.9% of constrained sequences in the human genome. Data from the pilot ENCODE project; figure taken from [19]. UTR, untranslated region; other ENCODE experimental annotations refers to the fraction of the genome that has been identified using a variety of experimental techniques for transcription, histone modifications, chromatin structure, sequence specific factors, and DNA replication. More information on these experiments is included in Table 1 of [19].

![Fig. 2.9](image)

**Fig. 2.9** Depicts some basic characteristics of the classes of interspersed repeats in the human genome. For more explanations, see text. (From [84])
Alu sequences are transcriptionally inactive, and are GC-rich. SINEs can retrotranspose in a non-autologous way, since they use the LINE machinery for transposition. Because of their abundance, they could mediate deletion events in the genome that result in human disorders [37]. SINEs are more abundant in G-light bands of human chromosomes (see Sect. 3.2.4).

- Retrovirus like (LTR transposons) are elements flanked by long terminal repeats. Those that contain all the essential genes are theoretically capable of transposition, but that has not happened in the last several million years. Collectively they account for 8% of the genome. Most are known as HERV (human endogenous retroviral sequences) and are transposition defective. Transcription from the HERV genes may modulate the transcriptional activity of nearby protein-coding genes [22].

- DNA transposon fossils [127] have terminal inverted repeats and are no longer active; they include two main families, MER1 and MER2, and comprise 3% of the genome.

More update information about repeats can be found in http://www.girinst.org/server/RepBase/.

### 2.1.2.1 Segmental Duplications

Approximately 5.2% of the human genome consists of segmental duplications or duplicons, i.e., regions of more than 1 kb, with greater than 90% identity, that are present more than once in the genome. Segmental duplications are either intrachromosomal (on the same chromosome, 3.9%), or interchromosomal (on different chromosomes, 2.3%; Fig. 2.10). Most of the “duplicons” are in the pericentromeric regions.

Figure 2.11 shows the distribution of intrachromosomal duplicons in the human genome [16, 118]. These duplications are important in evolution and as risk factors for genomic rearrangements that cause human disorders because of unequal crossing-over in meiosis (pathogenic microdeletions and microduplications). Some examples of these include cases of α-thalassemia [65] on chromosome 16p, Charcot–Marie–Tooth syndrome [104] on chromosome 17p, and velo-cardiac-facial syndrome [96] on chromosome 22q. Williams–Beuren syndrome [107] on chromosome 7q, and Smith–Magenis syndrome [29] on chromosome 17p.

### 2.1.2.2 Special Genomic Structures Containing Selected Repeats

#### 2.1.2.2.1 Human Centromeres

Human centromeres consist of hundreds of kilobases of repetitive DNA, some chromosome specific and some nonspecific [114, 122, 124]. Actually, most of the remaining sequence gaps in the human genome are mapped near and around centromeres. The structure of human centromeres is unknown, but the major repeat component of human centromeric DNA is an α-satellite or alphoid sequence [30] (a tandem repeat unit of 171 bp that contains binding sites for CENP-B, a centromeric-binding protein; see also Chap. 3, Sect. 3.2.3). Figure 2.12 shows an example of the structure of two human centromeres [3].

#### 2.1.2.2.2 Human Telomeres

Human telomeres [109] consist of tandem repeats of a sequence (TTAGGG)$_n$ that spans about 3–20 kb, beyond which at the centromeric side there are about 100–300 kb of subtelomeric-associated repeats [3] before any unique sequence is present.
Fig. 2.13 schematically shows the sequence organization of six human subtelomeric regions.

2.1.2.2.3 Short Arms of Human Acrocentric Chromosomes

The finished sequence of the human genome does not include the short arms of acrocentric chromosomes (13p, 14p, 15p, 21p, and 22p). Cytogenetic data show that the p arms contain large heterochromatic regions of polymorphic length [35, 138]. Molecular analysis revealed that they are composed mainly of satellite and other repeat families, including satellites I (AT-rich repeat of a monomer of 25–48 bp [73]), II (monomer repeat 5 bp [68]), III (monomer repeat also 5 bp [31]), β-satellite (a tandem repeat unit of 68 bp of the Sau3A family [94, 146]), and repeats ChAB4 [36], 724 [83], and D4Z4-like [89]. These repeats have a complex pattern and are often organized in subfamilies shared in addition, several other repetitive elements border the alphoid sequences. The length of these regions is also polymorphic in different individuals.

Fig. 2.11 Schematic representation of the intrachromosomal segmental duplications (from [16]). In each chromosome a blue line links a duplication pair. For example, on chromosome 21 there is only one duplicon shown; in contrast, on chromosome 22 there is a considerable number of duplications. Richly blue areas are considered susceptible to microduplication/microdeletion syndromes.
between different acrocentric chromosomes. The p arms encode the ribosomal (RNR) gene [53, 82] but may also encode other genes [88, 130]. Currently there is an initiative to sequence the short arm of chromosome 21 and thus extrapolate on the structure of the additional p arms of the other acrocentrics [88].

The most common chromosomal rearrangements in humans are Robertsonian translocations (~1 in 1,000 births), which involve exchanges between acrocentric p arms. Three to five percent of these translocations are associated with phenotypic abnormalities [143].

### 2.1.3 Mitochondrial Genome

In human cells there is also the mitochondrial genome, which is 16,568 nucleotides long and encodes for 13 protein-coding genes, 22 tRNAs, one 23 S rRNA, and one 16 S rRNA ([140–142]; http://www.mitomap.org). The mitochondria genome-encoded genes are all essential for oxidative phosphorylation and energy generation in the cell. Each cell has hundreds of mitochondria and thousands (10^7–10^9) of mitochondria DNA (mtDNA) copies. Human mtDNA has a mutation rate ~20 times higher than nuclear DNA. The inheritance of mtDNA is exclusively maternal (the oocyte contains 10^5 mtDNA copies). Several human phenotypes are due to pathogenic mutations in the mitochondrial genome [140] (Fig. 2.14).

### 2.2 Genomic Variability

The human genome is polymorphic, i.e., there are many DNA sequence variants among different individuals. These variants are the molecular basis of the genetic individuality of each member of our species. In addition, this genetic variability is the molecular substrate of the evolutionary process. Finally, this variability causes disease phenotypes or predispositions to common complex or multifactorial phenotypes and traits.

#### 2.2.1 Single Nucleotide Polymorphisms

The majority of the DNA variants are single nucleotide substitutions commonly known as SNPs (single nucleotide polymorphisms). The first SNPs were identified in 1978 in the laboratory of Y.W. Kan 3’ to the β-globin
gene [74] (at the time, these DNA polymorphisms were detected by restriction endonuclease digestion of DNAs and were called RFLPs, restriction fragment length polymorphisms). These polymorphic sites have two alternative alleles. In the example shown in Fig. 2.15, the depicted SNP has two alleles in the population: the blue C allele and the red T allele. The frequency of each allele could vary in different populations.

There is on average one SNP in ~1,000 nucleotides between two randomly chosen chromosomes in the population. Many of these SNPs are quite common. A common SNP is that in which the minor allele frequency (MAF) is more than 5%. On average two haploid genomes differ in ~3,000,000 SNPs. In addition, there is a large number of rare (MAF < 1%) or near-rare (MAF between 1% and 5%) SNP variants that could be identified by the genome sequencing of various individuals. The majority of heterozygous SNPs in the DNA of a given individual are relatively common in the population; on the other hand, most of the SNPs discovered in a population are more likely to be rare. The NCBI SNP database contains 25 million common and rare SNPs (http://www.ncbi.nlm.nih.gov/}

---

Fig. 2.14 Schematic representation of the circular mtDNA, its genes, its clinical relevant mutations, and certain polymorphic markers. Letters within the ring depict the genes encoded. Letters on the outside indicate amino acids of the tRNA genes. CR, the control of replication region that contains promoters for the heavy and light strands. Arrows outside show the location of pathogenic mutations. (From [142])

---

Fig. 2.15 Schematic representation of a single nucleotide polymorphism. Allele 1 has a C in the sequence, while allele 2 contains a T in the same position.
SNP/snp_summary.cgi; version 130; July 2009; Fig. 2.16). Of those, ~301,000 are in the protein-coding regions of genes, and ~188,000 result in amino acid substitutions (nonsynonymous substitutions). An international project known as HapMap (http://www.hapmap.org/) [6, 34, 50] has completed the genotyping of ~4,000,000 common SNPs in individuals of different geo-ethnic origins (4,030,774 SNPs in 140 Europeans; 3,984,356 in 60 Yoruba Africans; 4,052,423 in 45 Japanese and 45 Chinese; http://www.hapmap.org/downloads/index.html.en). Additional samples from further populations have been added recently.

The information content of SNPs (and polymorphic variation in general) is usually measured by the number of heterozygotes in the population (homozygotes are individuals that contain the same variant in both alleles; heterozygotes are individuals that contain two different variants in their alleles). The number of heterozygotes is a function of MAF based on Hardy-Weinberg principles (see Chap. 10). The pattern of DNA polymorphisms in a single chromosome is called haplotype (a contraction of “haploid genotype”; allelic composition of an individual chromosome). In the example shown in Fig. 2.17 the haplotype of polymorphic sites for the paternal (blue) chromosome is CGAATC while for the maternally inherited red chromosome it is GACGAT.

SNPs are shown as letters interrupting the lines. The haplotype is defined as the combination of SNP alleles per haploid genome.

**Fig. 2.16** The genomic region of Chr11: 5,194,075–5,214,074 is shown. For each of the nine SNPs shown in the bottom, the frequency of the two alternative alleles is shown in different populations. For example, for SNP rs11036364 that maps between the HBB and HBD globin genes, the allele frequencies are shown in the callout. The four original populations of the HapMap project were EUR, YRI, JPT, and CHB, while the other populations were added in a later stage. Modified from http://www.hapmap.org/

**Fig. 2.17** Schematic representation of haplotype of polymorphic variants in a segment of the genome. The parental origin is shown as the blue (paternally-inherited) and red (maternally-inherited) lines. SNPs are shown as letters interrupting the lines. The haplotype is defined as the combination of SNP alleles per haploid genome.
2.2.2 Short Sequence Repeats

Short sequence repeats (SSRs) are polymorphic variations due to a different number of short sequence repeat units, first described by Wyman and White [150] (then called VNTRs, variable number of tandem repeats[99]), and further elucidated by Jeffreys [72]. Most common are the dinucleotide repeats (described after the introduction of polymerase chain reaction amplification), but SSRs could be tri-, tetra-, or penta- repeats (often called microsatellites where the repeat unit \( n = 1–15 \) nucleotides). SSRs with longer repeat units (\( n = 15–500 \) nucleotides) are often termed minisatellites. These sequences comprise ~3% of the genome and there is ~1 SSR per 5 kb [84]. The most frequent dinucleotide SSR is the \((GT)_{n}\) with an occurrence in the genome of ~28 times per megabase, followed by the \((AT)_{n}\) SSR with ~19 times per megabase. The most common trinucleotide SSR is the \((TAA)_{n}\) that occurs approximately four times per megabase. The major advantage of SSRs (or microsatellites) is that there are more than two alleles per polymorphic site, and a large fraction of the human population is heterozygous for each SSR. Therefore, SSRs are extremely useful in linkage mapping and subsequent positional cloning for monogenic disorders [12, 17, 33] and other marking studies of the genome including the development of genomic linkage maps [43, 144]. In addition, SSRs are extensively used in forensic studies [15]. Figure 2.18 shows an example of an SSR with three alleles in the population.

2.2.3 Insertion/Deletion Polymorphisms (Indels)

This variation is due to the presence or absence of certain sequences. These sequences could be a few nucleotides, but they could also be transposons or interspersed repeats such as LINE or SINE elements [18, 112, 149]; alternatively, they could be pseudogenes [8] or other elements. Note that this category of variants is not completely separate from the next one; the arbitrary distinction is just the size of the variation in terms of base pairs. There are usually biallelic polymorphisms, which are not as common as SNPs but are useful for evolutionary studies and for the understanding of the dynamic structure of the human genome. In the example shown in Fig. 2.19, the blue sequence was inserted in the DNA and created a variant with two alleles: the blue allele 1 with insertion and the black allele 2 without.

---

Fig. 2.18 An example of a dinucleotide SSR with three alleles in the population: the blue allele with \((CA)_{13}\), repeats, the red allele with \((CA)_{16}\), and the green allele with \((CA)_{7}\).
Copy number variant (CNV) refers to large-scale structural variation of our genome in which there are large tandem repeats of 50 kb to 5 Mb long that are present in a variable number of copies. This type of polymorphic variant includes large-scale duplications and deletions [123] (see also Chap. 3, Sect. 3.4.4). These have been known since studies of the α-globin genes in humans [54]. In the Fig. 2.20 example, allele 1 contains three copies and allele 2 five copies of a large repeat. The phenotypic consequences of some of these variants that may contain entire genes is unknown. A CNV map of the human genome in 270 individuals has revealed a total of 1,440 such CNV regions which cover some 360 Mb (~12% of the genome [79, 108]). More recent estimates using more accurate methods for precise mapping of the size of CNVs suggest that ~6% of the genome contains CNVs. A list of these variants can be found at http://projects.tcag.ca/variation/. The extent of CNV in the human genome is certainly underestimated since there are numerous additional CNVs of less than 50 kb. The current methodology for the detection of CNVs is using comparative genomic hybridization (CGH) on DNA microarrays [25]. A further improvement of this method will allow us to detect small CNVs. The most detailed currently available CNV map of the human genome was recently established by the Genome Structural Variation Consortium. This consortium conducted a CNV project to identify common CNVs greater than 500 bp in size in 20 female CEU (European ancestry) and 20 female YRI (African ancestry) samples of the HapMap project. By employing CGH arrays that tile across the assayable portion of the genome with ~42 million probes from the company NimbleGen, this consortium could map 8,599 copy number variant events. Parts of these data have been provisionally released to the scientific community and can be viewed at http://www.sanger.ac.uk/humgen/cnv/42mio/.

### 2.2.5 Inversions

Large DNA segments could have different orientation in the genomes of different individuals. These inversion polymorphisms (Fig. 2.21) predispose for additional genomic alterations [9]. An example of a common inversion polymorphism involves a 900-kb segment of chromosome 17q21.31, which is present in 20% of European alleles but is almost absent or very rare in other populations [129]. These variants are difficult to identify and most of them have been detected by sequencing the ends of specific DNA fragments and comparing them with the reference sequence [79, 134].

### 2.2.6 Mixed Polymorphisms

There are combinations of repeat size variants and single nucleotide variants. Figure 2.22 depicts such an example; the repeat units of an SSR contain a SNP and, thus, even alleles with the same repeat number...
could be distinguished based on their exact DNA sequence [71]. These highly polymorphic systems could serve as “recognition barcodes” in humans.

2.2.7 Genome Variation as a Laboratory Tool to Understand the Genome

DNA variants, besides their functional importance in health and disease, are very useful in human genetics research because they serve as genomic markers for a variety of studies. Some of the uses of DNA variants are to:

1. Create linkage (genetic) maps of human chromosomes [1, 148]. This has allowed the initial mapping of the human genome and it was a prerequisite for the sequence assembly.
2. Map the genomic location of monogenic phenotypes to human chromosomes by linkage analysis [58, 81]. A large number of such phenotypes have been mapped to small genomic intervals because of the genotyping of members of affected families. Positional cloning of pathogenic mutations was subsequently possible.
3. Map the genomic location of polygenic phenotypes to human chromosomes by genomewide linkage and association studies [4, 20, 119].
4. Allow fetal diagnosis and carrier testing by linkage analysis of the cosegregation of a polymorphic marker and the phenotype of interest [10, 21].
5. Perform paternity and forensic studies [52]. A whole field was developed mainly with the use of microsatellite SSR variants [49, 51].
6. Study genome evolution and origin of pathogenic mutations [115, 116].
7. Study the recombination rate and properties of the human genome [28, 93].
8. Study the instability of the genome in tumor tissues [5].
9. Identify loss-of-heterozygosity in human tumors [27, 47].
10. Study uniparental disomy and thus help with understanding genomic imprinting [100, 128].
11. Study parental and meiotic origin, and decipher the mechanisms of nondisjunction [11, 13, 14].
12. Study population history and substructure [110, 132].

The chapters that follow include further discussions on different aspects (including evolution, phenotypic consequences, and disease susceptibility) related to the most precious human genome variability.

Acknowledgments I thank the members of the laboratory, past and present, for discussions, ideas, debates, and data. I also thank my mentors and my students for the learning process.

References

4. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14, 000 cases of seven
common diseases and 3,000 shared controls. Nature 447:661–678
sites and discovery of a large number of additional exons in ENCODE regions. Genome Res 17:746–759
106. Roy-Engel AM, Carroll ML, Vogel E, Garber RK, Nguyen SV et al (2001) Alu insertion polymorphisms for the...
study of human genomic diversity. Genetics 159: 279–290


factor VIII gene and further evidence of open reading frames in functional L1 elements. Genomics 4:290–296