Chapter 4

The Human Genome Project

P. Gross & T. Oelgeschläger
Eukaryotic Gene Regulation Laboratory, Marie Curie Research Institute, UK.

Abstract

The Human Genome Project, the most ambitious biological research project to date, was inaugurated in the 1980s with the aim to decipher the precise DNA sequence of the entire human genetic material and culminated in the publication of two human genome sequence drafts in February 2001. These drafts represent a major milestone in biological research as they provide the first panoramic view of the genomic landscape of a vertebrate. Among the results of the sequencing efforts was the surprising finding that only 1.4% of the human DNA encodes instructions for the assembly of proteins. The estimated total number of just over 30,000 human protein-coding genes is insufficient to explain human complexity compared to other organisms simply on the basis of gene number. Additional levels of complexity must exist, both in the coordinated temporal and spatial read-out of genetic information and in the functional interplay of expressed gene products. This insight has ushered in functional genomics, the genome-wide analysis of cell-specific gene expression patterns and protein interaction networks. The human genome sequence project has led to the emergence of novel technologies that are expected to promote research into virtually all aspects of life science. Medical sciences are expected to benefit in particular, as the availability of human genome sequence information has paved the way for the development of novel diagnostics and advanced therapeutics. These may ultimately provide the means for the prevention and targeted treatment of complex genetic disorders such as cancer.

1 Introduction

1.1 Genes

The elucidation of the precise mechanisms underlying heredity, the accurate transfer of biological information from one generation to the next, has been a key problem in biological sciences over several centuries. Gregor Mendel, an Austrian monk, concluded in 1865 from genetic crossing experiments that hereditary information is passed from parents to offspring in discrete packets [1]. Later on, these units of heredity were called genes. The biological nature of genes was not revealed until 1944, when it was demonstrated that genetic information is carried by a biological substance

WIT Transactions on State of the Art in Science and Engineering, Vol 27, © 2006 WIT Press
www.witpress.com, ISSN 1755-8336 (on-line)
doi:10.2495/978-1-85312-853-0/04

with the chemical designation deoxyribonucleic acid (DNA) [2]. DNA is a polymer composed of deoxyribonucleotides, each of which typically contains one of four different bases: adenine (A), guanine (G), thymine (T), or cytosine (C). The discovery of the double-helical structure of DNA by James Watson and Francis Crick in 1953, certainly one of the most significant scientific discoveries of the 20th century, provided the first clues on how DNA can be accurately duplicated in order to transmit genomic information from one generation to the next. Within the DNA double helix structure, two DNA strands closely interact via hydrogen bonds between their nucleotide bases to form specific inter-strand base pairs: adenine pairs with thymine and guanine pairs with cytosine [3]. Hence, the two strands of the DNA double helix are complementary to each other and can both serve as templates during replication, the process by which cells copy their genetic information before cell division. Complete replication of a DNA molecule results in two identical DNA duplexes, each consisting of one parental and one newly synthesised DNA strand [4–6].

Genetic information is encoded within the nucleotide sequence of the DNA strands. The read-out of genetic information occurs through a process called transcription, the synthesis of RNA, a single-stranded nucleic acid consisting of ribonucleotides, from a DNA template. Transcription is catalysed by RNA polymerases and results in a RNA molecule with the identical nucleotide sequence as the coding DNA strand of a gene. This RNA transcript can either have biological activity by itself or, as in the case of protein-coding genes, can be used as a template to synthesise a specific polypeptide chain from different amino acid components.

RNA molecules that serve as templates for protein synthesis are called messenger RNAs (mRNAs). The genetic code that is used to translate mRNAs into a protein polypeptide chain with specific amino acid sequence had been mainly worked out by 1964 [5–7]. A triplet of nucleotide bases, a codon, encodes one specific amino acid and the sequence of codons in mRNA is co-linear to the amino acid sequence of the resulting protein. There are $4^3 = 64$ possible codons, 61 of which encode the 20 naturally occurring amino acids in humans. The remaining three codons are used as signals that cause the termination of protein synthesis [4–6].

In 1977, it was discovered that genes could be much longer than would be predicted from the amino acid sequence of their protein products [8, 9]. Additional DNA sequences were found that that were not present in the corresponding mRNA used for protein synthesis, and these sequences were interspersed with coding sequences. Coding regions of a gene present in mRNA were subsequently called exons, and the non-coding regions were called introns. Transcription of a protein-coding gene consisting of exons and introns gives rise to a primary RNA molecule (pre-mRNA) from which the intron regions have to be removed to yield mature mRNA that then can be used for protein synthesis. This process of RNA splicing involves the deletion of intron sequences followed by the precise joining of exons so that the co-linearity of gene and protein is maintained between the individual exons and the corresponding parts of the protein chain. In the majority of human protein-encoding genes, exons are usually shorter than introns and the number of introns per gene can be higher than 10. Depending on their exon/intron structure protein-coding genes can vary considerably in length, ranging from 1 kb to several millions of base pairs (Mb) [5, 6, 10].

Gene expression (i.e. the transcription of a gene and its translation into a specific protein molecule) has to be strictly regulated to ensure the proper functioning of a cell. In multicellular organisms (metazoans), only a subset of the total genetic material is actively expressed at a given time in a given cell type. Perturbations in the coordinated expression of genes in only a single cell can have disastrous consequences for the whole organism, it can, for example, result in the development of cancer. Gene expression is primarily regulated at the level of transcription. The information for the precise control of transcription is encoded in specific DNA sequence elements called gene promoters. Core promoter elements serve to assemble the transcription machinery

and define the start site of transcription. The activity of core promoter elements is controlled by regulatory DNA sequences that are typically recognized by DNA-binding gene- and cell type-specific transcription activators or repressors. Regulatory promoter regions vary considerably in length and typically contain binding sites for several regulatory proteins. They can be found either close to the start site of transcription of a given gene (basal promoter elements) or at a distance of several thousand base pairs (kb) (enhancers) [5, 6].

1.2 Genome organisation

A genome is the entirety of all DNA within an organism. In addition to genomic DNA present in the nucleus, there is also DNA present in the mitochondria, semi-autonomous organelles found in most eukaryotic cells that serve as energy factories. The mitochondrial genome contributes to less than 1% of the total cellular DNA in mammalians and will not be discussed in this article.

In the nuclei of eukaryotic cells, DNA is organised in a set of chromosomes. Human cells contain 22 pairs of autosomes; one chromosome of each pair is inherited from the mother and the other from the father. Since cells contain two copies of each autosome, they also contain two copies of each gene, the so-called alleles. In addition to the autosomes, there are the sex chromosomes X and Y. Females possess two X chromosomes, one from each parent, whereas males possess an X chromosome inherited from their mother and a Y chromosome inherited from their father.

The DNA within the 23 human chromosomes contains a total of 3.2×10^9 nucleotide bases (3.2 Gb) and, if completely stretched, would be approximately 2 m long. To fit the genomic DNA into the cell nucleus, which is only a few micrometres in diameter, an enormous level of compaction has to be achieved. Chromosomes represent the most compact form of nuclear DNA and are only observed during cell division. In non-dividing cells, so-called interphase cells, the DNA material occupies the cell nucleus as chromatin without distinguishable chromosomes. In chromatin, DNA is organised into nucleosomes; $\sim$200 bp of DNA are wrapped around a protein disc containing a histone protein octamer [11]. Nucleosomes are arranged like beads on a string to form a 10 nm chromatin fibre that can be further compacted to a 30 nm fibre by incorporating a specific linker histone protein, H1 [5, 6]. As a 30 nm fibre, a human chromosome would still span the nucleus more than a 100 times. The mechanism(s) underlying further condensation of the 30 nm fibre into higher order structures such as chromosomes are not yet fully understood.

Two major types of chromatin can be distinguished based on the level of DNA condensation. Euchromatin occupies most of the nucleus and the underlying DNA fibres are much less densely packed as compared to heterochromatin, which exhibits a level of DNA compaction comparable to chromosomes. Of the 3.2 Gb of the human genome, 2.95 Gb or 92% are euchromatic and only 0.35 Gb or 8% are heterochromatic [5].

1.3 Genome contents

Much progress has been made in understanding the molecular structure of DNA and its organisation into chromatin. However, the function of the plethora of genomic DNA sequences remains poorly understood. Evidently, the pivotal function of DNA sequences in genes is to provide a blueprint for biologically active RNAs and proteins required for cell growth, differentiation, and development of multicellular organisms. However, the majority of the human genome (53%) consists of repeated DNA sequences of various types that do not confer essential cellular functions. These are sometimes referred to as 'junk' sequences although this implicative negative role is not justifiable a priori.

The largest proportion of repeated sequences, about 45% of the human genome, are parasitic DNA sequences, which can provide valuable clues about evolutionary events and forces [10, 12]. These DNA elements are discussed in detail in Chapter 3.1. Segmental duplications of 10 to 300 kb that have been copied from one region of the genome to another contribute to about 5% of the human genome [10, 12]. Blocks of tandemly repeated sequences are found in centromeres, constricted regions of the chromosome that include the site of attachment of the mitotic spindle, and in telomeres, the DNA regions at the chromosome ends. A different class of repeated DNA elements present in the human genome contains only small stretches of repeated DNA. Depending on the size of the repeat unit these sequence elements are either called microsatellites (1–13 bp) or minisatellites (14–500 bp) [5]. Satellite sequences have been extremely useful in human genetic studies, as there is considerable variation between individuals. The precise mapping of satellite positions on individual chromosomes has provided a comprehensive catalogue of gene markers [13].

Only a very small proportion of the human genome (~1.4%) contains protein-coding genes [10, 12]. However, the simple rule 'one gene one protein' does not always apply and the total number of distinct cellular proteins is significantly greater than the total number of protein-encoding genes. One mechanism by which a single gene can give rise to several distinct protein products is alternative splicing, which involves the differential assignment or usage of intros and exons within pre-mRNAs [5, 6]. Alternative splicing has been found to be more pronounced in humans than in any other species and it has been estimated that two to three alternative splicing products exist for each gene [14, 15]. Another process by which the informational content of RNA can be altered is RNA editing, which involves the introduction of changes at individual nucleotide positions or the addition of nucleotide bases within a mRNA [6]. Furthermore, protein synthesis can start or terminate at different positions within the mature mRNA, giving rise to protein products that are identical in amino acid sequence but different in length and therefore potentially different in function. In addition, proteins are subject to post-translational modifications that add to the complexity of their role and function within the cell. These include chemical modifications (e.g. phosphorylation, acetylation, methylation) or conjugation to other proteins such as ubiquitin or ubiquitin-related proteins [5, 6, 16, 17].

Finally, the human genome contains a class of genes that, although primary transcripts can be found in some cases, do not give rise to products with cellular function. These pseudogenes contain DNA sequences typical for functional genes (i.e. promoter and coding regions) but are rendered inactive by mutations in the DNA sequence that affect transcription, splicing, or translation. Most pseudogenes contain deletions of one or several DNA segments. Since there is no further selection against the accumulation of additional mutations once gene expression is abolished, the time of inactivation can be estimated by comparing the DNA sequences of pseudogenes with those of the original genes. Pseudogenes have been found to be several ten million years old [5].

2 The Human Genome Project

2.1 History of the Human Genome Project

The Human Genome Project was first proposed in the 1980s with the aim to decipher the entire human genetic material and to identify the complete set of human genes [18]. It was evident from the start that a project of this scale would require a communal effort in infrastructure building unlike any other previously attempted biomedical enterprise. The first official programme to sequence the human genome was announced in April 1990 as a joint effort of the Department of Energy and the National Institutes of Health in the US. This programme featured a broad approach that included

the construction of human genetic maps, which identify the relative position of particular genes on a chromosome and which provide starting points for the assembly of genomic DNA sequences. In parallel, efforts were directed to sequence key model organisms such as bacteria, worm, fly, and mouse. Furthermore, research into the ethical, legal and social issues raised by human genome research was intended.

In October 1990, the Human Genome Project was officially launched and genomic centres in the participating countries US, UK, France, Germany, Japan, and China, were created. In addition, the Human Genome Organisation was founded to provide a forum for the coordination of international genomic research. The main sequencing centre in the US was established at the Whitehead Institute in Cambridge, Massachusetts. In the UK, the Wellcome Trust and the MRC opened the Sanger Centre close to Cambridge where later one third of the entire sequencing effort would be taken on. Rapid progress was made. The first human genetic maps were developed and refined [19, 20] and genome sequencing of the first free-living organism, the bacterium *Haemophilus influenzae*, was completed in 1995 [21]. In the following years the genome sequences of two key model organisms in biological research, the yeast *Saccharomyces cerevisiae* [22] and the worm *Caenorhabditis elegans* [23], were released.

In 1998 the international collaboration fell apart over disputes regarding the strategic approach for sequencing the human genome. Craig Venter at the Institute for Genomic Research in Rockville, US, argued that the progress on the human genome could be accelerated considerably by using whole-genome shotgun sequencing. He and his colleagues had successfully used whole-genome shotgun sequencing to determine the *H. influenzae* genome sequence in record time [21]. However, Francis Collins, head of the National Human Genomic Research Institute, insisted on a more methodical and conservative sequencing strategy which he considered to give the highest possible quality of sequence data. Craig Venter finally left the publicly funded consortium to set up a biotechnology company, Celera Genomics. He announced the complete sequencing of the human genome by 2003, 2 years earlier than the completion date projected by the Human Genome Project. While Celera Genomics promised free access to raw DNA sequence data, they proposed to perform analyses of their genome sequence database on a commercial basis. This announcement created an uproar amongst the scientific community who argued that human genome sequence information was fundamental and that it should be freely accessible. Additional concerns regarded the future of the public programme after the long and difficult groundwork that had been done [18]. In response, the leaders of the public consortium announced new goals for the public project in order to beat Celera Genomics to the finish line. The pace of sequencing was to be increased to produce a first 'rough' draft covering ~90% of the human genome by spring 2001. This was the starting point for a race between the two groups that was punctuated by duelling press releases of respective milestones over the following two years [24]. Finally, the international consortium (Human Genome Project) and Celera Genomics jointly announced working drafts of the human genome sequence in a ceremony at the White House in Washington on 26 June 2000. A simultaneous publication of the results, however, collapsed over controversies regarding the amount of data Celera Genomics was willing to publicise and two separate reports on the human genome sequence were released. The publicly funded Human Genome Project published in the British journal *Nature* on 15 February 2001 [10], and Celera Genomics published in the American journal *Science* on 16 February 2001 [25].

2.2 Strategy of the Human Genome Project

The basis of any effort to decipher genomic information is the ability to determine the exact nucleotide sequence of any given DNA segment. Two research groups led by Frederick Sanger

at MRC Laboratory of Molecular Biology in Cambridge, UK [26], and by Walter Gilbert at Harvard University in Boston, US [27], independently published different strategies to sequence DNA segments with high accuracy in 1977. Sanger and Gilbert were awarded the Nobel Price for this groundbreaking work in 1980. Until the mid-80s DNA sequencing technologies were not significantly advanced and even state-of-the-art laboratories could only sequence around 500 nucleotide bases in one experiment. From a technical point of view, sequencing an entire genome seemed a virtually impossible task. This limitation was overcome with the development of automated sequencing machines, first introduced in 1986, that made daily outputs of over one million bases (1 Mb) possible [28].

Very large DNA molecules such as chromosomes cannot be sequenced directly. The DNA material needs to be fragmented into smaller and more manageable pieces that, after sequence determination, have to be re-assembled in the correct order and orientation *in silico*. This approach, called shotgun sequencing, was first introduced in 1981 and later more refined and extended to increase efficiency and accuracy [29]. Two different shotgun sequencing strategies were employed to decipher the human genome: the public consortium (Human Genome Project) used 'hierarchical shotgun sequencing' [10], whereas Celera Genomics utilised 'whole-genome shotgun sequencing' [21].

In hierarchical shotgun sequencing the target genome is first fragmented into pieces of about 100–200 kb in length. These DNA fragments are then randomly inserted into bacteria using bacterial artificial chromosome (BACs) vectors [10]. Bacteria populations carrying only one specific BAC construct, so-called clones, are isolated and propagated to multiply the DNA inserts. In order to ensure a complete representation of the entire genome in the resulting BAC library a very large number of individual BAC clones has to be isolated; the BAC library that was constructed to sequence the human genome consisted of more than 1.5 million different BAC clones. Next, BAC clones with overlapping sequences are identified and the order of their DNA inserts in the target genome is determined. An assembly of BAC clones that covers the entire genome is selected and sequenced. Around 30,000 BAC clones were sequenced by the publicly funded Human Genome Project. It is important to note that DNA fragments in BAC clones are still far too large to be sequenced directly. They are fragmented randomly into small pieces that can be sequenced with high accuracy. The DNA sequences obtained for random DNA fragments are first aligned to reconstruct the sequence of the source BAC clone before the whole sequence of the target genome is assembled.

Whole-genome shotgun sequencing circumvents the construction of a BAC library. The entire target genome is randomly fragmented into DNA pieces between 2 and 50 kb in length, which are sequenced directly [25]. Given the size of the human genome of 3.2 Gb, one can easily apprehend the enormous collection of single fragments required to warrant a complete coverage of the genome. In contrast to the hierarchical shotgun strategy, there is no initial information about the genomic location of individual DNA sequences. Therefore, correct alignment of the vast amounts of DNA fragments is the pivotal point of the whole-genome shotgun sequencing approach. Whole-genome shotgun sequencing assembly can be compared to an immense jigsaw puzzle with millions of pieces that need to be arranged in the right order and orientation. Powerful computers are employed to search for overlapping DNA sequences, to link DNA fragments of known sequence in the right order and orientation, and to compare the resulting sequence assemblies with known sequences or genetic markers in human genetic maps.

Both sequencing strategies described above have limitations. The major problem in deciphering the human genome is its high content of repetitive DNA sequences. This is in stark contrast to the genomes of simpler organisms that had been sequenced previously. Because repetitive DNA sequences do not occupy specific locations within the genome their positioning relative

to unique sequences is very difficult. This problem is more pronounced with the whole-genome sequencing strategy because the genomic origin of small DNA fragments containing repetitive sequences is not known. In the hierarchical shotgun strategy, repetitive DNA sequences can be traced back to their source BAC clones, for which additional information about their genomic vicinity is available. The major problem that resides in the hierarchical shotgun sequencing technique concerns the construction of the BAC library. Some regions of the genome are difficult to clone and will therefore be underrepresented. These cloning biases are difficult to overcome [10]. Finally, both sequencing strategies rely on computational methods to assemble the genome from sequenced DNA fragments. Complex mathematical algorithms have been developed but the computer programs employed for sequence assembly are far from being without errors. During sequence assembly, *in silico* DNA fragments might be misplaced or their orientation might be inverted. Some regions of the genome are less well resolved as others, leaving gaps in the genome sequence. Additional data have to be obtained to confirm these preliminary genomic DNA sequence assemblies, to define problematic regions of the genome, and to close the remaining sequence gaps.

3 The human genome sequence draft

The recently published human genome sequence drafts contain a number of gaps and uncertainties but have, despite their partial preliminary nature, already provided a number of important insights. The competing teams sequenced a comparable number of DNA nucleotide bases, the Human Genome Project sequenced 2.7×10^9 bases, and Celera Genomics sequenced 2.9×10^9 bases. The resulting human genome sequence drafts cover ~90% of the euchromatic portion of the genome and indicate a very similar genome composition. Celera Genomics reported that only 1.1% of the genome consists of exons and that 24% of the genome are introns. The remaining 74% of the genome are intergenic [25]. The Human Genome Project determined a protein-coding content of the genome of 1.1–1.4%. As illustrated in Fig. 1 more than half of the human genome was found to consist of different types of repeated sequence elements [10]. Further sequencing efforts by both the Human Genome Project consortium and Celera Genomics are ongoing and

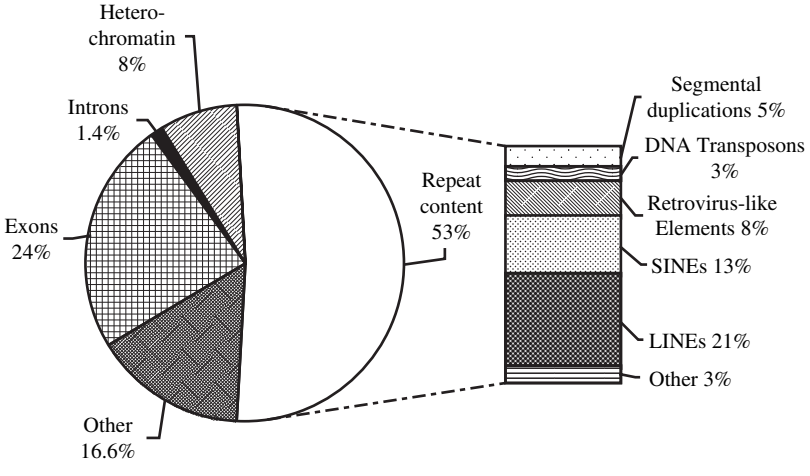

Figure 1: Human genome content.

Table 1: A selection of publicly accessible sequence databases on the World Wide Web.

http://www.ensembl.org	EBI/Sanger Centre; access to DNA and protein sequences
http://www.ncbi.nlm.nih.gov	NCBI; views of chromosomes and maps and loci with links to other NCBI resources
http://www.celera.com	Celera Genomics; central site for public access to data and tools
http://genome.ucsc.edu/	University of California at Santa Cruz; assembly of the draft genome sequence and updates
http:/genome/wustl.edu/gsc/human/Mapping/	Washington University; links to clone and accession maps of human genome
http://hgrep/ims.u-tokyo.ac.jp/	RIKEN and the University of Tokyo; overview over entire human genome structure
http://snp.cshl.org	The SNP Consortium; human SNP maps
http://www.ncbi.nlm.nih.gov/Omim	OMIM, Online Mendelian Inheritance in Man; information on human genes and disease
http://cgap.nci.nih.gov/	Cancer Genome Anatomy Project; annotated index of cancer related genes
http://www.ebi.ac.uk/swissprot/hpi/hpi.html	HPI; annotated human protein data
http://www.nhgri.nih.gov/ELSI/	NHGRI; information, links and articles on a wide range of social, ethical and legal issues

publicly accessible human genome resources are updated on a daily basis (Table 1). The complete human genome sequence is expected to be available by 2003.

3.1 Transposable DNA elements

About 45% of the human genome consists of transposable or mobile DNA elements, the most abundant class of repeat DNA sequences. These sequence elements have been acquired during our palaeontological past and have left a permanent mark in our genome. Transposable DNA elements can be divided into four types: (i) long interspersed elements (LINEs), (ii) short interspersed elements (SINEs), (iii) long terminal repeat (LTR) retrotransposons, and (iv) DNA transposons. In contrast to DNA-transposons, LINEs, SINEs, and LTR retrotransposons are transposed through RNA intermediates [30, 31].

LINEs are one of the most ancient components of the eukaryotic genome. There are about 850,000 LINE copies in the human genome, accounting for 21% of the genome sequence [10]. LINEs are $\sim$6 kb in length and harbour a promoter specific for RNA polymerase II and two open reading frames (i.e. DNA sequences that are potentially translated into proteins). LINE transcripts associate with their own protein products, one of which is a reverse transcriptase, an enzyme that catalyses the reverse transcription of RNA into DNA. Reverse transcription of

LINE RNAs yields complementary DNA that can be integrated back into the genome. LINEs are thought to be responsible for most of the reverse transcription activity in the human genome [31].

SINEs are shorter than LINEs and are typically between 100 and 400 bases in length. SINEs do not contain protein-coding sequences, but instead harbour a promoter specific for RNA polymerase III that is typical for tRNA genes. There are about 1.5 millions SINEs in the human genome accounting for 13% of human genomic DNA [10]. SINEs share sequences elements with LINEs and are thought to make use of the LINE reverse transcriptase machinery for transposition [31].

The most common SINE element is the Alu element, which is also the only known active SINE element in the human genome [31]. Alu elements were initially dismissed as parasitic and non-functional DNA sequences. However, in many species Alu elements are transcribed under conditions of stress. Alu RNAs can specifically bind double-stranded RNA-induced kinase (PKR), which under normal conditions inhibits translation of mRNA into proteins. Thus, binding of Alu RNA to PKR counteracts inhibition of translation and may therefore serve to stimulate protein production under stress. Consistent with the idea of positive Alu element function, Alu elements are preferentially found in regions of the genome that are rich in actively transcribed genes [30].

The third class of transposable elements is LTR retrotransposons that resemble retroviruses, RNA viruses that make use of reverse transcription to integrate into the host genome. There are about 450,000 copies of retrovirus-like elements in the human genome, accounting for 8% of human genomic DNA [10]. LTR retrotransposons encode for a number of proteins, including a reverse transcriptase [31].

The fourth class of transposable elements is DNA transposons, which encode a DNA transposase that mobilises the transposon element by a 'cut-and paste' mechanism [31]. The human genome contains about 300,000 DNA transposons that account for 3% of the genome sequence [10].

The sequence divergence within different classes of mobile DNA elements can be analysed to estimate their approximate age. Most DNA transposons found in the human genome are older than 100 million years and appear to have been inactive during the past 50 million years. LINE1, the predominant member of the LINE class and the only known active human LINE, is estimated to be more than 150 million years old. Alu elements are thought to exist in the human genome for at least 80 million years. LTR retrotransposons appear to be on the brink of extinction and only one LTR retrotransposon family appears to be active since humans diverged from chimpanzees [10].

While the human genome is filled with ancient transposons, the genomes of other organisms have been found to contain transposable elements of a more recent origin. The relatively old age of human mobile DNA elements are testament for an extremely slow rate by which non-functional elements are cleared from vertebrate genomes. The half-life of non-functional DNA elements is estimated to be 800 million years in humans but only 12 million years in the fruitfly (*Drosophila*) [10].

DNA sequence comparisons between the human and mouse genomes have led to some striking observations. The distribution of various classes of transposons is similar in both genomes. However, the activity of transposons in the mouse has not declined in the same way as in humans. The mouse genome contains a number of active transposon families that might contribute to a higher mutation rate compared to humans. It is estimated that 10% of all mutations found in the mouse genome are due to DNA transpositions into genes and this number appears to be 60 times lower in humans. Thus, evolutionary forces appear to affect the persistence of active transposable elements in humans and mice differentially [10].

Clues to the possible role or function of transposable DNA elements can be gained from their relative distribution within genomes. LINEs accumulate in regions of the genome that are rich in the nucleotide bases A and T, whereas SINEs, and especially Alu elements, are enriched in G/C-rich regions [30]. G/C-rich DNA sequences coincide with a high density of genes but their exact

biological functions have yet to be determined [32]. The preference of LINEs for A/T-rich regions that have a low gene content may reflect a mechanism by which parasitic DNA elements can persist in a genome without causing damage to the host organism. In this regard, the accumulation of SINEs in G/C-rich genome regions is quite puzzling, especially since SINEs rely on the LINE machinery for their transposition. A possible explanation might be that SINEs initially insert in A/T-rich regions with the same preference as LINEs but that evolutionary forces subsequently change the relative distribution of SINEs and LINEs within the genome. In light of their proposed positive function in stimulating protein production under stress, Alu elements may be regarded as genomic symbionts [30]. This may help to explain the preferential positioning of Alu sequences close to actively transcribed genes in both G/C-rich and A/T-rich regions of the genome.

Mobile DNAs are also responsible for innovations in the host genome, e.g. by introducing novel regulatory DNA elements or even novel genes. An example for acquired regulatory DNA elements are transcription terminator regions that are thought to originate from LTR retrotransposons. Twenty human genes are currently recognised to originate from transposons [30]. These include the genes *RAG1* and *RAG2* that encode the lymphocyte-specific proteins of the V(D)J recombination system responsible for antigen-specific immunity. Our immune system therefore may have originated from an ancient transposon insertion [33]. Another example for the utilisation of a transposable element is the *BC200* gene. BC200 is a brain-specific RNA located in the dendrites of all higher primates and most likely derived from an Alu element about 50 million years ago [30]. Further examples of human genes derived from transposable elements include the gene for telomerase, the enzyme responsible for the proper maintenance of chromosome ends [34], and CENPB, the major centromere-binding protein [35]. Preliminary analyses of the human genome sequence revealed 27 additional candidate genes that are suspected to originate from mobile elements and that will have to be characterised in the future [10].

Side effects of transposon activity are also observed. For example, reverse transcription of genic mRNAs by LINEs, which generally results in non-functional pseudogenes, can occasionally give rise to functional processed genes. It is believed that many intron-less genes have been created in this manner [10]. In addition, active transposable elements have been recognised as the cause of human disease. Haemophilia A, a disorder of blood coagulation that is linked to the X chromosome, is caused by disruption of the gene for a protein called factor VIII through insertion of LINE1 transposable elements. Examples for Alu elements involved in human disease include insertions into the factor IX gene as cause for haemophilia, insertions into the *APC* gene as cause for desmoid tumours, and insertions into the *BRCA*2 gene as cause for breast cancer [36].

3.2 Gene content

The ultimate aim in deciphering the human genome sequence was to compile a list of all human genes. This is a daunting task because genes represent only a very small proportion of the human genome and computer programs employed in gene finding are confronted with a significant signal-to-noise problem. Gene finding and gene prediction employs three basic approaches [10, 25]: (i) the identification of genes on the basis of known mRNAs [37], (ii) the identification of genes on the basis of sequence similarities to previously identified genes or proteins in other species [38], and (iii) ab initio recognition using DNA sequence databases and statistical information [25, 39–41]. All approaches to find or predict genes hold sources of errors. On one hand, genes might be missed because they are expressed only in a subset of cells or at very low levels and their mRNAs are thus undetectable [10]. On the other hand, the total number of genes in a genome tends to be overestimated because the different parts of long and complex genes can be misinterpreted as several distinct genes. Furthermore, the set of predicted genes identified solely on the basis of DNA

sequences characteristics typical for active genes might include pseudogenes. The development of advanced computer algorithms that allow gene finding with high accuracy will require a far more detailed understanding of the cellular mechanisms by which genes are recognised within the bulk of nuclear DNA.

The Human Genome Project reported the identification of about 15,000 known genes and predicted the existence of another 17,000 genes [10]. Celera Genomics reported strong evidence for about 26,000 human genes and weak evidence for 12,000 additional genes [25]. In view of earlier predictions ranging from 60,000 to 100,000 human genes, these numbers are unexpectedly low [42]. By comparing 30,000–40,000 human genes to 13,000 genes in the fruitfly *Drosophila melanogaster*, 18,000 genes in the primitive worm *C. elegans*, and 26,000 genes in the mustard weed *Arabidopsis thaliana*, it becomes evident that increased complexity of multicellular organisms is not simply achieved by using many more genes [43]. The human genome is the first vertebrate genome sequenced and more suitable sequence comparisons will be possible once the genomes of more closely related species become available. Of particular interest is the genome sequence of the chimpanzee, our closest relative, since it may hold the answer to the question whether the most significant advancements in humans, such as the origin of speech and the ability of abstract reasoning, are actually manifested in the genome sequence itself or whether they evolved from more subtle changes, for example, in specific gene expression patterns [44].

There is considerable variation in the size of the exons and introns and, consequently, in the size of protein-coding genes. On average, human genes are ~27 kb long, but many genes exceed 100 kb in length. The longest known human gene is the *dystrophin* gene (DMD) which spans over 2.4 Mb. The *titin* gene contains the largest number of exons, 178 in the Human Genome Project draft [10], and 234 in the Celera Genomics draft [25]. It also contains the longest known coding sequence at 80,780 bases and the longest single exon at 17,106 bases [10]. The average length of exons in humans is comparable to those found in the fruitfly or in the worm. Most exons in human genes are between 50 and 200 bases long. However, the intron size distribution differs significantly between fly, worm and humans. In humans, introns tend to be much longer with an average size of 3.3 kb and this variation in intron length results in a larger average gene size [10].

The distribution of the four different DNA nucleotide bases A, G, C, and T over the human genome is not uniform. Large genome segments with either high or low G/C content can be distinguished. Earlier studies had indicated that G/C-rich genome regions contain a higher gene density compared to regions that are low in G/C [32]. Results of the detailed analysis of the human genome sequence draft are broadly consistent with these observations. However, the actual proportion of genes located in genome segments that are relatively poor in G/C was found to be significantly higher than previously predicted [25].

The human genome can be described as oases of genes in a desert of non-coding DNA sequences. About 20% of the genome consist of very long gene-less DNA segments. The distribution of these gene deserts varies across the genome. Chromosomes 17, 19 and 22 contain the highest density of genes and only a small percentage of DNA sequences reside in gene deserts. The situation is reversed on chromosomes 4, 13, 18, and the sex chromosomes that contain a low gene density and a high percentage of gene deserts. It is noteworthy that genes deserts are not necessarily devoid of biological function.

Of particular interest is the distribution of CpG islands. The dinucleotide 5′-CpG-3′ ('p' designates the phosphate residue that connects the nucleotide residues) is underrepresented in the human genome because most CpGs become methylated at the cytosine base. This results in a spontaneous chemical reaction that ultimately leads to mutation of the CpG dinucleotide to TpG. CpG islands are regions of the genome containing unmethylated, and therefore stable, CpG dinucleotides. CpG islands are often associated with active genes, in particular with DNA regions

close to the start site of transcription [45]. Consistent with this observation, CpG islands have been shown to play important roles in the regulation of gene transcription and in gene imprinting, a process that determines gene activity in cell lineages during development and differentiation [46]. The Human Genome Project reports 28,890 CpG islands, and Celera Genomics counts a total of 28,519 CpG islands [10]. Both numbers are remarkably close to the number of predicted genes.

3.3 Single nucleotide polymorphism

Only 0.1% of the genome contribute to phenotypic variation amongst humans and, with the exception of identical twins, the genomes of two individuals are about 99.9% identical. Most of the DNA sequence variation between humans can be attributed to changes in DNA sequences at a single base pair, so-called single nucleotide polymorphisms (SNPs). SNPs occur on average every 1.9 kb when two chromosomes are compared. The International SNP Map Working Group presented a map of 1.42 million SNPs distributed throughout the entire genome [47]. Celera Genomics assigned 2.1 million SNPs to the genome [25]. Both groups reported that the genomic distribution of SNPs is markedly heterogeneous. About 75% of all SNPs are found outside genes. Within genes, the SNP rate is highest in introns and less than 1% of all SNPs are found in DNA sequence regions coding for proteins and therefore potentially affect protein structure and/or function. Nevertheless, this low percentage comprises thousands of candidate SNPs that may significantly contribute to the diversity of human proteins.

The identification of specific SNPs and their functional consequences is one of the major objectives of future genetic studies. It is well established that genetic variations affect the susceptibility to disease, the age of disease onset, the severity of illnesses, and the way the human body responds to medical treatment [48–50]. For example, single base differences in the *ApoE* gene have been implicated in Alzheimer's disease. The *ApoE* gene exists in three variants, *ApoE2*, *ApoE3*, and *ApoE4*, and individuals carrying the *ApoE4* version of the gene are the most likely to develop Alzheimer's disease [51]. Large-scale studies of SNP patterns in patients and healthy individuals will help to identify the molecular basis of many other diseases in the future.

A complete map of SNPs will also be prerequisite for detailed studies into the molecular basis for human phenotypic variation. SNP patterns entail a snapshot of the actions of evolutionary forces that are operative in human population genetics. For example, it could be demonstrated that our genes carry the signature of an expansion from Africa within the past 150,000 years [52]. A complete map of human SNPs is expected to fuel future research aimed to explore our evolutionary past and to discover the origin of our present diversity.

3.4 The human proteome

Analogous to the genome, the proteome represents the complete set of all proteins within an organism. Proteins typically consist of several discrete structural or functional domains that are conserved during evolution. More than 90% of protein domains in humans have counterparts in the fruitfly or the worm. About 60% of the human proteins that are predicted to exist based on the human genome sequence draft show sequence similarities to other organisms whose genomes have been sequenced. Also, 61% of the fly proteome, 41% of the worm proteome, and 46% of the yeast proteome have sequence similarities to predicted human proteins [53].

The draft of the human genome brought to light about 1200 protein families. Only 94 protein families, or 7%, appear to be vertebrate-specific suggesting that only a small number of novel protein domains were introduced into the vertebrate lineage. Vertebrate-specific protein families

reflect important physiological differences between vertebrates and other metazoans. A large proportion of these proteins exhibit functions in the immune and nervous systems [10].

Although there is only a small number of novel human protein families, there is substantial innovation in the creation of new proteins. New proteins can be created by rearrangement, insertion, or deletion of protein domains, resulting in new domain architectures. This mechanism is especially prominent in human proteins involved in extracellular structures and transmembrane structures where the total number of human domain architectures is more than twice of those found in the worm or the fruitfly [10]. A genome-wide analysis of domain architectures will be extremely helpful in resolving the evolutionary history of different species. About 40% of proteins predicted by the human genome sequence draft are of unknown function and cannot be assigned to known protein families [25]. A large proportion of proteins with known functions are either enzymes that play a crucial role in the cell metabolism, or proteins involved in signal transduction processes that are essential for intra- or inter-cellular communication.

The most common molecular function of human proteins is nucleic acid binding, employing 13.5% of the human proteome [25]. Nucleic acid-binding proteins include sequence-specific DNA-binding factors responsible for the regulation of gene transcription, and enzymes that participate in nucleic acid metabolism. Given the crucial importance of establishment and maintenance of cell type-specific gene expression patterns in multicellular organisms, it is not surprising that are a significant part of the proteome is engaged in gene regulation. A search of the human genome sequence revealed more than 2000 hypothetical genes that encode potential transcriptional activators [54]. These transcription factors need to be verified, biochemically characterised, and their target genes identified, before mechanisms of genome-wide transcription regulation processes can be fully elucidated.

Remarkably, a set of around 200 human proteins has significant amino acid sequence similarities to bacterial proteins, but not to any proteins found in yeast, worm, or the fruitfly [10]. These proteins appear to be of bacterial origin and were possibly acquired by gene transfer. Some of these genes are involved in metabolism and stress response, suggesting that they may have provided a selective advantage for the host organism during evolution.

The greater complexity of the human proteome as compared to the worm or the fruitfly is achieved only in part by the invention of novel proteins and novel protein architectures on the DNA level [10]. Additional levels of complexity arise from mechanisms such as alternative splicing and post-translational modifications. In addition, there are a bewildering number of potential interactions between individual cellular proteins that might affect their activity or function. The regulation of protein–protein interactions within the cell is considered to contribute significantly to the functional complexity of the human proteome.

A major objective of future research will be to decipher the human proteome and to ultimately identify the protein networks and functional pathways that give rise to complex multicellular organisms such as ourselves.

4 Functional genomics: assigning function to the genome

Biological functions are generally not evident from raw genome sequence data. For about 40% of the human genes, DNA sequence analysis has led to no prediction of function. In addition, the inferred functions of most of the remaining genes have yet to be confirmed [10, 25]. Functional genomics aims to determine the biological function(s) of individual genes or genome segments and comprises different areas including comparative genomics, gene expression studies, proteomics, and structural genomics. The combined data obtained from these areas will be required

to understand both the individual and collective function of genes, and will be prerequisite for a complete biochemical comprehension of cell biology.

4.1 Comparative genomics

Genome sequences comparisons between species are extremely valuable to elucidate innovations during evolution and to determine the timing of the divergence of species. DNA segments with important functions are more likely to retain their sequences during evolution than non-functional segments. Thus, conserved sequences between species are likely to point to important functions in key biological processes. Biological studies over the last century have made use of a number of key model organisms, including protozoans such as the yeasts *S. cervisiae* and *Schizosaccharomyces pombe*, metazoans such as the fruitfly (*D. melanogaster*) and the worm (*C. elegans*), and vertebrates such as zebrafish (*Brachydanio rerio*) and mouse (*Mus musculus*). There are two principal experimental approaches to identify and functionally characterise genes in animal models: forward and reverse genetics. Forward genetics starts with a mutant phenotype, which identifies the function of a gene. However, the identification of the corresponding DNA sequence within the genome by conventional gene-mapping techniques is a very time-consuming and laborious process. In contrast, reverse genetics starts from the DNA sequence of a known or predicted gene and attempts to gain insights into its function by obtaining phenotypic changes in model organisms upon gene mutation or gene deletion (knock-out). Complementary to both strategies, genetic crossing experiments are employed to examine the functional interplay of different genes. These studies were instrumental in the dissection of a number of fundamental metabolic and signalling pathways that are evolutionary conserved between species [55].

Genetically well characterised organisms such as the yeast *S. cerevisiae*, the fruitfly *D. melanogaster*, and the worm *C. elegans* were initially chosen for complete genome sequencing [22]. The genome sequences of these organisms proved to be invaluable both for the identification of human genes and the assignment of human gene function. In addition, detailed comparisons of the human genome sequence draft with the genome sequences of these distantly related organisms led to the identification of a number of vertebrate-specific candidate genes. However, confirmation of the vertebrate-specific nature of these genes and further elucidation of their function will require genome sequence comparisons with more closely related species such as the mouse. The mouse genome sequencing is well under way and is carried out by the publicly funded Mouse Sequencing Consortium and by Celera Genomics. Databases containing draft sequences obtained by the Mouse Sequencing Consortium are already accessible and completion of the mouse genome project is imminent. The available data have already revealed striking similarities between the human and mouse genomes. More than 180 cases of synteny, the presence of conserved DNA sequence segments that contain the same genes in the same order, have been found. Almost all genes on human chromosome 17 are found on mouse chromosome 4, and human chromosome 20 appears to be almost completely orthologous to mouse chromosome 2. The average length of the conserved segments is 14.5 Mb. The largest contiguous conserved segment found so far spans about 90 Mb on human chromosome 4 and corresponds to mouse chromosome 5 [10]. The completion of the mouse genome project is greatly anticipated and will further enhance our understanding of gene function in fundamental biological processes. Many human diseases with complex genetic background have counterparts in the mouse or rat. Therefore, knowledge of the mouse genome sequence will be instrumental for the diagnosis, prevention and treatment of human disease. In a number of instances, a conserved genome region containing a locus that contributes to a complex genetic disease involving several genes, a so-called quantitative trait

locus (QTL), could already be identified. Prominent examples for known human disease QTLs identified in animal models are cardiovascular disorders such as hypertension and atherosclerosis [56–58]. Large-scale genome-wide mutagenesis projects in mice have been set up in the UK [59] and in Germany [60] with the aim to screen thousands of mice mutants for links between DNA sequences and function.

To understand the controlled and coordinated read-out of genetic information, the identification of regulatory DNA sequences is of crucial importance. However, the identification of regulatory DNA sequences within the complex genomes of higher eukaryotes is extremely difficult. Transgenic studies have shown that human genes when introduced in mice are expressed in a manner that mimics their expression in their natural host. This observation suggests that the instructions for regulated gene transcription are evolutionary conserved [61]. Inter-species DNA sequence comparisons will therefore greatly facilitate the identification and functional characterisation of regulatory DNA elements.

Understanding coordinate gene regulation at the genome-wide level requires the identification of gene regulatory networks. Co-expression of certain genes may reflect their regulation by common sequence-specific transcription factors. Sequence comparisons between genes can therefore be used to identify common DNA elements that might be responsible for their coordinated expression. This approach is becoming increasingly feasible with the availability of genome-wide expression profiling data, in particular for smaller genomes. Genome-wide expression profiling in yeast has already been successfully used to identify regulatory networks involved in the coordinate expression of gene clusters during sporulation [62] and cell cycle progression [63]. Similar approaches to identify regulatory DNA sequence elements in mammals face challenges that do not exist in simpler organisms such as yeast. While regulatory sequences in yeast are typically located in close proximity to the transcription start site, regulatory elements in mammals are frequently found in great distances from their target genes. The size and complexity of mammalian genomes and their high content of non-coding sequences further complicates the identification of regulatory elements. Detailed DNA sequence analyses between closely related species may help to overcome these obstacles. Indeed, a comparison of the DNA sequences of mouse and human genes that are up-regulated in skeletal muscle revealed novel muscle-specific regulatory elements [64].

4.2 Proteomics

Comprehension of the genome at the proteome level will be prerequisite for a complete understanding of the functioning of a human cell. As outlined above, mechanisms such as alternative splicing of primary gene transcripts and post-translational modifications of gene products can considerably increase the complexity of the proteome over the genome. Indeed, the total number of different protein molecules expressed by the ~30,000 genes in the human genome is estimated to be in the order of 10^6. In addition, the activity and/or function of individual proteins may be subject to regulation, e.g. by specific protein–protein interactions, by targeting specific cellular compartments, by covalent modifications, and by protein degradation. This notion has led to independent efforts in proteomics, the study of protein function, subcellular localisation, and protein–protein interactions on a genome-wide scale [65]. Combinations of well-established biochemical and molecular biology methods, e.g. the combination of two-dimensional protein electrophoresis and mass spectrometry, are employed.

In order to combine and extract as much relevant information as possible a number of database resources were integrated into the human proteomics initiative (HPI), a joint effort between the Swiss Institute of Bioinformatics and the European Bioinformatics Institute. The HPI has

two phases. The first phase aims to annotate the protein products of all known human genes. The second phase is a long-term commitment to rapidly incorporate well-annotated protein data from ongoing research efforts and to provide the scientific community with free access to continuously updated databases. The HPI database contains currently over 6000 annotated human sequences as well as relevant information such as literature references, predicted post-translational modifications, splice variants, and known SNPs. Databases are being developed in order to allow automated annotations of predicted proteins from genome DNA sequences, and to facilitate a classification of proteins into cell type-specific proteome subsets [66]. These computational methods will provide meaningful tools to exploit the full potential of human genome sequence for basic and medical research by integrating biological and human genome data.

4.3 Structural genomics

The results of genome sequencing and recent advances in structure determination have ushered in structural genomics, a new field focused on the large-scale analysis of protein structures and functions. Three-dimensional high-resolution structures of proteins are required for understanding the molecular chemistry underlying their biological action. Protein structure determination can be a difficult and time-consuming process and the two major techniques used, X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy, possess certain drawbacks. X-ray crystallography requires the preparation of stable protein crystals, generally a laborious task, and NMR spectroscopy is limited to small and medium-sized molecules. However, technical advances are made and speed and productivity of macromolecular structure determination is continuously improved [67].

Structural genomics aims to provide structural information for all known cellular proteins by employing automated high-throughput methods. To achieve this goal, proteins are grouped into protein families based on their amino acid sequences and the structure of at least one family member is solved. Structure predictions for the remaining family members can then be made based on the known structures. Estimates for the total number of protein structure families range from 30,000 to 50,000, several orders of magnitude below the estimated total number of proteins in nature. Obviously, an initiative of this scale requires coordinated efforts of different disciplines such as biology, chemistry, physics, and bioinformatics.

The next pivotal step is to assign biological function to particular protein structures. High-resolution structures rarely give immediate insight into the function of a protein, but a rigorous analysis, for example, of key residues or protein surfaces may give functional clues. In addition, proteins with a low degree of amino acid sequence similarity may reveal unanticipated structural similarities. This may provide evidence for evolutionary links between protein families unrelated by amino acid sequence. The accuracy of the prediction of protein function based on structural information is expected to increase proportionally with the number of high-resolution protein structures available. However, the biological relevance of functional predictions based on structure has to be validated by independent biochemical or genetic studies.

There are a number of ongoing structural genomics projects. Japan was one of the first countries to recognise the importance of structural studies on a genome-wide scale and founded the RIKEN Structural Genomics Initiative in 1995 [68]. In the US, the National Institute of General Medical Sciences inaugurated its 'Protein Structure Initiative' in 1998, which is designed to organise a large cooperative effort in structural genomics and to produce a database resource which links sequence, structural, and functional information. Pilot projects in the US and in Canada have been set up to determine the structure of 10,000 proteins within the next 10 years [69]. Each project consists of a collaboration between different research groups and focuses on proteins

from different species or from different proteins families. There is at present little coordination between ongoing national efforts in Europe [70]. However, the Wellcome Trust in the UK is considering setting up a Structural Genomics Consortium, modelled on the highly successful SNP Consortium, including publicly accessible up-to-date protein structure/function databases.

5 Applications of the human genome sequence in medical sciences

Acquisition of the human genome sequence has led to the emergence of new technologies that promise to accelerate the identification and functional characterisation of genes involved in human disease and to facilitate major steps of drug development including drug target identification and validation, optimising drug efficacy, and reducing drug toxicity.

5.1 Genes and human disease

Many diseases have their origin in gene mutations. A prominent example is Haemophilia A, an inherited disease that affected the male lines of royal dynasties in Europe for several centuries. Haemophilia A is inherited recessively on the X chromosome. Heterozygotic females carrying only one disease variant of the gene are healthy, whereas males who carry the disease gene on their only X chromosome are haemophilic [71]. At present, more than 12,000 mutations associated with human diseases have been mapped to specific loci in the genome. The catalogue of these disorders is listed in a public database, Online Mendelian Inheritance in Man, (OMIM, http://www.ncbi.nlm.nih.gov/omim; see also Table 1). OMIM is updated almost daily, with new genes and mutations being discovered at a rapid pace.

Genetic disorders are classified by the number of mutated genes involved. Monogenic diseases are caused by mutations in a single gene, whereas polygenic diseases involve several mutated genes. Monogenic diseases are very rare compared to more complex polygenic diseases that affect several million people worldwide. Prominent examples of monogenic genetic diseases are cystic fibrosis (CF) and Huntington's disease (HD). CF is the most common potentially lethal genetic disease, affecting one in 2500 newborns in Northern Europe. The predominant symptoms of CF are obstructions of the airways by viscous mucus and, as a consequence, frequent bacterial infections of the respiratory tract. CF is inherited in a recessive fashion; heterozygote children carrying one mutant allele and one wild-type allele of the CF gene are healthy, whereas homozygote children carrying two mutant CF alleles develop CF [72]. The gene involved in the occurrence of CF was identified in 1989 by positional cloning [73–75]. Initially, the CF locus was mapped to chromosome 7 [73]. In an intense collaborative effort, sequential refinement of the genetic analysis yielded finally a candidate gene encoding a protein product that was called CFTR, for cystic fibrosis transmembrane conductance regulator [74]. It was subsequently found that 70% of all CF cases carried a three base pair deletion in the *CFTR* gene that caused the loss of a single amino acid residue in the CFTR protein product, resulting in its malfunction [75]. Cloning of the *CFTR* gene made biochemical studies possible and raised hopes for a gene therapy approach to CF [72]. The concept and strategies implicated in gene therapy are discussed in Section 5.6. HD is a neurodegenerative disease which typically affects individuals about 40 years of age and which progresses inexorably to death within 15 to 20 years [76]. Symptoms include alterations in mood, loss of memory, and involuntary movements. HD is caused by the insertion of CAG triplets into the gene *Huntingtin* (*Htt*). These additional DNA sequences lead to an extended stretch of glutamine amino acid residues in the corresponding protein product (Htt^{ex}). HD is inherited as a dominant trait; inheritance of only one Htt^{ex} allele is sufficient for the development of the disease.

It is not yet understood how the Htt^{ex} protein causes the degeneration of neurons, but aggregation of Htt^{ex} within cell nuclei appears to be, at least in part, responsible for its toxicity. Clearance of the mutant Htt^{ex} aggregates is therefore viewed as the most promising therapeutic strategy and medical research on HD is focussing on this approach [76].

Polygenic diseases represent traits that are caused by interactions between different genes, but are also frequently influenced by environmental factors such as smoking, the type of diet, exercise habits, and childhood exposure to infections [55]. Prior to the past decade, a connection between complex diseases and genetic factors was often not considered because genetic traits in patients did not follow an obvious pattern of inheritance. Only recently, a number of complex disorders with hitherto unknown origin could be linked to inherited genetic variations. While some progress has been made in identifying genes that contribute to complex diseases, the elucidation of underlying molecular mechanisms remains extremely difficult. The following findings exemplify earlier genomic research that has provided important insights into the pathogenesis of complex genetic diseases and that has paved the way for the development of novel therapeutics.

Hypertension (elevated blood pressure) has been associated with mutations in eight different genes and nine specific genes were found to contribute to hypotension (lowered blood pressure). The corresponding gene products act in a common regulatory pathway operative in the kidney that controls the net salt balance in the human body [77]. Heart failure is a predominant health-problem in industrialised countries and affects more than four million people in the United States alone. A common clinical type of heart failure is hypertrophic cardiomyopathy in which the mass of the heart is increased by a thickening of the left heart chamber wall. Ten different genes have been identified that, when mutated, can cause hypertrophic cardiomyopathy. The protein products of these genes are involved in different steps of the heart muscle contraction process [78]. Finally, cardiac arrhythmia is the failure of the heart to sustain a precise rhythm. It affects more than 300,000 individuals per year in the United States alone and can lead to sudden death. In recent years, six arrhythmia susceptibility genes have been discovered. Mutations in these genes affect ion channel proteins of heart cells [79].

5.2 Genetic basis of cancer

Cancer affects one in three people in the western world and is responsible for 20% of all deaths [80]. Throughout life, the DNA in human cells is exposed to mutagens, such as the ultraviolet component of sunlight and ionising radiation. This causes a progressive decay of genetic information that can occasionally lead to functional alterations in genes critical for cell proliferation. Gene mutations in only a single cell can be sufficient to give rise to cancer – the emergence of a cell population in which the mechanisms that control normal cell division and cell death are suspended. In addition, cancer cells can also acquire the ability to invade different tissues and to metastasise [81].

Tumorigenesis, the development of cancer, is a multistep process. It has been proposed that four to seven distinct events are required for the development of common epithelial cancer [82]. These events reflect genetic alterations that drive a progressive conversion from normal human cells into cancer cells. Events leading to tumorigenesis can either be the result of environmental influences (somatic), or inherited. Patients with inherited forms of cancer carry transforming mutations in every cell of their body, whereas in patients with somatic cancers, these mutations are found only in tumour cells. The vast majority of mutations in cancer are somatic and inherited forms of cancers contribute only to about one percent of all cancer cases [83]. Family history is the cardinal feature by which an inherited predisposition for cancer is recognised. However, inherited mutations in a cancer gene do not necessarily indicate similar probabilities for different

individuals to develop cancer; strong and weak predisposition can be distinguished. For certain cancers, e.g. breast cancer, mutations that result in a high cancer risk have been identified [84]. Women carrying mutations in the genes *BRCA-1* and *BRCA-2* have a lifetime breast cancer risk of 60–80%. However, only 15–20% of all inherited breast cancer cases are attributable to mutations in either gene. For the remaining 80%, mutations in different genes must be responsible for increased breast cancer susceptibility. A number of additional candidate genes have been proposed but require conformation by independent analyses [85].

Patterns of genetic alteration differ between different cancer types, probably because different tissues impose different constraints that need to be overcome. However, genetic alterations leading to cancer are not random, suggesting that cancers evolve along particular pathways [84]. This notion has lead to the optimistic view that determining the genetic alterations specific for a particular type of tumour cells, so-called molecular profiling, will provide information of clinical value, such as the future malignancy potential. It is hoped that this information can be used to tailor therapeutic approaches to individual cases [86]. More than 100 distinct types of human cancers with subtypes have been found within specific organs and all cancer types can be attributed to six essential alterations in cell physiology that collectively dictate malignant growth. Among these cancer-induced physiological changes are the capacity to sustain prolonged cell growth, the evasion of controlled cell death (apoptosis), and the ability to invade different tissues. Each acquisition of a cancer capability successfully breaches an anticancer defence mechanism hardwired into cells and tissues. However, the paths that different cells can take on their way to malignancy are highly variable and, in a given cancer type, specific gene mutations may be found only in a subset of tumours. Furthermore, mutations in critical genes resulting in the acquisition of capabilities that override normal cellular mechanisms may occur at different times during cancer development. Finally, genetic changes may affect various tumours differentially. In some tumours, a specific mutation might lead to the acquisition of only one cancer capability, whereas in other tumours the same genetic event could facilitate the simultaneous acquisition of several distinct capabilities. Despite the evident complexity of these pathways, it is believed that all tumours ultimately reach common biological endpoints, the acquisition of cancer capabilities [81].

Two major types of cancer genes can be distinguished. Oncogenes typically harbour mutations in their DNA sequence and the corresponding mutated protein products can cause the development of cancer. Oncogenes were first isolated from viruses capable to transform cultured human cells into cancer cells. Subsequently, it was discovered that human cells contain homologues to viral oncogenes that are involved in normal cellular functions [87]. These cellular genes were termed proto-oncogenes and their mutation or aberrant activation can promote transformation of normal cells into cancer cells [88]. A number of proto-oncogenes, e.g. oncogenes of the *Ras* family, could be identified in studies in which 'normal' cells were transfected with DNA isolated from animal tumours.

Tumour-suppressor genes serve to prevent the formation of cancer and their inactivation contributes to tumorigenesis [89]. Tumour suppressor genes were first proposed in studies of cancer in children [90]. Retinoblastoma, a cancer affecting one or both eyes in young children is in 35–40% of all cases inherited. Statistical analysis of inherited retinoblastoma cases revealed a requirement for one transforming event that occurred with constant probability over time. In contrast, the appearance of the sporadic form of retinoblastoma, which occurs usually much later in life, was consistent with a requirement for two such events. These observations suggested that two events were necessary for both the somatic and inherited forms of retinoblastoma and that in the case of the inherited form one event was already present in the germ line. Subsequently, a deletion in a region of chromosome 13 was found in some cases of inherited retinoblastomas and further studies confirmed that tumorigenesis required the loss of function of both copies of this specific

chromosomal region [91]. The *Rb* gene was finally identified to be responsible for retinoblastomas and is now regarded as the prototype of tumour-suppressor genes. Many more tumour-suppressor genes have been identified since, especially in inherited forms of cancer. Oncogenic mutations in inherited cancers are rare, presumably because they promote cancer during early stages in development and are therefore embryonic lethal [84]. Tumour-suppressor genes can be classified into different types. 'Classical' tumour-suppressor genes such as *Rb*, of which loss-of-function is required for cancer development, are termed 'gatekeepers'. Another class of tumour suppressor genes termed 'caretakers' contains genes involved in functions outside the actual pathway of cancer development. These genes are important for normal DNA repair and genome integrity and their mutation accelerates the acquisition of cancer capabilities [92].

A particularly complex mechanism by which loss-of-function or gain-of-function of genes can be acquired is the perturbation of epigenetic regulation. Epigenetic regulation of gene expression is an important process, which insures the expression of certain genes only at specific stages in development. In many cases, developmental genes are permanently switched off after they have fulfilled their function. Permanent silencing involves an inheritable marking of genes, e.g. the methylation of cytosine residues within promoter regions [93]. Perturbations in epigenetic gene regulation can lead to either expression of genes that need to be silenced or, conversely, to the silencing of genes, such as tumour-suppressor genes, that need to be active. It is not clear whether silencing of particular genes in cancer occurs through a stochastic process or whether certain promoters are predisposed. It is also unclear what exactly determines the particular molecular mechanism by which loss-of-function events occur since the frequency of distinct mechanisms can differ considerably between tumour types [94].

5.3 Identification of disease genes and disease pathways

Application of the human genome sequence is anticipated to accelerate both medical genetics and its application, the targeted treatment of genetic diseases. The identification of candidate genes involved in human disease had been extremely laborious and time-consuming. For example, identification and characterisation of the human CTFR gene involved in CF took several years [73]. In contrast, about 30 disease genes could be identified and their chromosomal location determined during the period in which the human genome was sequenced [10]. The human sequence database is further used to identify paralogues, genes that arose as a consequence of gene duplication within the genome and therefore contain closely related DNA sequences. Finding paralogues of disease genes is important for two reasons. First, paralogues can give rise to related genetic diseases. For example, the *CNBG3* gene has been identified as a paralogue for the *CNGA3* gene that, when mutated, causes colour blindness [95, 96]. Another example is the discovery of the *Presenilin-1* and *Presinilin-2* genes that can cause the early onset of Alzheimer's disease [97]. Second, paralogues may provide the means for therapeutic intervention, as exemplified by attempts to reactivate foetally expressed haemoglobin genes in individuals suffering from sickle cell disease or β-thalassaemia, caused by mutations in the *β-Globin* gene [98]. A complete scan of the human genome sequence revealed more than 200 potential paralogues that will have to be experimentally confirmed and characterised [10].

The identification of disease genes, and in particular the identification of sets of genes underlying complex polygenic disorders, by genome-wide DNA sequence comparisons between healthy individuals and patients is extremely complex due to the large number of single nucleotide polymorphisms present within the human genome sequence (see Section 3.3) [47]. SNPs responsible for human disease have to be distinguished from sequence variations that have no apparent detrimental effect. In addition, some SNPs may not affect disease immediately but may nevertheless

increase the susceptibility of individuals towards specific diseases. A large number of genomic sequences from healthy individuals and patients will have to be compared to establish a comprehensive catalogue of SNPs.

Alterations in cellular gene expression patterns that coincide with the manifestation of disease can be analysed using DNA microarray (DNA chip) technologies. DNA microarrays containing many thousand DNA molecules, each specific for a single gene, can be used to measure the levels of individual RNA species throughout the entire cellular RNA population [99, 100]. By comparing cellular gene expression profiles of healthy individuals with those of patients on a genome-wide level, specific genes that may be misregulated, and therefore contribute to the disease, may be identified. In addition, specific alterations in cellular gene expression patterns can be used to diagnose a particular disease. In addition to its application in genome-wide expression profiling, DNA microarray technology is currently developed to identify genetic variations (e.g. SNPs) on a genome-wide scale.

Once candidate disease genes have been identified, their protein products need to be placed within the specific cellular pathways in which they exert their function. The human genome sequence allows for the first time a comprehensive analysis of cellular protein networks, hitherto accessible only through classical biochemical and genetic studies. Information on known genetic networks and cellular pathways in well-characterised model organisms can be used as a starting point to identify protein orthologues in humans. Searching the human genome sequence database may in addition reveal the existence of protein paralogues that may function in identical or related cellular pathways. Furthermore, the presence of common DNA regulatory elements and common expression behaviour under a range of conditions are taken as good indicators for genes that operate in networks. In this regard, DNA microarray technology has been proven an extremely powerful tool [55, 100–102].

Two large-scale projects that make use of the Human Genome Project sequence data have been initiated to complete a catalogue of genetic changes related to cancer. The Cancer Genome Anatomy Project (CGAP) was launched in 1997 by the National Cancer Institute in the US as a collaborative network providing an up-to-date and publicly accessible annotated index of genes linked to cancer. CGAP databases are available through a series of web sites (http://cgap.nci.nih.gov/; [103]). A tumour gene index has been established that contains cancer genes arranged by tissues, stages of cancer, and specificity of gene expression. Furthermore, a database for chromosomal rearrangements related to cancer was created using more than 30,000 different patient cases with different types of tumours. The CGAP has developed into an important genetics and genomics tool and studies performed with the aid of CGAP have already started to yield important results. For example, new tumour markers, genes that are specifically expressed in tumour endothelium cancers, ovarian cancers, and glioblastomas, were identified [103, 104]. A large-scale project with a similar aim, the Cancer Genome Project, funded by the Wellcome Trust, was launched at the Sanger Centre in the UK. DNA microarray technology has been successfully employed to identify cancer genes and to decipher cancer pathways and several studies indicate that genome-wide expression profiling can be used both for cancer diagnosis and prognosis. Genome-wide profiling of gene transcription has been obtained for patients with leukaemia [105] and breast cancer [106]. In addition, three genes involved in metastasis, the most disastrous attribute of cancer, could be identified. At least one candidate gene, *Rho C*, might represent an attractive drug target to prevent the spread of cancer [107]. DNA microarray technology further allowed the characterisation of two different classes of B-cell lymphoma that are indistinguishable by tumour histology. Importantly, only one tumour type reacted to chemotherapy [108]. Finally, multidrug resistance (MDR) is a major problem associated with cancer treatment [109, 110]. Because the genes and their corresponding proteins responsible for MDR are known, DNA microarray technology might be used

to screen patient samples for appropriate cancer drugs. These results have important implications for the development of customised cancer therapies.

Another priority is the development of tumour cell-specific drugs. Most cancer drugs lack specificity and affect normal cells as well as cancer cells. This causes serious side effects and toxicities and thus limits their therapeutic value. The first example of a drug tailored for a specific type of cancer is Glivec® (Novartis Pharma AG) which has been approved for use in the United States and in Switzerland in 2001. Glivec® specifically targets chronic myeloid leukaemia (CML), a blood cancer caused when part of chromosome 9 is exchanged with chromosome 22 in a white blood cell. This rearrangement forms a new structure, the Philadelphia chromosome, and creates an active *Bcr-Abl* gene by gene fusion. The Bcr-Abl protein product is a protein kinase that affects cell growth and cell differentiation [111]. Glivec® specifically blocks the action of the Bcr-Abl kinase and therefore specifically inhibits the growth of CML cells whereas normal white blood cells are unaffected. The process of drug discovery is significantly enhanced by the use of human genome sequence data and further targeted drug therapies are within scope. Using automated systems, up to 100,000 substances per day can be screened for their ability to affect expression of a particular gene [112]. Once gene mutations responsible for cancer predisposition are identified, drugs can be designed that prevent tumour formation. For example, a combination of drugs directed to different cancer pathways has been shown to reduce tumour formation in a specific strain of mice that has increased susceptibility to colon cancer [113]. These drugs are now in clinical trials. It is anticipated that additional anti-cancer drugs will be identified within a short period of time. However, because these drugs have to be administered over long periods of time, extensive clinical trials involving a large number of individuals have to be carried out in order to guarantee their safe application [84]. For this reason, the full impact of the human genome sequence on the development of novel cancer therapies will not be seen for many years.

5.4 Drug target identification

A recent survey lists 483 drug targets employed by the pharmaceutical industry that account for virtually every drug on the market. Of these drug targets 45% are G protein-coupled cell membrane receptors and 28% are enzymes [114]. Knowledge of the entire set of human protein-coding genes increases the number of potential drug targets to the order of thousands and this prospect has led to a massive expansion of genomic research towards pharmaceutical applications. One important example for potential drug targets is the significant number of ligand-binding domains in proteins [115].

Access to the human genome sequence draft has already led to the proposal of a number of new candidate drug targets. The authors of the human genome sequence draft discussed potential drug targets for molecular pathways important in schizophrenia and in mood disorders [10]. They further reported the discovery of a new receptor molecule that may constitute a promising target for the development of drugs against asthma. Finally, the initial search of the human genome sequence for paralogues of classic drug targets has led to the identification of 16 novel drug target candidates. These include paralogue genes coding for receptors of neurotransmitters and of insulin-like growth factors [10].

Once proteins important for a particular biochemical pathway are identified, demonstrating that affecting their function has therapeutic utility is the pivotal point [50]. This process has been time-consuming and expensive in the past. Post-genomic technologies such as genome-wide gene expression profiling in cultured human cells promise considerable improvements in cost-effectiveness and accuracy of drug target validation. The application of functional genomics

in pharmaceutics will significantly facilitate the identification of appropriate candidate drugs, thereby reducing failures in the clinical phase of drug development [50].

5.5 Pharmacogenetics

Pharmacogenetics investigates how genetic variations in patients affect the therapeutic value of a particular drug [50, 116]. Adverse reactions of patients to a particular drug have already been correlated with amino acid variations in drug-metabolising enzymes, such as plasma cholinesterase and glucose 6-phosphate dehydrogenase, more than 50 years ago [117]. Since then, SNPs in more than 30 drug-metabolising enzymes as well as in a number of drug transporters have been linked to compromised levels of drug efficacy or drug safety. This information is already been used to prevent drug toxicities by screening patients for specific SNPs prior to drug treatment [116].

Currently, only a small percentage of drug toxicities can be explained with genetic variation in specific genes. A systematic research into the genetic basis of adverse drug reaction has been hampered by the fact that severe reactions of patients are rare and difficult to trace. With the availability of the human genome sequence, a genomic approach to pharmacogenetics is the method of choice to establish a comprehensive catalogue of gene products involved in the binding, metabolism, or transport of specific drugs [116]. DNA samples from patients with poor or adverse effects to a specific drug can be compared to samples from patients that respond well to the treatment using as markers the set of SNPs available in the human genome database. Complementary to this approach, the human sequence database can be used to identify paralogues of genes encoding known regulators of drug kinetics or drug dynamics. Finally, human genome sequence information may also be used to conduct pharmacogenetic research retrospectively [118]. Comparative genome-wide analysis of stored patient DNA samples may be undertaken after the completion of clinical trials and even after a drug has been introduced into the market. The ultimate vision is to prevent drug prescription by trial and error and to match appropriate therapies to the specific constitution of individual patients.

5.6 Gene therapy

Gene therapy involves the targeted delivery of genes in order to replace or to compensate for malfunctioning genes responsible for genetic diseases. Genetic disorders that include enzyme deficiencies such as cystic fibrosis require long-term and regulated expression of the transgene. Treatment of other genetic diseases such as cancer may require the delivery and expression of a transgene only during a short-term period, e.g. to induce cell death. For clinical applications that require only a short-term presence of the therapeutic gene product, the possibility of protein transduction, the delivery of the gene product instead of the gene, has been raised recently [119].

The first human genetic engineering project was initiated in 1989 in the US. Tumour-infiltrating lymphocytes that contained certain marked genes were transferred into patients with advanced cancer. These experiments had two major objectives, to demonstrate safe delivery of a transgene into patients and to demonstrate its presence in patient cells [120]. Subsequently, additional clinical trials were initiated that addressed different genetic disorders such as malignant melanomas, neuroblastomas, haemophilia B, and cystic fibrosis. Cystic fibrosis (CF) has long been seen as the most promising candidate for human gene therapy [72]. First, the *CFTR* gene mutated in CF patients has been intensively characterised. Second, a successful gene transfer of the *CFTR* gene into cultured cells was demonstrated. Finally, gene therapy of CF appeared to be particularly

feasible since the affected lung tissue is readily accessible through the airways, allowing the use of aerosols for targeted gene delivery.

Most gene delivery systems rely on modified viruses that release their genome containing the transgene upon cell infection. Different virus types have been used in human therapy trials [121]. Adenovirus, a naturally occurring pathogen that causes mild infections in human airways and eyes, is an attractive vector system for short-term gene delivery [122]. The biology of adenovirus is well understood and foreign genes can be readily inserted into the virus genome. In addition, adenovirus infects both dividing and non-dividing cells and adenovirus gene expression does not require integration into the host cell genome. Finally, adenovirus infection does not pose a cancer risk as opposed to other viruses that are known to induce tumours. Inflammatory reactions to viral gene products can be circumvented by using modified viruses lacking the corresponding genes. These 'stealth viruses' are not readily detected by the immune system and therefore have increased chances to deliver the therapeutic gene to the target tissue.

For long-term gene delivery retroviruses, which integrate their genetic material into the host cell's genome, are the preferred vectors. For example, a Moloney retrovirus vector was successfully used in gene therapy of severe combined immunodeficiency-X1 (SCID-X1), a lethal immune disorder [123]. As an alternative to viral vectors, non-viral delivery systems that make use of liposomes or protein-DNA complexes are being explored [124].

Over the last decade, more than 4000 patients have been enrolled in gene transfer experiments, but only a few unambiguous results have been produced [121]. Documented cases for successful gene therapy include the transfer of the gene for the blood coagulation factor IX into three patients with haemophilia B in the US in March 2000 [125] and the treatment of two children with SCID-X1 in France in April 2000 [123]. Tragically, a young volunteer, Jesse Gelsinger, died in September 1999 from a massive immune response to adenovirus during a gene therapy trial in the US. His death led to a temporary halt of gene therapy trials and spurred a thorough investigation into the safety of viral gene delivery systems. Thus, while the recently acquired knowledge of our entire library of genetic information significantly increases the prospects for a cure of inherited disorders, the general applicability of gene therapy remains to be demonstrated.

6 Concluding remarks

The human genome sequence draft is a remarkable achievement that has provided important novel insights into the structure and function of our genome. However, as pointed out by many during the course of the human sequencing project: 'the sequence is just the beginning'. Over the next few years, we will certainly witness a number of important developments based on the published human genome sequence drafts.

Completion of the human genome sequence projected for 2003, together with improved analyses of genomic data, will help to finally answer the crucial question of the exact number of genes and proteins and will provide the basis for a comprehensive mapping of all genetic variations such as SNPs. This information will ultimately allow a complete elucidation of cellular pathways, both on the genomic and on the proteomic level, and will help to identify all genes involved in human disease. Functional genomics will provide valuable information regarding the cell type-specific expression and function of specific genes and place them into complex regulatory networks. Medical and pharmaceutical sciences will benefit greatly from novel insights into the molecular basis of disease. Additional genome sequences of a variety of species will become available and will add to our understanding of evolutionary forces and mechanisms. The comparison of human and chimpanzee genome sequences is of particular interest; it may reveal if and to what extend

specific human features such as conscious thinking and speech are manifested in the genome. The daunting task of biology in the post-genomic era is to understand how genes orchestrate and maintain life both in single cells and in complicated organisms. As eloquently pointed out by Craig Venter and colleagues: 'the real challenge will lie ahead as we seek to explain how our minds have come to organise thoughts sufficiently well to investigate our own existence' [25].

References

[1] Mendel, G., Experiments in plant hybridization. *Verh. Naturforsch. Ver. Brünn*, **4**, pp. 3–47, 1865.

[2] Avery, O.T., MacLeod, M.C. & McCarthy, M., Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *J. Exp. Med.*, **98**, pp. 451–460, 1944.

[3] Watson, J.D. & Crick, F.H., A structure for DNA. *Nature*, **171**, pp. 737–738, 1953.

[4] Saenger, W., *Principles of Nucleic Acid Structure*, Springer Verlag: New York, 1984.

[5] Lewin, B., *Genes*, Oxford University Press: Oxford, 1997.

[6] Alberts, B. *et al.*, *Molecular Biology of the Cell*, Garland Publishing Inc.: New York & London, 1994.

[7] Nirenberg, M. & Leder, P., RNA codewords and protein synthesis. *Science*, **145**, pp. 1399–1407, 1964.

[8] Berget, S.M., Morre, C.S. & Sharp, P., Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. *Proc. Nat. Acad. Sci. USA*, **74**, pp. 3171–3175, 1977.

[9] Chow, L.T., Gelinas, R.E., Broker, T.R. & Roberts, R.J., An amazing sequence arrangement at the 5′ end of adenovirus 2 mRNA. *Cell*, **12**, pp. 1–8, 1977.

[10] International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome. *Nature*, **409**, pp. 860–921, 2001.

[11] Richmond, T.J., Finch, J.T., Rushton, B., Rhodes, D. & Klug, A., Structure of the nucleosome core particle at 7 Å resolution. *Nature*, **311**, pp. 532–537, 1984.

[12] Li, W.-H., Gu, Z., Wang, H. & Nekrutenko, A., Evolutionary analyses of the human genome. *Nature*, **409**, pp. 847–849, 2001.

[13] Dib, C. *et al.*, A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature*, **380**, pp. 152–154, 1996.

[14] Brett, D. *et al.*, EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett.*, **474**, pp. 83–86, 2000.

[15] Galas, D.J., Making Sense of the Sequence. *Science*, **291**, pp. 1257–1260, 2001.

[16] Hershko, A. & Ciechanover, A., The ubiquitin system. *Annu. Rev. Biochem.*, **67**, pp. 425–479, 1998.

[17] Hochstrasser, M., Evolution and function of ubiquitin-like protein-conjugation systems. *Nat. Cell Biol.*, **2**, pp. E153–157, 2000.

[18] Roberts, L., Controversial from the start. *Science*, **291**, pp. 1182–1188, 2001.

[19] Gyapay, G. *et al.*, The 1993–94 Genethon human genetic linkage map. *Nat. Genet.*, **7**, pp. 246–339, 1994.

[20] Hudson, T.J. *et al.*, An STS-based map of the human genome. *Science*, **270**, pp. 1945–1954, 1995.

[21] Fleischmann, R.D. *et al.*, Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, pp. 496–512, 1995.

[22] Mewes, H.W. *et al.*, Overview of the yeast genome. *Nature*, **387(suppl.)**, pp. 7–65, 1997.

[23] The *C. elegans* Sequence Consortium, Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, **282**, pp. 2012–2018, 1998.
[24] Marshall, E., Sharing the glory, not the credit. *Science*, **291**, pp. 1189–1193, 2001.
[25] Venter, J.C. *et al.*, The sequence of the human genome. *Science*, **291**, pp. 1304–1351, 2001.
[26] Sanger, F., Nicklen, S. & Coulson, A.R., DNA sequencing with chain-terminating inhibitors. *Proc. Nat. Acad. Sci. USA*, **74**, pp. 5463–5467, 1977.
[27] Maxam, A. & Gilbert, W., A new method for sequencing DNA. *Proc. Nat. Acad. Sci. USA*, **74**, pp. 560–564, 1977.
[28] Roberts, L., A history of the Human Genome Project. *Science*, **291**, pp. 1195–1200, 2001.
[29] Edwards, A. *et al.*, Automated DNA sequencing of the human HPRT locus. *Genomics*, **6**, pp. 593–608, 1990.
[30] Smit, A.F.A., Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.*, **9**, pp. 657–663, 1999.
[31] Smit, A.F.A., The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev.*, **6**, pp. 743–748, 1996.
[32] Gardiner, K., Base composition and gene distribution: critical pattern in mammalian genome organisation. *Trends Genet.*, **12**, pp. 519–524, 1996.
[33] Agrawal, A., Eastman, Q.M. & Schatz, D.G., Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature*, **394**, pp. 744–751, 1998.
[34] Malik, H.S., Burke, W.D. & Eickbush, H., The age and evolution of non-LTR retrotransposable elements. *Mol. Biol. Evol.*, **16**, pp. 793–805, 1999.
[35] Csink, A.K. & Henikoff, S., Something from nothing: the evolution and utility of satellite repeats. *Trends Genet.*, **14**, pp. 200–204, 1998.
[36] Kazazian Jr, H.H. & Moran, J.V., The impact of L1 retrotransposons on the human genome. *Nat. Genet.*, **19**, pp. 19–24, 1998.
[37] Ewing, B. & Green, P., Analysis of expressed sequence tags indicates 35,000 human genes. *Nat. Genet.*, **25**, pp. 232–234, 2000.
[38] Pruitt, K.D. & Maglott, D.R., RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, pp. 137–140, 2001.
[39] Burge, C. & Karlin, S., Prediction of complex gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, pp. 78–94, 1997.
[40] Kulp, D., Haussler, D., Reese, M.G. & Eckmann, F.H., A generalized hidden Markov model for the recognition of human genes in DNA. *ISMB*, **4**, pp. 134–142, 1996.
[41] Guigo, R., Agrawal, P., Abril, J.F., Burset, M. & Fickett, J.W., An assessment of gene prediction accuracy in large DNA sequences. *Genome Res.*, **10**, pp. 1631–1642, 2000.
[42] Fields, C., Adams, M.D., White, O. & Venter, J.C., How many genes in the human genome? *Nat. Genet.*, **7**, pp. 345–346, 1994.
[43] Baltimore, D., Our genome unveiled. *Nature*, **409**, pp. 814–816, 2001.
[44] Pääbo, S., The human genome and our view of ourselves. *Science*, **291**, pp. 1219–1220, 2001.
[45] Bird, A., CpG islands as gene markers in the vertebrate nucleus. *Trends Genet.*, **3**, pp. 342–347, 1987.
[46] Mann, J.R., Szabo, P.E., Reed, M.R. & Singer-Sam, J., Methylated DNA sequences in genomic imprinting. *Crit. Rev. Eukaryot. Gene Expr.*, **10**, pp. 241–247, 2000.

[47] The International SNP Map Working Group, A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**, pp. 928–933, 2001.
[48] Brookes, A.J., The essence of SNPs. *Gene*, **234**, pp. 177–186, 1999.
[49] Lander, E.S., The new genomics: global views of biology. *Science*, **274**, pp. 536–539, 1996.
[50] Roses, A.D., Pharmacogenetics and the practice of medicine. *Nature*, **405**, pp. 857–865, 2000.
[51] Roses, A.D., Apoliprotein E affects the rate of Alzheimer disease expression: beta-amyloid burden is a secondary consequence dependent on APOE genotype and duration of the disease. *J. Neuropathol. Exp. Neurol.*, **53**, pp. 429–437, 1994.
[52] Stoneking, M., From the evolutionary past ... . *Nature*, **409**, pp. 821–822, 2001.
[53] Rubin, G.M., Comparing species. *Nature*, **409**, pp. 820–821, 2001.
[54] Tupler, R., Perini, G. & Green, M., Expressing the human genome. *Nature*, **409**, pp. 832–833, 2001.
[55] Peltonen, L. & McKusick, V.A., Dissecting human disease in the postgenomic era. *Science*, **291**, pp. 1224–1229, 2001.
[56] Stoll, M. *et al.*, New target regions for human hypertension via comparative genomics. *Genome Res.*, **10**, pp. 473–482, 2000.
[57] Lusis, A.J., Atherosclerosis. *Nature*, **407**, pp. 233–241, 2000.
[58] Kreutz, R. *et al.*, Dissection of a quantitative trait locus for genetic hypertension on rat chromosome 10. *Proc. Nat. Acad. Sci. USA*, **92**, pp. 8778–8782, 1995.
[59] Nolan, P.M. *et al.*, A systematic, genome-wide, phenotype-driven mutagenesis programme for gene function studies in the mouse. *Nat. Genet.*, **25**, pp. 440–443, 2000.
[60] Hrabe de Angelis, M. *et al.*, Genome-wide, large-scale production of mutant mice by ENU mutagenesis. *Nat. Genet.*, **25**, pp. 444–447, 2000.
[61] Pennacchio, L.A. & Rubin, E.M., Genomic strategies to identify mammalian regulatory sequences. *Nat. Rev. Genet.*, **2**, pp. 100–109, 2001.
[62] Chu, S. *et al.*, The transcriptional program of sporulation in budding yeast. *Science*, **282**, pp. 699–705, 1998.
[63] Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. & Church, M.J., Systematic determination of genetic network architecture. *Nat. Genet.*, **22**, pp. 281–285, 1999.
[64] Wassermann, W.W., Palumbo, M., Thompson, W., Fickett, J.W. & Lawrence, C.E., Human-mouse genome comparisons to locate regulatory sites. *Nat. Genet.*, **26**, pp. 225–228, 2000.
[65] Pandey, A. & Mann, M., Proteomics to study genes and genomes. *Nature*, **405**, pp. 837–846, 2000.
[66] O'Donovan, C., Apweiler, R. & Bairoch, A., The human proteomics initiative (HPI). *Trends Biotechnol.*, **19**, pp. 178–181, 2001.
[67] Russell, R.B. & Eggleston, D.S., New roles for structure in biology and drug discovery. *Nat. Struct. Biol.*, **7**, pp. 928–930, 2000.
[68] Yokoyama, S. *et al.*, Structural genomics in Japan. *Nat. Struct. Biol.*, **7**, pp. 943–945, 2000.
[69] Terwillinger, T.C., Structural genomics in North America. *Nat. Struct. Biol.*, **7**, pp. 935–939, 2000.
[70] Heinemann, U., Structural genomics in Europe: slow start, strong finish? *Nat. Struct. Biol.*, **7**, pp. 940–942, 2000.

[71] Gitschier, J. *et al.*, Characterization of the human factor VIII gene. *Nature*, **312**, pp. 326–330, 1984.
[72] Collins, F.S., Cystic fibrosis: molecular biology and therapeutic implications. *Science*, **256**, pp. 774–779, 1992.
[73] Rommens, J.M. *et al.*, Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science*, **245**, pp. 1059–1065, 1989.
[74] Riordan, J.R. *et al.*, Identification of the cystic fibrosis gene: cloning and characterization of complimentary DNA. *Science*, **245**, pp. 1066–1073, 1989.
[75] Kerem, B. *et al.*, Identification of the cystic fibrosis gene: genetic analysis. *Science*, **245**, pp. 1073–1080, 1989.
[76] Tobin, A.J. & Signer, E.R., Huntington's disease: the challenge for cell biologists. *Trends Cell Biol.*, **10**, pp. 531–536, 2000.
[77] Lifton, R.P., Gharavi, A.G. & Geller, D.S., Molecular mechanisms of human hypertension. *Cell*, **104**, pp. 545–556, 2001.
[78] Seidman, J.G. & Seidman, C., The genetic basis for cardiomyopathy: from mutation identification to mechanistic paradigms. *Cell*, **104**, pp. 557–567, 2001.
[79] Keating, M. & Sanguinetti, M.C., Molecular and cellular mechanisms of cardiac arrhythmia. *Cell*, **104**, pp. 569–580, 2001.
[80] Futreal, P.A. *et al.*, Cancer and genomics. *Nature*, **409**, pp. 850–852, 2001.
[81] Hanahan, D. & Weinberg, R.A., The hallmarks of cancer. *Cell*, **100**, pp. 57–70, 2000.
[82] Renan, M.J., How many mutations are required for tumorigenesis? Implications from human cancer data. *Mol. Carcinogenesis*, **7**, pp. 139–146, 1993.
[83] Fearon, E.R., Human cancer syndromes: clues to the origin and nature of cancer. *Science*, **278**, pp. 1043–1050, 1997.
[84] Ponder, B.A.J., Cancer genetics. *Nature*, **411**, pp. 336–341, 2001.
[85] Nathanson, K.N., Wooster, R. & Weber, B.L., Breast cancer genetics: what we know and what we need. *Nat. Med.*, **7**, pp. 552–556, 2001.
[86] Liotta, L. & Petricoin, E., Molecular profiling of human cancer. *Nat. Rev. Genet.*, **1**, pp. 48–56, 2000.
[87] Parada, L.P., Tabin, C.J., Shih, C. & Weinberg, R.A., Human EJ bladder carcinoma oncogene is a homologue of Harvey Sarcoma virus ras gene. *Nature*, **297**, pp. 474–477, 1982.
[88] Bishop, J.M., Enemies within: the genesis of retrovirus oncogenes. *Cell*, **23**, pp. 5–6, 1981.
[89] Weinberg, R.A., Tumor suppressor genes. *Science*, **254**, pp. 1138–1145, 1991.
[90] Knudson, A.G., Mutation and cancer: statistical study of retinoblastoma. *Proc. Nat. Acad. Sci. USA*, **68**, pp. 820–823, 1971.
[91] Cavenee, W.K. *et al.*, Expression of recessive alleles by chromosomal mechanisms in retinoblastoma. *Nature*, **305**, pp. 779–784, 1983.
[92] Kinzler, K.W. & Vogelstein, B., Gatekeepers and caretakers. *Nature*, **386**, pp. 761–763, 1997.
[93] Surani, M.A., Imprinting and the initiation of gene silencing in the germ line. *Cell*, **93**, pp. 309–312, 1998.
[94] Baylin, S.B. & Herman, J.G., DNA hypermethylation in tumorigenesis. *Trends Genet.*, **16**, pp. 168–174, 2000.
[95] Kohl, S. *et al.*, Mutations in the CNGB3 gene encoding the beta-subunit of the cone photoreceptor cGMP-gated channel are responsible for achromatopsia (ACHM3) linked to chromosome 8q21. *Hum. Mol. Genet.*, **9**, pp. 2107–2116, 2000.

[96] Sundin, O.H. *et al.*, Genetic basis of total colourblindness among the Pingelapese islanders. *Nat. Genet.*, **25**, pp. 289–293, 2000.
[97] Sherrington, R. *et al.*, Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. *Nature*, **375**, pp. 754–760, 1995.
[98] Olivieri, N.F. & Wheatherall, D.J., The therapeutic reactivation of fetal haemoglobin. *Hum. Mol. Genet.*, **7**, pp. 1655–1658, 1998.
[99] Lockhart, D.J. & Winzeler, E.A., Genomics, gene expression and DNA arrays. *Nature*, **405**, pp. 827–836, 2000.
[100] Young, R., Biomedical discovery with DNA arrays. *Cell*, **102**, pp. 9–15, 2000.
[101] Rubin, E.M. & Tall, A., Perspectives for vascular genomics. *Nature*, **407**, pp. 265–269, 2000.
[102] Friddle, C.J., Koga, T., Rubin, E.M. & Bristow, J., Expression profiling reveals distinct sets of genes altered during induction and regression of cardiac hypertrophy. *Proc. Nat. Acad. Sci. USA*, **97**, pp. 6745–6750, 2000.
[103] Strausberg, R.L., Buetow, K.H., Emmert-Buck, M.R. & Klausner, R.D., The Cancer Genome Anatomy Project. *Trends Genet.*, **16**, pp. 103–106, 2000.
[104] Riggins, G.J. & Strausberg, R.L., Genome and genetic resources from the Cancer Genome Anatomy Project. *Hum. Mol. Genet.*, **10**, pp. 663–667, 2001.
[105] Golub, T.R. *et al.*, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, pp. 531–537, 1999.
[106] Perou, C.M. *et al.*, Molecular portraits of human breast cancer tumours. *Nature,* **406**, pp. 747–752, 2000.
[107] Clark, E.A., Golub, T.R., Lander, E.S. & Hynes, R.O., Genomic analysis of metastasis reveals an essential role for Rho C. *Nature*, **406**, pp. 532–535, 2000.
[108] Alizadeh, A.A. *et al.*, Distinct types of diffuse B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, pp. 503–511, 2000.
[109] Liscovitch, M. & Lavie, Y., Multidrug resistance: a role for cholesterol efflux pathways? *Trends Biochem. Sci.*, **25**, pp. 530–534, 2000.
[110] Szabo, D., Keyzer, H., Kaiser, H.E. & Molnar, J., Reversal of multidrug resistance of tumour cells. *Anticancer Res.*, **20**, pp. 4261–4274, 2000.
[111] Thijsen, S., Schuurhuis, G., van Oostveen, J. & Ossenkoppele, G., Chronic myeloid leukemia from basics to bedside. *Leukemia*, **13**, pp. 1646–1674, 1999.
[112] Marchant, J., Know your enemy. *New Scientist*, pp. 46–50, 2000.
[113] Torrance, C.J. *et al.*, Combinatorial chemoprevention of intestinal neoplasia. *Nat. Med.*, **6**, pp. 1024–1028, 2000.
[114] Drews, J., Drug discovery: a historical perspective. *Science*, **287**, pp. 1960–1964, 2000.
[115] Bailey, D., Zanders, E. & Dean, P., The end of the beginning for genomic medicine. *Nat. Biotechnol.*, **19**, pp. 207–211, 2001.
[116] Rothberg, B.E.G., Mapping a role for SNPs in drug development. *Nat. Biotechnol.*, **19**, pp. 209–211, 2001.
[117] Evans, W.E. & Relling, M.V., Pharmacogenomics: translating functional genomics into rational therapeutics. *Science*, **286**, pp. 487–491, 1999.
[118] McCarthy, A., Pharmacogenetics. *BMJ*, **322**, pp. 1007–1008, 2001.
[119] Ford, K.G., Souberbielle, B.E., Darling, D. & Farzaneh, F., Protein transduction: an alternative to genetic intervention? *Gene Therapy*, **8**, pp. 1–4, 2001.
[120] Anderson, W.F., Human gene therapy. *Science*, **256**, pp. 808–813, 1992.
[121] Marshall, E., Gene therapy on trial. *Science*, **288**, pp. 951–957, 2000.

[122] Benihoud, K., Yeth, P. & Perricaudet, M., Adenovirus vectors for gene delivery. *Curr. Opin. Biotechnol.*, **10**, pp. 440–447, 1999.
[123] Cavazzana-Calvo, M. *et al.*, Gene therapy of human severe combined immunodeficiency (SCID)-X1 disease. *Science*, **288**, pp. 669–672, 2000.
[124] Cristiano, R.J. *et al.*, Viral and nonviral gene delivery vectors for cancer gene therapy. *Cancer Detect. Prev.*, **22**, pp. 445–454, 1998.
[125] Kay, M.A. *et al.*, Evidence for gene transfer and expression of factor IX in haemophilia B patients treated with an AAV vector. *Nat. Genet.*, **24**, pp. 257–261, 2000.