Genetics and Genomics

Peter A. Kopp
Division of Endocrinology, Metabolism and Molecular Medicine, Northwestern University, Chicago, IL, USA

The understanding of the molecular basis of many endocrine and non-endocrine disorders has grown during the last decade (see OMIM, Online Mendelian Inheritance in Man, a comprehensive catalog of human genes and genetic disorders: http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim) (Table 1.1). With the exception of simple trauma, every disease has a genetic component. In monogenic disorders, for example, congenital adrenal hyperplasia (CAH), the genetic component is the major etiologic factor. In complex disorders, multiple genes in conjunction with environmental and lifestyle factors contribute to the pathogenesis; hence their designation as polygenic or multifactorial disorders. In other instances, genetic factors influence the manifestation of disease indirectly by defining the host’s susceptibility and resistance as, for example, in an environmental disease such as infection.

Genetics can be defined as the science of heredity and variation. Medical genetics, the clinical application of genetics, has historically focused on chromosomal abnormalities and inborn errors of metabolism, because of readily recognizable phenotypes and techniques to diagnose the conditions. Analysis of the transmission of human traits and disease within families, together with the study of the underlying molecular basis, has culminated in understanding many monogenic or Mendelian disorders, which has led to a significant modification in the diagnostic process for an increasing number of them. Many major health care problems, such as diabetes mellitus type 2, obesity, hypertension, heart disease, asthma and mental illnesses, are complex and we are at the early stages of unraveling the genetic alterations predisposing to these disorders, which are significantly influenced by exogenous factors. It is important to recognize that phenotype can also be influenced by genetic and environmental modifiers in monogenic disorders. For example, the expression of the phenotype in monogenic forms of diabetes mellitus due to mutations in the MODY (Maturity Onset of Diabetes in the Young) genes is influenced by factors such as diet and physical activity.

Cancer can also be viewed as a genetic disease, because somatic mutations in genes controlling growth and differentiation are key elements in its pathogenesis. Many cancers are associated with a predisposition conferred by hereditary germline mutations.

The term genome, introduced before the recognition that DNA is the genetic material, designates the totality of all genes on all chromosomes in the nucleus of a cell. Genomics refers to the discipline of mapping, sequencing and analyzing genomes. Because of the rapidly growing list of sequenced genomes in numerous organisms, genomics is currently undergoing a transition with increasing emphasis on functional aspects.

Genome analysis can be divided into structural and functional genomics. The analysis of differences among genomes of individuals of a given species is the focus of comparative genomics. The complement of messenger RNAs (mRNAs) transcribed by the cellular genome is called the transcriptome and the generation of mRNA expression profiles is referred to as transcriptomics. Epigenetic alterations, chemical modifications of DNA or chromatin proteins, influence gene transcription. The sum of all epigenetic information defines the epigenome and is a current focus of high-throughput analyses (epigenomics).

The term proteome has been coined to describe all the proteins expressed and modified following expression by the entire genome in the lifetime of a cell. Proteomics refers to the study of the proteome using techniques of large-scale protein separation and identification. The field of metabolomics aims at determining the composition and alterations of the metabolome, the complement of low-molecular-weight molecules. The relevance of these analyses lies in the fact that proteins and metabolites function in modular networks rather than linear pathways. Hence, any physiological or pathological alteration may have many effects on the proteome and metabolome. Metagenomics refers to the analysis of the genomes of the microorganisms present in a specific compartment (e.g. the gut flora).

The growth of biological information has required computerized databases to store organize, annotate and index the data. This has led to the development of bioinformatics, the application of informatics to (molecular) biology. Computational and
CHAPTER 1

Table 1.1 Selected databases relevant for genomic medicine.

<table>
<thead>
<tr>
<th>Site</th>
<th>Content</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>National Center for Biotechnology Information (NCBI)</td>
<td>Portal with extensive links to genomic databases, PubMed, OMIM. Links to educational online resources including guidelines for the use of genomic databases</td>
<td><a href="http://www.ncbi.nlm.nih.gov/">http://www.ncbi.nlm.nih.gov/</a></td>
</tr>
<tr>
<td>National Human Genome Research Institute</td>
<td>Information about the human genome sequence, genomes of other organisms and genomic research</td>
<td><a href="http://www.genome.gov/">http://www.genome.gov/</a></td>
</tr>
<tr>
<td>European Bioinformatics Institute (EBI)</td>
<td>Portal to numerous databases and tools for the analysis of sequences and structures</td>
<td><a href="http://www.ebi.ac.uk">http://www.ebi.ac.uk</a></td>
</tr>
<tr>
<td>DNA Database of Japan</td>
<td>Portal to numerous databases and tools for the analysis of sequences and structures</td>
<td><a href="http://www.ddbj.nig.ac.jp/">http://www.ddbj.nig.ac.jp/</a></td>
</tr>
<tr>
<td>University of California, Santa Cruz (UCSC) Genome Bioinformatics</td>
<td>Reference sequence of the human and other genomes. Multiple tools for sequence analysis</td>
<td><a href="http://genome.ucsc.edu/">http://genome.ucsc.edu/</a></td>
</tr>
<tr>
<td>Swiss-Prot</td>
<td>Protein sequence database with description of protein function, domains structure, post-translational modifications and variants</td>
<td><a href="http://www.ebi.ac.uk/swissprot/index.html">http://www.ebi.ac.uk/swissprot/index.html</a></td>
</tr>
<tr>
<td>Protein Structure Database</td>
<td>Portal to Biological Macromolecular Structures</td>
<td><a href="http://www.rcsb.org/pdb/home/home.do">http://www.rcsb.org/pdb/home/home.do</a></td>
</tr>
<tr>
<td>American College of Medical Genetics</td>
<td>Access to databases relevant for the diagnosis, treatment and prevention of genetic disease</td>
<td><a href="http://www.acmg.net/">http://www.acmg.net/</a></td>
</tr>
<tr>
<td>Genecards</td>
<td>A database of human genes, their products and involvement in diseases</td>
<td><a href="http://bioinformatics.weizmann.ac.il/cards/">http://bioinformatics.weizmann.ac.il/cards/</a></td>
</tr>
<tr>
<td>Gene Ontology</td>
<td>The Gene Ontology project provides a controlled vocabulary to describe gene and gene product attributes in any organism</td>
<td><a href="http://www.geneontology.org/">http://www.geneontology.org/</a></td>
</tr>
<tr>
<td>Chromosomal Variation in Man</td>
<td>Catalog of chromosomal disorders</td>
<td><a href="http://www.wiley.com/legacy/products/subject/life/borgaonkar/access.html">http://www.wiley.com/legacy/products/subject/life/borgaonkar/access.html</a></td>
</tr>
<tr>
<td>Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources (DECIPHER)</td>
<td>Catalog of genomic and clinical information of patients with chromosomal disorders</td>
<td><a href="http://www.sanger.ac.uk/PostGenomics/decipher">http://www.sanger.ac.uk/PostGenomics/decipher</a></td>
</tr>
<tr>
<td>Mitochondrial disorders, DNA repeat sequences &amp; disease</td>
<td>Catalog of disorders associated with mtDNA mutations and DNA repeats</td>
<td><a href="http://neuromuscular.wustl.edu/">http://neuromuscular.wustl.edu/</a></td>
</tr>
<tr>
<td>National Organization for Rare Disorders</td>
<td>Catalog of rare disorders including clinical presentation, diagnostic evaluation and treatment</td>
<td><a href="http://www.rarediseases.org/">http://www.rarediseases.org/</a></td>
</tr>
</tbody>
</table>

Mathematical tools are essential for the management of nucleotide and protein sequences, the prediction and modeling of secondary and tertiary structures, the analysis of gene and protein expression and the modeling of molecular pathways, interactions and networks. Numerous continuously evolving databases provide easy access to the expanding information about the genome of humans and other species, genetic disease and genetic testing (Table 1.1). The integration of data generated by transcriptomic, proteomic, epigenomic and metabolomic analyses through informatics, systems biology, is an emerging discipline aimed at understanding phenotypic variations and creating comprehensive models of cellular organization and function. These efforts are based on the expectation that an understanding of the complex and dynamic changes in a biological system may provide insights into pathogenic processes and the development of novel therapeutic strategies and compounds.

**DNA, genes and chromosomes**

**Structure of DNA**

The recognition in 1944 that DNA carries the genetic information was followed by the deduction of its structure in 1953. DNA is a double-stranded helix. Each strand consists of a backbone formed...
by a deoxyribose-phosphate polymer (Fig. 1.1). Four different nitrogen-containing bases are attached to the sugar ring: the purines, adenine (A) and guanine (G), and the pyrimidines, cytosine (C) and thymidine (T). The two strands of DNA are complementary and held together by hydrogen bonds pairing adenine with thymidine and guanine with cytosine. The double-stranded nature of DNA and its strict base pair complementarity permit faithful replication during cell division, as each strand can serve as a template for the synthesis of a new complementary strand referred to as semi-conservative replication. The complementary structure of the two strands is also of importance as a defense against DNA damage. Damage or loss of a base on the opposite strand can be repaired using the intact strand as a template.

The presence of four different bases provides surprising genetic diversity. In the protein-coding regions of genes, the DNA bases are arranged into codons, triplets of bases that encode one of the 20 different amino acids or a stop codon. Combinatorial arrangement of the four bases creates 64 different triplets ($4^3$). Many amino acids, as well as the stop of translation, can be specified by several different codons. Because there are more codons...
than amino acids, the genetic code is said to be degenerate. Arranging the codons in different combinations of various length permits the generation of a tremendous diversity of polypeptides.

The human genome

The Human Genome Project, launched in the 1980s, first led to the creation of genetic and physical maps. A genetic map describes the order of genes and defines the position of a gene relative to other loci on the same chromosome. It is constructed by assessing how frequently two markers are inherited together, i.e. linked, by linkage studies. Distances of the genetic map are expressed in recombination units or centimorgans (cM). One centimorgan corresponds to a recombination frequency of 1% between two polymorphic markers and corresponds to approximately 1 Mb of DNA. Physical maps indicate the position of a locus or gene in absolute values. Sequence-tagged sites (STSs), any site in a chromosome or genome that is identified by a known unique DNA sequence, have been widely used for physical mapping and, after cloning of DNA fragments, they have served as landmarks for arranging overlapping cloned DNA fragments in the same order as they occur in the genome. These overlapping clones then allowed the characterization of contiguous DNA sequences (contigs). This approach led to high-resolution physical maps by cloning the whole genome into overlapping fragments. The complete DNA sequence of each chromosome provides the highest resolution physical map and, after publication of a first draft of the whole genome in 2000, its sequence analysis was largely completed in 2003.

Human DNA consists of about 3 billion base pairs (bp) of DNA per haploid genome contained in the 23 chromosomes. The smallest chromosome (chromosome 21) contains approximately 47 million bp, the largest (chromosome 1) 247 million bp. The human genome is estimated to contain about 30,000–40,000 genes. This number is smaller than the original estimates (up to 100,000 genes), which were derived from the large diversity of proteins. This observation indicates that alternative splicing of genes and the use of various promoters are important mechanisms generating protein diversity (Fig. 1.1).

Historically, genes were identified because they conferred specific traits that are transmitted from one generation to the next. Genes can be defined as functional units that are regulated by transcription and encode RNA (Fig. 1.1). The majority of RNA transcripts consist of mRNA which is subsequently translated into protein. Other RNA transcripts exert specialized functions, such as transfer of amino acids for polypeptide synthesis (tRNA), contribute to ribosome structure (rRNA) or regulate transcription. MicroRNAs (miRNAs) are small non-coding RNAs that regulate gene expression by targeting miRNAs of protein coding genes or non-coding RNA transcripts. They have an important role in developmental and physiologic processes and can act as tumor suppressors or oncogenes in cancer development.

Genes account for 10–15% of the genomic DNA. Much of the remaining DNA consists of highly repetitive sequences, the function of which remains incompletely understood. These repetitive DNA regions, along with non-repetitive sequences that do not encode genes, are, in part, involved in the packaging of DNA into chromatin and chromosomes or in the regulation of gene expression. Genes are unevenly distributed across the various chromosomes and vary in size from a few hundred to more than 2 million base pairs. The vast majority of genes are located in nuclear DNA but a few are found in mitochondrial DNA (mtDNA).

A major goal of human genetics aims at understanding the role of common genetic variants in susceptibility to common disorders. This involves identifying, cataloging and characterizing gene variants, followed by performing association studies. The variants include short repetitive sequences in regulatory or coding regions and single-nucleotide polymorphisms (SNPs), changes in which a single base in the DNA differs from the usual base at that position (Fig. 1.2). SNPs occur roughly every 300 bp and most are found outside coding regions. SNPs within a coding sequence can be synonymous (i.e. not altering the amino acid code) or non-synonymous. There are roughly 3 million differences between the DNA sequences of any two copies of the human genome. The identification of approximately 10 million SNPs that occur commonly in the human genome through the International HapMap Project is of great relevance for genome-wide association studies (GWAS).

Structure and function of genes

The structure of a typical gene consists of regulatory regions followed by exons and introns and downstream untranslated regions (Fig. 1.1). The regulatory regions controlling gene expression most commonly involve sequences upstream (5′) of the transcription start site, although there are examples of control elements located within introns or downstream of the coding region of a gene. Exons designate the regions of a gene that are eventually spliced together to form the mature mRNA. Introns refer to the intervening regions between the exons that are spliced out of precursor RNAs during RNA processing. A gene may generate various transcripts through the use of alternative promoters and/or alternative splicing of exons (Fig. 1.1). These mechanisms contribute to the diversity of proteins and their functions.

The regulatory DNA sequences of a gene, which are typically located upstream of the coding region, are referred to as the promoter. The promoter region contains specific sequences, response elements that bind transcription factors. Some transcription factors are ubiquitous; others are cell-specific. Gene expression is controlled by additional regulatory elements, enhancers and locus control regions, which may be located far away from the promoter region. The transcription factors that bind to the promoter and enhancer sequences provide a code for regulating transcription that is dependent on developmental state, cell type and endogenous and exogenous stimuli. Transcription factors interact with other nuclear proteins, co-activators and co-repressors and generate large regulatory complexes that ultimately activate or repress transcription.
In the eukaryotic cell nucleus, DNA is packaged by histones into nucleosomes. This packaging inhibits transcription by impeding the binding of transcriptional activators to their cognate DNA sites. Therefore, alterations in chromatin structure typically precede gene transcription. Repression is often associated with histone deacetylation. Conversely, activation of transcription may involve histone acetylation, which results in the remodeling of chromatin and subsequent binding of trans-acting factors to DNA (Fig. 1.3). Once bound to DNA, the transcription factor complexes recruit proteins that form the basal transcription complex including RNA polymerase. Gene transcription occurs with the synthesis of RNA from the DNA template by RNA polymerase. mRNA is encoded by the coding strand of the DNA double helix and is translated into proteins by ribosomes. The transcriptional termination signals reside in the 3′ region of a gene. A polyadenylation signal encodes a poly-A tail, which influences mRNA export to the cytoplasm, stability and translation efficiency.

Transcription factors account for about 30% of all expressed genes. Mutations in transcription factors cause a significant number of endocrine and non-endocrine genetic disorders. Because a given set of transcription factors may be expressed in various tissues, it is not uncommon to observe a syndromic phenotype. The mechanism by which transcription factor defects cause disease often involves haploinsufficiency, a situation in which a single copy of the normal gene is incapable of providing sufficient protein production to assure normal function. Biallelic mutations in such a gene may result in a more pronounced phenotype. For example, monoallelic mutations in the transcription factor HESX1 (RPX) result in various constellations of pituitary hormone deficiencies and the phenotype is variable among family members with the same mutation. Inactivating mutations of both alleles of HESX1 cause familial septo-optic dysplasia and combined pituitary hormone deficiency.

Gene expression is also influenced by epigenetic events, such as X-inactivation and imprinting, i.e. a marking of genes that results in monoallelic expression depending on their parental origin. In this situation, DNA methylation leads to silencing, i.e. suppression of gene expression on one of the chromosomes. Genomic imprinting has an important role in the pathogenesis of several genetic disorders [e.g. Prader–Willi syndrome and Albright hereditary osteodystrophy (AHO)].

**Chromosomes**

The normal diploid number of chromosomes in humans is 46, consisting of two homologous sets of 22 autosomes (chromosomes 1 to 22) and a pair of sex chromosomes. Females have two X chromosomes (XX), whereas males have one X and one Y chromosome (XY). As a consequence of meiosis, germ cells – sperm or oocytes – are haploid and contain one set of 22 autosomes and one of the sex chromosomes. At the time of fertilization, the pairing of the homologous chromosomes from the mother and father results in reconstitution of the diploid genome. With each cell division, i.e. mitosis, chromosomes are replicated, paired, segregated and divided into two daughter cells.

The normal human genome contains large blocks (>1 kb) of DNA sequences, often containing numerous genes, that can be duplicated one or several times or missing on a given chromosome. These copy number variants (CNV) tends to vary in a specific manner among different populations. CNVs are associated with hot spots of chromosomal rearrangements. Because CNVs can result in differential levels of gene expression, they are thought...
to contribute significantly to normal phenotypic variation. Genomic imbalances resulting from CNVs are also causally involved in numerous disorders; for example, Williams syndrome (short stature, elfin facies, supravalvular aortic stenosis and hypercalcemia).

### Replication of DNA, mitosis and meiosis

Genetic information in DNA is transmitted to daughter cells during two different types of cell division: mitosis and meiosis. Somatic cells divide by mitosis, allowing the diploid (2n) genome to replicate itself during cell division. The formation of germ cells, sperm and ova, requires meiosis, a process that leads to the reduction of the diploid (2n) set of chromosomes to the haploid state (1n).

Before mitosis, cells exit the resting or G0 state and enter the cell cycle. After traversing a critical checkpoint in G1, cells undergo DNA synthesis (S phase), during which the DNA in each chromosome is replicated, yielding two pairs of sister chromatids (2n to 4n). The process of DNA synthesis requires strict fidelity in order to avoid transmitting errors to subsequent generations of cells. Therefore, genetic abnormalities of enzymes that are involved in DNA mismatch repair predispose to neoplasia because of the rapid acquisition of additional mutations (e.g. xeroderma pigmentosa, Bloom syndrome, ataxia telangiectasia and hereditary non-polyposis colon cancer). After completion of DNA synthesis, cells enter G2 and progress through a second checkpoint before entering mitosis. Subsequently, the chromosomes condense and are aligned along the equatorial plate at metaphase. The two identical sister chromatids, held together at the centromere, divide and migrate to opposite poles of the cell. After the formation of a nuclear membrane around the two separated sets of chromatids, the cell divides forming two daughter cells with a diploid (2n) set of chromosomes.

Meiosis occurs only in germ cells of the gonads. It involves two steps of cell division that reduce the chromosome number to the haploid state. Recombination, the exchange of DNA between homologous paternal and maternal chromosomes during the first cell division, is essential for generating genetic diversity. Each chromosome pair forms two sister chromatids (2n to 4n). This is followed by an exchange of DNA between homologous chromosomes through the process of crossover. In most instances, there is at least one crossover on each chromosomal arm. This recombination process occurs more frequently in female meiosis than in male meiosis. Subsequently, the chromosomes segregate randomly. As there are 23 chromosomes, this can generate 223 (>8 million) possible combinations of chromosomes. Together with the genetic exchanges that occur during recombination through crossover, chromosomal segregation generates tremendous diversity and therefore each gamete is genetically unique. The processes of recombination and independent segregation of chromosomes provide the foundation for performing linkage analyses, in which the inheritance of linked genes is correlated with the presence of a disease or genetic trait. After the first meiotic division, which results in two daughter cells (2n), the two chromatids of each chromosome separate during a second meiotic division to yield four gametes with a haploid chromosome set (1n). Through fertilization of an egg by a sperm, the two haploid sets are combined, thereby restoring the diploid state (2n) in the zygote.
Analysis of chromosomes, DNA and RNA

Analyses of large alterations in the genome are possible using cytogenetics, fluorescence in situ hybridization (FISH), Southern blotting, high-throughput genotyping and sequencing. More discrete sequence alterations rely heavily on the use of the polymerase chain reaction (PCR). PCR permits rapid genetic testing and mutational analysis with small amounts of DNA extracted from solid tissues, nucleated blood cells, leukocytes, buccal cells or hair roots. Reverse transcription PCR (RT-PCR) transcribes RNA into a complementary DNA strand, which can then be amplified by PCR. RT-PCR can be used for sequence analyses of the coding regions and to detect absent or reduced levels of mRNA expression resulting from a mutated allele.

Screening for point mutations can be performed by numerous methods, such as sequencing of DNA fragments amplified by PCR, recognition of mismatches between nucleic acid duplexes or electrophoretic separation of single- or double-stranded DNA. Most traditional diagnostic methods are gel-based and focus on single genes. Novel techniques for the analysis of mutations, genetic mapping and mRNA expression profiles are rapidly evolving. Chip techniques allow hybridization of DNA or RNA to hundreds of thousands of probes simultaneously. Microarrays are being used clinically for mutational analysis of several human disease genes, for the identification of viral sequence variations and for large-scale analyses of mRNA transcripts. Comprehensive genotyping of SNPs can be performed with microarray and beadarray technologies or mass spectrometry. These technologies are widely used for genotyping in GWAS, analyses of copy number variations and characterization of genomic DNA methylation. While traditional sequencing technologies are still of importance and widely used, particularly for the sequencing of PCR products, high-throughput sequencing technologies are rapidly evolving and several platforms have become commercially available in the recent past. These techniques provide the foundation to expand from a focus on single genes to analyses at the scale of the genomes of prokaryotes and eukaryotes. It is anticipated that sequencing of the whole human genome of an individual for a cost of $1000 or less will soon become a reality. In addition to sequencing DNA, high-throughput sequencing technologies can be used for the characterization of RNA expression, non-coding and microRNAs, protein–DNA interactions, epigenomic alterations and metagenomic analyses.

The availability of comprehensive individual sequence information is expected to have a significant impact on medical care and preventative strategies but it also raises ethical and legal concerns how such information may be used by insurers and employers. Protection against discrimination based on genetic information for health insurance and employment is needed; for example, the recently introduced Genetic Information Nondiscrimination Act (GINA) in the USA is an important first step to avoid misuse of genetic information.

Concerns that the exclusive protection of genetic risks results in an increasing discrimination against lifestyle risks persist. The impact of genetic testing on health care costs has not been addressed in detail and probably depends on the specific disorder and the availability of effective therapeutic modalities. In certain instances it can be cost-effective; for example, in carrier detection in family members of individuals affected by multiple endocrine neoplasia type 2 (MEN2). The marketing of genetic testing directly to consumers (consumer genomics) through the Internet by commercial companies raises numerous questions about the accuracy and confidentiality of the information, how the results should be handled and how to ensure appropriate regulatory oversight.

Genetic linkage and association

There are two primary strategies for mapping genes that cause or increase susceptibility to human disease: linkage and association studies.

Genetic linkage refers to the fact that genes and polymorphic DNA markers such as microsatellites and SNPs are physically connected, i.e. linked, to one another along the chromosomes (Fig. 1.4). Two principles are essential for understanding the concept of genetic linkage. First, when two genes are close together on a chromosome, they are usually transmitted together, unless a recombination event separates them. Secondly, the odds of a crossover or recombination event between two linked genes are proportional to the distance that separates them. Thus, genes that are further apart are more likely to undergo a recombination event than genes that are very close together. The detection of chromosomal loci that segregate with a disease by linkage has been widely used to identify the gene responsible for the disease by positional cloning, a technique of isolating a gene from the knowledge of its map location. It has also been used to predict the odds of disease gene transmission in genetic counseling.

Polymorphisms are essential for linkage studies because they provide a means to distinguish the maternal and paternal chromosomes in an individual. On average, one out of every 300 bp varies from one person to the next. Although this degree of variation seems low (99.9% identical), it means that more than 3 million sequence differences exist between any two unrelated individuals and the probability that the sequence at such loci will differ on the two homologous chromosomes is high (often, >70–90%). These sequence variations include a variable number of tandem repeats (VNTRs), microsatellites [also referred to as short tandem repeats (STRs)] and SNPs. Most microsatellite markers consist of di-, tri- or tetranucleotide repeats that can be measured readily using PCR and primers that reside on either side of the repeat sequences (Fig. 1.4). Automated analysis of SNPs with microarrays, beadarrays or mass spectrometry are now the methods of choice for determining genetic variation for linkage and association studies.

In order to identify a chromosomal locus that segregates with a disease, it is necessary to determine the genotype or haplotype of DNA samples from one or several pedigrees. A haplotype
Figure 1.4 Analysis of polymorphic microsatellite markers and linkage analysis. Upper panel: The example depicts a CAC trinucleotide repeat with three alleles in a nuclear family. PCR with primers flanking the polymorphic region results in products of variable length, depending on the number of CAC repeats. After characterization of the alleles in the parents, transmission of the paternal and maternal alleles can be determined in the offspring. The gel on the right shows the concomitant analysis of multiple microsatellites. The PCR products reflecting the different alleles can be distinguished by differences in length and fluorescent labels. The red marker (arrowed) included in every lane is a size standard. Lower panel: Determination of polymorphic microsatellite markers flanking the arginine vasopressin (AVP) gene located on chromosome 20p13 in a family with autosomal dominant neurohypophyseal diabetes insipidus. The parental origin of the alleles can be determined in generation II (p, paternally inherited; m, maternally inherited). The three affected individuals II-1, II-4, II-6 share the same alleles for the markers D20S95/115/189/186. In individual II-4, a recombination has occurred between markers D20S117 and 95 of the paternal chromosome. The unaffected individual II-3 has inherited the alternate paternal alleles. Although individual II-5 is homozygous for the alleles segregating with the phenotype in this family, she is not related to I-1 and she does not have diabetes insipidus. The haplotype of these markers is only associated with the phenotype in the original family but not in the general population.
designates a group of alleles that are closely linked, i.e. in close proximity on a chromosome and that are usually inherited as a unit. After characterizing the alleles, one can assess whether certain marker alleles co-segregate with the disease. Markers closest to the disease gene are less likely to undergo recombination events and therefore receive a higher linkage score. Linkage is expressed as a logarithm of odds (lod) score, i.e. the ratio of the probability that the disease and marker loci are linked rather than unlinked. Lod scores of +3 (1000:1) are generally accepted as supporting linkage.

Allelic association refers to a situation in which the frequency of an allele is significantly increased or decreased in a particular disease. Linkage and association differ in several respects. Genetic linkage is demonstrable in families or sibships. Association studies compare a population of affected individuals with a control population. Association studies are often performed as case–control studies that include unrelated affected individuals and matched controls or as family-based studies that compare the frequencies of alleles that are transmitted to affected children. Allelic association studies are useful for identifying susceptibility genes in complex disorders. When alleles at two loci occur more frequently in combination than would be predicted based on known allele frequencies and recombination fractions, they are said to be in linkage disequilibrium (Fig. 1.5).

The HapMap project and its impact on association studies
After the identification of the approximately 10 million SNPs that are commonly found in the human genome, the International HapMap Project has generated a catalog of common genetic variants that occur in individuals from distinct ethnic backgrounds (http://www.hapmap.org/). SNPs that are in close proximity are inherited together as blocks referred to as haplotypes, hence the name HapMap. These blocks can be identified by genotyping selected SNPs, so called Tag SNPs, an approach that greatly reduces cost and workload (Fig. 1.6). This permits one to char-
acterize a limited number of SNPs in order to identify the set of haplotypes present in an individual and greatly facilitates performing GWAS aiming at the elucidation of the complex interactions among multiple genes and lifestyle factors in multifactorial disorders.

### Medical genetics

#### Mutations and human disease

**Structure of mutations**

Mutations are an important cause of genetic diversity as well as disease. A mutation can be defined as any change in the nucleotide sequence of DNA regardless of its functional consequences (Fig. 1.7). Mutations are structurally diverse. They can affect one or a few nucleotides, consist of gross numerical or structural alterations in individual genes or chromosomes or involve the entire genome. Mutations can occur in all domains of a given gene. Large deletions may affect a portion of a gene or an entire gene or, if several genes are involved, they may lead to a contiguous gene syndrome. Occasionally, mispairing of homologous sequences leads to unequal crossover. This results in gene duplication on one of the chromosomes and gene deletion on the other chromosome.

For example, a significant fraction of growth hormone (GH) gene deletions involves unequal crossing-over. The GH gene is a member of a large gene cluster that includes a growth hormone variant gene as well as several structurally related chorionic somatomammotropin genes and pseudogenes, which are highly homologous but functionally inactive relatives of a normal gene. Because such gene clusters contain multiple homologous DNA sequences arranged along the same chromosome, they are particularly prone to undergo recombination and, consequently, gene duplication or deletion.

Unequal crossing-over between homologous genes can result in fusion gene mutations, as illustrated, for example, by glucocorticoid-remediable aldosteronism (GRA). GRA is caused by a rearrangement involving the genes that encode aldosterone synthase (CYP11B2) and steroid 11β-hydroxylase (CYP11B1), normally arranged in tandem on chromosome 8q. Because these two genes are 95% identical, they are predisposed to undergo unequal recombination. The rearranged gene product contains the regulatory regions of 11β-hydroxylase upstream to the coding sequence of aldosterone synthetase. The latter enzyme is then expressed in the adrenocorticotropic hormone (ACTH) dependent zona fasciculata of the adrenal gland, resulting in overproduction of mineralocorticoids and hypertension.

Gene conversion refers to a non-reciprocal exchange of homologous genetic information by which a recipient strand of DNA receives information from another strand having an allelic difference. The original allele on the recipient strand is converted to the new allele as a consequence of this event. These alterations may range from a few to several thousand nucleotides. Gene conversion often involves exchange of DNA between a gene and a related pseudogene. For example, the 21-hydroxylase gene (CYP21A1) is adjacent to a non-functional pseudogene. Many of the nucleotide substitutions found in the CYP21A gene in patients with CAH correspond to sequences present in the pseudogene, suggesting gene conversion as the underlying mechanism of mutagenesis. In addition, mitotic gene conversion has been suggested as a mechanism to explain revertant mosaicism in which an inherited mutation is “corrected” in certain cells.

**Trinucleotide repeats** may be unstable and expand beyond a critical number. Mechanistically, the expansion is thought to be caused by unequal recombination and slipped mispairing. A pre-mutation represents a small increase in trinucleotide copy number. In subsequent generations, the expanded repeat may increase further in length. This increasing expansion is referred to as dynamic mutation. It may be associated with an increasingly severe phenotype and earlier manifestation of the disease (anticipation). Trinucleotide expansion was first recognized as a cause of the fragile X syndrome, one of the most common causes of mental retardation. Malignant cells are also characterized by

---

**Figure 1.7** Examples of mutations. The coding strand is shown with the encoded amino acid sequence in the one-letter code and the three-letter code.
Mutations involving single nucleotides are referred to as point mutations (Fig. 1.7). Substitutions are called transitions if a purine is replaced by another purine base (A to G) or if a pyrimidine is replaced by another pyrimidine (C to T). Changes from a purine to a pyrimidine or vice versa are referred to as transversions. Certain DNA sequences, such as successive pyrimidines or CG dinucleotides, are particularly susceptible to mutagenesis. Therefore, certain types of mutations (C to T or G to A) are relatively common. Moreover, the nature of the genetic code results in overrepresentation of certain amino acid substitutions. If the DNA sequence change occurs in a coding region and alters an amino acid, it is called a missense mutation. Depending on the functional consequences of such a missense mutation, amino acid substitutions in different regions of the protein can lead to distinct phenotypes. Small deletions and insertions alter the reading frame if they do not represent a multiple of three bases. Such “frameshift” mutations lead to an entirely altered carboxy-terminus. Mutations may also be found in the regulatory sequences of genes and result in reduced gene transcription. Mutations in intron sequences or in exon junctions may destroy or create splice donor or splice acceptor sites.

Some mutations are lethal, some have less deleterious yet recognizable consequences and some confer evolutionary advantage. Mutations occurring in germ cells can be transmitted to the progeny. Alternatively, mutations can occur during embryogenesis or in somatic tissues. Mutations that occur during development lead to mosaicism, a situation in which tissues are composed of cells with different genetic constitutions, as illustrated by Turner syndrome or McCune–Albright syndrome. If the germline is mosaic, a mutation can be transmitted to some progeny but not others, which sometimes leads to confusion in assessing the pattern of inheritance. Other somatic mutations are associated with neoplasia because they confer a growth advantage to cells by activating (proto)oncogenes or inactivating tumor suppressor genes. Epigenetic events, heritable changes that do not involve changes in gene sequence (e.g. altered DNA methylation), may also influence gene expression or facilitate genetic damage.

Polymorphisms are sequence variations that have a frequency of at least 1% and do not usually result in an overt phenotype. Often they consist of single base pair substitutions that do not alter the protein coding sequence because of the degenerate nature of the genetic code, although some might alter mRNA stability, translation or the amino acid sequence. Silent base substitutions and SNPs are encountered frequently during genetic testing and must be distinguished from true mutations that alter protein expression or function. However, some SNPs or combinations of SNPs may have a pathogenic role in complex disorders by conferring susceptibility for the development of the disease.

**Functional consequences of mutations**

Mutations can broadly be classified as gain-of-function and loss-of-function mutations. The consequences of an altered protein sequence often need experimental evaluation in vitro to determine that the mutation alters protein function. The appropriate assay depends on the properties of the protein and may, for example, involve enzymatic analyses, electromobility shift experiments or reporter gene assays (Fig. 1.8).

Gain-of-function mutations are typically dominant and result in phenotypic alterations when a single allele is affected. Inactivating mutations are usually recessive and an affected individual is homozygous or compound heterozygous (i.e. carrying two different mutant alleles) for the disease-causing mutations. Mutation in a single allele can result in haploinsufficiency, a situation in which one normal allele is not sufficient to maintain a normal phenotype. Haploinsufficiency is a commonly observed mechanism in diseases associated with mutations in transcription factors. For example, monoallelic mutations in the transcription factor TTF1 (NKX2.1) are associated with transient congenital hypothyroidism, respiratory distress and ataxia.

The clinical features among patients with an identical mutation in a transcription factor often vary significantly. One mechanism underlying this variability consists of the influence of modifying genes. Haploinsufficiency can affect the expression of rate-limiting enzymes. For example, in MODY2 heterozygous glucokinase mutations result in haploinsufficiency with a higher threshold for glucose-dependent insulin release and mild hyperglycemia.

Mutation of a single allele can result in loss-of-function due to a dominant-negative effect. In this case, the mutated allele interferes with the function of the normal gene product by several different mechanisms. The mutant protein may interfere with the function of a multimeric protein complex, as illustrated by Liddle syndrome, which is caused by mutations in the β- or γ-subunit (SCCN1B, SCCN1G) of the renal sodium channel. In thyroid hormone resistance, mutations in the thyroid hormone receptor β (TRβ, THRα) lead to impaired T3 binding; the receptors cannot release co-repressors and they silence transcription of target genes. The mutant protein can be cytotoxic, as in autosomal-dominant neuronal hypohyphalgal diabetes insipidus, in which abnormal folding leads to retention in the endoplasmic reticulum and degeneration of neurons secreting arginine vasopressin (AVP).

An increase in dosage of a gene product may also result in disease. For example, duplication of the DAX1 (NR0B1) gene results in dosage-sensitive sex reversal.

**Genotype and phenotype**

An observed trait is referred to as a phenotype. The genetic information defining the phenotype is called the genotype. Alternative forms of a gene or a genetic marker are referred to as alleles, which may be polymorphic variants of nucleic acids that have no apparent effect on gene expression or function. In other instances, these variants may have subtle effects on gene expression, thereby conferring adaptive advantages or increased susceptibility. Commonly occurring allelic variants may reflect mutations in a gene that clearly alter its function, as illustrated, for example, by the ΔF508 deletion in the cystic fibrosis conductance regulator (CFTR).
Figure 1.8  Functional analysis of a mutation in the transcription factor PAX8 found in a child with thyroid hypoplasia and congenital hypothyroidism.

Upper panels: The point mutation 119A > C leads to a substitution of glutamine 40 by proline (Q40P) in the DNA binding domain of the transcription factor.

Left lower panel: Gel shift experiment. A radiolabeled DNA response element migrates very fast through the gel in the absence of a protein-DNA interaction (lanes 1–3). The wild type PAX8 protein binds to the responses element and this leads to an electromobility shift (lane 4). Cold oligo in excess can compete for the labeled oligo documenting that the interaction is specific (lane 5). The mutated protein is unable to bind to this response element (lane 6).

Right lower panel: Plasmid vectors encoding wild type or mutated PAX8 were transfected into embryonic kidney cells together with a luciferase reporter gene. The reporter gene consists of a plasmid containing a PAX8 response element upstream of the coding sequence for luciferase. The transcriptional stimulation of the luciferase gene can be determined by measuring the light emission of cell lysates incubated with the substrate luciferin. The example shows that the wild type protein stimulates transcription of the luciferase reporter gene (pCMX = control vector). In contrast, there is no significant induction by the mutant. Co-transfection of wild type and mutant plasmids in different ratios shows that the mutant does not have a dominant negative effect. (After Congdon et al. 2001 with permission.)

Because each individual has two copies of each chromosome, an individual can have only two alleles at a given locus. However, there can be many different alleles in the population. The normal or common allele is usually referred to as wild type. When alleles at a given locus are identical, the individual is homozygous. Inheriting such identical copies of a mutant allele occurs in many autosomal-recessive disorders, particularly in circumstances of consanguinity. If the alleles are different, the individual is heterozygous at this locus. If two different mutant alleles are inherited at a given locus, the individual is referred to as a compound heterozygote. Hemizygous is used to describe males with a mutation in an X-chromosomal gene or a female with a loss of one X-chromosomal locus.

A haplotype refers to a group of alleles that are closely linked together at a genomic locus. Haplotypes are useful for tracking the transmission of genomic segments within families and for detecting evidence of genetic recombination, if the crossover event occurs between the alleles.

Allelic and phenotypic heterogeneity

Allelic heterogeneity refers to the fact that different mutations in the same genetic locus can cause an identical or similar pheno-
type. **Phenotypic heterogeneity** occurs when more than one phenotype is caused by allelic mutations. For example, different mutations in the androgen receptor can result in a wide phenotypic spectrum. In some cases, the receptor is deleted or mutated in a manner that inactivates it completely. In a karyotypic male, this leads to testicular feminization. In contrast, the phenotype may be milder if the androgen receptor is only partially inactivated. In these patients, the phenotype may include infertility, gynecomastia or episidias. Allelic heterogeneity is explained by the fact that many different mutations are capable of altering protein structure and function. Allelic heterogeneity creates a significant problem for genetic testing because one must often examine the entire genetic locus for mutations, because these can differ in each patient.

**Locus or non-allelic heterogeneity and phenocopies**

**Non-allelic or locus heterogeneity** refers to the situation in which a similar disease phenotype results from mutations at different genetic loci. This often occurs when more than one gene product produces different subunits of an interacting complex or when different genes are involved in the same genetic cascade or physiological pathway. For example, congenital hypothyroidism associated with dysshormonogenesis can arise from mutations in several genes (NIS, TG, TPO, PDS/SLC26A4, DUOX2, DUOXA2, DEHAL1) located on different chromosomes. The effects of inactivating mutations in these genes are similar because the protein products are all required for normal hormone synthesis. Similarly, the genetic forms of diabetes insipidus can be caused by mutations in several genes. Mutations in the AVP-NPII gene cause autosomal-dominant or -recessive forms of neurohypophyseal diabetes insipidus. The nephrogenic forms can be caused by mutations in the X-chromosomal AVPR2 receptor gene, whereas mutations in the aquaporin 2 (AQP2) gene cause either autosomal-recessive or -dominant nephrogenic diabetes insipidus.

Recognition of non-allelic heterogeneity is important because the ability to identify disease loci in linkage studies is reduced by including patients with similar phenotypes but different genetic disorders. Genetic testing is more complex because several different genes need to be considered along with the possibility of different mutations in each of the candidate genes.

**Phenocopies** designate a phenotype that is identical or similar but results from non-genetic or other genetic causes. For example, obesity may be caused by several rare Mendelian defects, the result of a complex disorder or have a primarily behavioral origin. As in non-allelic heterogeneity, the presence of phenocopies has the potential to confound linkage studies and genetic testing. Patient history, subtle differences in clinical presentation and rigorous clinical testing are key in assigning the correct phenotype.

**Variable expressivity and incomplete penetrance**

**Penetrance and expressivity** are two different yet related concepts which are often confused. Penetrance is a qualitative notion designating whether a phenotype is expressed for a particular genotype. Expressivity is a quantitative concept describing the degree to which a phenotype is expressed. It is used to describe the phenotypic spectrum in individuals with a particular disorder. Thus, expressivity is dependent on penetrance.

Penetrance is complete if all carriers of a mutant express the phenotype, whereas it is said to be incomplete if some individuals do not have any features of the phenotype. Dominant conditions with incomplete penetrance are characterized by skipping of generations with unaffected carriers transmitting the mutant gene. For example, hypertrophic obstructive cardiomyopathy (HOCM) caused by mutations in the myosin-binding protein C gene is a dominant disorder with clinical features in only a subset of patients who carry the mutation. Incomplete penetrance in some individuals can confound pedigree analysis. In many conditions with postnatal onset, the proportion of gene carriers affected varies with age. Therefore, it is important to specify age when describing penetrance. Variable expressivity is used to describe the phenotypic spectrum in individuals with a particular disorder.

Some of the mechanisms underlying expressivity and penetrance include modifier genes (**genetic background**), gender and environmental factors. Thus, variable expressivity and penetrance illustrate that genetic and/or environmental factors do not influence only complex disorders but also “simple” Mendelian traits. This has to be considered in genetic counseling, because one cannot always predict the course of disease, even when the mutation is known.

**Sex-influenced phenotypes**

Certain mutations affect males and females quite differently. In some instances, this is because the gene resides on the X or Y sex chromosomes. As a result, the phenotype of mutated X-linked genes will usually be expressed fully in males but variably in heterozygous females, depending on the degree of X-inactivation and the function of the gene. Because only males have a Y chromosome, mutations in genes such as SRY (which causes male-to-female sex reversal) or DAZ (deleted in azoospermia), which causes abnormalities of spermatogenesis, are unique to males.

Other diseases are expressed in a sex-limited manner because of the differential function of the gene product in males and females. Activating mutations in the luteinizing hormone receptor (LHR) cause dominant male-limited precocious puberty in boys. The phenotype is unique to males because activation of the receptor induces testosterone production in the testis, whereas it is functionally silent in the immature ovary. Homozygous inactivating mutations of the follicle-stimulating hormone (FSH) receptor cause primary ovarian failure in females because the follicles do not develop in the absence of FSH action. Affected males have a more subtle phenotype, because testosterone production allowing sexual maturation is preserved and spermatogenesis is only partially impaired. In congenital adrenal hyperplasia, most commonly caused by 21-hydroxylase deficiency, cortisol production is impaired and ACTH stimulation of the adrenal gland leads to increased production of androgenic precursors. In females, the increased androgen concentration
causes ambiguous genitalia, which can be recognized at birth. In males, the diagnosis may be made on the basis of adrenal insufficiency at birth because the increased adrenal androgen level does not alter sexual differentiation or later in childhood because of the development of precocious puberty.

**Approach to the patient**
Clinical and biochemical evaluation is the first step in any attempt to unravel underlying pathogenic mechanisms. The family history is important to recognize the possibility of a hereditary component. For this purpose, it is extremely useful to draw a pedigree of the nuclear and, in some cases, of the extended family. This should include information about ethnic background, age, health status and deaths, particularly deaths in infancy which may have been forgotten. The physician should explore whether other individuals within the family are affected by the same or a related illness as the index patient. This should be followed by a survey for the presence of commonly occurring disorders.

Because of the possibility of age-dependent expressivity and penetrance, the family history may need updating on subsequent encounters. If the family history or other findings suggest a genetic disorder, the clinician has to assess whether some of the patient’s relatives may be at risk of carrying or transmitting the disease. This information may become of practical relevance for carrier detection, genetic counseling or early intervention and prevention of a disease in relatives.

Where a diagnosis at the molecular level may be available, the physician faces several challenges. Genetic testing in children poses distinct ethical issues. In general, it should be limited to situations in which it has an immediate impact on the medical management of that child; it requires informed consent by the parents. If there is no apparent benefit, testing should usually be deferred until the patient can consent independently. This is particularly relevant in devastating disorders that manifest only later in life, such as Huntington disease.

If genetic testing is considered an option, the physician will have to identify an appropriate laboratory to perform the test. The GeneTests website (http://www.genetests.org/servlet/access), a publicly funded medical genetics resource, contains an International Laboratory Directory which is useful for identifying approved laboratories offering testing for inherited disorders (Table 1). For rare disorders, the test may only be available through research laboratories.

If a disease-causing mutation is expected in all cells as a result of germline transmission, DNA can be collected from any tissue, most commonly nucleated blood cells or buccal cells, for cytogenetic and mutational analyses. In the case of somatic mutations, which are limited to neoplastic tissue, an adequate sample of the lesion will serve for the extraction of DNA or RNA. For the detection of pathogens, the material to be analyzed will vary and may include blood, cerebrospinal fluid, solid tissues, sputum or fluid obtained through bronchoalveolar lavage.

New findings on the genetic basis of endocrine disorders are published in numerous scientific journals, books and databases. The continuously updated OMIM catalog lists several thousand genetic disorders and provides information about the clinical phenotype, molecular basis, allelic variants and pertinent animal models (Table 1). Hyperlinks to other electronic resources (e.g. PubMed, GenBank or databases compiling gene mutations) provide access to useful information that is relevant for both clinicians and basic scientists.

**Chromosomal disorders**
Chromosomal (cytogenetic) disorders are caused by numerical or structural aberrations in chromosomes. Large duplications and deletions are well recognized as cause of specific genetic disorders. Molecular cytogenetics has led to the identification of more subtle chromosome abnormalities such as microdeletions and duplications, imprinting syndromes and genomic imbalances brought about by CNVs.

Errors in meiosis and early cleavage divisions occur frequently. Ten to 25% of all conceptions harbor chromosomal abnormalities, which often lead to spontaneous abortion in early pregnancy. Numerical abnormalities, especially trisomy, which is found in about 25% of spontaneous abortions and 0.3% of newborns, are more common than structural defects. Trisomy 21, the most frequent cause of Down syndrome, occurs in 1:600–1000 live births. Trisomies 13 and 18 are also frequent.

Numerical abnormalities in sex chromosomes are relatively common. Males with a 47,XXY karyotype have Klinefelter syndrome and females with trisomy 47,XXX may be subfertile. Autosomal monosomies are usually incompatible with life but 45, XO is present in 1–2% of all conceptuses but leads to spontaneous abortion in 99% of cases. Mosaicism (e.g. 45, XO/45,XX, 45,XO/45,XXX), partial deletions, isochromosomes and ring chromosomes can also cause Turner syndrome. Sex chromosome monosomy usually results from loss of the paternal sex chromosome. 47,XXX can result from maternal or paternal nondisjunction, while the autosomal trisomies are most commonly caused by maternal non-disjunction during meiosis I, a defect that increases with maternal age. Trisomies are typically associated with alterations in genetic recombination.

Structural rearrangements involve breakage and reunion of chromosomes. Rearrangements between different chromosomes, translocations, can be reciprocal or Robertsonian. Reciprocal translocations involve exchanges between any of the chromosomes; Robertsonian rearrangements designate the fusion of the long arms of two acrocentric chromosomes. Other structural defects include deletions, duplications, inversions and the formation of rings and isochromosomes. Deletions affecting several tightly clustered genes result in contiguous gene syndromes, disorders that mimic a combination of single gene defects. They have been useful for identifying the location of new disease-causing genes. Because of the variable size of gene deletions in different patients, a systematic comparison of phenotypes and locations of deletion breakpoints allows the positions of particular genes to be mapped within the critical genomic region. Structural chromosome defects can be present in a “balanced” form without an abnormal
phenotype. However, they can be transmitted in an “unbalanced” form to offspring and thus cause an hereditary form of chromosome abnormality.

Paternal deletions of chromosome 15q11-13 cause Prader–Willi syndrome (PWS), while maternal deletions are associated with Angelman syndrome. The difference in phenotype results from the fact that this chromosomal region is imprinted, i.e. differentially expressed on the maternal and paternal chromosome.

Traditional karyotype analysis usually identifies chromosomal rearrangements and/or aberrations of 3–5 Mb and larger. Comparative genomic hybridization and other techniques now permit the detection of more subtle, submicroscopic chromosomal imbalances such as CNVs. The clinical relevance of these alterations is not always known. The Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources (DECIPHER, http://www.sanger.ac.uk/PostGenomics/decipher) catalogs pertinent genomic and clinical information of such patients and can assist in the interpretation of genome-wide, high-resolution tests.

Acquired somatic abnormalities in chromosome structure are often associated with malignancies and are important for diagnosis, classification and prognosis. Deletions can lead to loss of tumor suppressor genes or DNA repair genes. Duplications, amplifications and rearrangements, in which a gene is put under the control of another promoter, can result in gain-of-function of genes controlling cell proliferation. For example, rearrangement of the 5′ regulatory region of the parathyroid (PTH) gene located on chromosome 11q15 with the cyclin D1 gene from 11q13 creates the PRAD1 oncogene, resulting in overexpression of cyclin D1 and the development of parathyroid adenomas.

Monogenic Mendelian disorders
Monogenic human diseases are often called Mendelian disorders because they obey the rules of genetic transmission defined by Gregor Mendel. The mode of inheritance for a given phenotype or disease is determined by pedigree analysis. About 65% of human monogenic disorders are autosomal dominant, 25% are autosomal recessive and 5% are X-linked. Genetic testing now available for many of these disorders has an increasingly important role in clinical medicine.

Autosomal-dominant disorders
In autosomal-dominant disorders, mutations in a single allele are sufficient to cause the disease; recessive disorders are the consequence of biallelic loss-of-function mutations. Various disease mechanisms are involved in dominant disorders, which include gain-of-function, a dominant-negative effect, and haploinsufficiency. In autosomal-dominant disorders, individuals are affected in successive generations and the disease does not occur in the offspring of unaffected individuals. Males and females are affected with equal frequency because the defective gene resides on one of the 22 autosomes. Because the alleles segregate randomly at meiosis, the probability that an offspring will be affected is 50%. Children with a normal genotype do not transmit the disorder. The clinician must be aware that an autosomal-dominant disorder can be caused by de novo germ-line mutations, which occur more frequently during later cell divisions in gametogenesis, explaining why siblings are rarely affected. New germline mutations occur more frequently in fathers of advanced age. The clinical manifestations of autosomal-dominant disorders may be variable as a result of differences in penetrance or expressivity. Because of these variations, it is sometimes difficult to determine the pattern of inheritance.

Autosomal-recessive disorders
The clinical expression of autosomal-recessive disorders is usually more uniform than in autosomal-dominant disorders. Most mutated alleles lead to a partial or complete loss-of-function. They frequently involve receptors, proteins in signaling cascades or enzymes in metabolic pathways. The affected individual, who can be of either sex, is homozygous or compound heterozygous for a single gene defect. In most instances, an affected individual is the offspring of heterozygous parents. In this situation, there is a 25% chance that the offspring will have a normal genotype, a 50% probability of a heterozygous state and a 25% risk of homozygosity for the recessive alleles. In the case of one unaffected heterozygous and one affected homozygous parent, the probability of disease increases to 50% for each child. In this instance, the pedigree analysis mimics an autosomal-dominant mode of inheritance (pseudodominance). In contrast to autosomal-dominant disorders, new mutations in recessive alleles usually result in an asymptomatic carrier state without apparent clinical phenotype.

Many autosomal-recessive diseases are rare and occur more frequently in isolated populations in the context of parental consanguinity. A few recessive disorders, such as sickle cell anemia, cystic fibrosis and thalassemia, are relatively frequent in certain populations, perhaps because the heterozygous state may confer a selective biological advantage. Although heterozygous carriers of a defective allele are usually clinically normal, they may display subtle differences in phenotype that become apparent only with more precise testing or in the context of certain environmental influences.

X-linked disorders
Because males have only one X chromosome, a female individual always inherits her father’s X chromosome in addition to one of the two X chromosomes of her mother. A son inherits the Y chromosome from his father and one maternal X chromosome. The characteristic features of X-linked inheritance are therefore the absence of father-to-son transmission and the fact that all daughters of an affected male are obligate carriers of the mutant allele. The risk of developing disease caused by a mutant X-chromosomal gene differs in the two sexes. Because males have only one X chromosome, they are hemizygous for the mutant

Genetics and Genomics
allele. Consequently, they are more likely to develop the mutant phenotype, regardless of whether the mutation is dominant or recessive. A female may be either heterozygous or homozygous for the mutant allele, which may be dominant or recessive, and the terms X-linked dominant or X-linked recessive apply only to the expression of the mutant phenotype in women. In females, the expression of X-chromosomal genes is influenced by X chromosome inactivation. This can confound the assessment because skewed X-inactivation may lead to a partial phenotype in female carriers of an X-linked recessive defect, such as inactivating mutations of the AVPR2 receptor, the cause of X-linked nephrogenic diabetes insipidus.

Y-linked disorders

The Y chromosome harbors relatively few genes. Among them, the sex region-determining Y factor (SRY), which encodes the testis-determining factor (TDF), is essential for normal male development. Because the SRY region is closely adjacent to the pseudoautosomal region, a chromosomal segment on the X and Y chromosomes with a high degree of homology, crossing-over can occasionally involve the SRY region. Translocations can result in XY females, with the Y chromosome lacking the SRY gene, or XX males harboring the SRY gene on one of the X chromosomes. Point mutations in the SRY gene may result in individuals with an XY genotype and an incomplete female phenotype. Men with oligospermia or azoospermia frequently have microdeletions of the AZF (azoospermia factor) regions on the long arm of the Y chromosome, which contain several genes involved in the control of spermatogenesis. They may have point mutations in the transcription factor DAZ (deleted in azoospermia), which is located in this chromosomal region.

Exceptions to simple Mendelian inheritance

Mitochondrial disorders

Mendelian inheritance refers to the transmission of genes encoded by DNA in nuclear chromosomes but each mitochondrion contains several copies of a circular chromosome. mtDNA is small (16.5 kb) and encodes transfer and ribosomal RNAs and 13 proteins that are part of the respiratory chain involved in oxidative phosphorylation and ATP generation. In contrast to the nuclear chromosomes, the mitochondrial genome does not recombine and is inherited through the maternal line because sperm does not contribute significant cytoplasmic components to the zygote. The D-loop, a non-coding region of the mitochondrial chromosome, is highly polymorphic. This property, together with the absence of recombination of mtDNA, makes it a helpful tool for studies tracing human migration and evolution and for specific forensic applications.

Inherited mitochondrial disorders are transmitted in a matrilineal fashion. All children from an affected mother inherit the disease but it will never be transmitted from an affected father to his offspring except by intracytoplasmic sperm injection (ICSI). Alterations in the mtDNA affecting enzymes required for oxidative phosphorylation lead to reduction of ATP supply, generation of free radicals and induction of apoptosis. Several syndromic disorders arising from mutations in the mitochondrial genome are known in humans and they affect both protein-coding and tRNA genes. The pleiotropic clinical spectrum often involves (cardio)myopathies and encephalopathies because of the high dependence of these tissues on oxidative phosphorylation.

Many may present with endocrine features. For example, the mitochondrial DIDMOAD syndrome consists of diabetes insipidus, diabetes mellitus, optic atrophy and deafness. The age of onset and the clinical course are variable because of the unusual mechanisms of mtDNA replication. mtDNA replicates independently from nuclear DNA and, during cell replication, the proportion of wild type and mutant mitochondria can drift among different cells and tissues. The resulting heterogeneity in the proportion of mitochondria with and without a mutation is referred to as heteroplasmia and underlies the phenotypic variability that is characteristic of mitochondrial diseases. Nuclear genes that encode proteins that are important for normal mitochondrial function can cause mitochondrial dysfunctions associated with autosomal-dominant or -recessive forms of inheritance.

Acquired somatic mutations in mitochondrial genes are thought to be involved in several age-dependent degenerative disorders involving muscle and the peripheral and central nervous systems. Because of the high degree of polymorphisms in mtDNA and the phenotypic variability of these disorders, it is difficult to establish that an mtDNA alteration is causal for a clinical phenotype.

Trinucleotide expansion disorders

Several diseases are associated with an increase in the number of trinucleotide repeats above a certain threshold. In some instances, the repeats are located within the coding region of the genes. For example, an expansion in a CAG repeat in the androgen receptor, which encodes a polyglutamine motif in its amino-terminus, leads to the X-linked form of spinal and bulbar muscular atrophy (SBMA, Kennedy syndrome) and can be associated with partial androgen insensitivity. Similarly, an expansion in the huntingtin (HD) gene is the cause of Huntington disease. In other instances, the repeats are located in regulatory sequences. If an expansion is present, the DNA fragment is unstable and tends to expand further during cell division; hence the designation dynamic mutation. The length of the nucleotide repeat often correlates with the severity of the disease. When repeat length increases from one generation to the next, disease manifestations may worsen or appear at an earlier age, a phenomenon referred to as anticipation. In Huntington disease, for example, there is a correlation between age of onset and length of the triplet codon expansion.

Mosaicism

Mosaicism refers to the presence of two or more genetically distinct cell lines in the tissues of an individual. It results from a mutation that occurs during embryonic, fetal or extraterine development. The developmental stage at which the mutation arises will determine whether germ cells and/or somatic cells are
involved. Chromosomal mosaicism results from non-disjunction at an early embryonic mitotic division, leading to the persistence of more than one cell line, as exemplified by some patients with Turner syndrome. Somatic mosaicism is characterized by a patchy distribution of genetically altered somatic cells that occurs early in development. This is best illustrated by the McCune–Albright syndrome, which is caused by activating mutations in the GNAS1 gene encoding the stimulatory G-protein α (Gαs). The clinical phenotype varies depending on the tissue distribution of the mutation. Manifestations include ovarian cysts that secrete sex steroids and cause precocious puberty, polyostotic fibrous dysplasia, café-au-lait skin pigmentation, growth hormone-secreting pituitary adenomas and, among others, hypersecreting autonomous thyroid nodules.

Epigenetic modifications, X-inactivation, imprinting and uniparental disomy

According to traditional Mendelian principles, the parental origin of a mutant gene is irrelevant for the expression of the phenotype, although there are important exceptions to this rule. X-inactivation prevents the expression of most genes on one of the two paternal chromosomes (Angelman syndrome). The inheritance of either two maternal chromosomes 15 (PWS) or uniparental disomy prevents the expression of most genes on one of the autosomes. This phenomenon, genomic imprinting, leads to preferential expression of an allele depending on its parental origin. It is of importance in disorders in which the transmission of disease is dependent on the sex of the transmitting parent and has an important role in the expression of certain genetic disorders.

The two classic examples are the Prader–Willi and Angelman syndromes. PWS is characterized by diminished fetal activity, obesity, hypotonia, mental retardation, short stature and hypogonadotropic hypogonadism. Deletions in PWS occur exclusively on the paternal chromosome 15. Patients with Angelman syndrome present with mental retardation, seizures, ataxia and hypotonia and have deletions at the same site of chromosome 15 but they are located on the maternal chromosome 15. These two syndromes may also result from uniparental disomy, i.e. by the inheritance of either two maternal chromosomes 15 (PWS) or two paternal chromosomes (Angelman syndrome).

Another example of importance for pediatric endocrinology concerns the GNAS1 gene encoding the Gαs subunit. Heterozygous loss-of-function mutations in the GNAS1 gene lead to Albright hereditary osteodystrophy (AHO) with its characteristic features including short stature, obesity, round face, brachydactyly, subcutaneous ossifications and mental deficits. Paternal transmission of GNAS1 mutations leads to the AHO phenotype alone (pseudopseudohypparathyroidism), while maternal transmission leads to AHO in combination with resistance to several hormones such as PTH, thyroid-stimulating hormone (TSH) and gonadotropins, which act through transmembrane receptors coupling to Gαs (pseudohypparathyroidism type IA). These phenotypic differences result from a tissue-specific imprinting of GNAS1, which is expressed primarily from the maternal allele in tissues such as the proximal renal tubule and the thyroid. In most other tissues, however, it is expressed biallelically. Disrupting mutations in the maternal allele lead to loss of Gαs expression in proximal tubules and loss of PTH action in the kidney, while mutations in the paternal allele have little effect on PTH action. In patients with isolated renal resistance to PTH (pseudohypparathyroidism type IB), an imprinting defect of GNAS1 leads to decreased Gαs expression in the proximal renal tubules.

Somatic mutations

Acquired mutations that occur in somatic rather than germ cells are called somatic mutations. This creates a chimeric situation and, if the cells proliferate, a neoplastic lesion. Therefore, cancer can be defined as a genetic disease at the cellular level. Cancers are monoclonal, indicating that they have arisen from a single precursor cell that has acquired one or several mutations in genes controlling growth and/or differentiation. These mutations are somatic, i.e. restricted to the tumor and its metastases, but not found in surrounding normal tissue. The molecular alterations include dominant gain-of-function mutations in oncogenes, recessive loss-of-function mutations in tumor suppressor genes and DNA repair genes, gene amplification and chromosome rearrangements. Rarely, a single mutation in certain genes may be sufficient to transform a normal cell into a malignant cell but the development of a malignant phenotype in most cancers requires several genetic alterations for the gradual progression from a normal to a cancerous cell, a process termed multistep carcinogenesis.

In many cancer syndromes, there is an inherited predisposition to tumor formation. In these instances, a germline mutation is inherited in an autosomal-dominant fashion. This germline alteration affects one allele of an autosomal tumor suppressor gene. If the second allele is inactivated by a somatic mutation in a given cell, this will lead to neoplastic growth (Knudson two-hit model). In this instance, the defective allele in the germline is transmitted in a dominant way, whereas the tumorigenic mechanism results from a recessive loss of the tumor suppressor gene in affected tissues.

The classic example to illustrate this phenomenon is retinoblastoma, which can occur as a sporadic or hereditary tumor. In sporadic retinoblastoma, both copies of the retinoblastoma (RB) gene are inactivated through two somatic events. In hereditary retinoblastoma, one mutated or deleted RB allele is inherited in an autosomal-dominant manner and the second allele is inactivated by a subsequent somatic mutation. This “two-hit” model applies to other inherited cancer syndromes, such as multiple endocrine neoplasia type 1 (MEN1), which is caused by mutations in the tumor suppressor gene menin.

Inherited defects in enzymes involved in DNA replication and repair can lead to a significant increase in mutations and are associated with several disorders predisposing to cancer.

Complex disorders

Many disorders have a complex etiology involving multiple genes (polygenic disorders), often in combination with environmental
and lifestyle factors (multifactorial disorders). The major health care problems, cardiovascular disease, hypertension, diabetes, obesity, asthma and psychiatric disorders, fall into this category but it also includes certain developmental abnormalities, such as cleft palate, congenital heart defects and neural tube defects.

Compared with single gene defects, complex disorders have a low heritability and do not fit a Mendelian pattern of inheritance. Twin studies are particularly helpful in demonstrating the importance of genetic and environmental factors. For example, first-degree relatives of patients with diabetes mellitus type 1 are about 15 times more likely to develop diabetes. The concordance rate for developing diabetes is about 50% in monozygotic twins and about 8% in dizygotic twins. The discordance rate in monozygotic twins illustrates the significant requirement for environmental factors. In addition, some of the susceptibility genes, a designation indicating that the carrier is susceptible to develop the disease, have a low penetrance. Susceptibility genes or loci can be mapped using several methods, including linkage analyses, association studies and affected sib-pair analyses. Current efforts aim to identify these genes by establishing correlations between SNPs or SNP haplotypes and complex disorders in large populations through GWAS. The HapMap data are significantly facilitating this type of study because they allow genotyping a reduced number of tag SNPs reflecting certain haplotypes (Fig. 1.6). The results of GWAS may, in part, depend on ethnicity and ascertainment criteria.

The study of rare monogenic diseases may also provide insights into genetic and molecular mechanisms important for the understanding of complex disorders. For example, the identification of the genetic defects underlying the various autosomal-dominant forms of MODY have defined them, in part, as candidate genes contributing to the pathogenesis of diabetes mellitus type 2.

**Genomics and post-genomic techniques**

Broadly defined, genomics designates the discipline of mapping, sequencing and analyzing genomes. The completion of the structural analysis of the human genome (structural genomics) has been followed by a rapid emergence of “postgenomic” disciplines focusing on biological function of the gene products (functional genomics). These disciplines are concerned with analyses of gene transcripts (transcriptomics), proteins and their secondary modifications and interactions (proteomics), epigenetic modifications of DNA and chromatin proteins (epigenomics), metabolites and their networks (metabolomics) and comprehensive analyses of the genomes of microorganisms populating specific compartments (metagenomics). The ultimate goal is the integration of these complementary data into a systems biology that permits a comprehensive definition of the phenotype and pathophysiologic perturbations.

What can be expected from these developments? Genotyping may become important for stratifying patients according to disease risk and for predicting the response to certain drugs. Gene expression studies can be used for the assessment of prognosis and for guiding therapy. Proteomic studies may allow diagnosis of early stages of malignancy. Because most drug targets are proteins, proteomics will be important for drug discovery and development. These technologies, individually and in combination, permit first insights into the pathogenesis of complex disorders. Genomic approaches may have further impact on health care as a result of a thorough understanding of the genomes and proteins of infectious agents (e.g. Plasmodium falciparum or Mycobacterium tuberculosis), which may lead to the development of novel therapeutic strategies and compounds. Comprehensive genomic analyses of the metagenome are expected to provide new insights in the interactions between the host and the microbial environment in health and disease.

**Comparative genomics**

Comparative genomics involves the analysis of two or more genomes to identify the extent of similarity or large-scale screening of a genome to identify sequences present in another genome. Applications involve comparisons of prokaryotic and eukaryotic genomes to infer evolutionary relationships. The detection of high evolutionary conservation can be used as a screen for regulatory elements within otherwise poorly conserved non-coding DNA. Electronic screening of expressed sequence tags (EST) databases can identify homologs of genes in other species. For example, systematic screening of the dbEST database of ESTs has revealed many relevant human homologs of Drosophila genes known to be loci for mutant phenotypes.

**Pharmacogenomics**

Broadly defined, the scope of pharmacogenetics and pharmacogenomics is to define how the genome influences the response of an individual to a drug. Although many non-genetic factors influence the effects of medications, genetic polymorphisms in receptors, transporters, channels and enzymes can result in variable absorption, distribution, metabolism and excretion of a drug that ultimately lead to differential response or toxic concentrations. For example, a polymorphism in thiopurine methyltransferase (TMT) inactivates the enzyme and is associated with hematopoietic toxicity of mercaptopurine. Determination of the TMT genotype is therefore important for choosing a safe dose of the medication. In other instances, drug effects will be influenced by polymorphisms in multiple genes.

Further development of genotyping and pharmacogenomics may improve the safety of medical therapy by choosing appropriate medications and dose levels, thereby decreasing the number of adverse drug reactions. It is also expected that this discipline will have an impact on drug development because screenings of SNPs can be used to enroll or exclude subgroups of patients.

**Transcriptomics**

mRNA expression of one or a few genes has usually been determined by Northern blot analysis, a technique that has been largely replaced by semiquantitative RT-PCR, but both techniques permit
only the analysis of the expression pattern of a limited number of genes. Paralleling the characterization of the genomic sequences of humans and other organisms, as well as the genes that they encode, various expression profiling techniques have been developed. These analyses enable surveys of gene expression patterns for thousands of genes in a single assay (Fig. 1.9). Such profiles are useful for the understanding of gene regulation and interactions in normal and pathologic tissues. As the complement of mRNAs transcribed by the cellular genome is also referred to as the transcriptome, the generation of mRNA expression profiles is now also referred to as transcriptomics.

The most widely used techniques for expression profiling include microarrays and serial analysis of gene expression (SAGE). After hybridization with the labeled probes, the microarrays are scanned and special software allows analysis of the fluorescence intensities for each spot (Fig. 1.9). The limitations of microarray technology include relatively high cost, special equipment and the inability to detect novel transcripts but SAGE is a powerful tool that allows a comprehensive analysis of gene expression patterns without the requirement of pre-existing probes or sophisticated equipment.

The simplest way to identify genes of potential interest by expression profiling is to search for those that are consistently upregulated or downregulated. The identification of patterns of gene expression and regulated classes of genes may, however, provide more informative insights into their biological function and relevance. Genes that are part of a particular pathway or that respond to a common endogenous or exogenous stimulus are expected to be co-regulated and should consequently show similar patterns of expression. Several computational techniques, such as hierarchical clustering, self-organizing maps and mutual information, are used for the analysis of gene expression data. The choice of the appropriate algorithm(s) for these analyses is a crucial element of the experimental design and the methods that are used to analyze the data can have a profound influence on the interpretation of the results.

**Epigenomics**

Eukaryotic genomes contain modifications of DNA or chromatin proteins. The totality of these epigenetic marks is designated as an epigenome. In contrast to the genome, the epigenome is highly variable between cells and changes within a single cell over time. Epigenetic modifications vary among different regions throughout the genome and result in alterations in levels of gene transcription. Thus, epigenetic changes and inheritance result in phenotypic consequences without changing the DNA sequence. Importantly, epigenetic marks can be heritable. They are propagated to the daughter cells during cell division and certain loci within the genome are paternally or maternally imprinted.

Epigenetic modifications consist of methylation of DNA, acetylation or methylation of histone proteins and interaction of proteins with histones. In eukaryotes, DNA methylation is found exclusively at cytosine residues. DNA methylation is involved in the repression of genes in the inactivated X chromosome (increased methylation in promoter regions, decreased methylation in intragenic regions) and it has an important role in imprinting, the establishment and maintenance of the allele-specific expression of maternally or paternally inherited genes. In mammalian genomes, the majority of the genome is methylated and all categories of DNA sequences, such as genes, intergenic regions and transposons (segments of DNA capable of independent replication and insertion of a copy into a new position within the same or another chromosome), can be the target of methylation. Unmethylated domains account for only about 2% of the genome and often consist of so-called CpG islands, DNA regions of ∼1000 bp with a high occurrence of the dinucleotide CG. CpG islands are very frequently found in promoters and they are usually unmethylated, at least in germline DNA.

Although most CpG islands remain unmethylated, independent of expression state, a minority become methylated during development resulting in silencing of the associated gene. During mitosis, the DNA replication results in the generation of a methylated and an unmethylated strand; the hemimethylated DNA is subsequently converted to a fully methylated state by methyltransferases and the methylation status is therefore maintained.
The importance of conserving the methylation status is illustrated by the fact that mutations in methyltransferases result in disease. Mutations in the MECP2 result in Rett syndrome (autism, dementia, ataxia) and mutations in the DNMT3B cause the immunodeficiency–centromeric instability–facial anomalies syndrome.

On the inactive X chromosome, a large number of CpG islands become heavily methylated and the associated genes are no longer expressed. Without inactivation of one X chromosome in every cell of a female, the differences in the number of X chromosomes between females and males would result in the expression of the double amount of each gene product in females. In order to avoid this, dosage compensation is required. The inactivation of one X chromosome is dependent on the X inactivation center (Xic). The Xic locus is necessary to count the number of X chromosomes and it ensures that all but one X chromosome are inactivated. It also contains a gene called Xist. The Xist gene codes for an RNA that is not translated into a protein. The Xist RNA binds to one of the X chromosomes, which results in secondary events that silence the chromosome through methylation and deacetylation of histones. Once the inactivation of one of the X chromosomes has occurred, the pattern is passed on to descendant cells during mitosis.

Differential methylation of paternal and maternal alleles results in imprinting and results in a difference in the expression between the alleles inherited from each parent. Methylation is usually associated with inactivation of the imprinted gene. Remarkably, the imprinting pattern is inherited. Because one chromosome is imprinted, a lack of the other chromosome or the inheritance of two copies from the same parent can result in disease if it involves an imprinted chromosome or locus (e.g. PWS and Angelman syndrome). The faithful inheritance of the methylation pattern in germ cells involves a two-step process during gametogenesis (Fig. 1.10). First, the methylation is erased by a genome-wide demethylation and then the pattern specific for each sex is imposed on both alleles. This ensures that the pattern in the zygote is identical to the one originally present in the paternal and maternal somatic cells (Fig. 1.10).

Overall, the understanding of how epigenetic variation impacts health and disease is in its early stages. It is established that aberrant DNA methylation is involved in the development of malignancies. Deletion of imprinted loci or uniparental disomy are well recognized as a cause of a series of syndromic phenotypes. Significant differences in DNA methylation have also been reported in monozygotic twins. Epigenetic phenomena may also be of importance in the pathogenesis of complex disorders. Further insights into the contribution of epigenetic modifications in the development of disease are expected to be gained in the near future through comprehensive analyses at the level of the whole genome with the newly available high-throughput technologies.

**Proteomics**

The term proteome designates the complete set of proteins expressed by the genome. Proteomics includes studies focusing on the expression and function of the proteome, as well as aspects of structural biology. The study of the proteome is difficult because it is so dynamic. Apart from the expression of isoforms, proteins undergo a plethora of secondary modifications and protein–protein interactions (interactome) and they form higher order complexes.

Comprehensive analyses of the proteome relied initially on protein separation by two-dimensional gel electrophoresis with subsequent mass spectrometric identification of protein spots. This approach is constrained to the most abundant proteins in the sample. Studies of the proteome are now performed predominantly with direct mass spectrometric analyses, which have undergone technological refinement and can identify ever smaller amounts of proteins from complex mixtures.

The classic methodology for studying protein–protein interactions was the yeast two-hybrid system. Currently, various protein- and antibody-based arrays are emerging to study protein activities, secondary modifications and interactions. Structural proteomics has the ambitious goal of systematically understanding the structural basis for protein interactions and function.

Clinical proteomics aims to use proteomic patterns for disease detection and surveillance. In this approach, high-throughput mass spectrometry generates a proteomic fingerprint of a diagnostic sample, such as serum, fine-needle or nipple fluid aspirate. Bioinformatic pattern recognition algorithms can then be applied to identify patterns of protein alterations that can discriminate benign from malignant tissues. Importantly, the specific pattern itself could be diagnostic and the underlying identities of the proteins that comprise the patterns do not need to be known. As well as identifying novel diagnostic and prognostic biomarkers for human cancer, proteomics is also expected to have an impact on drug discovery and action, given that most drugs target proteins and subsequently modify intracellular networks.

Inspired by the success of the Human Genome Project, the Human Proteome Organization (HUPO) aims at coordinating proteomic research (http://www.hupo.org/). Major current goals in proteomics include definition of the plasma proteome, analyses of specific cell types, mapping of organelle compositions, generation of antibodies to all human proteins, generation of protein interaction maps and analyses of important model and pathogenic organisms.

**Metabolomics**

The metabolome can be defined as the quantitative complement of all the low-molecular-weight molecules present in cells in a particular physiological or developmental state. While metabolomics is complementary to transcriptomics and proteomics, there are several attractive reasons for analyzing the metabolome. Relative to alterations in the transcriptome and the proteome, changes in the metabolome are often amplified. Moreover, alterations in metabolic fluxes are not regulated by gene expression alone and, reflecting the activities of the cell at a functional level, affect the concentrations of numerous individual metabolites. High-throughput analyses of metabolites can be performed with tools such as nuclear magnetic resonance spectroscopy and mass spec-
Figure 1.10 Imprint reset. Certain genomic regions are imprinted in a parent-specific fashion. The unmethylated chromosomal region is actively expressed, the methylated region is silenced and not expressed. In the germline, the imprint is reset. In this case, both chromosomes are unmethylated in the maternal germline and methylated in the paternal germline. In the zygote, the imprinting pattern is again identical with the situation in the somatic cells of the parents.

Bioinformatics

The enormous amounts of diverse biological data generated by recent biotechnological advances have led to the development and evolution of bioinformatics, in which biology and information technology converge. Initially, bioinformatics focused on the development and creation of nucleotide and protein databases and methods for the analysis of the deposited sequences. The application of bioinformatics for analysis of nucleotide and polypeptide sequences is well established and widely used. The largest of these sequence databases include GenBank at the National Center for Biotechnology Information (NCBI), Ensembl at the European Molecular Biology Laboratory (EMBL), the DNA Data Bank of Japan and SwissProt, among others (Table 1.1). They permit rapid retrieval of sequence information of genomic DNA, mRNA, ESTs, SNPs or polypeptides for a rapidly growing number of species. The evolution of these databases has been accompanied by expanding capabilities to annotate sequences, linking the data with other electronic resources and more sophisticated tools for analysis of nucleotide and protein sequences. It is crucial that bioinformatics software development is linked at an early stage...
through agreed documentation, standardized rules for structuring web forms [eXtensible Markup Language (XML)] and controlled vocabularies that allow different tools to exchange primary data sets.

Navigation in this web of continuously evolving databases is often intimidating but the high degree of interconnection between the multitude of databases permits relatively easy exploration of this knowledge. User guides and the online NCBI handbook (Table 1.1) provide helpful instructions to questions such as the following:

1. How does one find a gene of interest and determine the structure of this gene?
2. How can one find information about SNPs?
3. How can one find all the members of a human gene family?
4. For a given protein, how can one determine whether it contains any functional domains of interest?
5. What other proteins contain the same functional domains as this protein?
6. How can one determine whether there is a similarity to other proteins, not only at the sequence level but also at the structural level?

Sequence alignments are performed most easily with the Basic Local Alignment Search Tool (BLAST), which compares a DNA or polypeptide sequence of interest with nucleotide or protein databases. In addition to determining the identity of an isolated nucleotide or protein fragment, this approach can detect similar related sequences in one or several organisms. This type of comparison is the basis for the development of gene and protein families and unearls evolutionary relationships. The sensitivity and specificity of a BLAST search can be modified in such a way that even discrete homologies can be unraveled. By referring to databases of known regulatory element sequences, computer programs can inspect genomic sequences for the presence of regulatory elements. Motifs (i.e. specific amino acid patterns associated with defined functions) can be identified by submitting a polypeptide to computational analysis. This may permit assignment of a protein to functional and structural families and making predictions on the functional role of newly isolated proteins.

As more protein structures are identified, the relationship between structure and function becomes easier to predict. The development of more accurate algorithms for predicting and modeling secondary and tertiary structures is, in part, moving out of the laboratory and into the hands of bioinformaticists.

The field of bioinformatics is now challenged to integrate the data generated by the various “-omic” techniques with the hope of elucidating the functional relationships between genotype and observed phenotype, thereby permitting a system-wide analysis from genome to phenome.

Acknowledgments

This work has, in part, been supported by 1R01DK63024-01 from NIH/NIDDK.

Bibliography


