The Human Genome: An Introduction

Article *in* The Oncologist · February 2001

DOI: 10.1634/theoncologist.2001-0100 · Source: PubMed

The Oncologist
Fundamentals of Cancer Medicine

The Human Genome: An Introduction

JEROEN AERSSENS, MARTIN ARMSTRONG, RON GILISSEN, NADINE COHEN

Department Pharmacogenomics, Janssen Research Foundation, Beerse, Belgium

INTRODUCTION

During the past two decades, tremendous progress has been made in genetics and genomics. Diseases that run in families have been recognized for many centuries, but it was only in the early 1980s that the first mutations in a gene responsible for a disease could be identified. Subsequently, numerous discoveries of disease-related mutations in other genes have been found, initially in rare single-gene disorders, but more recently also in common disorders such as Alzheimer's disease and cancer. Applications of this newly discovered information provide new opportunities for progression in medical science, including the design of genetic tests to diagnose or predict (subtypes of) diseases, the redefinition of diseases and the understanding of their pathogenesis based on the molecular mechanisms behind them, and the selection of new target molecules for drug discovery. The time has now come that such applications will transfer further toward the day-to-day practice of the clinician and transform the practice of clinical medicine. For this to happen, an appropriate education and understanding of the basic concepts of genetics and genomics by clinicians is needed. This article aims to provide a general introduction of these concepts for clinicians not familiar with these fields. The list of references and the websites indicated in the manuscript should encourage the reader to get a broader appreciation of this research area.

WHAT GENETICISTS ARE ALWAYS TALKING ABOUT: DNA

Our inherited information is encoded in a macromolecule called "DNA" (deoxyribonucleic acid). Basically, our DNA is like a bar code—it encrypts information. The vast majority of DNA molecules are stored within the nucleus of the cell and covered with proteins, together forming chromosomes. DNA completely dissolves in aqueous solutions (and thus in its natural environment). When precipitated in alcoholic solvents (ethanol, isopropanol) during extraction procedures, the DNA becomes visible as a white viscous clot.

All multicellular organisms, including humans, start their life as a single cell (the fertilized egg), and almost all cells of the organism developed from this single cell contain a full and identical copy of the DNA of this single cell. In humans, each somatic cell contains approximately 0.006 nanogram of DNA which harbors all the genetic information needed to develop and function normally. For an adult human body, this makes a total of about 600 grams of DNA. For clinical genetic analyses, genomic DNA is usually isolated from leukocytes, although identical DNA could be isolated from virtually any other cell of an individual, as well. From 5 to10 ml whole blood, approximately 100-200 micrograms of DNA can be extracted, which is in most cases more than sufficient to perform genetic analyses.

THE DNA IS IN THE CHROMOSOMES: TWO COPIES OF A LIBRARY

The DNA in our cells is divided over a constant number of chromosomes (46 in humans), each of them with a specific size and form, as can be observed under the microscope using specific coloring techniques (e.g., Giemsa staining). Two sex chromosomes (X and Y) determine the gender of an individual (XX for females, XY for males); the other 44 so-called autosomal chromosomes are not different in males and females. Together, these 46 chromosomes comprise two nearly identical copies of the whole genome: one copy of the genome is inherited from the father (via a set of 22 autosomal

Correspondence: Jeroen Aerssens, Ph.D., Department of Pharmacogenomics, Janssen Research Foundation, Turnhoutseweg 30, Beerse B-2340, Belgium. Telephone: 32-14-606146; Fax: 32-14-607162; e-mail: jaerssen@janbe.jnj.com Received September 29, 2000; accepted for publication November 28, 2000. ©AlphaMed Press 1083-7159/2001/$5.00/0

chromosomes and an X or Y chromosome) and the other copy from the mother (via another set of 22 autosomal chromosomes and one X chromosome).

The autosomal chromosomes derived from the father and mother are two-by-two homologous: they look similar under the microscope and comprise the same genes, or eventually, variants of the same genes. These chromosomes are numbered from 1 to 22, mainly based on size (1 being the largest and 22 the smallest chromosome). Thus, each somatic cell contains two copies of each of these 22 chromosomes, and thus two copies of each of the genes located on these chromosomes. One could compare this with a library which contains two copies of each book, although there might sometimes be different editions of each specific book. In women, the two X chromosomes are also homologous: one copy is inherited from their mother and the other copy from their father. Men, on the contrary, have one X chromosome inherited from their mother and one Y chromosome inherited from their father. The Y chromosome is much smaller than the X chromosome and contains many fewer genes as well.

In germ cells, only one copy of each homologous chromosome is present (thus in total, 22 autosomal and one sex chromosome). During reproduction, two germ cells (one egg cell and one sperm cell) combine their genetic information so that the offspring will contain two copies of each chromosome: one from the father and one from the mother (Fig. 1). Thus, the gender of the offspring, determined by the combination of the sex chromosomes of the mother (always X) and the father (X or Y) is completely dependent on whether the sperm cell contains an X or Y chromosome.

The Size and Structure of DNA—A Double Helix

From a structural point of view, the DNA looks like a long chain of connected letters without any spaces or punctuation marks (Fig. 2A). The total physical length of all the DNA chains in each of our cells is approximately two meters, with a diameter of 0.000002 mm. In order to write the DNA text, the body has at its disposal four different but related building blocks (called nucleotides or, more precisely, deoxyribonucleotides): A, C, G, and T, representing respectively adenine, cytosine, guanine, and thymine. These nucleotides are connected by a deoxyribose-phosphate backbone. Each phosphate links the hydroxyl group on the 3′ carbon atom of a deoxyribose of one nucleotide to the hydroxyl group on the 5′ carbon atom in the deoxyribose group of the adjacent nucleotide. Importantly, all the information content encrypted within the DNA is in the specific sequence of these nucleotides. The information stored within the DNA code can be used for translation into functional activity (i.e., production of proteins) only in one orientation on the backbone, namely the 5′-to-3′ direction. Therefore, nucleotide sequences are usually displayed from the 5′-to-3′ end (from left to right).

Attached to this DNA strand is a second DNA strand which is the exact complement of the first one. This is possible because of the complementary chemical structure of the DNA building blocks. Two kinds of base pairs, often referred to as complementary base pairs, exist in all DNA: the As on one strand always pair with Ts on the other strand (via two hydrogen bonds) while Cs pair with Gs (via three hydrogen bonds). Thus, if the sequence of one strand is known, the sequence of the complementary strand can easily be

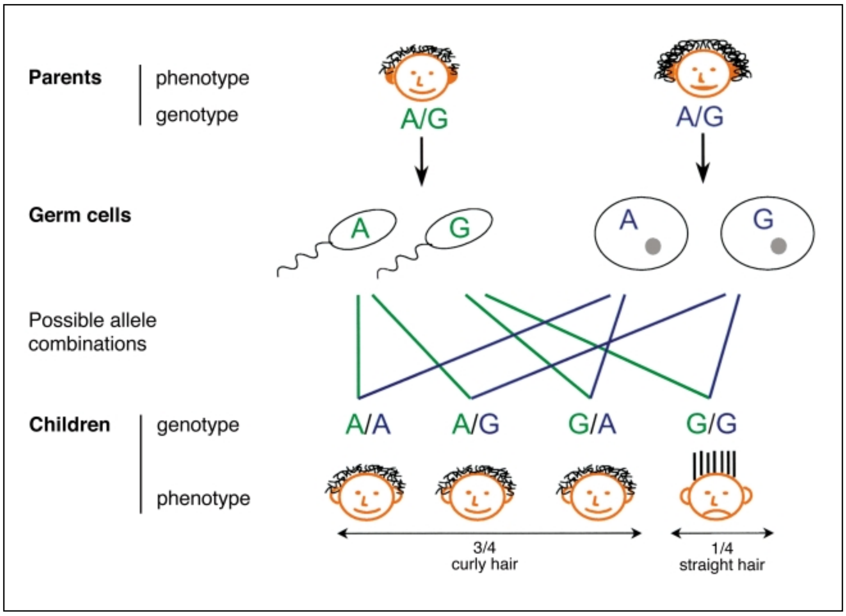

Figure 1. Schematic overview of the inheritance of our genetic information. *Each individual has two copies of each chromosome (each harboring one copy of the genes on the chromosome). The germ cells comprise only one of these copies. Through combination of the genetic material from the sperm cell and egg cell, the offspring inherits one copy of each chromosome from the father and one from the mother. Assume a particular gene in the DNA which determines the phenotype hair style, and for which two variants exist (A and G). The genotype, which is the combination of the variants on the two inherited homologous chromosomes in an individual (e.g., A/G), will determine the phenotype. In the example, the G/G genotype is linked with straight hair, while individuals with the other genotypes (A/G and A/A) have curly hair (indicating that the G variant, associated with straight hair, is a recessive characteristic).*

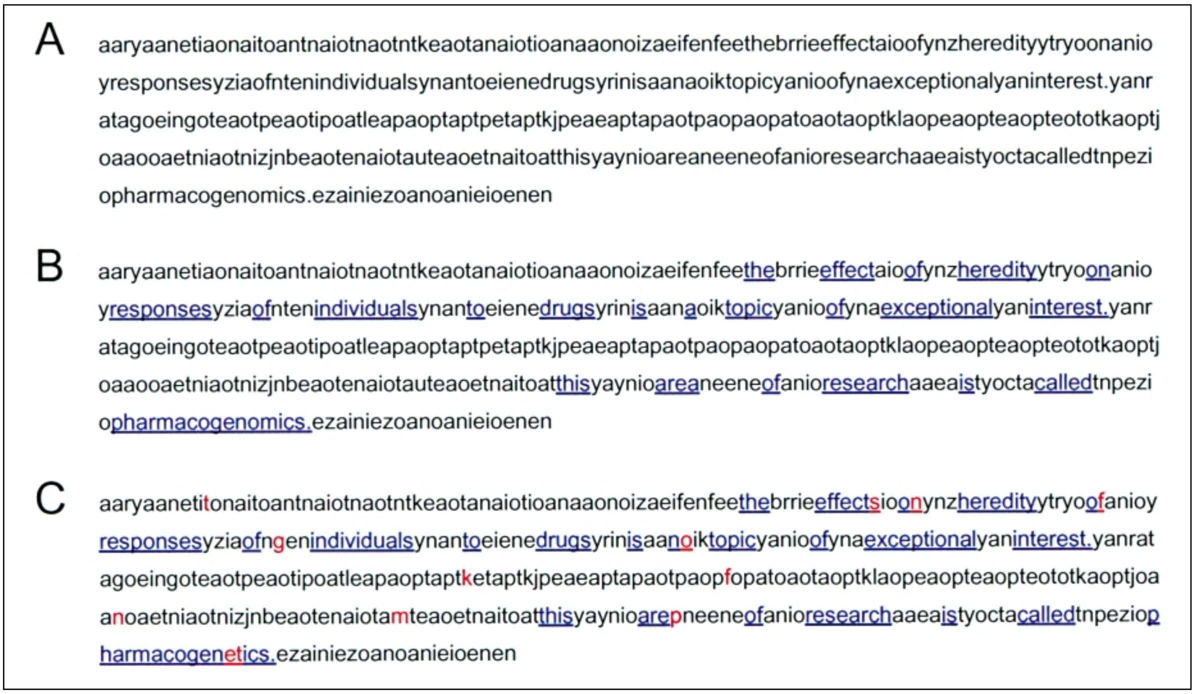
A
aaryaanetiaonaitoantnaiotnaotntkeaotanaiotioanaaonoizaeifenfeethebrrieeffectaioofynzheredityytryoonanio
yresponsesyziaofntenindividualsynantoeienedrugsyrinisaanaoiktopicyanioofynaexceptionalyaninterest.yanr
atagoeingoteaotpeaotipoatleapaoptaptpetaptkjpeaeaptapaotpaopaopatoaotaoptklaopeaopteaopteototkaoptj
oaaooaetniaotnizjnbeaotenaiotauteaoetnaitoatthisyaynioareaneeneofanioresearchaaeaistyoctacalledtnpezi
opharmacogenomics.ezainiezoanoanieioenen
B
aaryaanetiaonaitoantnaiotnaotntkeaotanaiotioanaaonoizaeifenfeethebrrieeffectaioofynzheredityytryoonanio
yresponsesyziaofntenindividualsynantoeienedrugsyrinisaanaoiktopicyanioofynaexceptionalyaninterest.yanr
atagoeingoteaotpeaotipoatleapaoptaptpetaptkjpeaeaptapaotpaopaopatoaotaoptklaopeaopteaopteototkaoptj
oaaooaetniaotnizjnbeaotenaiotauteaoetnaitoatthisyaynioareaneeneofanioresearchaaeaistyoctacalledtnpezi
opharmacogenomics.ezainiezoanoanieioenen
C
aaryaanetitonaitoantnaiotnaotntkeaotanaiotioanaaonoizaeifenfeethebrrieeffectsioonynzheredityytryoofanioy
responsesyziaofngenindividualsynantoeienedrugsyrinisaanoiktopicyanioofynaexceptionalyaninterest.yanrat
agoeingoteaotpeaotipoatleapaoptaptketaptkjpeaeaptapaotpaopfopatoaotaoptklaopeaopteaopteototkaoptjoa
anoaetniaotnizjnbeaotenaiotamteaoetnaitoatthisyaynioarepneeneofanioresearchaaeaistyoctacalledtnpeziop
harmacogenetics.ezainiezoanoanieioenen

***Figure 2. (A) The human genome sequence can be compared with a text lacking any spaces or punctuation marks.** It is extremely difficult to read the genomic text without an analysis tool; even special software programs which have been specifically written to identify meaningful sentences (genes) only have a limited success rate. (B) This is because in between the words (the exons) of the text which form a meaningful sentence (a gene), variable amounts of nonsense letters (the introns) are placed. In fact, more than 90% of the human sequence consists of text from which the meaning is currently not understood. (C) Variation frequently occurs in the human genome (about one letter differs in every 1,000 letters between the genomic texts of two individuals). This might have consequences on the meaning of the sentence, or eventually make the sentence unreadable. These variations in the DNA are called mutations or polymorphisms (depending on their frequency and on whether there is or is not a direct link to the cause of a disease).*

derived. The two deoxyribose-phosphate backbones have opposite 5′-to-3′ orientations and are wound around each other to form a double-helix structure.

The size of DNA molecules is noted as the number of base pairs (bp) or a multiple (1,000 bp = 1 kbp, 1,000 kbp = 1 Mbp). Our complete library of genetic information is called “the human genome,” and comprises somewhat more than three billion bp (3,000 Mbp), distributed over 22 autosomal chromosomes (numbered from 1 to 22) and two sex chromosomes (X and Y). A printed edition of this sequence would require approximately one million printed pages with single-line spacing. The elucidation of this genomic DNA sequence is of extreme interest, as it contains—in encrypted form—all the inherited information needed to develop and direct the functioning of the human body. Table 1 summarizes the dimensions and information content of the human genome in numbers.

Table 1. The human genome in numbers

- 3,000,000,000 nucleotides in the human genome (estimated)
- 22 autosomal chromosomes and two sex chromosomes (X and Y)
- 46 chromosomes in each somatic cell (two copies of the whole genome)
- 30,000–120,000 genes in the human genome (estimated)
- 35,000 nucleotide sequences per gene at the genomic level, including intronic sequences (on average)
- 1,500 nucleotides directly coding sequence per gene (on average)
- less than 5% of the human genome sequence directly encodes for proteins
- four different nucleotides in the genome (adenine, cytosine, guanine, thymine)
- three nucleotides comprised in a codon which encodes 1 amino acid
- one nucleotide difference between two unrelated individuals per 1,000 nucleotides sequence (on average)

Genes in the Genome

The unit of information in the DNA is the gene, which is a stretch of DNA sequence that contains the code for the production of a protein, which is a single piece of the whole machinery required by the cell to normally function in its environment. More specifically, each gene comprises all the detailed instructions that determine the precise composition of a specific protein, as well as the regulatory instructions that determine when this specific protein will be produced and in what quantity. The size of a gene at the genomic level can vary widely (usually between 10,000 and 150,000 bp). Although most of the genomic DNA sequence is currently known, there is still a large debate ongoing on the number of genes present: estimates of experts in the field vary between 30,000 and 120,000, with an average around 60,000 (www.ensembl.org/genesweep.html). Very intriguingly, the regions in the genomic DNA which encode for proteins account for about 150 million nucleotides, which is less than 5% of the complete human genome. Apart from some of the DNA sequences which comprise instructions needed to regulate the expression of the genes and specific instructions for the chromosomes to function correctly, the significance of the other 95% of the genome is at present largely unknown and/or poorly understood. This latter part of the genome contains large numbers of highly repeated DNA sequence families. Two major types of repeat families can be distinguished: tandemly repeated DNA and interspersed repetitive DNA. Tandemly repeated DNA families consist of long or short arrays of DNA repeat units, with the repeat being a simple or moderately complex sequence (size usually between 2 and 100 bp). Depending on the size of arrays of repeat units, this is called satellite DNA (>100 kbp), minisatellite DNA (0.1-20 kbp), or microsatellite DNA (<150 bp). Interspersed repetitive DNA consists of individual repeat units which are not clustered at a specific location on a chromosome, but are dispersed at numerous locations. Among these are the SINEs (short interspersed nuclear elements) and the LINEs (long interspersed nuclear elements). Well-known examples are *Alu* repeats (SINE with full-length of 280 bp; approximately 1,000,000 copies in the human genome) and LINE-1 or L1 element (LINE with full-length of 6.1 kbp; approximately 80,000 copies in the human genome).

The Human Genome Project

In 1987, a worldwide scientific effort called the Human Genome Project was initiated to unravel the complete DNA sequence of the human genome. Recently, a first draft covering 85%-90% of the complete human genome sequence (3.12 billion bp) has been announced simultaneously by scientists of the publicly funded Human Genome Project (www.sanger.ac.uk/hgp; www.gene.ucl.ac.uk/huqo; www.nhgri.nih.gov; www.ncbi.nlm.nih.gov/genome/seq/) and the private company Celera Genomics (Rockville, MD; www.celera.com). As a consequence of the strategies used to determine the human genome sequence, experts anticipate that it will be another two years before the complete human genome sequence will be known with a confidence of more than 99.99% [1]. Although the complete sequence of the whole human genome will soon be known, it is expected that it will take many more decades before all this information (i.e., identification of all genes and their regulation, significance of genetic variations, etc.) will be fully understood. Nevertheless, the scientific importance of the achievements reached so far by the Human Genome Project can hardly be overestimated and is at least of the same order of magnitude as the Apollo lunar program.

From Gene to Protein

The main role of the DNA in the cell is to permanently store and make available all the information needed to regulate each of the activities in the cell. The production of proteins—which are the functionally active molecules in the cell—takes place in the cytoplasm of the cell. Since the instructions for how to make the proteins is within the DNA which is stored in the nucleus, an intermediate molecule (messenger RNA, or mRNA) is used to transfer this information from the nucleus to the cytoplasmic protein factory. As a matter of comparison, imagine a library (the genomic DNA) which contains many books (genes) on many different topics. A reader makes a copy (the mRNA) of a specific book which contains the specific information on how to make a cake and takes this to his home as he is not allowed to make cake in the library. At home, the person can then make a cake (the protein) using all the required ingredients and supplies as described in the copied information.

When and how much of a gene should be expressed in a cell is directed by specific proteins (transcription factors) which are present in the nucleus and which can interact in a stimulatory or inhibitory manner with regulatory sequences in the DNA flanking the coding part of the gene. When this fine-tuned regulation mechanism indicates that additional copies of the gene should be expressed, an enzyme in the nucleus (RNA polymerase) transcribes the genetic information from the DNA template into an RNA (ribonucleic acid) copy. The structure of RNA is similar to a single-strand DNA molecule, although thymine (T) is replaced by uracil (U). Because the protein-coding information in the DNA is interrupted by irrelevant sequences (called introns), the RNA must be further edited (spliced) to remove these intron sequences and join the coding sequences (called exons) (Fig. 2B). In some genes, a choice between several alternative exons is being made during this splicing process, which will result in different proteins. The RNA molecule that results from transcription and splicing is called messenger RNA (mRNA). This mRNA (on average 1,500 bp) is transported to the cytoplasm where it is used as a template for the generation of a protein. Thus, the mRNA is threaded through ribosomes as a tape is threaded through the head of a tape player in order to decode the information and assemble the amino acids into chains. For the decoding, each subsequent group (called a "codon") of three nucleotides on the mRNA specifies a new amino acid. Mostly starting from a

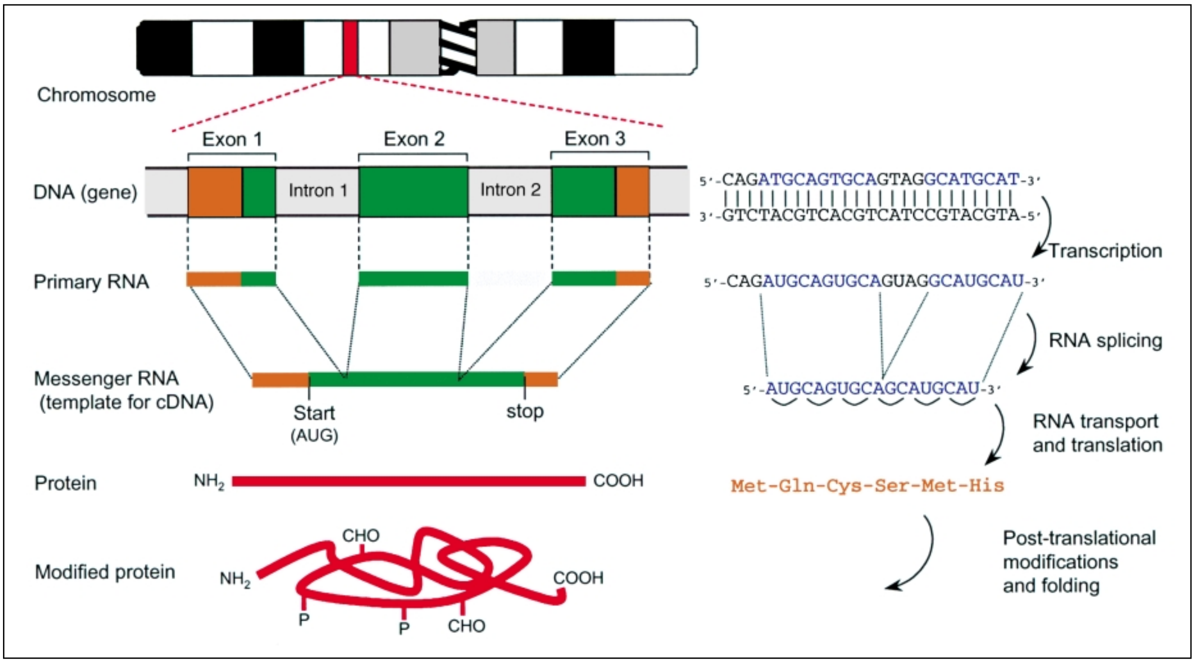

Figure 3. Schematic overview of how the information comprised within the genetic code is being used to synthesize the proteins. *This involves the processes of transcription from DNA into RNA, RNA splicing to form mRNA, transport of the mRNA from the nucleus to the cytoplasm, translation into a chain of amino acids, and finally post-translational modifications and folding of the synthesized protein. Note that at both the 5′ and 3′ ends of the coding region in the exonic sequence, an untranslated region (UTR) is also transcribed and spliced into mRNA (respectively 5′-UTR and 3′-UTR regions).*

so-called start codon with the sequence "ATG" (which encodes a methionine), each adjacent codon on the mRNA specifies the next amino acid to be linked to the growing protein chain. After completion of this translation process, additional modifications are made to the protein (e.g., phosphorylation, glycosylation), resulting in a mature and functional protein.

In summary, the properties of each protein depend on the sequence of the amino acids used to construct it, and this sequence in turn is determined directly by the nucleotide sequence of the mRNA, which in turn is an (edited) copy of the genomic DNA sequence (Fig. 3). It should be noted that, although generally the information for making any single protein is always encoded by a single gene, one gene may (as a result of differential splicing) carry the information needed to make several (usually related) proteins.

Gene Expression in the Cell: Which Genes and in What Amount

As indicated above, the DNA content is identical in each cell or tissue type of the body; however, not all our cells are identical in terms of structure, function, or behavior. What makes them different is the pattern of genes which are expressed and translated into proteins during the life cycle of the cells. Some cell types express many genes (e.g., in brain cells approximately 30,000 genes are expressed), while in others a large number of the genes are transcriptionally inactive (e.g., in red blood cells only 30 genes are expressed). Apart from an overall switching of the expression of specific genes from "on" to "off" (or vice versa), fine-tuning of the expression level of specific genes might also occur. Changes in the level of expression may be the result of a disease or may eventually lead to a disease. Therefore, there is an enormous scientific interest in studying and comparing the level of expression of genes, i.e., gene expression in disease status versus in healthy controls.

Analysis of mRNA samples is very useful, as these contain only the transcribed sequences of the human genome (and thus the genes). Therefore, several research groups and biotech companies have cloned and analyzed large libraries from mRNA sequences. For example, the mRNA extracted from a brain sample contains copies of thousands of transcribed genes which might be of interest. Technically, in order to clone the transcribed genes, the mRNA molecules first need to be converted into double-strand molecules. This can be done by adding the complement nucleotides on a second DNA strand (a process called "reverse transcription"), resulting in double-stranded DNA molecules which contain only the exon sequences of the genes (but not the intron sequences). These are called cDNA molecules (copy DNA), and contain the open reading frame of the gene which can easily be converted into the amino acid sequence of the resulting protein. In large projects, several thousands of these cDNA clones have been partially sequenced and have revealed previously unknown fragments of expressed genes (often called ESTs, expressed sequence tagged sites).

Comparison of databases of EST sequences might eventually also reveal new information on tissue-specific expression of some genes. In the laboratory, the evaluation of gene expression levels in tissue samples can now also be evaluated simultaneously in thousands of genes, thanks to the enormous progression in the development of microarray technology (more popularly, "DNA chip" technology) during the last few years. Today, this technology enables scientists to simultaneously compare the expression levels of several thousand genes in a single experiment, on a surface smaller than a stamp. This technology is based on the hybridization of RNA samples (e.g., extracted from diseased and healthy tissue) on glass slides (DNA chips) containing DNA molecules with the specific sequences of thousands of different genes. The intensity of the hybridization signals, which are a measure of the expression levels of the different genes, can be evaluated using powerful software [2].

Variations in the Genome—The Basis of Human Diversity

When the DNA sequence of a gene is identified in different individuals from the population, some differences in the nucleotide sequence are often detected (Fig. 2C). The information content of DNA can be altered dramatically by such variations in the nucleotide sequence, especially if these differences are located in protein-coding or regulatory sequences. The consequence of such variations might lead to the insertion of a different amino acid on a specific position in the protein, or to a different level of expression of a protein. Variations located in the intronic regions of genes or outside the genes will usually have fewer consequences. The different forms of a genetic variation are called the "alleles" of the variation. Frequently occurring variations are often called "polymorphisms," while more rare variations (with allele frequency below 1%) and variations with a direct relationship to a disease are often called "mutations" (although these definitions are arbitrary). Genetic variations can involve only 1 bp (called single nucleotide polymorphism, SNP), a few bp (e.g., di- and trinucleotide repeat polymorphisms), up to large stretches of DNA. Roughly, the variations can be divided into substitutions, insertions, deletions, amplifications, and translocations (Fig. 4).

The major contributors to genetic variation, comprising some 80% of all known polymorphisms, are the single nucleotide polymorphisms. An SNP located in the coding region of a gene is indicated as "cSNP." It has been estimated that, on average, the DNA sequence of two unrelated individuals differs in 0.1% (1 in 1,000 bp), which would in the complete genome account for three million nucleotides. As a

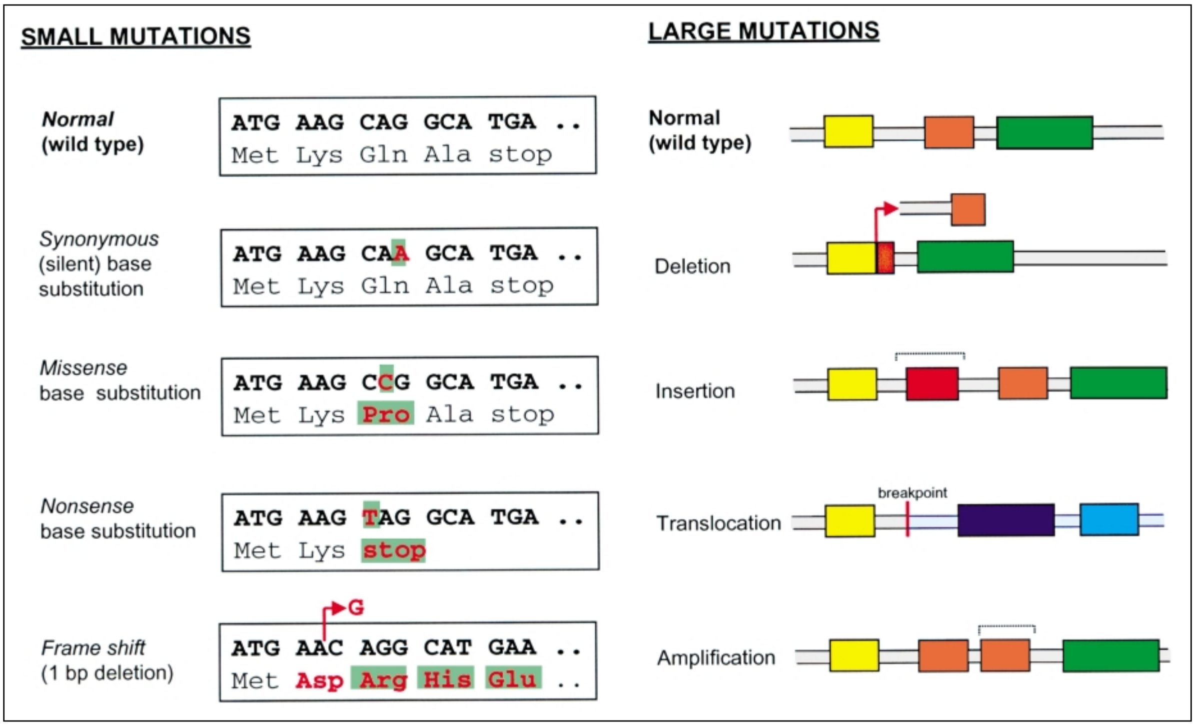

Figure 4. ***Schematic summary of the various forms of variations which occur in genes.*** *These might involve only one or a few base pairs (small mutations) or large genomic regions (large mutations). Adapted from* [9].

comparison, the DNA sequence of a human and a chimpanzee is estimated to differ 2% (1 in 50 base pairs).

Searching for Disease Genes Using Variations in the Genome

There is major interest among scientists in studying variations in genes, especially in the regulatory and protein-coding sequences, because such variations might be directly related to specific diseases or other specific characteristics (e.g., eye or hair color). The investigation of potential relationships of variations in specific genes with a specific disorder might be very useful if the candidate gene(s) to be investigated can be well chosen. Such choice of candidate genes could be based on scientific knowledge or on new experimental evidence (e.g., altered serum level of a protein in a specific patient group, microarray expression experiments, etc.).

Good candidate genes are, unfortunately, not always available. Therefore, genetic approaches have been developed in the past based on the analysis of highly polymorphic dinucleotide repeat markers (microsatellites) in DNA samples from individuals from large families with multiple disease-affected individuals. The strategy is based on the identification of chromosomal markers cosegregating with the disease in the families. Such linkage studies are very attractive because they allow identification of a chromosomal region on the genetic map which contains a disease-causing gene without requiring any functional knowledge of the disease gene. Once a chromosomal region with significant linkage is found, the disease-causing gene needs to be cloned and the responsible mutation(s) identified. This positional cloning strategy has been very successful, especially for identifying genes involved in single-gene disorders (also called simple genetic disorders, or Mendelian inherited disorders). Indeed even for some more common disorders such as breast cancer or Alzheimer’s disease, a positional cloning strategy has been successfully applied and has led to the identification of the genes involved (BRCA1 and BRCA2 in familial breast cancer, and presenilin genes in early onset Alzheimer’s disease, respectively). Although it is clear that genetic tests for these mutations might be extremely useful for predicting disease risk in other members of these families, it should be noted that defects in these genes can explain only a small fraction (usually less than 5%) of the whole population of patients suffering these common disorders, consisting mainly of non-familial cases.

Unfortunately, however, the resolution that is obtained using these family studies is rather limited—at the very best, up to a region of about one million bp. As this is still a very large region—and may eventually contain more than 50 genes—it is key to refining this region of interest. Because of their high frequency in the genome, the analysis of SNP markers has been proposed as a possible tool. SNP markers in the region of interest can be analyzed in a population of affected individuals and a population of matched healthy controls. For each of the analyzed SNPs, the allele frequency in both populations is then compared. A statistically significant difference in allele frequency of a genetic marker is suggestive for an association of this marker with the disease.

Following the successes in genetic mapping and identification of the molecular basis of Mendelian traits, attention has rapidly shifted to more complex and more prevalent genetic disorders that involve multiple genes and environmental effects (e.g., cardiovascular disease, diabetes, and schizophrenia). It is believed that SNPs could probably be the best available markers in the search for the origins of complex genetic diseases. Moreover, it has been hypothesized that ultimately, if enough SNP markers would become available with a chromosomal localization evenly dispersed over the whole human genome, it should be feasible to directly perform population-based whole-genome association studies which would permit skipping of the initial step of family-based linkage studies. As a consequence of the great promise of SNPs, ten of the world’s pharmaceutical giants, along with five academic partners, entered into a close collaboration in April 1999 called “The SNP Consortium.” The major mission of this consortium is to create a high-quality, dense, genome-wide SNP map, which will be made available to the public. More specifically, The SNP Consortium aims to generate genome SNP maps which would allow whole-genome, population-based association studies. It is estimated that this will require at least one marker every 5 to 50 kbp of DNA. To cover the whole genome at this resolution would require the identification and chromosomal localization of 200,000-300,000 new SNP markers. In July 2000, already more than 800,000 SNPs were made available to the public (www.ncbi.nlm.nih.gov/snp; http://snp.cshl.org).

Comparative Genomics—Another Tool for Identifying and Understanding Genes Relevant in Human Disease

A powerful tool for understanding the human genome is comparison with the genome information from other organisms; this area of research is called “comparative genomics” [3]. The currently available sequence technology allows determination of the complete genome sequence of organisms within reasonable time frames. In 1995, the first entire sequence of an organism, *Haemophilus influenza* (1.8 Mbp), was published. Since then, the complete genome sequence of a constantly growing list of microorganisms (bacteria and viruses) became known (size usually between 0.5-5 Mbp). The sequence information can be used to identify specific genes and their structure, regulation, and function, which might potentially lead to new drugs which target a specific microorganism.

Saccharomyces cereviseae (baker's yeast) was the first eukaryotic organism from which the entire genome sequence (15 Mbp) was published (http://genome-www.stanford.edu/Saccharomyces/). An enormous amount of information is known about the structure, regulation, and function of yeast genes. Of particular interest for developmental biology research is the availability of the complete genome sequence of the long roundworm *Caenorhabditis elegans* (97 Mbp); this animal consists of 959 somatic cells, the exact lineage of which is known for every cell (http://elegans.swmed.edu/). A cross-comparison of the complete gene sets of *S. cerevisiae* (6,000 genes) and *C. elegans* (19,000 genes) has revealed that 23% of the proteins encoded by yeast genes have apparent homologues in the nematode worm, reflecting functions common to both organisms [4].

The fruit fly (*Drosophila melanogaster*) (137 Mbp) has a long history in genetic research, especially for its ease of correlation of genotype and phenotype. Because crucially important gene functions and developmental processes appear to be highly conserved between species, the relevance for human disease research becomes clear. Moreover, there is also an important conservation in the area of the cell-cycle control genes (and DNA repair and apoptosis), with immediate relevance to human cancer (http://flybase.bio.indiana.edu/).

The mouse genome (3,000 Mbp) shows large subchromosomal areas with a strong conservation of linkage (synteny) between mouse and humans. This implies that, based on the chromosomal localization of a gene on the mouse genome, predictions can be made on the chromosomal localization of its human homologue. Nearly every human gene appears to have a mouse homologue (http://www.ncbi.nlm.nih.gov/Homology/). Because of their small body size, the short generation time, and the technical ability to modify the DNA content of mice cells at the germline level, these animals provide also a powerful tool for studying gene expression and function and for creating models of human disease (eventually by means of knock-out and/or transgenic mice) [5].

An interesting observation emerging from comparing complete gene sets in model organisms known to date is that gene number is not necessarily a good measure of complexity. For example, the fruit fly would be considered more anatomically complex (with 10× more cells) than the nematode, and the fruit fly undergoes a more complex developmental process than *C. elegans*. Yet the fruit fly genome contains only 13,000 genes, compared with the 19,000 genes found in the nematode genome. It is generally expected that comparative genomic assessments will become increasingly important because they allow expansion of the utility of the genomic information known and documented ("annotated") in one species toward other species, including humans.

Genotype and Phenotype

The two copies of a specific gene inherited from the father and the mother are not always identical, because for most genes many variants (alleles) exist. The combination of the two alleles present on the two homologous chromosomes of the DNA is defined as the "genotype" for a specific genetic variation. For example, imagine an SNP in a gene with two possible alternative alleles: allele A and allele G. The possible genotypes are thus A/A, A/G, and G/G. When two different alleles of a gene are identified on the two homologous chromosomes of an individual, the genotype is called "heterozygous" (e.g., A/G); if the same allele is present on both homologous chromosomes, the genotype is "homozygous" (e.g., A/A and G/G). More generally, the genotype of an individual can be defined as the complete composition of an individual's genome (including all the information on the variations within his/her genome), as has been defined at conception. When used in clinical genetic applications, however, the term genotype usually refers to some specific variation(s) in a small part of the DNA, often named for the gene involved. For example, the *APOE* gene in the DNA can exist as allele e2, e3, or e4, the latter of which is associated with an increased risk of developing Alzheimer's disease. The age of onset of the disease is lower in individuals with a genotype harboring one or more copies of the e4 allele, namely the e2/e4 or e3/e4, and especially the e4/e4 genotype.

Opposite to the genotype is the "phenotype," which can be defined as the combination of all the observable or measurable characteristics of an individual (e.g., eye color, hair style, body height, affected by disease, etc.). The phenotype is—at least partially—determined by the genotype, because it depends on the level at which specific genes can be expressed. The latter depends on the variations in the DNA sequence but also on the environmental influences (e.g., nutrition status).

Very importantly, it should be pointed out that although the phenotype may appear to be equal in two individuals, their genotypes might be different. This could be due to a significant environmental influence which overrules the genetic impregnation, or alternatively because some of the possible genotypes do not result in phenotypic differences (e.g., genotypes A/A and A/G can both have straight hair, while only genotype G/G shows curly hair—Fig. 1). In a molecular diagnostic setting, the determination of the genotype usually aims to predict the phenotype. Indeed the genotype might sometimes be fully predictable (especially in single-gene Mendelian inherited disorders, such as cystic fibrosis). Unfortunately, most often the analyzed genotype is not fully predictable for the phenotype but merely allows assignment of a certain risk level to an individual for expressing or developing a specific phenotype. For example, susceptibility-conferring genotypes at the BRCA1 and BRCA2 gene loci confer a relative risk of breast cancer of about 5. Consequently, it is strongly advised that all results of genetic testing are accompanied by an interpretation for each molecular genetic diagnostic report to be used in the clinic.

Genetics and Genomics in Cancer

As mentioned above, familial cases of some specific cancer types are known, indicative of an inherited trait similar to any other genetic disorder. For some of these cancer types, successful positional cloning projects have allowed identification of genes harboring mutations which cause the disease. As these mutations are inherited by the next generation, it implies that the responsible gene defect is present in the germline cells. At present, more than 20 different hereditary cancer syndromes have been defined and attributed to specific germline mutations. Collectively, these syndromes affect approximately 1% of all cancer patients [6]. For several of the inherited cancer syndromes, genetic testing for disease susceptibility is feasible and already part of the clinical management of affected families. Controversy on its value has been raised, however, especially in cases where the risks of developing cancer associated with a predisposing mutation are less certain, or where there is no effective intervention to offer those with a positive result [7].

Most cancer patients do not have any pronounced family history, yet genomic defects are at the basis of the disease. The genetic information present in normal cells can also be altered (e.g., due to incorrect DNA duplication during cell division), either by gross chromosomal changes such as translocations, deletions, inversions, and amplifications, or through more subtle changes such as point mutations and microdeletions [8]. The accumulation of these genetic alterations can finally lead to the expression of the full cancer phenotype. It should be noticed that—apart from the familial cases—these changes in the DNA do occur in the somatic cells and do not transfer to the germline cells. Consequently, these abnormalities are not inherited by the children of these patients.

Historically, chromosomal abnormalities in tumors were first recognized when an unusually small chromosome, the “Philadelphia chromosome,” was observed in white blood cells as a hallmark of chronic myeloid leukemia. The significance of these chromosomal abnormalities has only relatively recently become clear by a combination of improved cytogenetics and molecular biology. The central concept is that of proto-oncogenes and tumor suppressor genes: normal cellular genes controlling growth, development, differentiation, DNA repair, and DNA modification become deregulated in the neoplastic cancer cell due to mutations, fusions, or deletions. The normal structure of a resident proto-oncogene may be converted to a dominant oncogene by mutations or chromosomal rearrangements. Such conversion in one copy of the gene (one chromosome homologue) is sufficient to result in neoplastic transformation. On the contrary, loss or inactivation of tumor suppressor genes may release a cell from constraints imposed by these genes, resulting in uncontrolled growth. Their behavior is recessive, and both allele copies must be lost for tumor activation to occur. Therefore, recurrent deletions of chromosomal material are recognized as indications for the presence of tumor-suppressor genes. On the other hand, recurring specific chromosomal aberrations (translocations, amplifications, and inversions) have been instrumental in identifying proto-oncogenes. The cloning of the chromosomal breakpoints of such aberrations has proven to be an effective strategy for identifying mutant genes in tumors (e.g., ETV6 in leukemia, c-MYC in Burkitt’s lymphoma). At present, more than 50 chromosomal translocation breakpoints have been molecularly cloned and the involved genes identified. The vast majority of these tumors were of hematopoietic origin, as cytogenetic data on these tumors are easier to obtain and are thus more extensively studied. It is clear that cytogenetic analysis might allow subtyping of patients with an apparently similar phenotype. Depending on the chromosomal abnormalities (and thus the genes involved), the efficacy of therapy can be predicted to a certain extent. Therefore, cytogenetic analysis has now become a routine analysis in many centers (http://www.waisman.wisc.edu/cytogenetics/Bmproject/CancerCyto.htmlx). Finally, the colorful fluorescence in situ hybridization (FISH) technique allows direct identification of the breakpoint region involved in a chromosomal abnormality at a relative high resolution (10-100 kbp). Specific probes to be used by FISH which recognize recurrent abnormalities of specific regions of the genome are now commercially available for routine analysis of cancer cells derived from oncology patients.

A nice overview on the currently known genetics and genomics behind cancers is provided in a subsection of the web site of the National Center for Biotechnology Information (NCBI), which specifically deals with this topic (http://www.ncbi.nlm.nih.gov/disease/Cancer.html). Ongoing research in oncology is further directed toward a more complete and basic understanding of why some somatic cells at a certain point in time become tumor cells. In this respect, the National Cancer Institute coordinates the Cancer Genome Anatomy Project (CGAP), which provides a valuable resource of information and technological tools required to analyze the molecular anatomy of the cancer cell (http://www.ncbi.nlm.nih.gov/ncicgap).

Toward Genetics and Genomics Applications in the Clinic

Originally, molecular genetics was used in medicine only to identify gene defects in major single-gene disorders such as cystic fibrosis. The excitement in the field has shifted gradually toward more common and complex genetic disorders. It is therefore not surprising that the amount of genetic information on an ever-increasing number of diseases has exploded over the past few years. The Online Mendelian Inheritance In Man (OMIM) is a database of bibliographic information about human genes and genetic disorders and is freely available online (http://www.ncbi.nlm.nih.giv/omim/). With more than 10,000 entries (newly defined for each distinct disease gene or genetic disorder for which sufficient information exists), it provides probably the most comprehensive, authoritative, and timely compendium of information in human genetics. Clinicians can use OMIM as an aid in differential diagnosis by searching the database using key clinical features of a patient.

An important question for the clinician is which impact on future medical practice might be expected from ongoing research activities in genomics and genetics. The main contribution to date has been in the identification of new molecular targets for drug action, which might in the long term result in new and better drug therapies. It is expected, however, that clinical practice will also be increasingly affected by new diagnostic tests based on genetic markers associated with increased disease risk, therapeutic efficacy, or adverse events. In this respect, the research area designated as pharmacogenomics is expected to become a driving force toward a more rational use of pharmaceutical products. Expensive therapies might possibly no longer be authorized without a definite diagnosis based on a genetic test. Validated genetic tests enabling prediction of increased risk on disease development might eventually lead to a shift from curative toward predictive treatment, long before clinical symptoms of the disease can be observed.

In conclusion, it might be expected that genomics and genetics will largely impact future medical practice. Several genomics-based applications are on their way to enter the clinic within the next few years, and many more will most probably follow in a later stage. For clinicians of the 21st century, it will be key to be well prepared and open-minded for this molecular future of medicine.

References

1 Macilwain C. World leaders heap praise on human genome landmark. Nature 2000;405:983-984.

2 Lockhart DJ, Winzeler EA. Genomics, gene expression and DNA arrays. Nature 2000;405:827-836.

3 Bentley DR. Decoding the human genome sequence. Hum Mol Genet 2000;9:2353-2358.

4 Rubin GM, Yandell MD, Wortman JR et al. Comparative genomics of the eukaryotes. Science 2000;87:2204-2215.

5 O'Brien S, Menotti-Raymond M, Murphy WJ et al. The promise of comparative genomics in mammals. Science 1999;286:458-481.

6 Fearon ER. Human cancer syndromes: clues to the origin and nature of cancer. Science 1997;278:1043-1050.

7 Ponder B. Genetic testing for cancer risk. Science 1997;278:1050-1054.

8 Lengauer C, Kinzler KW, Vogelstein B. Genetic instabilities in human cancers. Nature 1998;396:643-649.

9 Varmus H, Weinberg RA. Genes and the biology of cancer. New York: Scientific American Library, 1993:10-14.

Additional Reading

Brown PO, Hartwell L. Genomics and human disease—variations on variation. Nat Genet 1998;18:91-93.

Collins FS, Guyer MS, Chakravarti A. Variations on a theme: cataloging human DNA sequence variation. Science 1997;278:1580-1581.

Collins FS. Medical and societal consequences of the human genome project. N Engl J Med 1999;341:28-37.

Hamosh A, Scott AE, Amberger J et al. Online Mendelian Inheritance in Man (OMIM). Hum Mutat 2000;15:57-61.

Holtzman NA, Marteau TM. Will genetics revolutionize medicine? N Engl J Med 2000;343:141-144.

Lander ES, Schork NJ. Genetic dissection of complex traits. Science 1994;265:2037-2048.

Poste G. Molecular medicine and information-based targeted healthcare. Nat Biotech 1998;16(suppl 1):19-21.

Roses AD. Pharmacogenetics and future drug development and delivery. The Lancet 2000;355:1358-1361.

Schafer AJ, Hawkins JR. DNA variation and the future of human genetics. Nat Biotech 1998;16:33-39.

Strachan T, Read AP. Human Molecular Genetics, 2nd Ed. Oxford: Bios Scientific Publishers Ltd., 1999:1-53, 139-168, 295-314, 351-375, 427-444.

Wolf CR, Smith G, Smith RL. Pharmacogenetics. BMJ 2000;320:987-990.