The human genome. Part I

History of the Human Genome Project. The structure of the human genome: protein-coding genes and non-coding DNA. Satellite DNA. Tandem repeats. Single nucleotide polymorphisms (SNP). Transposed elements of the genome: transposons, retrotransposons.

1. Describe the structure of the human genome: protein-coding genes, intergenic regions (spacers), satellites, tandem repeats, single nucleotide polymorphisms (SNPs).
2. Explain the role of non-coding DNA in the human genome.
3. Discuss the prospects for applying knowledge about the human genome in medicine and pharmaceuticals.
4. Describe DNA transposons, retrotransposons, retroviral integration.
5. Provide examples of human diseases triggered by transposable elements.
6. Discuss the usage of transposable elements in medicine.

Part II

Methods for the study of nucleic acids and proteins. Proteomic methods of analysis. Bioinformation databases. DNA diagnostics: polymerase chain reaction, restriction analysis, FISH hybridization. Linked immunosorbent assay. Bioethics of genetic experiments with humans.

1. Give the definitions of genomics, proteomics and bioinformatics, describe their research methods.
2. Explain the Sanger, Maxam-Gilbert, NGS (New Generation Sequencing) and other methods of genome sequencing.
3. Characterize and analyze the main methods of protein research: two-dimensional gel electrophoresis, mass spectrometry, chromatography, X-ray structural analysis, nuclear magnetic resonance.
4. Describe EMBL-EBI, DDJB, NCBI, PIR, MIPS, NBRF, SwissProt, UniProt and other bioinformatical databases.
5. Give the definition of molecular diagnostics and describe its various methods.
6. Explain the reasons for choosing different methods of molecular diagnostics to detect different types of hereditary diseases (gene, chromosomal and genomic), infectious diseases and metabolic diseases, give specific examples.
7. Discuss the ethics of conducting genetic and molecular biological experiments on humans.

Human Genome Project (HGP), an international collaboration that successfully determined, stored, and rendered publicly available the sequences of almost all the genetic content of the chromosomes of the human organism, otherwise known as the human genome.

The Human Genome Project (HGP) refers to the international 13-year effort, formally begun in October 1990 and completed in 2003, to discover all the estimated 20,000-25,000 human genes and make them accessible for further biological study.

Project goals

- *identify* all the approximately 20,000-25,000 genes in human DNA,
- *determine* the sequences of the 3 billion chemical base pairs that make up human DNA,
- *store* this information in databases,
- *improve* tools for data analysis,
- *transfer* related technologies to the private sector, and
- *address* the ethical, legal, and social issues (ELSI) that may arise from the project.

As part of the HGP, parallel studies were carried out on selected model organisms such as the bacterium *E. coli* and the mouse to help develop the technology and interpret human gene function.

In 1988, Congress appropriated funds to the Department of Energy (DOE) and the National Institutes of Health (NIH) to begin planning the Human Genome Project. Planners set a 15-year time frame, estimated that the price tag would be $3 billion, and laid out formal goals to get the job done.[1] On October 1, 1990, the Human Genome Project officially began.[2] According to early plans, the human race would witness its own blueprint in fine detail in the year 2005.

In the fall of 1998, however, improvements in technology, success in achieving early mapping goals, emerging research opportunities, and a growing demand for the human DNA sequence prompted project leaders in the United States and abroad to promise the blueprint — the complete DNA sequence of the human genome — two years ahead of schedule, in 2003.

Indeed, only six months later, in March 1999, 15 percent of the sequence was in a finished or nearly finished state. The largest centers participating in the Human Genome Project received new grants to begin full-scale sequencing of the human genome, and the timetable was moved up yet again. Pilot sequencing projects had been so successful that the planners of the Human Genome Project now felt confident that at least 90 percent of the human sequence could be completed in "working draft" form by the spring of 2000, considerably earlier than expected.

Until the entire human genome sequence has been completed, in 2003 or perhaps earlier, the working-draft sequence will be very useful, especially for finding genes, exons, and other genomic features. But because the working draft will contain gaps and will not be entirely accurate, it will not be as useful as the finished sequence for studying DNA features that span large regions or require a high degree of accuracy over long stretches of the sequence. The HGP was further intended to improve the technologies needed to interpret and analyze genomic sequences, to identify all the genes encoded in human DNA, and to address the ethical, legal, and social implications that might arise from defining the entire human genomic sequence.

The final product must have four characteristics — the four A's of the Human Genome Project.

First, the sequence must be *accurate* — that is, the DNA spellings must have an accuracy of 99.99 percent or better. The HGP was further intended to improve the technologies needed to interpret and analyze genomic sequences, to identify all the genes encoded in human DNA, and to address the ethical, legal, and social implications that might arise from defining the entire human genomic sequence. Second, large-scale sequencing requires that the shorter lengths of sequenced DNA be accurately *assembled* into longer, genomic-scale pieces that reflect the original genomic DNA. Third, the human DNA sequence must also be *affordable*, and technology development will aim to reduce the cost as much as possible. Finally, the high-quality, finished human DNA sequence should be *accessible* within 24 hours through public data bases.

Implications for Understanding Genetic Illness

Maps and other forms of genome technology provide the tools for a gene-isolation technique known as positional cloning.**[9]** This technique allows a researcher to confirm the genetic basis of a disease and identify the responsible gene, even when little is known about the gene's function. So far, over 100 disease-linked genes have been isolated with the use of the positional-cloning technique. Whereas gene discovery by this route once took years to decades, an investigator using these powerful tools can now sometimes map and isolate a gene in a matter of weeks. Increasingly, gene hunters are combining positional-cloning techniques with information in EST data bases to narrow their gene searches to rational candidates. This method, called positional candidate cloning,**[10]** has been used to isolate many altered genes associated with human disease.

Gene isolation provides the best hope for understanding human disease at its most fundamental level (. Knowledge about genetic control of cellular functions will underpin future strategies to prevent or treat disease phenotypes. The recent isolation of genes for Parkinson's disease,[11-14] for example, has greatly advanced molecular research on this baffling disease. In one study of families with early-onset Parkinson's disease, gene hunters mapped a suspect gene to a region of chromosome 4.[11] The region contained approximately 100 genes, among which was 1 known to encode the protein α-synuclein. Earlier research had shown that α-synuclein accumulates in brain cells of people with Alzheimer's disease,[15] and people with Parkinson's disease have similar deposits in the substantia nigra. In just a few months, the researchers showed conclusively that a missense mutation in the α-synuclein gene caused Parkinson's disease in the study families. Further research has shown that a mutation in a gene encoding a protein critical to the breakdown of α-synuclein and other proteins also results in the Parkinson's disease phenotype in a different family.[14] Although most cases of Parkinson's disease appear to have limited heritability, studying rare families of this sort provides crucial clues to the pathways involved — α-synuclein is found in the Lewy bodies in virtually all cases of Parkinson's disease. An understanding of the genetic control of the proteolytic processes of brain proteins may provide new targets for interventions in a number of related neurodegenerative disorders characterized by the accumulation of protein deposits, including Alzheimer's disease, Huntington's disease, and spinocerebellar ataxia.

Even before a gene's role in disease is fully understood, diagnostic applications can be useful in preventing or minimizing the development of health consequences. DNA tests that look for the presence of disease-linked mutations, for example, are proving to be the most immediate commercial application of gene discovery and the one now used most frequently by clinicians. These tests may help establish the diagnosis of a genetic disease, foreshadow the development of disease later in life, or identify healthy heterozygote carriers of recessive diseases. Genetic tests can be performed at any stage of the human life cycle, and the sampling procedures are becoming less invasive. Whereas genetic testing was once sought almost exclusively by couples with a family history of early-onset disease, for the purpose of family planning, information about genetic status is increasingly sought by persons who wish to learn about their own predisposition to adult-onset illness.

In a growing number of instances, strategies can be implemented to reduce or prevent illness when a genetic cause or predisposition is known. Successes in reducing disease through treatment have been achieved for the hereditary disorders hemochromatosis, phenylketonuria, and familial hypercholesterolemia, among others. Risk reduction through early detection and lifestyle changes may be possible in the case of disorders associated with predisposing mutations, such as some cancers. As therapies build on knowledge gained about the molecular basis of disease, increasing numbers of illnesses that are now refractory to treatment may yield to molecular medicine in the future.

The recent discovery of an altered gene (*HFE*) that leads to hereditary hemochromatosis,[16] a common disorder of iron metabolism, provides an interesting example of the potential for using information about mutations to prevent an adult-onset disease phenotype. A recessive condition, hereditary hemochromatosis affects about 1 in 300 persons of northern European descent and is easily treatable if diagnosed early. Its major symptoms — liver cirrhosis, heart failure, diabetes, arthritis, and other organ damage — do not occur until midlife and are easily misdiagnosed. Untreated, the disease causes early death, but treatment by phlebotomy to remove excess iron allows people with hereditary hemochromatosis to live a normal life span. A single substitution of the amino acid tyrosine for cysteine at codon 282 accounts for the majority of cases.[17]

At first glance, hereditary hemochromatosis seems to be an ideal target for public health approaches to the prevention of hereditary disease: the disorder is common, the number of disease-linked mutations in the gene are few, and effective treatment can minimize or eliminate the effects of the disease. But closer examination reveals a number of complexities that have so far militated against the rapid introduction of this genetic test as a tool for disease prevention.[16] Because the penetrance of the altered *HFE* gene is reduced, especially in females, clinical signs can range from none that are detectable to severe organ damage from iron overload. At the moment, simple detection of the mutation does not predict the most likely clinical course. Before population testing for *HFE* mutations is considered, further research is needed to explain the variations in phenotype among mutation carriers and to correlate the genotype more closely with health outcomes.

The rather straightforward mendelian rules that govern the inheritance of disease traits have been worked out for many rare disorders that result from highly penetrant changes in a single gene. But teasing out the genetic components of the so-called complex disorders — diabetes, heart disease, most common cancers, autoimmune disorders, and psychiatric disorders — that result from the interplay of environment, lifestyle, and the small effects of many genes remains a formidable task. Most of the successful efforts to identify genes associated with common diseases have focused on highly heritable subgroups, including the *BRCA1* [18] and *BRCA2* [19] genes in breast cancer, the gene for hepatocyte nuclear factor 4α (*HNF-4* α) in maturity-onset diabetes of the young (MODY) type 1,[20] the gene for glucokinase (*GCK*) in MODY type 2,[21] the gene for hepatocyte nuclear factor 1α (*HNF-1*α) in MODY type 3,[22] the gene for human Mut S homologue 2 (*hMSH2*)[23,24] and the gene for human Mut L homologue 1 (*hMLH1*)[25] in hereditary nonpolyposis colon cancer, and the gene for α-synuclein in Parkinson's disease.[11] Linkage analysis and positional-cloning techniques are well suited to discovering genes with such strong influences. But these strategies are not as easily applied to the multiple, low-penetrance variants, which in the aggregate account for a larger percentage of illnesses. Identification of weakly penetrant alleles that contribute to common disorders requires new and more powerful approaches.

To assist in these efforts, the Human Genome Project is initiating new studies of genetic variation in the human population to provide a dense map of common DNA variants. DNA sequence variations include insertions and deletions of nucleotides, differences in the copy number of repeated sequences, and single-nucleotide polymorphisms, or SNPs (pronounced “snips”), which occur most frequently throughout the human genome. About 1 in every 300 to 500 bases in human DNA may be a SNP.

SNPs can be used as markers in whole-genome linkage analysis of families with affected members, as well as in association studies of individuals in a population. Association studies may directly test a variant with potential functional importance or may take advantage of the phenomenon of linkage disequilibrium — in which a marker and a gene are inherited together — to map gene variants associated with disease. Because the human species consists of relatively few generations, recombination events have not disrupted linkage disequilibrium over distances of 3000 to 100,000 bases in most populations. Consequently, association studies view large human populations as evolutionary families and do not rely on studies of nuclear families for gene mapping.[26,27]

Some SNPs may contribute directly to a trait or disease phenotype by altering function. Though most SNPs are located outside protein-coding sequences, those within coding sequences, called cSNPs, are of particular interest because they are more likely to affect gene function. A large, well-characterized collection of SNPs will become increasingly important for the discovery of DNA sequence variations that affect biologic function. Work is already under way with NIH support to develop a catalogue of 60,000 or more SNPs. A recently formed pharmaceutical consortium will support the production of 300,000 more, with the work being done at the publicly funded genome centers and all the data deposited in the public domain. This is a wonderful example of a public–private partnership to develop a powerful set of research tools that all can use.

New Forms of Technology for Genetic Analysis and Risk Assessment

The transition from genetics to genomics marks the evolution from an understanding of single genes and their individual functions to an understanding of the actions of multiple genes and their control of biologic systems. Whereas the tools of the Human Genome Project initially advanced research on single genes, they are now forming the basis for genomic-scale analysis of the human organism.

The so-called DNA chip currently provides one promising approach to genome-scale studies of genetic variation,[28] detection of heterogeneous gene mutations,[29] and gene expression.[30] The result of an adaptation of dot blot hybridization techniques, DNA chips, also called microarrays, generally consist of a thin slice of glass or silicon about the size of a postage stamp on which threads of synthetic nucleic acids are arrayed. Sample probes are added to the chip, and matches are read by an electronic scanner. As with semiconductors, the capacity of DNA chips has doubled about every two years, so chips that held a few hundred arrays not so long ago now hold hundreds of thousands.

Microarray technology has been applied to the detection of DNA variations as well as expression of messenger RNA in individual cells and tissues. Microarrays are used clinically to detect human immunodeficiency virus sequence variations, p53 gene mutations in breast tissue, and expression of cytochrome P450 genes. In the laboratory, microarray technology has also been applied to genomic comparisons across species,[31] genetic recombination,[32] and large-scale analysis of gene copy number and expression, as well as protein expression, in cancerous tissues.[33]

Use of microarrays and other new technologies to detect DNA variations holds promise, along with family histories and data from large population studies, for establishing a person's risk of contracting common, adult-onset disorders. A base-line genome scan could provide helpful information about a person's risk profile and point to the prevention strategies — if available — that should be used.

Genetic Knowledge and Individualized Medicine

Identifying human genetic variations will eventually allow clinicians to subclassify diseases and adapt therapies to the individual patient. There may be large differences in the effectiveness of medicines from one person to the next. Toxic reactions can also occur and in many instances are likely to be a consequence of genetically encoded host factors. That basic observation has spawned the burgeoning new field of pharmacogenomics, which attempts to use information about genetic variation to predict responses to drug therapies.

For example, researchers discovered that patients with Alzheimer's who have the ε4 subtype of the gene for apolipoprotein E (*APOE ε4*), which affects cholinergic function in the brain, are less likely to benefit from the cholinomimetic drug tacrine than are patients without this subtype.[34] This finding will help in the analysis of data from clinical trials of Alzheimer's therapies and will promote the development of new therapies specifically designed for *APOE ε4* carriers.

In another example, cholesteryl ester transfer protein (CETP) plays an important part in the metabolism of high-density lipoprotein, a lipoprotein associated with lowered susceptibility to atherosclerosis. A certain genetic variant of the *CETP* gene is correlated with higher plasma CETP levels and lower levels of plasma high-density lipoprotein. One study showed that in men who carried this genetic variant, treatment with pravastatin slowed the progression of coronary atherosclerosis.[35] This finding may allow physicians to predict which patients with coronary artery disease will benefit from treatment with pravastatin.

In a third example, the formation of venous blood clots in the brain and legs is a rare but serious side effect of birth-control pills. One study has shown a dramatically increased risk of cerebral-vein thrombosis among women taking oral contraceptives who also carry the blood-clotting variant factor V Leiden.[36] The risk of other venous thrombotic events is also increased in this group. Foreknowledge of the presence of this variant and consideration of alternative forms of birth control might be useful in minimizing the risk of thrombosis in these women.

Not only will genetic tests predict responsiveness to drugs on the market today, but also genetic approaches to disease prevention and treatment will include an expanding array of gene products for use in developing tomorrow's drug therapies. Since the Food and Drug Administration's approval of recombinant human insulin in 1982, over 50 additional gene-based drugs have become available for clinical use. These include drugs for the treatment of cancer, heart attack, stroke, and diabetes, as well as many vaccines.[37]

Not all therapeutic advances for gene discovery will be genes or gene products. In other instances, molecular insights into a disorder, derived from gene discovery, will suggest a new treatment. Sodium phenylbutyrate, for example, which is approved for the regulation of blood ammonia levels, is being tested in clinical trials for the treatment of cystic fibrosis.[38] The main clinical phenotype in people with cystic fibrosis results from a mutation in the gene for cystic fibrosis transmembrane conductance regulator (CFTR) protein. The mutation prevents normal amounts of CFTR protein from crossing the cell membrane, diminishing the ability of chloride and water to enter and exit the cell. Sodium phenylbutyrate apparently stimulates expression of CFTR protein, allowing more of it to reach the correct location.

Ethical, Legal, and Social Implications

One of the most active areas of the ELSI program has been policy development related to the privacy and fair use of genetic information, particularly in health insurance, employment, and medical research. Debates in this area focus largely on the potential of genetic information to predict an increased likelihood of the eventual development of a disease phenotype in a currently healthy person.

Although many states have attempted to address “genetic discrimination” in health insurance and employment, federal legislation would provide the most comprehensive protection. Concern about the confidentiality of genetic information may make people reluctant to volunteer for studies involving disease-linked gene mutations or genetic therapy, for fear that the results could result in the loss of a job or the loss of insurance coverage.

Largely on the basis of recommendations formulated in workshops held by the Human Genome Project and the National Action Plan on Breast Cancer,[39,40] the Clinton administration endorsed the need for congressional action to protect against genetic discrimination in health insurance and employment. In 1996, Congress enacted the Health Insurance Portability and Accountability Act, which represented a large step toward protecting access to health insurance in the group-insurance market but left several serious gaps in the individual-insurance market that must still be closed.

In the area of workplace discrimination, the Equal Employment Opportunity Commission has interpreted the Americans with Disabilities Act as covering on-the-job discrimination based on “genetic information relating to illness, disease or other disorders.”[41] But no claims of genetic discrimination have been brought to the commission, and the guidance has yet to be tested in court, so the degree of protection actually provided by the act remains uncertain.

In the area of privacy, as part of the partnership between the National Action Plan on Breast Cancer and the Human Genome Project, medical researchers, policy makers, and representatives of law, government, the insurance industry, and public health have recently assessed current policies and practices designed to protect confidentiality in genetics research and have identified areas where new or modified policies or practices might enhance the protection of privacy and promote the conduct of research. The group is developing a set of principles for researchers, research institutions, state and federal agencies, and policy makers to consider in formulating measures to protect the privacy of research information.

Other important steps have been taken to ensure the responsible integration of genetic tests into clinical practice. For the most part, genetic testing in the United States has developed successfully, providing options for avoiding, preventing, and treating inherited disorders. But the rapid pace of test development combined with the rush to market new products may create an environment in which the tests are made available before they have been adequately validated. On the recommendation of the Task Force on Genetic Testing,[42] assembled by the Human Genome Project's NIH–DOE ELSI Working Group, the Secretary of the Department of Health and Human Services has established an advisory panel to ensure the safe introduction of genetic tests into clinical practice.

Completion of the first human-genome sequence and the expansion of human genetic research to include studies of genetic variation among subpopulations have raised new questions about ethical, legal, and social issues. The 1998–2003 plan includes an examination of these issues as well as the integration of genetic technology and information into health care and public health activities; the use of knowledge about genomics and gene–environment interactions in nonclinical settings; examination of a variety of philosophical, theological, and ethical perspectives on new genetic knowledge; and consideration of the ways in which racial, ethnic, and socioeconomic factors affect the use, understanding, and interpretation of genetic information, the use of genetic services, and the development of policy.

Genome structure

Humans have two genomes, nuclear and mitochondrial. Normal diploid cells contain two copies of the nuclear genome and a much larger but variable number of copies of the mitochondrial genome.

The human haploid genome consists of about 3 x 10^9 base pairs of DNA. Although the sequence of the human genome has been (almost) completely determined by DNA sequencing, it is not yet fully understood. Most (though probably not all) genes have been identified by a combination of high throughput experimental and bioinformatics approaches, yet much work still needs to be done to further elucidate the biological functions of their protein and RNA products. Recent results suggest that most of the vast quantities of noncoding DNA within the genome have associated biochemical activities, including regulation of gene expression, organization of chromosome architecture, and signals controlling epigenetic inheritance.Genomic DNA exists as single linear pieces of DNA that are associated with a protein called a nucleoprotein complex. The DNA-protein complex is the basis for the formation of chromosomes, virtually all of the genomic DNA is distributed among the 23 chromosomes that reside in the cellular nucleus.

Human genomes include both protein-coding DNA genes and noncoding DNA. Protein-coding sequences account for only a very small fraction of the genome (approximately 1.5%), and the rest is associated with non-coding RNA genes, regulatory DNA sequences, LINEs, SINEs, introns, and sequences for which as yet no function has been determined

The content of the human genome is commonly divided into coding and noncoding DNA sequences. Coding DNA is defined as those sequences that can be transcribed into mRNA and translated into proteins during the human life cycle; these sequences occupy only a small fraction of the genome (<2%). Noncoding DNA is made up of all of those sequences (ca. 98% of the genome) that are not used to encode proteins.

Some noncoding DNA contains genes for RNA molecules with important biological functions (noncoding RNA, for example ribosomal RNA and transfer RNA). The exploration of the function and evolutionary origin of noncoding DNA is an important goal of contemporary genome research, including the ENCODE (Encyclopedia of DNA Elements) project, which aims to survey the entire human genome, using a variety of experimental tools whose results are indicative of molecular activity.

Because non-coding DNA greatly outnumbers coding DNA, the concept of the sequenced genome has become a more focused analytical concept than the classical concept of the DNA-coding gene.

One of the main open questions in the field of the origin of life is the biogenesis of proteins and nucleic acids as ordered sequences of monomeric residues, possibly in many identical copies. The days of peering down the microscope to detect chromosome abnormalities are gone, replaced by chromosome analysis at the genomic level.

- Polymerase chain reaction (PCR)
- Sanger sequencing
- Southern blotting
- Multiplex ligation probe amplification (MLPA)
- Array comparative genomic hybridisation (array CGH)
- Karyotyping
- Fluorescent in situ hybridisation (FISH)
- Quantitative fluorescent PCR (QF-PCR)
- Single nucleotide polymorphism (SNP) genotyping and genome wide association studies (GWAS)

The extraction and analysis of cell free fetal DNA, including non-invasive prenatal testing (NIPT).

Literature:

1. Brown TA (2002). The Human Genome (2nd ed.). Oxford: Wiley-Liss.

2. Francis S. Collins. Medical and Societal Consequences of the Human Genome Project // N Engl J Med 1999; 341:28-37. DOI: 10.1056/NEJM199907013410106

3. Alberts et al., pp. 287-292.