The Human Genome Project

54

Robert K. Murray, MD, PhD

BIOMEDICAL SIGNIFICANCE

The information deriving from determination of the sequences of the human genome and those of other organisms will change biology and medicine for all time. For example, with reference to the human genome, new information on our origins, on disease genes, on diagnosis, and possible approaches to therapy are already flooding in. Progress in fields such as genomics, proteomics, bioinformatics, biotechnology, and pharmacogenomics is accelerating rapidly.

The aims of this chapter are to briefly summarize the major findings of the Human Genome Project (HGP) and indicate their implications for biology and medicine.

THE HUMAN GENOME PROJECT HAS A VARIETY OF GOALS

The HGP, which started in 1990, is an international effort whose principal goals were to sequence the entire human genome and the genomes of several other model organisms that have been basic to the study of genetics (eg, *Escherichia coli, Saccharomyces cerevisiae* [a yeast], *Drosophila melanogaster* [the fruit fly], *Caenorhabditis elegans* [the roundworm], and *Mus musculus* [the common house mouse]). Most of these goals have been accomplished. In the United States, the National Center for Human Genome Research (NCHGR) was established in 1989, initially directed by James D. Watson and subsequently by Francis Collins. The NCHGR played a leading role in directing the United States effort in the HGP. In 1997, it became the National Human Genome Research Institute (NHGRI). The international collaboration—involving groups from the USA, UK, Japan, France, Germany, and China—came to be known as the International Human Genome Sequencing Consortium (IHGSC). Initially, a number of short-term goals were established for the United States effort—eg, producing a human genetic map with markers 2–5 centimorgans (cM) apart and constructing a physical map of all 24 human chromosomes (22 autosomal plus X and Y) with markers spaced at approximately 100,000 base pairs (bp). Figure 54–1 summarizes the differences between a genetic map, a cytogenetic map, and a physical map of a chromosome. These and other initial goals were achieved and surpassed by the mid-nineties. In 1998, new goals for the United States wing of the HGP were announced. These included the aim of completing the entire sequence by the end of 2003 or sooner. Other specific objectives concerned sequencing technology, comparative genomics, bioinformatics, ethical considerations, and other issues. By the fall of 1998, about 6% of the human genome sequence had been completed and the foundations for future work laid. Further progress was catalyzed by the announcement that a second group, the private company Celera Genomics, led by Craig Venter, had undertaken the objective of sequencing the human genome. Venter and colleagues had published in 1995 the entire genome sequences of *Haemophilus influenzae* and *Mycoplasma genitalium,* the first of many species to have their genomic sequences determined. An important factor in the success of these workers was the use of a shotgun approach, ie, sonicating the DNA, sequencing the fragments, and reassembling the sequence, based on overlaps. For comparison, a variety of approaches that have been used at different times to study normal and disease genes are listed in Table 54–1.

A Draft Sequence of the Human Genome Was Announced in June 2000

In June 2000, leaders of the IHGSC and the personnel at Celera Genomics announced completion of working drafts of the sequence of the human genome, covering more than 90% of it. The principal findings of the two groups were published separately in February 2001 in special issues of Nature (the IHGSC) and Science (Celera). The draft published by the Consortium was the product of at least 10 years of work involving 20 sequencing centers located in six countries. That published by Celera and associates was the product of some 3 years or less of work; it relied in part on data obtained by the IHGSC. The combined achievement has been hailed, among other descriptions, as providing a Library of Life, supplying a Periodic Table of Life, and finding the Holy Grail of Human Genetics.

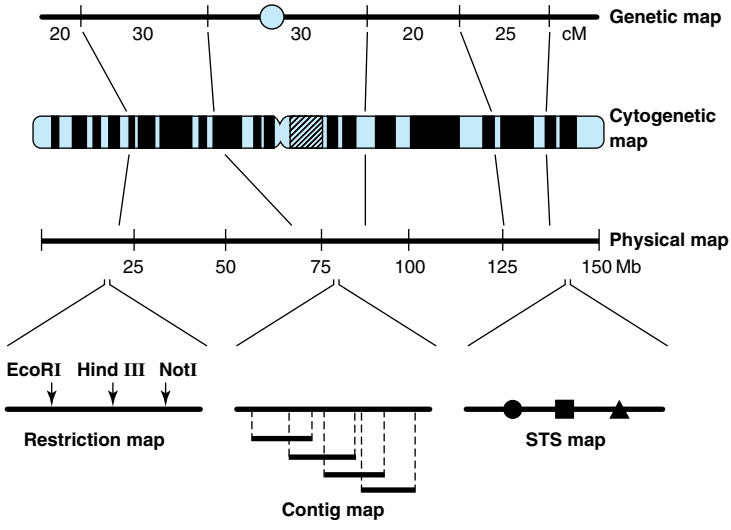

Figure 54–1. Principal methods used to identify and isolate normal and disease genes. For the genetic map, the positions of several hypothetical genetic markers are shown, along with the genetic distances in centimorgans between them. The circle shows the position of the centromere. For the cytogenetic map, the classic banding pattern of a hypothetical chromosome is shown. For the physical map, the approximate physical positions of the above genetic markers are shown, along with the relative physical distances in megabase pairs. Examples of a restriction map, a contig mark, and an STS map are also shown. (Reproduced, with permission, from Green ED, Waterston RH: The Human Genome Project: Prospects and implications for clinical medicine. JAMA 1991;266: 1966. Copyright © 1991 by the American Medical Association.)

Different Approaches Were Used by the Two Groups

We shall summarize the major findings reported in the two drafts and comment on their implications. While there are differences between the drafts, they will not be dwelt on here, as the areas of general agreement are much more extensive. It is worthwhile, however, to summarize the different approaches used by the two groups. Basically, the IHGSC employed a **map first, sequence later approach.** In part, this was because sequencing was a slow process when the public project started, and the strategy of the Consortium evolved over time as advances were made in sequencing and other techniques. The overall approach, referred to as **hierarchical shotgun sequencing,** consisted of fragmenting the entire genome into pieces of approximately 100–200 kb and inserting them into bacterial artificial chromosomes (BACs). The BACs were then positioned on individual chromosomes by looking for marker sequences known as sequence-tagged sites (STSs), whose locations had been already determined. STSs are short (usually < 500 bp), unique genomic loci for which a PCR assay is available. Clones of the BACs were then broken into small fragments (shotgunning). Each fragment was then sequenced, and computer algorithms were used that recognized matching sequence information from overlapping fragments to piece together the complete sequence.

Celera used the **whole genome shotgun approach,** in effect bypassing the mapping step. Shotgun fragments were assembled by algorithms onto large scaffolds, and the correct position of these scaffolds in the genome was determined using STSs. A scaffold is a series of "contigs" that are in the right order but not necessarily connected in one continuous sequence. Contigs are contiguous sequences of DNA made by assembling overlapping sequenced fragments of a natural chromosome or a BAC. The availability of **high-throughput sequenators, powerful computer programs,** the element of competition, and other factors accounted for the rapid progress made by both groups from 1998 onward.

Table 54–1. Principal methods used to identify and isolate normal and disease genes.

Procedure	Comments
Detection of specific cytogenetic abnormalities	For instance, a small deletion of band Xp21.2 was important in cloning the gene involved in Duchenne muscular dystrophy.
Extensive linkage studies	Large families with defined pedigrees are desirable. Dominant genes are easier to recognize than recessives.
Use of probes to define marker loci	Probes identify STSs, RFLPs, SNPs,[1] etc; thousands, covering all the chromosomes, are now available. It is desirable to flank the gene on both sides, clearly delineating it.
Radiation hybrid mapping[2]	Now the most rapid method of localizing a gene or DNA fragment to a subregion of a human chromosome and constructing a physical map.
Use of rodent or human somatic cell hybrids	Permits assignment of a gene to one specific chromosome but not to a subregion.
Fluorescence in situ hybridization	Permits localization of a gene to one chromosomal band.
Use of pulsed-field gel electrophoresis (PFGE) to separate large DNA fragments	Permits isolation of large DNA fragments obtained by use of restriction endonucleases (rare cutters) that result in very limited cutting of DNA.
Chromosome walking	Involves repeated cloning of overlapping DNA segments; the procedure is laborious and can usually cover only 100–200 kb.
Chromosome jumping	By cutting DNA into relatively large fragments and circularizing it, one can move more quickly and cover greater lengths of DNA than with chromosomal walking.
Cloning via YACs, BACs, cosmids, phages, and plasmids	Permits isolation of fragments of varying lengths.
Detection of expression of mRNAs in tissues by Northern blotting using one or more fragments of the gene as a probe	The mRNA should be expressed in affected tissues.
PCR	Can be used to amplify fragments of the gene; also many other applications.
DNA sequencing	Establishes the highest resolution physical map. Identifies open reading frame. Facilities with many high throughput instruments could sequence millions of base pairs per day.
Databases	Comparison of DNA and protein sequences obtained from unknown gene with known sequences in databases can facilitate gene identification.

Abbreviations: STS, sequence tagged site; RFLP, restriction fragment linked polymorphism; SNP, single nucleotide polymorphism; YAC, yeast artificial chromosome; BAC, bacterial artificial chromosome; PCR, polymerase chain reaction.

[1]Many single nucleotide polymorphisms (SNPs) are being detected and catalogued. These are stable and frequent, and their detection can be automated. It is anticipated that they will be particularly useful for mapping complex traits such as diabetes mellitus.

[2]Radiation hybrid mapping (consult http://compgen.rutgers.edu/rhmap/ for a detailed bibliography of this technique) makes use of a panel of somatic cell hybrids, with each cell line containing a random set of irradiated human genomic DNA in a hamster background. Briefly, the radiation fragments the DNA into small pieces of variable length; if a gene is located close to another known gene, it is likely that the two will remain linked (compare genetic linkage) on the same fragment. An STS marker is typed against a radiation hybrid panel by using its two oligonucleotide primers to perform a PCR assay against the DNA from each hybrid cell line of the panel. If enough markers are typed on one panel, continuous linkage can be established along each arm of a chromosome, and the markers can be assembled into the map as a single linkage group.

DETERMINATION OF THE SEQUENCE OF THE HUMAN GENOME HAS PRODUCED A WEALTH OF NEW FINDINGS

Only a small fraction of the findings can be covered here. The interested reader is referred to the original articles. Table 54–2 summarizes a number of the highlights, which can now be described.

Most of the Human Genome Has Been Sequenced

Over 90% of the human genome had been sequenced by July 2000. This is by far the largest genome sequenced, with an estimated size of approximately 3.2 gigabases (Gb). Prior to the human genome, that of the fruit fly had been the largest (~180 Mb) sequenced. Gaps still exist, small and large, and the quality of some of the sequencing data will be refined since some of the findings are probably not exactly right.

The Human Genome Is Estimated to Encode About 30,000–40,000 Proteins

The greatest surprise provided by the results to date has been the apparently low number of genes encoding proteins, estimated to lie between 30,000 and 40,000. The higher number could increase as new data are obtained. This number is approximately twice that found in the roundworm (19,099) and three times that of the fruit fly (13,061). The figures suggest that the complexity of humans compared with that of the two simpler organisms must have explanations other than strictly gene number.

Table 54–2. Major findings reported in the rough drafts of the human genome.

- More than 90% of the genome has been sequenced; gaps, large and small, remain to be filled in.
- Estimated number of protein-coding genes ranges from 30,000 to 40,000.
- Only 1.1–1.5% of the genome codes for proteins.
- There are wide variations in features of individual chromosomes (eg, in gene number per Mb, SNP density, GC content, numbers of transposable elements and CpG islands, recombination rate).
- Human genes do more work than those of the roundworm or fruit fly (eg, alternative splicing is used more frequently).
- The human proteome is more complex than that found in invertebrates.
- Repeat sequences probably constitute more than 50% of the genome.
- Approximately 100 coding regions have been copied and moved by RNA-based transposons.
- Approximately 200 genes may be derived from bacteria by lateral transfer.
- More than 3 million SNPs have been identified.

Only 1.1–1.5% of the Human Genome Encodes Proteins

Analyses of the available data reveal that 1.1–1.5% of the genome consists of exons. About 24% consists of introns, and 75% of sequences lying between genes (intergenic). Comparisons with the data on the roundworm and fruit fly have shown that exon size across the three species is relatively constant (mean size of 145 bp in humans). However, intron size in humans is much more variable (mean size of over 3300 bp), resulting in great variation in gene size.

The Landscape of Human Chromosomes Varies Widely

There are marked differences among individual chromosomes in many features, such as gene number per megabase, density of single nucleotide polymorphisms (SNPs), GC content, number of transposable elements and CpG islands, and recombination rate. To take one example, chromosome 19 has the richest gene content (23 genes per megabase), whereas chromosome 13 and the Y chromosome have the sparsest content (5 genes per megabase). Explanations for these variations are not apparent at this time.

Human Genes Do More Work Than Those of Simpler Organisms

Alternative splicing appears to be more prevalent in humans, involving at least 35% of their genes. Data indicate that the average number of distinct transcripts per gene for chromosomes 22 and 19 were 2.6 and 3.2, respectively. These figures are higher than for the roundworm, where only 12.2% of genes appear to be alternatively spliced and only 1.34 splice variants per gene were noted.

The Human Proteome Is More Complex Than That of Invertebrates

Relatively few new protein domains appear to have emerged among vertebrates. However, the number of distinct domain architectures (~1800) in human proteins is 1.8 times that of the roundworm or fruit fly. About 90 vertebrate-specific families of proteins have been identified, and these have been found to be enriched in proteins of the immune and nervous systems.

The results of the two drafts are rich in information about protein families and classes. One example is shown in Table 54–3, in which the major classes of proteins encoded by human genes are listed. As can be seen, the largest class is "unknown." Identification of these unknown proteins will be a major focus of activity for many laboratories.

Repeat Sequences Probably Constitute More Than 50% of the Human Genome

Repeat sequences probably account for at least half of the genome. They fall into five classes: (1) transposon-derived repeats (interspersed repeats); (2) processed pseudogenes; (3) simple sequence repeats; (4) segmental duplications, made up of 10–300 kb that have been copied from one region of the genome into another; and (5) blocks of tandemly repeated sequences, found at centromeres, telomeres, and other areas. Considerable information on most of the above classes of repeat sequences—of great value in understanding the architecture and development of the human genome—is reported in the drafts. Only two points of interest will be mentioned here. It is speculated that Alu elements, the most prominent members (about 10% of the total genome) of the short interspersed elements (SINEs), may be present in GC-rich areas because of positive selection, implying that they are of benefit to the host.

Table 54–3. Major classes of proteins encoded by human genes.[1]

Class of Protein	Number (%)[2]
Unknown	12,809 (41%)
Nucleic acid enzymes	2,308 (7.5%)
Transcription factors	1,850 (6%)
Receptors	1,543 (5%)
Hydrolases	1,227 (4.0%)
Select regulatory molecules (eg, G proteins, cell cycle regulators)	988 (3.2%)
Protooncogenes	902 (2.9%)
Cytoskeletal structural proteins	876 (2.8%)
Kinases	868 (2.8%)

[1]Data from Venter JC et al: The sequence of the human genome. Science 2001;291:1304.

[2]The percentages are derived from a total of 26,383 genes reported in the rough draft by Celera Genomics. Classes containing more than 2.5% of the total proteins encoded by the genes identified when this rough draft was written are arbitrarily listed as major.

Segmental duplications have been found to be much more common than in the roundworm or fruit fly. It is possible that these structures may be involved in exon shuffling and the increased diversity of proteins found in humans.

Other Findings of Interest

The last three major points of interest listed in Table 54–2 will be briefly described together.

Approximately 100 coding regions are estimated to have been copied and moved by RNA-based transposons (retrotransposons). It is possible that some of these genes may adopt new roles in the course of time. A surprising finding is that over 200 genes may be derived from bacteria by lateral transfer. None of these genes are present in nonvertebrate eukaryotes. More than 3 million SNPs have been identified. It is likely that they will prove invaluable for certain aspects of gene mapping.

It is stressed that the findings listed here are only a few of those reported in the drafts, and the reader is urged to consult the original reports (see References, below).

FURTHER WORK IS PLANNED ON THE HUMAN & OTHER GENOMES

The IHGSC has indicated that it will determine the complete sequence, it is hoped, by 2003. The task involves filling in the gaps and identifying new genes, their locations, and functions. Regulatory regions will be identified, and the sequences of other large genomes (eg, of the house mouse; of *Rattus norvegicus,* the Norway rat; of *Danio rerio,* the zebra fish; of *Fugu rubripes,* the tiger puffer fish; and of one or more primates) will be obtained; indeed, a draft version of the genome of the tiger puffer fish was published in 2002. Additional SNPs will be identified; a complete catalog of these variants is expected to be of great value in mapping genes associated with complex traits and for other uses as well. Along with the above, existing databases will be added to as new information flows in, and new databases will probably be established to serve specific purposes. A variety of studies in functional genomics (ie, the study of genomes to determine the functions of all their genes and their products) will also be undertaken.

IMPLICATIONS FOR PROTEOMICS, BIOTECHNOLOGY, & BIOINFORMATICS

Many fields will be influenced by knowledge of the human genome. Only a few are briefly discussed here.

Proteomics (see Chapter 4) in its broadest sense is the study of all the proteins encoded in an organism (ie,

the proteome), including their structures, modifications, functions, and interactions. In a narrower sense, it involves the identification and study of multiple proteins linked through cellular actions—but not necessarily the entire proteome. With regard to humans, many individual proteins will be identified and characterized; their interactions and levels will be determined in physiologic and pathologic states, and the resulting information will be entered into appropriate databases. Techniques such as two-dimensional electrophoresis, a variety of modes of mass spectrometry, and antibody arrays will be central to expansion of this rapidly growing field. Overall, proteomics will greatly advance our knowledge of proteins at the basic level and will also nourish **biotechnology** as new proteins that are likely to have diagnostic, therapeutic, and other uses are discovered and methods for their economic production are developed. Specialists in **bioinformatics** will be in demand, as this field rapidly gears up to manage, analyze, and utilize the flood of data from genomic and proteomic studies.

IMPLICATIONS FOR MEDICINE

Practically every area of medicine will be affected by the new information accruing from knowledge of the human genome. The **tracking of disease genes** will be enormously facilitated. As mentioned above, SNP maps should greatly assist determination of genes involved in complex diseases. Probes for any gene will be available if needed, leading to **improved diagnostic testing** for disease susceptibility genes and for genes directly involved in the causation of specific diseases. The field of **pharmacogenomics** (see Chapter 53) is already expanding greatly, and it is possible that in the future drugs will be tailored to accommodate the variations in enzymes and other proteins involved in drug action and metabolism found among individuals. Studies of genes involved in **behavior** may lead to new insights into the causation and possible treatment of psychiatric disorders. Many **ethical issues**—eg, privacy concerns and the use of genomic information for commercial purposes—will have to be addressed. It will also be important that medical and economic benefits accrue to individuals in **Third World countries** from the anticipated effects on health services and the diagnosis and treatment of disease.

SUMMARY

- Determination of the complete sequence of the human genome, now almost completed, is one of the most significant scientific achievements of all time.
- Many important findings have already emerged. The one to date that has generated the most discussion is that the number of human genes may be only two to three times that estimated for the roundworm and the fruit fly.
- Information flowing from the Human Genome Project is having major influences in fields such as proteomics, bioinformatics, biotechnology, and pharmacogenomics as well as all areas of biology and medicine.
- It is hoped that the knowledge derived will be used wisely and fairly and that the benefits that will ensue regarding health, disease, and other matters will be made available to all people everywhere.

REFERENCES

Collins FS, McKusick VA: Implications of the Human Genome Project for medical science. JAMA 2001;285:540. (The February 7, 2001, issue describes opportunities for medical research in the 21st century. Many articles of interest.)

Hedges SB, Kumar S: Vertebrate genomes compared. Science 2002;297:1283. (The same issue—No. 5585, August 23—contains a draft version of the genome of the tiger puffer fish.)

McKusick VA: The anatomy of the human genome: a neo-Vesalian basis for medicine in the 21st century. JAMA 2001;286:2289. (The November 14, 2001, issue contains a number of other excellent articles—eg, on clinical proteomics, pharmacogenomics—relating to the Human Genome Project and its impact on medicine.)

Nature 2001;409(6822) (February 15), and Science 2001;291(5507) (February 16). (These two issues present the rough drafts prepared by the IHGSC and Celera, respectively, along with many other articles analyzing the meaning and significance of the findings.)

Science 2001;294(5540) (October 5). (This issue contains a number of articles under the title Genome: Unlocking Biology's Storehouse. They describe new ideas, approaches, and research related to genome information.)

APPENDIX

SELECTED WORLD WIDE WEB SITES

The following is a list of Web sites that readers may find useful. The sites have been visited at various times by one of the authors (RKM). Most are located in the USA, but many provide extensive links to international sites and to databases (eg, for protein and nucleic acid sequences) and online journals. RKM would be grateful if readers who find other useful sites would notify him of their URLs by e-mail (rmurray6745@rogers. com) so that they may be considered for inclusion in future editions of this text.

Readers should note that URLs may change or cease to exist.

ACCESS TO THE BIOMEDICAL LITERATURE

HighWire Press: http://highwire.stanford.edu

(Extensive lists of various classes of journals—biology, medicine, etc—and offers also the most extensive list of journals with free online access.)

National Library of Medicine: http://www.nlm.nih.gov/

(Free access to Medline via PubMed.)

GENERAL RESOURCE SITES

The Biology Project (from the University of Arizona): http://www.biology.arizona.edu/default.html

Harvard Department of Molecular & Cellular Biology Links: http://mcb.harvard.edu/BioLinks.html

SITES ON SPECIFIC TOPICS

American Heart Association: http://www.americanheart.org

(Useful information on nutrition, on the role of various biomolecules—eg, cholesterol, lipoproteins—in heart disease, and on the major cardiovascular diseases.)

Cancer Genome Anatomy Project (CGAP): http://www.ncbi.nlm.nih.gov/ncicgap

(An interdisciplinary program to generate the information and technical tools needed to decipher the molecular anatomy of the cancer cell.)

European Bioinformatics Institute: http://www.ebi.ac.uk/ebi_home.html

(Maintains the EMBL Nucleotide and SWISS-PROT databases as well as other databases.)

GeneCards: http://bioinformatics.weizmann.ac.il/cards/

(A database of human genes, their products, and their involvements in diseases. From the Weizmann Institute of Science.)

GeneTestsGeneClinics: http://www.geneclinics.org/

(A medical genetics information resource with comprehensive articles on many genetic diseases.)

Genes and Disease: http://www.ncbi.nlm.nih.gov/disease/

(Coverage of the genetic basis of many different types of diseases.)

The Glycoscience Network (TGN): http://www.vei.co.uk/TGN/tgn_side.htm

(TGN is an informal worldwide grouping of scientists who share an interest in carbohydrates. The site contains considerable information on carbohydrates and an extensive list of links to other sites dealing with sugar-containing molecules).

Howard Hughes Medical Institute: http://www.hhmi.org/

(An excellent site for following current biomedical research. Contains a comprehensive Research News Archive.)

The Human Gene Mutation Database: http://archive.uwcm.ac.uk/uwcm/mg/hgmd0.html

(An extensive tabulation of mutations in human genes from the Institute of Medical Genetics in Cardiff, Wales.)

Human Genome Project Information: http://www.ornl.gov/hgmis/

(From the Human Genome Program of the United States Department of Energy.)

The Institute for Genetic Research: http://www.tigr.org/

(Sequences of various bacterial genomes and other information.)

Karolinska Institute Nutritional and Metabolic Diseases: http://www.mic.ki.se/Diseases/c18.html

(Access to information on many nutritional and metabolic disorders.)

MITOMAP: http://www.mitomap.org/

(A human mitochondrial genome database.)

National Center for Biotechnology Information: http://ncbi.nlm.nih.gov/

(Information on molecular biology and how molecular processes affect human health and disease.)

National Human Genome Research Institute: http://www.genome.gov/

(Extensive information about the Human Genome Project.)

National Institutes of Health (NIH): http://www.nih.gov/

(Includes links to the separate Institutes and Centers that constitute NIH, covering a wide range of biomedical research.)

Neuroscience (Biosciences): http://neuro.med.cornell.edu/VL/

(A comprehensive list of neuroscience resources; part of the World-Wide Web Virtual Library.)