

# The Human Genome Project

## *Applications in the Diagnosis and Treatment of Neurologic Disease*

Glen A. Evans, MD, PhD

**T**he Human Genome Project (HGP), an international program to decode the entire DNA sequence of the human genome in 15 years, represents the largest biological experiment ever conducted. This set of information will contain the blueprint for the construction and operation of a human being. While the primary driving force behind the genome project is the potential to vastly expand the amount of genetic information available for biomedical research, the ramifications for other fields of study in biological research, the biotechnology and pharmaceutical industry, our understanding of evolution, effects on agriculture, and implications for bioethics are likely to be profound.

The human genome is composed of 3 billion nucleotides of DNA, organized as 23 chromosomes, and contains an estimated 60 000 to 70 000 gene-encoded proteins. The genes constitute only about 2% to 3% of the entire DNA sequence while 40% of the sequence consists of repetitive sequences of unknown significance. The remaining 58% of the sequences is unique, noncoding DNA responsible for functions as yet to be determined. It is believed that most genetic control elements, the promoters, enhancers, genetic switches, and control elements necessary for regulating gene expression in a complex organism, are located in this noncoding DNA.<sup>1</sup>

The HGP was initiated in the United States in 1990 at an estimated cost of \$3 billion (\$1 per nucleotide) and with a 15-year time frame.<sup>2</sup> The strategy was to develop the technology and biological background information during the first half of the project and to actually determine the majority of the sequence during the second half. Thus, the first 8 years of the HGP were spent in constructing basic genetic and physical maps of the human genome, developing the technology for high throughput DNA sequencing, and investigating the genomes of model organisms. The knowledge of sequences of model organisms is essential for decoding the highly complex human genome. The genetic maps of humans are ordered polymorphic DNA markers

spaced along chromosomes with the distance separating them being a measure of the frequency of recombination. The genetic maps form the basis of positional cloning,<sup>3</sup> the ability to isolate disease genes based on patterns of inheritance. The development of genetic maps rapidly led to the construction of physical maps, sets of ordered DNA markers known as sequence-tagged sites, and overlapping sets of yeast artificial chromosome clones.<sup>4</sup> These physical maps, where the separation between markers is about 50 to 100 kilobase (kb) and where distances separating the markers is measured in base pairs (bp), are the immediate precursors of DNA sequencing. The first half of the project also included sequencing of the genomes of a number of model organisms. These organisms included the bacteria *Escherichia coli*, the yeast *Saccharomyces cerevisiae*, the nematode *Caenorhabditis elegans*, and the fruit fly *Drosophila melanogaster*. Because of the value of the mouse as a genetic organism, physical and genetic maps of the mouse were constructed. Determining the complete DNA sequence of the mouse, which approximates humans' in size and complexity, falls outside the financial limits of the HGP. While these organisms are the official focus of the HGP, other projects in the scientific community have resulted in the complete DNA sequencing of a wide range of organisms (**Figure**).

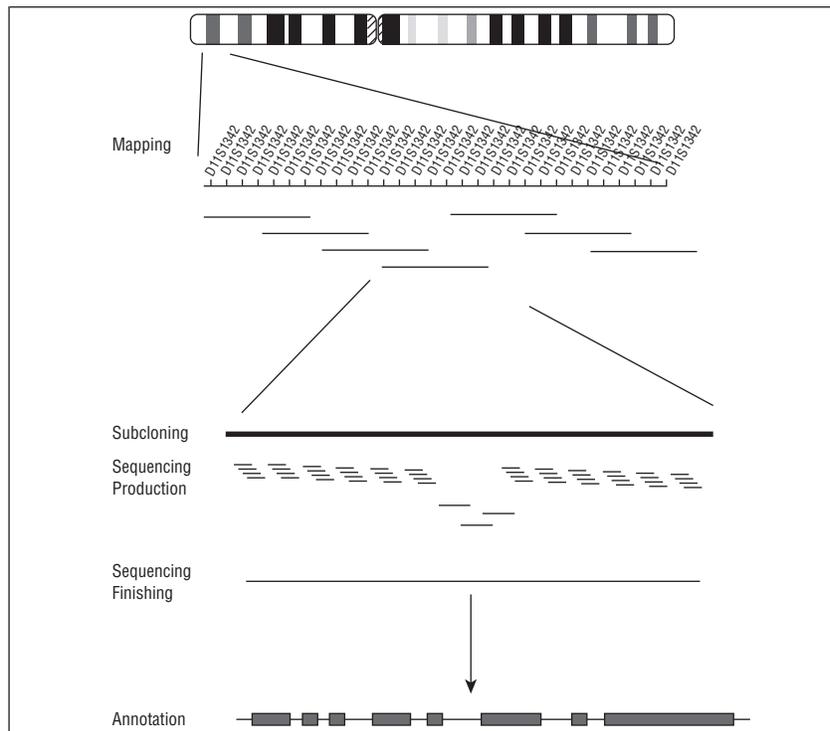
Since 1996, a substantial effort has been devoted to developing the technology and infrastructure to enable large-scale DNA se-

From the McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, Dallas.

quencing. The National Institute for Human Genome Research of the National Institutes of Health established a number of pilot sequencing projects, which so far have resulted in a relatively small amount of the human genomic sequence. These pilot projects are expected to “ramp-up” to a scale sufficient to enable completion of the sequence by 2005. Thus, while in 1998 the first half of the genome project is complete and only a small percentage of the genome has been sequenced, the project remains on track and the sequence is expected to be completed on or ahead of schedule.

## METHODS

The HGP is being conducted as a collaboration between large-scale sequencing centers coordinated by the National Institute for Human Genome Research and the Department of Energy. Following the developmental phase resulting in genetic and physical maps and technology for high throughput sequencing, the US project was organized into 9 large-scale genomic sequencing centers distributed among universities and research institutes. Each center has focused on a specific portion of the human genome, determined by scientific interest, and each center is applying a slightly different organization and technical approach to developing pilot scale sequencing efforts with capabilities of between 2 and 10 million bp. In addition, at least 5 international centers are similarly developing sequencing capacity sufficient to complete a substantial portion of the human genome sequence. The US centers are presently located at Baylor College of Medicine, Houston, Tex; University of Oklahoma, Norman; University of Texas Southwestern Medical Center, Dallas; Stanford University, Stanford, Calif; the Whitehead Institute/Massachusetts Institute of Technology, Boston; University of Washington, Seattle; Washington University, St Louis, Mo; the Institute for Genome Research, Baltimore, Md; and the Joint Genome Institute of the Department of Energy, Walnut Creek, Calif. The efforts remain highly interactive and collaborative while



*Genomes that have been sequenced or are being sequenced.*

competing for scale-up funds through a National Institutes of Health grant mechanism.

To complete the sequence of the human genome by 2005, at least 6 of the genome sequencing centers will need to reach an individual sequencing output of between 50 to 100 million bp and sustain that output for 5 years. About two thirds of the final sequence is expected to be completed in the United States and the remaining one third in international centers, predominantly the Sanger Center located in Hinxton, England, and funded by the Wellcome Trust, London, England. Other international sequencing centers are located in France, Germany, Japan, Canada, and Australia. The actual amount of sequencing done by each center will vary depending on the experience, size, efficiency, and funding of the center.

Large-scale human genomic sequencing combines science, technology, industrial management practices, and a factorylike production environment (**Table**). The actual process of sequencing in a large-scale DNA sequencing center is, in general, divided into a number of steps: (1) mapping; (2) shotgun cloning; (3) sequence production; (4) sequence finishing; and (5) an-

notation and publication. Mapping involves the use of detailed chromosome maps composed of ordered sequence-tagged sites that were constructed in the first half of the genome project. These sequence-tagged site markers from these maps are used to isolate the actual clones for sequencing from a human genome library. The source of the DNA used for the library construction was selected to ensure that the individual whose genome is being sequenced is anonymous and has given appropriate informed consent to have their genome sequenced and deposited in the public domain.

## CLONING

When clones are isolated from the genomic library and assembled into a sequence-ready map, clones that form the minimal path through the region are selected for sequencing. These are transferred to the cloning laboratory and shotgun cloned into a sequencing vector. For a 120-kb bacterial artificial chromosome clone, about 3000 clones or thirty 96-well plates of clones are picked for sequencing. This level of redundancy is necessary to achieve the high accuracy sequence demanded by the genome project.

### Genome Sequences Completed or in Progress

Organism	Genome Size, Megabase	Estimated Genes
<i>Saccharomyces cerevisiae</i>	12.1	6034
<i>Escherichia coli</i>	4.6	4288
<i>Bacillus subtilis</i>	4.2	~4000
<i>Synechocystis</i> species	3.6	3168
<i>Archaeoglobus fulgidus</i>	2.2	2471
<i>Pyrobaculum aerophilum</i>	2.2	Not applicable
<i>Haemophilus influenzae</i>	1.8	1740
<i>Methanobacterium thermoautotrophicum</i>	1.8	1855
<i>Helicobacter pylori</i>	1.7	1590
<i>Methanococcus jannaschii</i>	1.7	1692
<i>Aquifex aolicus</i>	1.5	1508
<i>Borrelia burgdorferi</i>	1.3	863
<i>Treponema pallidum</i>	1.1	1234
<i>Mycoplasma pneumoniae</i>	0.8	677
<i>Mycoplasma genitalium</i>	0.6	470
<i>Caenorhabditis elegans</i>	100	Not applicable
<i>Drosophila melanogaster</i>	120	Not applicable
<i>Mus musculus</i>	3000	60 000-70 000
<i>Homo sapiens</i>	3000	60 000-70 000

### SEQUENCE PRODUCTION

Production sequencing is carried out using automated DNA sequencing instruments operated in a production, or factorylike, environment. In many genome-sequencing centers, DNA sequencing is carried out 24 hours a day, 7 days a week, to get the most efficient use of the expensive instrumentation. The data are collected, the quality of the raw sequence is analyzed, and then individual sequences of 300 to 800 bp in length are assembled into longer sequences using automated assembly programs. Depending on the characteristics of the sequence, including the initial data quality and other parameters, the data will be assembled into a number of contigs or groups of overlapping sequences that together form a larger sequence. Generally, 3000 high-quality reads will assemble into between 2 and 20 or more contigs representing 100 to 150 kb.

### SEQUENCE FINISHING

While production sequencing is carried out in a semi-industrial process, finishing the sequence must be done individually by highly trained laboratory personnel. Completing the sequence, filling gaps between contigs, sequencing difficult regions or regions of complex content, ensuring high sequence quality, and checking for errors are all included in a process called se-

quence finishing. Individual experienced scientists and technicians called *finishers* visually examine the initial sequence assembly and then design and carry out additional sequencing reactions to link the contigs. The finisher also accesses the quality of the sequence and checks for errors made by humans or computers. High sequence quality comes through redundant sequencing and consistent overlaps of the same region. The assembly programs determine statistical accuracy for the entire sequence and the finisher checks it visually to determine which regions need to be repeated. The finisher then resequences any regions where a base call is questionable. The final finished sequence has an accuracy level of no more than 1 discrepant base for every 10 000 bp.

The final stage in genomic sequencing is annotation of the sequence for content and submission to the database. The final sequence is initially annotated by automated computer programs to identify important features such as known repetitive sequences, known genes, mapping markers such as sequence-tagged sites, expressed sequence tags, predicted genes based on computer predictions or similarities with sequences from other organisms, structural features such as CpG islands, or telomeric repeats. The annotated sequence is then submitted to GenBank, maintained at the National Center for Biotechnology Information at the Na-

tional Library of Medicine where it is immediately available to the scientific community through the internet (<http://ncbi.nlm.nih.gov>). Also, most large-scale sequence centers make all the preliminary, incomplete sequence available every 24 hours either through individual Web sites or through GenBank, where they are annotated as incomplete sequence. In addition, the National Center for Biotechnology Information maintains the Human Genome Sequencing Index,<sup>5</sup> a database listing regions of the genome being sequenced, or planned for sequencing, by each of the international sequencing centers. In this way, investigators interested in specific genes or specific regions of the genome may determine if that region has been sequenced or is under investigation as part of the genome project.

### RELEVANCE TO CLINICAL PRACTICE AND THE PRACTICING NEUROLOGIST

From the discovery of the DNA double helix by Watson and Crick in 1953 to the present day, our knowledge of human genetics is based on, at most, 2% to 3% of the entire sequence of the genome, and thus on about 2% to 3% of all human genes. The HGP was designed to complete the entire sequence before the year 2005, expanding our knowledge of human genes from 2% to 3% to 100% within the next 7 years. While it is hard to quantify the effect this is likely to have on medical practice in the short- and long-term, we may comfortably predict that this expansion in medical knowledge will move into the clinic with an increasing rate.<sup>6</sup> The practicing neurologist is likely to find an increasing dependence on DNA-based diagnostic and therapeutic approaches over the next 20 years.

Well before the HGP is completed, the resulting flood of information is expected to have major ramifications for clinical practice. This is already beginning to happen, at least with respect to the diagnosis of some hereditary diseases and cancers. In the case of diseases like Huntington or Alzheimer disease, patients frequently tell their physicians that they have the disease in the family and request ge-

netic testing. While only a handful of genetic tests for neurologic diseases are currently available, the rapid pace of genomics is likely to increase this number by many orders of magnitude in the next few years. Thus, in the near future, health care providers will need to have sufficient expertise in genetics to advise patients whether to investigate genetic susceptibilities and undergo testing. In addition to the armamentarium of neurologic studies, the practicing neurologist will need sufficient expertise and knowledge of basic Mendelian and non-Mendelian inheritance patterns and their implications to understand probabilistic statements of disease risk. The neurologist will need to have immediate accessibility to facts pertaining to a specific genetic disease. Such information is already available over the Internet, a source of information that is increasing exponentially.<sup>7</sup>

More commonly, physicians will hear accounts of breakthroughs in the genetics of neurologic disease or other genetic disease through the lay press or newspapers. Often, they may first be approached by patients themselves who may have the disease or who may have family members with the disease. These patients will be seeking expert advice on the value of these discoveries to their particular condition.

#### RELEVANCE TO THE STUDY OF NEUROSCIENCE

The human genome contains 60 000 to 70 000 genes. Our current knowledge of tissue-specific gene expression says at least 50% of these genes are expressed in the brain. At the present time, our knowledge of the molecular genetics of the nervous system is based only on a small fraction of these genes. The completion of the HGP promises to provide a wealth of raw material for the molecular neuroscientist including new receptors, neural membrane proteins, neuron-specific transcription factors, and many other components of the machinery of the brain.

In addition to the sequences of newly discovered genes, new technology for gene analysis that has been stimulated by the genome project

promises to increase the speed, accuracy, and depth of information that can be accumulated. As an example, the expression analysis of tens of thousands of genes is becoming feasible using "DNA chip" technology in which oligonucleotide probes are attached to a microelectronic device allowing near-instantaneous quantification of the expression levels of messenger RNAs in a specific cell or tissue. The DNA chip technology being developed for rapid DNA diagnostics in medical practice will have far-reaching implications for basic research as well allowing a genome-wide profiling of gene expression in specific cells and tissues. The complete sequence, coupled with technology for rapid gene tracing, will allow an incredibly detailed look into the genome function of the neuron.

The roundworm *C elegans* is an important model organism for human genome analysis and provides an important insight into the type of whole organism analysis being introduced by the HGP. *Caenorhabditis elegans* consists of 959 somatic cells, all of which have been identified, named, and their lineage determined. The complete genomic sequence of 100 million bp will be completed next year, enabling DNA chip expression analysis and the determination of the levels of expression of each of the 14 000 genes in each of the 958 cells at all points of development. In addition, the 302 neurons and 56 glial cells that form the nervous system, and the synapses that connect them, have been completely defined so that the complete wiring diagram of synapsis of every neuron cell with every other has been described.<sup>8</sup> Within the next 2 years, *C elegans* may come as near as possible to being completely described at the cellular and molecular level. The HGP, with the sequencing of the human DNA and its expression profiling in diverse cell types, will lead the way to perhaps the determination of the complete wiring diagram of the human brain, perhaps the next large-scale biology project following in the footsteps of human genomics. The advent of genomewide thinking among molecular geneticists, and soon among neuroscientists, may move neuroscience forward into the next level of understanding of hu-

man neurophysiology, development, and behavior.

#### CONCLUSIONS

In the future, it is clear that genetic discoveries, including the vast amount of information being made available by the HGP, will lead to opportunities for treatment. The progress in treatment will tend to reduce the dilemmas surrounding genetic testing. In some cases, the neurologist will still be left with the dilemma of a disease that can be accurately diagnosed but for which no effective therapy is yet available. However, the vast array of biological information made available by the genome project will also allow widespread approaches of emerging technologies such as gene therapy, in which defective gene sequences may be replaced. In the case of other disorders, the knowledge gained by genetic dissection may lead to new, more effective drug therapies or immediate approaches to improving health through altering lifestyle. Biomedicine in the next century is likely to take on an entirely new role in society based on the beginnings inherent in the HGP.

Accepted for publication April 24, 1998.

Reprints: Glen A. Evans, MD, PhD, McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center at Dallas, 5323 Harry Hines Blvd, Dallas, TX 75235.

#### REFERENCES

1. Gabor-Mikolos GL, Rubin GM. The role of the genome project in determining gene function: insights from model organisms. *Cell*. 1996;86:521-529.
2. Rowen L, Mahairas G, Hood L. Sequencing the human genome. *Science*. 1997;278:605-607.
3. Collins F. Positional cloning: let's not call it reverse anymore. *Nat Genet*. 1992;1:3-6.
4. Selleri L, Smith MW, Holmsen AL, et al. High-resolution physical mapping of a 250-kb region of human chromosome 11q24 by genomic sequence sampling (GSS). *Genomics*. 1995;26:489-501.
5. National Center for Biotechnology Information. Available at: <http://www.ncbi.nlm.nih.gov/hugo/>.
6. Collins F. Sequencing the human genome. *Hosp Pract*. 1997;32:35-50.
7. Lawrence S, Giles CL. Searching the World Wide Web. *Science*. 1998;280:98-99.
8. Chalfie M, White J. *The Nematode Caenorhabditis elegans*. New York, NY: CSH Press; 1988: 337-391.