Human Genome Project

From Wikipedia, the free encyclopedia

The Human Genome Project (HGP) was an international scientific research project with a primary goal to determine the sequence of chemical base pairs which make up DNA and to identify and map the approximately 20,000–25,000 genes of the human genome from both a physical and functional standpoint. The first available assembly of the genome was completed in 2000 by the UCSC Genome Bioinformatics Group, composed of Jim Kent (then a UCSC graduate student of molecular, cell and developmental biology), Patrick Gavin, Terrence Furey and David Kulp.

The project began in 1990, initially headed by James D. Watson at the U.S. National Institutes of Health. A working draft of the genome was released in 2000 and a complete one in 2003, with further analysis still being published. A parallel project was conducted outside of government by the Celera Corporation. Most of the government-sponsored sequencing was performed in universities and research centers from the United States, the United Kingdom, Canada, and New Zealand. The mapping of human genes is an important step in the development of medicines and other aspects of health care.

The HGP originally aimed to map the nucleotides contained in a haploid reference human genome (more than three billion). Several groups have announced efforts to extend this to diploid human genomes including the International HapMap Project, Applied Biosystems, Perlegen, Illumina, JCVI, Personal Genome Project, and Roche-454.

The "genome" of any given individual (except for identical twins and cloned organisms) is unique; mapping "the human genome" involves sequencing multiple variations of each gene. The project did not study the entire DNA found in human cells; some heterochromatic areas (about 8% of the total genome) remain un-sequenced.

Background

The project began with the culmination of several years of work supported by the United States Department of Energy, in particular workshops in 1984 and 1986 and a subsequent initiative of the US Department of Energy. This 1987 report stated boldly, "The ultimate goal of this initiative is to understand the human genome" and "knowledge of the human as necessary to the continuing progress of medicine and other health sciences as knowledge of human anatomy has been for the present state of medicine." Candidate technologies were already being considered for the proposed undertaking at least as early as 1985.

James D. Watson was head of the National Center for Human Genome Research at the National Institutes of Health (NIH) in the United States starting from 1988. Largely due to his disagreement with his boss, Bernadine Healy, over the issue of patenting genes, Watson was forced to resign in 1992. He was replaced by Francis Collins in April 1993, and the name of the Center was changed to the National Human Genome Research Institute (NHGRI) in 1997.

The $3-billion project was formally founded in 1990 by the United States Department of Energy and the U.S. National Institutes of Health, and was expected to take 15 years. In addition to the United States, the international consortium comprised geneticists in the United Kingdom, France, Germany, Japan, China, and India.

Due to widespread international cooperation and advances in the field of genomics (especially in sequence analysis), as well as major advances in computing technology, a 'rough draft' of the genome was finished in 2000 (announced jointly by then US president Bill Clinton and the British Prime Minister Tony Blair on June 26, 2000). Ongoing sequencing led to the announcement of the essentially complete genome in April 2003, 2 years earlier than planned. In

May 2006, another milestone was passed on the way to completion of the project, when the sequence of the last chromosome was published in the journal Nature.

State of completion

There are multiple definitions of the "complete sequence of the human genome". According to some of these definitions, the genome has already been completely sequenced, and according to other definitions, the genome has yet to be completely sequenced. There have been multiple popular press articles reporting that the genome was "complete." The genome has been completely sequenced using the definition employed by the International Human Genome Project. A graphical history of the human genome project shows that most of the human genome was complete by the end of 2003. However, there are a number of regions of the human genome that can be considered unfinished:

- First, the central regions of each chromosome, known as centromeres, are highly repetitive DNA sequences that are difficult to sequence using current technology. The centromeres are millions (possibly tens of millions) of base pairs long, and for the most part these are entirely un-sequenced.
- Second, the ends of the chromosomes, called telomeres, are also highly repetitive, and for most of the 46 chromosome ends these too are incomplete. It is not known precisely how much sequence remains before the telomeres of each chromosome are reached, but as with the centromeres, current technological restraints are prohibitive.
- Third, there are several loci in each individual's genome that contain members of multigene families that are difficult to disentangle with shotgun sequencing methods - these multigene families often encode proteins important for immune functions.
- Other than these regions, there remain a few dozen gaps scattered around the genome, some of them rather large, but there is hope that all these will be closed in the next couple of years.

In summary: the best estimates of total genome size indicate that about 92.3% of the genome has been completed and it is likely that the centromeres and telomeres will remain un-sequenced until new technology is developed that facilitates their sequencing. Most of the remaining DNA is highly repetitive and unlikely to contain genes, but it cannot be truly known until it is entirely sequenced. Understanding the functions of all the genes and their regulation is far from complete. The roles of junk DNA, the evolution of the genome, the differences between individuals, and many other questions are still the subject of intense interest by laboratories all over the world.

Goals

The sequence of the human DNA is stored in databases available to anyone on the Internet. The U.S. National Center for Biotechnology Information (and sister organizations in Europe and Japan) house the gene sequence in a database known as GenBank, along with sequences of known and hypothetical genes and proteins. Other organizations such as the University of California, Santa Cruz, and Ensembl present additional data and annotation and powerful tools for visualizing and searching it. Computer programs have been developed to analyze the data, because the data themselves are difficult to interpret without such programs.

The process of identifying the boundaries between genes and other features in raw DNA sequence is called genome annotation and is the domain of bioinformatics. While expert biologists make the best annotators, their work proceeds slowly, and computer programs are increasingly used to meet the high-throughput demands of genome sequencing projects. The best

current technologies for annotation make use of statistical models that take advantage of parallels between DNA sequences and human language, using concepts from computer science such as formal grammars.

Another, often overlooked, goal of the HGP is the study of its ethical, legal, and social implications. It is important to research these issues and find the most appropriate solutions before they become large dilemmas whose effect will manifest in the form of major political concerns.

All humans have unique gene sequences. Therefore the data published by the HGP does not represent the exact sequence of each and every individual's genome. It is the combined genome of a small number of anonymous donors. The HGP genome is a scaffold for future work in identifying differences among individuals. Most of the current effort in identifying differences among individuals involves single nucleotide polymorphisms and the HapMap.

Key findings of Genome Project:

1. There are approx. 24,000 genes in human beings, the same range as in mice and twice that of roundworms. Understanding how these genes express themselves will provide clues to how diseases are caused.
2. All human races are 99.99 % alike, so racial differences are genetically insignificant.
3. Most genetic mutation occurs in the male of the species and as such are agents of change. They are also more likely to be responsible for genetic disorders.
4. Genomics has led to advances in genetic archaeology and has improved our understanding of how we evolved as humans and diverged from apes 25 million years ago. It also tells how our body works, including the mystery behind how the sense of taste works.

Human Genome Project is called a Mega Project because of the following facts:

1. The human genome has approx. 3.3×10^9 base-pairs; if the cost of sequencing is US $3 per base-pair, then the approx. cost will be US $10 billion.
2. If the sequence obtained were to be stored in a typed form in books and if each page contains 1000 letters and each book contains 1000 pages, then 3300 such books would be needed to store the complete information.
3. The enormous quantity of data expected to be generated also necessitates the use of high speed computer hard-drives for data storage and super-computers for retrieval and analysis.

History

In 1976, the genome of the RNA virus Bacteriophage MS2 was the first complete genome to be determined, by Walter Fiers and his team at the University of Ghent (Ghent, Belgium). The idea for the shotgun technique came from the use of an algorithm that combined sequence information from many small fragments of DNA to reconstruct a genome. This technique was pioneered by Frederick Sanger to sequence the genome of the Phage Φ-X174, a virus (bacteriophage) that primarily infects bacteria that was the first fully sequenced genome (DNA-sequence) in 1977. The technique was called shotgun sequencing because the genome was broken into millions of pieces as if it had been blasted with a shotgun. In order to scale up the method, both the sequencing and genome assembly had to be automated, as they were in the 1980s.

Those techniques were shown applicable to sequencing of the first free-living bacterial genome (1.8 million base pairs) of *Haemophilus influenzae* in 1995 and the first animal genome (~100 Mbp). It involved the use of automated sequencers, longer individual sequences using

approximately 500 base pairs at that time. Paired sequences separated by a fixed distance of around 2000 base pairs which were critical elements enabling the development of the first genome assembly programs for reconstruction of large regions of genomes (aka 'contigs').

Three years later, in 1998, the announcement by the newly-formed Celera Genomics that it would scale up the shotgun sequencing method to the human genome was greeted with skepticism in some circles. The shotgun technique breaks the DNA into fragments of various sizes, ranging from 2,000 to 300,000 base pairs in length, forming what is called a DNA "library". Using an automated DNA sequencer the DNA is read in 800bp lengths from both ends of each fragment. Using a complex genome assembly algorithm and a supercomputer, the pieces are combined and the genome can be reconstructed from the millions of short, 800 base pair fragments. The success of both the public and privately funded effort hinged upon a new, more highly automated capillary DNA sequencing machine, called the Applied Biosystems 3700, that ran the DNA sequences through an extremely fine capillary tube rather than a flat gel. Even more critical was the development of a new, larger-scale genome assembly program, which could handle the 30–50 million sequences that would be required to sequence the entire human genome with this method. At the time, such a program did not exist. One of the first major projects at Celera Genomics was the development of this assembler, which was written in parallel with the construction of a large, highly automated genome sequencing factory. Development of the assembler was led by Brian Ramos. The first version of this assembler was demonstrated in 2000, when the Celera team joined forces with Professor Gerald Rubin to sequence the fruit fly Drosophila melanogaster using the whole-genome shotgun method. At 130 million base pairs, it was at least 10 times larger than any genome previously shotgun assembled. One year later, the Celera team published their assembly of the three billion base pair human genome.

The Human Genome Project was a 13 year old mega project, that was launched in the year 1990 and completed in 2003. This project is closely associated to the branch of biology called Bio-informatics. The human genome project international consortium announced the publication of a draft sequence and analysis of the human genome—the genetic blueprint for the human being. An American company—Celera, led by Craig Venter and the other huge international collaboration of distinguished scientists led by Francis Collins, director, National Human Genome Research Institute, U.S., both published their findings.

This Mega Project is co-ordinated by the U.S. Department of Energy and the National Institute of Health. During the early years of the project, the Wellcome Trust (U.K.) became a major partner, other countries like Japan, Germany, China and France contributed significantly. Already the atlas has revealed some starting facts. The two factors that made this project a success are:

1. Genetic Engineering Techniques, with which it is possible to isolate and clone any segment of DNA.
2. Availability of simple and fast technologies, to determining the DNA sequences.

Being the most complex organisms, human beings were expected to have more than 100,000 genes or combination of DNA that provides commands for every characteristics of the body. Instead their studies show that humans have only 30,000 genes – around the same as mice, three times as many as flies, and only five times more than bacteria. Scientist told that not only are the numbers similar, the genes themselves, baring a few, are alike in mice and men. In a companion volume to the Book of Life, scientists have created a catalogue of 1.4 million single-letter differences, or single nucleotide polymorphisms (SNP's) – and specified their exact locations in the human genome. This SNP map, the world's largest publicly available catalogue of SNP's,

promises to revolutionize both mapping diseases and tracing human history. The sequence information from the consortium has been immediately and freely released to the world, with no restrictions on its use or redistribution. The information is scanned daily by scientists in academia and industry, as well as commercial database companies, providing key information services to bio-technologists. Already, many genes have been identified from the genome sequence, including more than 30 that play a direct role in human diseases. By dating the three millions repeat elements and examining the pattern of interspersed repeats on the Y-chromosome, scientists estimated the relative mutation rates in the X and the Y chromosomes and in the male and the female germ lines. They found that the ratio of mutations in male Vs female is 2:1. Scientists point to several possible reasons for the higher mutation rate in the male germ line, including the fact that there are a greater number of cell divisions involve in the formation of sperm than in the formation of eggs.

Benefits

The work on interpretation of genome data is still in its initial stages. It is anticipated that detailed knowledge of the human genome will provide new avenues for advances in medicine and biotechnology. Clear practical results of the project emerged even before the work was finished. For example, a number of companies, such as Myriad Genetics started offering easy ways to administer genetic tests that can show predisposition to a variety of illnesses, including breast cancer, disorders of hemostasis, cystic fibrosis, liver diseases and many others. Also, the etiologies for cancers, Alzheimer's disease and other areas of clinical interest are considered likely to benefit from genome information and possibly may lead in the long term to significant advances in their management.

There are also many tangible benefits for biological scientists. For example, a researcher investigating a certain form of cancer may have narrowed down his/her search to a particular gene. By visiting the human genome database on the world wide web, this researcher can examine what other scientists have written about this gene, including (potentially) the three-dimensional structure of its product, its function(s), its evolutionary relationships to other human genes, or to genes in mice or yeast or fruit flies, possible detrimental mutations, interactions with other genes, body tissues in which this gene is activated, diseases associated with this gene or other datatypes.

Further, deeper understanding of the disease processes at the level of molecular biology may determine new therapeutic procedures. Given the established importance of DNA in molecular biology and its central role in determining the fundamental operation of cellular processes, it is likely that expanded knowledge in this area will facilitate medical advances in numerous areas of clinical interest that may not have been possible without them.

The analysis of similarities between DNA sequences from different organisms is also opening new avenues in the study of evolution. In many cases, evolutionary questions can now be framed in terms of molecular biology; indeed, many major evolutionary milestones (the emergence of the ribosome and organelles, the development of embryos with body plans, the vertebrate immune system) can be related to the molecular level. Many questions about the similarities and differences between humans and our closest relatives (the primates, and indeed the other mammals) are expected to be illuminated by the data from this project.

The Human Genome Diversity Project (HGDP), spinoff research aimed at mapping the DNA that varies between human ethnic groups, which was rumored to have been halted, actually did continue and to date has yielded new conclusions. In the future, HGDP could possibly expose

new data in disease surveillance, human development and anthropology. HGDP could unlock secrets behind and create new strategies for managing the vulnerability of ethnic groups to certain diseases (see race in biomedicine). It could also show how human populations have adapted to these vulnerabilities.

Ethical, legal and social issues

The project's goals included not only identifying all of the approximately 24,000 genes in the human genome, but also to address the ethical, legal, and social issues (ELSI) that might arise from the availability of genetic information. Five percent of the annual budget was allocated to address the ELSI arising from the project.

Debra Harry, Executive Director of the U.S group Indigenous Peoples Council on Biocolonialism (IPCB), says that despite a decade of ELSI funding, the burden of genetics education has fallen on the tribes themselves to understand the motives of Human genome project and its potential impacts on their lives. Meanwhile, the government has been busily funding projects studying indigenous groups without any meaningful consultation with the groups. (See Biopiracy.)

The main criticism of ELSI is the failure to address the conditions raised by population-based research, especially with regard to unique processes for group decision-making and cultural worldviews. Genetic variation research such as HGP is group population research, but most ethical guidelines, according to Harry, focus on individual rights instead of group rights. She says the research represents a clash of culture: indigenous people's life revolves around collectivity and group decision making whereas the Western culture promotes individuality. Harry suggests that one of the challenges of ethical research is to include respect for collective review and decision making, while also upholding the Western model of individual rights.

The distribution of genes in mammalian chromosomes is striking. It turns out that human chromosomes have crowded urban centres with many genes in close proximity to one another and also vast expanses of unpopulated desert where only non coding DNA can be found. This distribution of genes is in marked contrast to the genomes of many other organisms. The full set of proteins encoded by the human genome is more complex than those of the invertebrates because humans have rearranged old protein domains into a rich collection of new architectures. The sequence will serve as a foundation for a broad range of functional genomic tools to help biologists to probe the function of the genes in a more systematic manner. Comparative genomics will also offer scientists insights into important regions in the sequence that perform regulatory functions. The human genome sequence provides a great help to build the tools to conquer most of the illness that cause untold human sufferings and premature death. Already the genome has helped to detect more than 30 diseased genes, including some of the common diseases like breast cancer, colour blindness etc. There will be a lot more emphasis now on preventive medicines. The consortium's ultimate goal is to produce a completely 'finished' sequence with no gaps and 99.9 % accuracy. Although the near finished version is adequate for most biomedical research, the Human Genome Project has made a commitment to filling all gaps and resolving all uncertainty in the sequence by the year 2003 C.E. The draft genome sequence has provided an initial look at the human gene content, but many uncertainties remain. One of the Human Genome Project priorities will be to refine the data to accurately reflect every gene and every alternatively spliced form. Several steps are needed to reach this ambitious goal.