Human Genome Project

Student information

Introduction

The human genome contains more than three billion DNA base pairs and all of the genetic information needed to make us. The human genome was first mapped and sequenced over a period of 13 years from 1990 to 2003. The Human Genome Project (HGP) was a ground-breaking international initiative, considered to be one of the most ambitious scientific projects undertaken in the twentieth century.

What is a genome and why does it need to be sequenced?

All organisms, from bacteria to elephants have genomes. A genome contains the genetic information needed to make a living organism, written in the DNA four-letter code of bases, or nucleotides. Sequencing an organism's genome gives us a comprehensive view of this information, with which we can better understand their evolution, development and biological functions. Sequencing the human genome has helped researchers to identify important genes and genetic sequences, to better understand their role in disease, and to investigate our origins using variations in the DNA sequence.

What was the Human Genome Project?

The Human Genome Project aimed to sequence the entire human genome and provide the data free to the world. It was the first major global collaboration of its kind and the largest biological research project ever undertaken, involving thousands of staff in institutes across the globe.

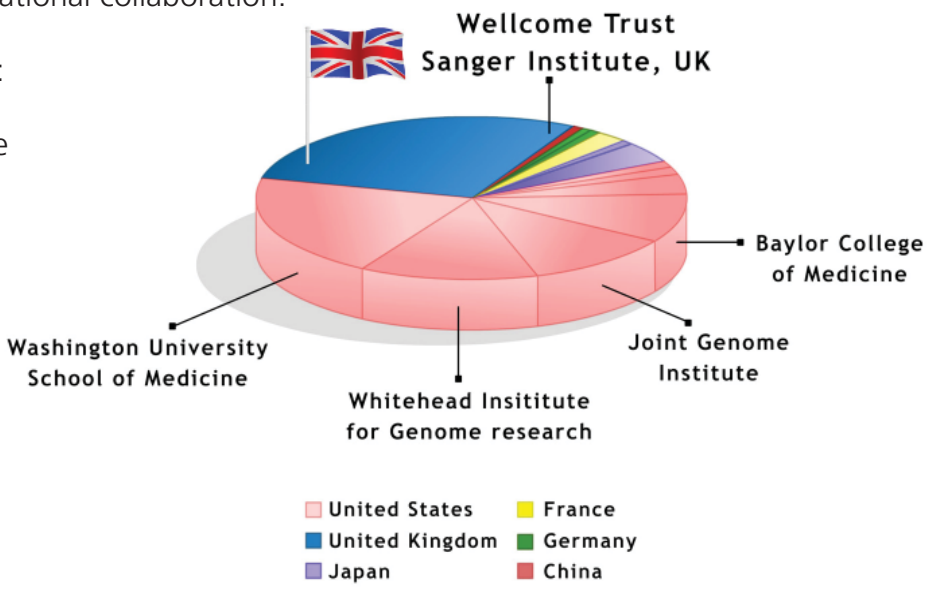
Before the HGP started in 1990 many scientists felt that sequencing the human genome was an impossible challenge, because of its sheer size. By assigning different portions of the genome to different research groups in a coordinated and efficient way, the HGP researchers were able to overcome this challenge.

Who took part in the project?

Twenty institutes from six different countries (China, France, Germany, Japan, UK and USA) were involved in the HGP making it a truly international collaboration.

The five biggest contributors were:

- Wellcome Trust Sanger Institute
- Washington University School of Medicine
- Whitehead Institute/MIT centre for Genome research
- The DOE's Joint Genome Institute
- Baylor College of Medicine

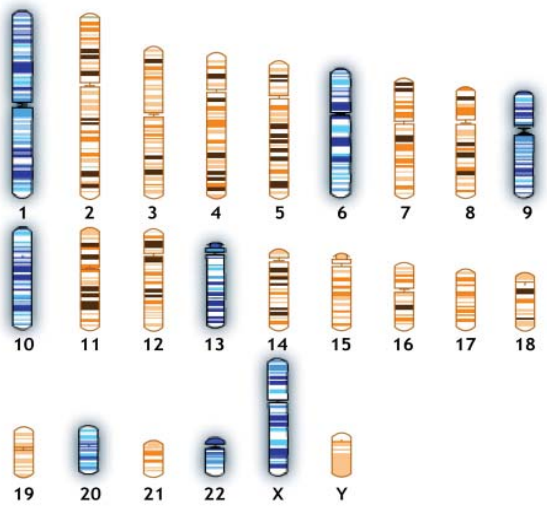


Human Genome Project

Student information

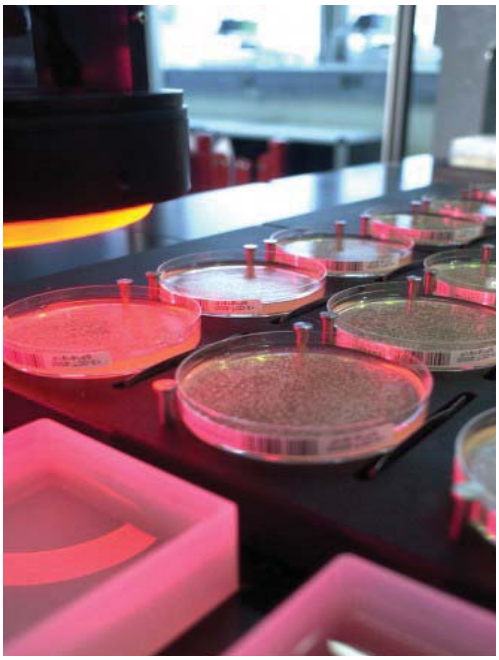
What was the Sanger Institute's contribution?

The Wellcome Trust Sanger Institute (WTSI) was the single largest contributor to the HGP. Initially funded to sequence 1/6th of the genome it acquired additional funding to sequence just under a third of the entire human genome (29%). The diagram on the right shows the chromosomes that were sequenced, or partially sequenced, by the Sanger Institute coloured in blue.



How was the human genome sequenced?

Sequencing technology can only sequence a few hundred base pairs of DNA at a time. This meant that the three billion base pair human genome had to be broken up into small pieces for sequencing, which were then reassembled like a giant jigsaw puzzle.



Petri dishes arranged so that a robot can select and pick off colonies of the bacteria, *E. coli*. The bacterial cells contain cloned DNA that will be prepared for sequencing. © Wellcome Images

The genome was first broken into 200,000 base pair sections (clones) and inserted into bacterial DNA, creating living libraries of the DNA clones. These libraries could be copied and shipped between collaborating institutes. The clones could then be broken into smaller pieces (4000-6000 base pairs), re-inserted into bacteria and cultured to make enough DNA for sequencing.

To sequence the DNA, bacterial colonies were transferred to tubes where the cells were lysed (split open) and the DNA extracted. These sections of DNA were then sequenced by machine using the Sanger sequencing method. The resulting data was pieced together by computers and researchers, to form the whole genome sequence.

For a more detailed explanation of how the genome was sequence watch the animations "How the human genome was sequenced", "DNA libraries", "Subcloning" and "Sanger sequencing method" at www.yourgenome.org

How long did the Human Genome Project take?

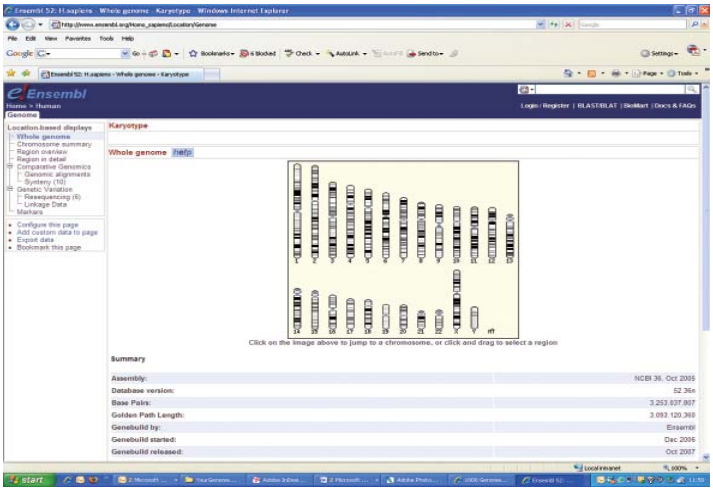
In total the HGP took 13 years; it was expected to take more than 15 years. The project officially started in October 1990, a first "draft" was announced in June 2000 and a "finished" sequence was completed in April 2003 which was published in 2004. Work continues to refine the "reference" human genome sequence and it is being continually updated as new data becomes available.

Human Genome Project

Student information

Who has access to the human genome data?

Put simply, everyone. One of the major principles of the HGP was to provide free and open access to the data for everyone in the scientific community and the public domain. The HGP sequence data was deposited in freely available, online public databases as soon as was possible, with the target of putting the data online 24 hours. Genome browsers such as www.ensembl.org are continually developing and expanding to act as repositories for genomic data which can be accessed and used by anyone.



Ensembl is just one of the genome browsers available to researchers worldwide. It provides access to more than 50 species' genomes

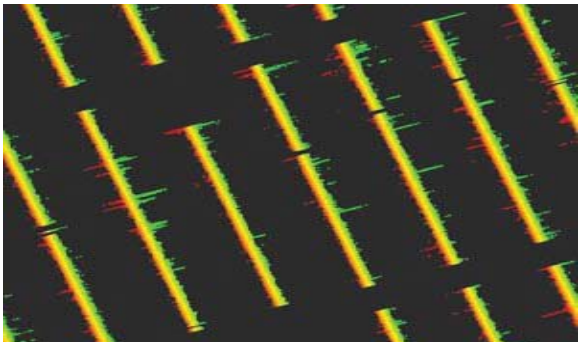
What happened after the human genome was finished?

Although the Human Genome Project has officially finished, research continues on the human reference sequence, identifying new features, functions and comparing variation between individuals, populations and other species. Researchers are also still working on filling in the "gaps" in the sequence. Several major international projects and collaborations have also been established since the completion of the HGP. These include:

International Hapmap project

Much of our genetic variation is caused by differences in single bases, or nucleotides within our DNA code: these are called single nucleotide polymorphisms, or SNPs. Each of us has a unique genetic code that typically differs in about three million nucleotides from every other person.

The HapMap project was a three year publicly and privately funded project that aimed to chart the patterns of genetic variation common in the world's population. The results released in 2005 showed patterns of association between SNPs studied in different populations from around the world and these can be used to simplify studies to understand how genetic variation contributes to health and disease. http://www.sanger.ac.uk/Info/Press/2005/051026.shtml; http://www.hapmap.org/



Copy Number Variation

As well as single base differences, studies using the human genome sequence revealed a new type of genetic variation, in which large regions of DNA are absent from or duplicated in different individuals. The discovery and publication of this 'Copy Number Variation' (CNV) in 2006 has changed our understanding and research of genetic factors in disease. http://www.sanger.ac.uk/Info/Press/2006/061122.shtml

Human Genome Project

Student information

Wellcome Trust Case Control Consortium (WTCCC)

This nine million pound study examined DNA samples from 17,000 people across the UK, to better understand the genetic factors underlying common diseases and was published in 2007. The study identified genetic variation across the genome associated with tuberculosis, coronary heart disease, type 1 diabetes, type 2 diabetes, rheumatoid arthritis, Crohn's disease, bipolar disorder and hypertension (high blood pressure). It was the largest study of its kind and brought together 50 leading research groups and 200 scientists in the field of human genetics from dozens of UK institutions including the Wellcome Trust Sanger Institute, Cambridge University and Oxford University.

Following the success of the WTCCC project, a follow up project (WTCCC2) gained £30 million funding in 2008 to sequence DNA samples from 120,000 people. The WTCCC2 project will perform genome-wide association studies for 13 disease conditions: Ankylosing spondylitis, Barrett's oesophagus and oesophageal adenocarcinoma, glaucoma, ischaemic stroke, multiple sclerosis, pre-eclampsia, Parkinson's disease, psychosis endophenotypes, psoriasis, schizophrenia, ulcerative colitis and visceral leishmaniasis.
http://www.sanger.ac.uk/Info/Press/2007/070606.shtml
http://www.sanger.ac.uk/Info/Press/2008/080414.shtml
http://www.wtccc.org.uk/

WTCCC2 will also investigate the genetics of reading and mathematics abilities in children and genetic influences on how patients respond to "statin" drugs. © Wellcome Images

1000 Genomes Project

1000 Genomes Project is an international research consortium that was launched in January 2008. The project aims to sequence the genomes of at least a thousand people from around the world, to identify very clearly those variations between individuals that are medically important and map these on the genome. The project is supported by the Wellcome Trust Sanger Institute, the Beijing Genomics Institute, Shenzhen (BGI Shenzhen) in China and the National Human Genome Research Institute (NHGRI), part of the National Institutes of Health (NIH).

As with other major human genome reference projects, data from the 1000 Genomes Project will be made swiftly available to the worldwide scientific community through freely accessible public databases.
http://www.sanger.ac.uk/Info/Press/2008/080122.shtml; www.1000genomes.org

Further reading & weblinks

http://www.yourgenome.org/hgp/
http://www.nature.com/nature/supplements/collections/humangenome/index.html
http://dnai.org
http://doegenomes.org