Steve Scott

Key words
genome
DNA
sequencing
sharing data

Ten years on

The Human Genome Project today

On Monday 26th June 2000, the US President, Bill Clinton, and UK Prime Minister, Tony Blair, announced a significant scientific landmark. Scientists had completed a first draft of the human genome, the DNA instructions for making a human being. The project had taken 10 years and promised a new age of genetic discovery and a revolution in medicine. But what has happened since the release of this draft sequence? Have there been any surprises? And has it really had much of an impact? ***Steve Scott*** *of the Wellcome Sanger Institute brings the story up-to-date.*

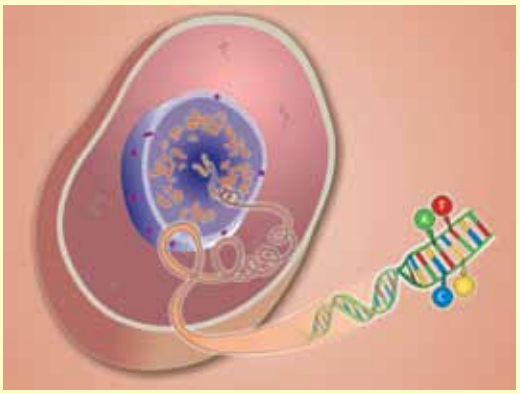
Big biology

The Human Genome Project (HGP) was the first 'Big Science' project in biology, a phrase usually used to describe large-scale government-funded science projects focusing on physics and astronomy. The HGP involved scientists from 20 centres in the USA, UK, Japan, France, Germany and China, working together to map the human genome (see Box 1). It introduced scientists to a new style of collaborative research involving the sharing of information, expertise and resources for a common goal.

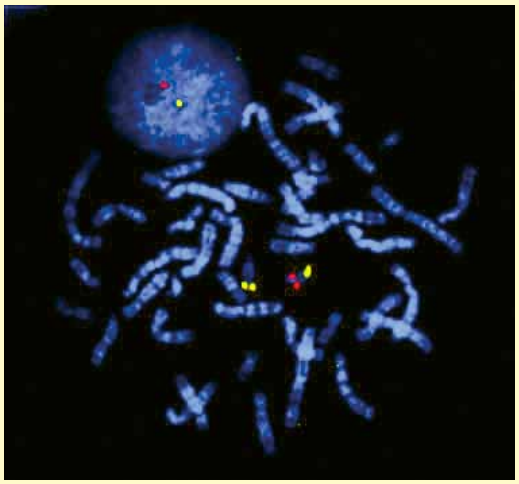
Since the HGP, many consortia have been established to study specific areas of biology. For example, the 1000 Genomes Project was launched

A small part of the Human Genome

Box 1: What is a genome?



A schematic picture showing the relationship between the structure of a typical cell, the nucleus, and the DNA stored in it. All the DNA in a nucleus is the genome.



A fluorescence in-situ hybridisation (FISH) image of chromosomes from a patient with Digeorge syndrome

All organisms, from bacteria to complex organisms like humans, have a genome. The genome is the genetic information needed to make the organism, written in a four-letter code of bases or nucleotides. It is made from a chemical called DNA, which looks like a twisted ladder with each rung of the ladder being made of a pair of bases. The Human Genome Project (HGP) succeeded in sequencing all 3 billion DNA bases of the human genome to create a list of all the genes in the genome. Sequencing the human genome has helped researchers to identify important genes and genetic sequences to better understand human evolution, development and disease.

in 2008 to sequence the genomes of 1000 people from Europe, East Asia, West Africa, South Asia and the Americas. It involves sequencing centres in the UK, China and the USA and will create a catalogue of human variation that will be used to identify genetic regions associated with human diseases. Sequencing technologies have improved and costs have fallen such that the study will now sequence the genomes of over 2000 individuals.

The International Cancer Genome Consortium (ICGC) involves researchers from 11 countries working together to sequence the genomes of more than 50 types of tumour. In 2009, the sequences of small-cell lung cancer, malignant melanoma (a form of skin cancer) and 24 breast cancer genomes were published. This knowledge will improve our understanding of the genetic mechanisms that control cancer development and identify potential targets for early diagnosis and treatments.

We knew from the outset that the human genome sequence was going to be an important resource to the scientific community. But how do you ensure all researchers have access to the information? A meeting in Bermuda in 1996 (Box 2) resulted in the drawing up of key principles to ensure that access to the human genome data was free and without restriction.

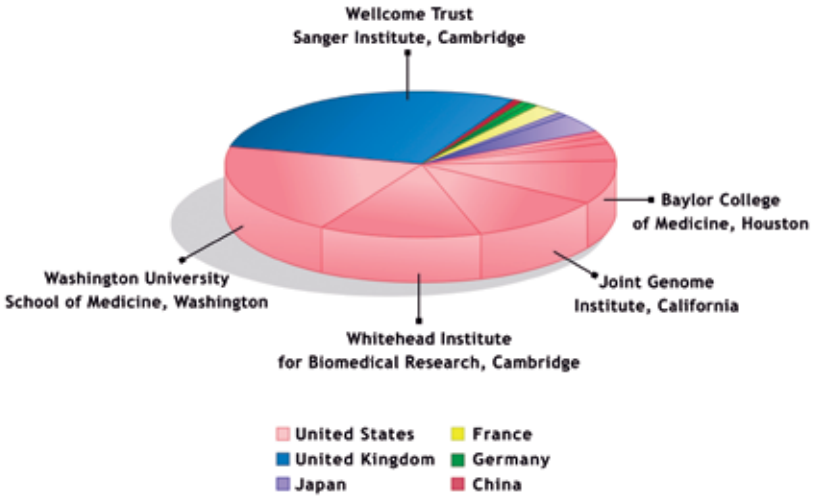
A member of the Sequencing Team at the Wellcome Trust Sanger Institute with Illumina sequencing machines

DNA sequence methods

In 1977, the first organism to have its whole DNA genome sequenced was that of a virus, bacteriophage phi X 174. This virus has just 5386 bases and only short pieces of DNA up to 50 letters long were sequenced at a time using radioactive molecules to label the DNA.

During the Human Genome Project pieces of DNA up to 1000 letters long were able to be sequenced using a faster, safer automated system that used fluorescent labels. However, even with these advances in technology it took several years to sequence and piece together the 3 billion letters of the human genome.

In the last six years, DNA sequencing technologies have taken further leaps forward. A human genome can now be sequenced within a day at a cost of less than $30 000 making it possible to compare and find differences between human genomes. Genome-wide association studies scan the genomes



The Human Genome Project was a collaboration between 20 centres in six countries

Box 2: The Bermuda Principles

On the 25th-28th February 1996 the key members of the Human Genome Project met in Bermuda. They discussed a strategy for sequencing the human genome and what should happen to the data generated. Free data release was high on the agenda of the meeting and discussions led to key principles being agreed. The Bermuda Principles have become the moral foundation for publicly-funded genome sequencing:

- All human genomic sequence information should be freely available in order to encourage research and development and to maximise its benefit to society.
- Sequence assemblies should be released as soon as possible.
- High quality or 'finished' annotated sequences should be submitted immediately to the public databases.

These principles have become fundamental in a drive for open access to data, information and published scientific articles. Today, genomic data are freely available on genome browsers such as Ensembl (www.ensembl.org) and UCSC (genome.ucsc.edu). Ensembl contains the genomes of over 50 vertebrate species from alpaca to zebrafish. You can access and download the entire human genome sequence from Ensembl. We wouldn't recommend you print it though, as you would need 150 000 sheets of A4 paper and a lot of ink cartridges for your printer!!

of many people for specific markers to find genetic variations associated with disease. The areas of the genome common in people with disease are investigated more closely to find out the role they play in disease. These studies have identified genes involved in diseases in cancer, diabetes, heart disease and obesity amongst others.

As the cost of sequencing continues to fall, the likelihood of you being able to afford to have your genome sequenced comes ever closer. This raises many questions. Would you want to get your genome sequenced? Who would you allow to have access to that information?

The issue of the scientific journal *Nature* in which the first sequencing of the complete human genome was reported

Fewer genes than expected

So how many genes do we have in our genomes? Surprisingly, it was revealed that the human genome contains only 23 000 protein-coding genes; a similar number to the worm and fly. So if only 2% of the human genome contains instructions for producing proteins, what does the rest do?

Approximately half of the genome is made up of repetitive DNA; regions containing a short sequence of DNA that is repeated many times and has no known function. Some of the remainder may have be regulatory, coding for RNA that switches genes on or off or coordinate the development. However, we are only just beginning to understand what some of this DNA does and more research is needed to uncover the mysteries of what all non-coding DNA in the genome is for.

Our DNA differs in many ways

The Human Genome sequence was derived from DNA samples from several anonymous donors. However, approximately 70% of the human genome sequence came from a single donor. As the project progressed it became clear that genomes from different individuals vary in many ways. Not only are there single letter differences throughout the DNA but sections and even big chunks are different.

Several projects have tried to compare genomes from different people. The International HapMap Project identified and catalogued differences using data from 270 different populations from African Asian and Europe. This information was then used to find genetic variants linked to human disease. DNA samples from individuals from the HapMap Project were also used to study variation in the human genome due to gains or losses of sections of DNA; so called copy number variation. This type of variation accounts for about 12% of the variation in the human genome, a similar amount to that due to single letter changes. But how important is this variation? How can we have such vast differences without any harmful effects?

The hunt for genetic variation in the human genome continues. The 1 000 Genomes Project was launched in 2008 and this year the UK10K project is beginning to sequence the whole or protein-coding parts of the genomes of 10 000 people from across the UK. UK10K will completely sequence the genomes of 4 000 people, including some twins, who have been studied for many years with their health and development extensively described. The remainder of the people being studied have a condition thought to have a genetic component such as severe obesity, autism, schizophrenia and congenital heart disease. With all this information we will gain a better understanding of the genetic variation between us. The challenge is to translate the knowledge we gain into improved healthcare.

The HGP legacy

The release of the draft human genome in June 2000 marked the dawn of a new era of biological discovery. However, there are still many unanswered questions about the human genome. How is the coordination of the development of the body controlled? What changes occur in cancer genomes and how can these be targeted to improve treatment? How do small and large changes to the sequence of DNA influence susceptibility to disease? There is still much to uncover and explore. Through these discoveries we will be able to advance our understanding of the human body, how it is affected in disease and how we can utilise genomics to improve healthcare.

Steve Scott works in the Communication and Public Engagement Programme, Wellcome Trust Sanger Institute, Cambridge, UK

The changing shape of DNA sequencing

The methods of sequencing of DNA have advanced during the last 30 years which has resulted in the visual outputs of sequence changing: (left) an autoradiograph of a radioactive sequencing gel (Image: Bart Barrell, Genome Research Limited); (centre) an automated sequencing output from the Human Genome Project with the four DNA bases represented by four colours (Image: Genome Research Limited); (right) DNA clusters on a next generation sequencing machine. Millions of DNA clusters are sequenced on a small flowcell (Image: Genome Research Limited).