

The Human Genome Project

Overview

In 1990, the Human Genome Project (HGP) was launched and a 15-year plan to sequence the human genome began. The 'human genome' is the complete set of DNA for a human being, found in every human cell, besides mature red blood cells. DNA consists of sets of paired bases and is arranged into 24 distinct chromosomes, each containing between 50 million and 250 million of these pairs. The human genome has around 3 billion DNA base pairs in total.

More than 2,000 scientists from over 20 institutes in six countries collaborated to produce a first draft of the genome in 2001. This first version had around 150,000 gaps and the order and orientation of many of the smaller segments had not been established.

By 2003 a second draft was ready. This time there were just 341 gaps in the published version and work continued as smaller teams sequenced individual chromosomes over the following three years. The final chromosome – and the largest – was published in the spring of 2006.

The entire collection of human chromosome DNA sequences is now freely available to the worldwide research community online.

The Human Genome Project since 2003

In April 2003, the US National Human Genome Research Institute (NHGRI) was heralding a revolution in biological research: the completion of the HGP meant that "the genomic era is now a reality."

The human genome sequence is essentially a detailed descriptive data set about the processes which make us human – it's like the source code for human software. The challenge now is to analyse the information in such a way that is useful and meaningful for scientific and medical research. Over the past four years, the scientific community has been adapting to this new reality. But researchers have not only set about utilising the raw information the HGP produced, they have also been drawing on the technological and research processes and practices which emerged over the course of its history. Unravelling the human genome has not only opened up new fields of scientific study; it has changed how we undertake the research itself.

Lessons learnt from the HGP

Francis Collins, a leading figure in the HGP, outlined in a *Nature* editorial four lessons for scientific research drawn from the experience of the project. Firstly, he emphasised the importance of making data freely and openly available to the public and scientific community to work on as soon as possible after its verification and prior to official publication. This policy underlies all of the HGP-derived research initiatives discussed below. It originated with the Bermuda Principles agreed by the international genome community in 1996 and lies at the ideological heart of the HGP's history.

In the mid-1990s, the public project to map the human genome was threatened by a private company offering to get the job done more quickly,

but to make a business out of selling the sequence under a proprietary licence. This provoked a backlash amongst the scientific community who balked at the idea that the code for human life might be sold as a commercial product. They argued that this was imposing dangerous restrictions on the field of genomics, even before this new area of biology had found its footing. In 1996 the Wellcome Trust called a conference in Bermuda to address the problem of individual research groups clinging to their parts of the genome and the threat of a proprietary sequence release. The resulting discussion produced the Bermuda Principles; a set of rules for the international genome community which stated that the big sequencing centres would automate the release of sequence data and make it available online for anyone to access and make use of.

In January 2003, The Wellcome Trust sponsored a second international meeting to again discuss pre-publication data release, this time at Fort Lauderdale in Florida. Participants made recommendation for high-throughput, large-genome projects, which the Wellcome Trust considers to be "community resource projects." These will typically be multi-centred, multi-funded and international projects. The Fort Lauderdale meeting reaffirmed the Bermuda Principles and extended them. The latter had affected only sequence assemblies of 2,000 bases or more, the 2003 agreement applied to all sequence data.

There are now three main genome browsers which contain all the HGP data: the National Centre for Biotechnology Information (NCBI) GenBank database; the University of California at Santa Cruz Genome Browser and the European Bioinformatics Institute Ensemble database. The International Sequencing Consortium (ISC) also has a website to provide a single, central site for the scientific community and the public to have access to up-to-date information about animal and other genomic sequencing projects. The ISC was established to provide a forum for genomic sequencing groups and their funding agencies to share information, coordinate research efforts and address common issues raised by genomic sequencing, such as data release and data quality. Other HGP-derived projects publish data on their own websites.

The second lesson learnt from the HGP according to Collins concerned the importance of international participation. Collaboration between research centres world-wide gives the practical advantage of a wide pool of resources and funding. It also has an ideological aspect: the human genome is our 'shared inheritance' – it is something which every person has in common and as such, it was felt that the project should involve researchers from across the globe. Collaboration also helps science in developing countries to progress, thus avoiding the polarisation of scientific knowledge between those in the 'first' world and scientists in less wealthy nations.

Other commentators have argued that the HGP also indicated that 'big science' in the future might be successfully funded by bringing industry in as a partner. Henry Lambright argues that it was possible for industry to work with the HGP because it played by 'the rules', meaning that companies involved had to agree to certain open data release policies. He believes that the pattern in science, reflected in the HGP, is going to be towards increasingly, large-scale efforts that cross agency lines, involve public-private ventures and stretch beyond the nation-state.

The final two lessons to be drawn from the HGP, Collins argues, are about the importance of setting high standards for data quality which are rigorously tested with external assessment. And lastly, an understanding that the raw data alone will not provide the scientific community with maximum benefit: carrying out in-dept data analyses as part of a production project is both an opportunity and a responsibility.

The HGP drew together an international community focused on one common goal and in doing so it changed the nature of biological research. Traditionally, this was a field very focused on individualistic enterprise with researchers pursuing their own projects more or less independently. Necessity drove the HGP to try a different approach: technological change and the massive amount of financial investment required pushed them to assemble interdisciplinary teams, encompassing engineering and informatics as well as biology; automate procedures wherever possible; and concentrate research in major centres to maximize economies of scale. Post-HGP projects have followed a similar course. As the NHGRI says, "the era of team-oriented research in biology is here."

The 2003 vision for the future of genomics

Recognising that the goals of the HGP would be achieved earlier than expected (the project had originally been due for completion in 2005) the NHGRI convened a series of meetings from 2001 to 2003 to generate a vision for the future of genome research. They outlined the conclusions of this consultation process in April 2003 in 'A vision for the future of genomics research'. The paper sets out a series of 'grand challenges' for the future of genomics, all resting on the foundation of the HGP and crossing three thematic areas: genomics in biological research to further our scientific understanding of the genome; genomics for use in medical research for the improvement of human health and finally, the importance of understanding the relationship between genomics and society, including policy options and the social and ethical consequences of the knowledge created by increased genetic understanding.

The NHGRI also outlines six cross-cutting elements which were to work concurrently with these aims. Firstly, any post-HGP work should continue to produce large, publicly available data sets such as genomic maps and sequences. Secondly, the HGP aided various technological developments which were scaled up and made efficient for research purposes and it was felt that this kind of development should continue. During the HGP, computational methods became intrinsic to modern biological research and, according to the NHGRI, this must also be developed further as large-scale methods for data generation improve and the complexity of the data increases. Scientists must also be trained to work in this developing environment. The NHGRI felt that since our understanding of genetics will increasingly have a social impact beyond the scientific community – for example, how should we utilise our abilities to select and manipulate genetic material in human reproduction? - the risks and opportunities which this presents must be understood. The Ethical, Legal and Social Implications (ELSI) Research Program - established in 1990 as an integral part of the HGP - still continues its work in this field today. Lastly, the NHGRI highlights the importance of educating both the general public and health professionals about developments within the

genomics field.

This vision has informed the post-HGP initiatives which utilise and build on the genomic data. Some examples are detailed below.

Comparative genomics

Comparing the human genome data to the genomic structure of other living things is an important step in understanding the functionality of our genes. The advanced draft or finished genome sequences have now been published for five mammals: human, mouse, rat, chimpanzee and dog. This provides an important comparative background to the HGP and is a continuing area of research interest.

Sequencing techniques will continue to be important. Various projects are working on increasing the efficiency for constructing genomes for individual patients. The Large-Scale Genome Sequencing Program is responsible for the administration and support for this research. Technical advances already mean that the cost of DNA sequencing has declined dramatically: from \$10 in 1990 to less than \$0.09 per base pair in 2002. The NHGRI and others are pursuing the development of new technologies to sequence any individual's genome for \$1,000 or less.

The five largest NHGRI-supported genome-sequencing centres can generate 150 billion base pairs of sequence each year between them. Additional capacity is provided by the Joint Genome Institute of the Department of Energy in the US and in other countries, especially the UK, Japan, France, Germany and China.

Other projects are harnessing the power of large-scale sequencing programmes to reach the long-term objective of making human DNA sequencing a tool for both research and medical practice. These include the Medical Sequencing Program and the Cancer Sequencing Project. Another project, the Cancer Genome Atlas, for example, aims to identify all the abnormalities in 10,000 or more tumour specimens derived from 50 different cancer types. The Knockout Mouse Project has similar uses. It aims to generate a comprehensive and public resource comprised of mouse embryonic stem cells containing a 'knockout' in every gene in the mouse genome. This means an existing mouse gene has been inactivated or 'knocked out' in order to observe how the absence of this genetic material affects the animal. While individual laboratories have been generating knockouts for years, many of these are still not publicly accessible and numerous genes remain to be subjected to this strategy. As humans share many genes with mice, the results of these knockout tests can help researchers study how similar genes may cause or contribute to diseases in humans such as cancer, obesity, heart disease and arthritis.

These projects are both driving improvements in technology and producing new findings. Their large-scale production methods echo that of the HGP and the ethic of early public disclosure is at their heart.

The International HapMap Project

The HapMap project was aimed at developing a haplotype map of the human genome based on the original HGP. The haplotype map, or 'HapMap', allows researchers to find genes and genetic variations that affect health and disease.

Although the DNA sequence of any two people is 99.9% identical, the variations crucially affect an individual's disease risk. The points where the sequence differs at a single DNA base are called single nucleotide polymorphisms (SNPs). Sets of SNPs on the same chromosome are inherited in blocks called haplotypes. The HapMap project is an attempt to identify these blocks and study the common patterns in genetic variation. The purpose is to enable the study of genetic associations with disease and the genetic factors contributing to different people's responses to environmental factors, susceptibility to infection and the effectiveness and adverse responses to drugs and vaccines.

The project was launched in 2002 with \$100 million worth of public and private funding and involved nine research groups and more than 200 researchers in six countries; Canada, China, Japan, Nigeria, the UK and the US. It analysed blood samples from people in Nigeria, Japan, and China and from those with northern and western European ancestry living in the US. Different parts of the genome were assigned to various investigators from across the participating countries. They mapped the entire genome of 269 people to identify tiny differences in key areas of DNA.

The HapMap makes possible searching for common variants expected to play a role in risk for common diseases. The Wellcome Trust has initiated a control consortium which aims to do this for eight common diseases. In the US, a public-private partnership, the Genetic Association Information Network (GAIN) aims at a similar analysis of seven common diseases. The US Genes, Environment and Health Initiative was awarded \$40 million in 2007 by the National Institute of Health for work which will include both genome-wide association analysis of common diseases and the development of better tools for the assessment of environment exposure, dietary intake and physical activity. The HapMap is publicly accessible.

ENCODE

The ENCyclopaedia of DNA Elements (ENCODE) project, established in September 2003, brings together several dozen laboratories that aim to identify comprehensively the functional elements in the genome. It is intensively exploring a carefully chosen 1% of the genome. All the data is placed on a public browser as soon as it is verified.

ENCODE is a NHGRI-led project, intended to help scientists mine and fully utilize the human sequence, gain a deeper understanding of human biology, predict potential disease risk and develop new strategies for prevention and treatment of disease.

The project is organised as an open consortium. ENCODE project participants come from across the world and study a range of functional elements based on the genome, utilising a number of different technologies.

Part of the project is to develop technology to produce new high throughput methods to identify functional elements. By initially concentrating on a limited portion of the human genome, the NHGRI hopes that all of those who have experience and insight into the problem will be willing to participate, whether or not their approaches are proprietary or have already generated proprietary data. The ENCODE Consortium is open to all academic, government and private sector scientists interested in participating in an open process to facilitate the comprehensive interpretation of the human genome sequence and who agree to the criteria for participation for the project.

The Structural Genomics Consortium

Structural genomics is the generation of the three-dimensional structure of proteins. The goal for studying the structural genomics of any organism is the complete structural description of all proteins encoded by the genome of that organism. This is crucial for drug design, diagnosis and treatment of disease and advancing our understanding of basic biology.

The Wellcome Trust has committed £18 million to the Structural Genomics Consortium (SGC), an international collaboration aiming to unravel the structures of proteins of medical relevance and place them in the public domain without restriction. The end result will be structural information to stimulate the development of new and improved drugs and other healthcare projects. The Wellcome Trust established the SGC in 2003, in partnership with GlaxoSmithKline and four of Canada's leading research funding agencies. In 2005, a consortium of Swedish sponsors provided additional funding.

The goal of this undertaking was to develop the infrastructure and technologies necessary for rapid data production, with the aim of having the capacity to determine 200 protein structures per year. Over the first four years, the SGC has been targeting 375 proteins that have relevance to human health and disease, such as proteins associated with diabetes, cancer and infectious diseases such as malaria. Targets are also chosen based on interest from the academic and pharmaceutical communities, expertise within the Consortium and scientific impact. The SGC deposited its 400th structure into the Protein Data Bank in March 2007 and is currently operating at a pace of 200 structures a year at a cost of \$125,000 per structure.

Systems biology

One of the greatest impacts of having whole-genome sequences and powerful new genomic technologies may be an entirely new approach to conducting biological research. In the past, researchers studied one or a few genes or proteins at a time. Yet biological processes do not operate in isolation and so this can produce incomplete or misleading results. Researchers now can approach questions systematically, on a much grander scale and are able to look at biological functions in their systemic environments.

President of the Institute for Systems Biology (ISB) and the biologist who created a DNA sequencer to map the human genome, Leroy Hood argues that the HGP has catalyzed the emergence of this new approach to biology. 'Systems biology' analyzes all the interrelationships of the elements in a biological system, rather than studying them in isolation. Interactions

between different components in a system are crucial for an organism's form and function. For example, the human immune system is not the result of a single gene or mechanism, but an interaction between many different genetic and external factors.

Systems biology has been made possible by both the information and the technologies developed through the HGP. The ISB gives four contributing factors to the growth of this area of science: firstly the data from the HGP and the resulting increasingly understanding of how genes function; secondly the communication of massive amounts of data between scientists, made possible by the internet; thirdly, the development of powerful new research technologies, and lastly the contributions of scientists from many different disciplines, including computing, engineering and mathematics.

It is hoped that by gaining greater knowledge of the effects of interaction between different aspects of biological systems, we will be better placed to understand and develop treatments for a wide range of diseases. It has been argued that the application of systems approaches to medicine will lead to the rise of predictive, preventive, and personalized health care over the next 15-20 years and will totally transform how medicine is practiced. Founded in 2000, the ISB is one example of a research institute based on this new understanding of where biological science disciplines could be headed.