Genomics and the Human Genome Project

Learning goals

By the end of this material you should have learnt about:

- The genome of an organism i.e. the structure and function of an organism's genetic material
- The purpose and function of the human genome project

Graphical representations of the chromosomes that contain the human genome

Genomics

Genomics is the study of the **genome** of an organism - its entire genetic material in the form of RNA, DNA, genes and chromosomes. It concentrates on understanding the structure and function of an organism's genetic material from the molecular level upwards, including interactions between genes, interactions between genes and the proteins they produce, and interactions between genes and environmental factors. It has obvious links to the related science of **proteomics**, which focuses on understanding the structure and function of the proteins produced by the genome (the **proteome**).

The starting point of genomics is to **sequence** the genome of an organism: that is, to analyse the order of the four nucleotides - adenine, thymine, cytosine and guanine - in the DNA (or RNA) molecules that represent the organism's heritable genetic information. This forms the basis of an attempt to identify the structural genes that code for proteins and RNA, as well as the regulatory sequences that control the expression of genes and the non-coding sequences. The efficiency and accuracy of modern automated sequencing technologies, in conjunction with sophisticated data analysis tools, have enabled the sequences of the complete genomes of a number of species to be identified and recorded.

**Functional genomics** looks at the way in which the molecular information contained within the nucleotide sequences of the genome is expressed in the production of specific proteins and RNA products. It moves beyond the sequence of nucleotide bases in the genome to examine the biological role of specific genes, and the ways in which sets of genes - and their protein products - interact during development, health and disease. It investigates the factors...
influencing patterns of gene expression in specific species under specific conditions. The tools involved include microarrays, to investigate levels of gene expression and the interactions between genes and their products, and bioinformatics, to manage and store the large amount of complex data involved.

**Comparative genomics** is a new field in which the genome sequences of different species are compared. The similarities and differences between the function of parts of the genome in different species provide information about gene expression, gene regulation, the processes of evolutionary change and the evolutionary links between species. The organisms involved, apart from human beings, tend to be those of commercial or research interest - and the genomes of some 200 organisms have been sequenced so far. The amount of information involved means that much of the work involves innovative data analysis techniques and extensive electronic databases.

The sequencing of the genome sits alongside work to sequence the protein products produced by the genome. The field of proteomics involves determining the amino acid sequences of all the proteins produced by a given genome (known collectively as the proteome), and the project to sequence the human proteome is coordinated by the [Human Proteome Organisation (HUPO)](https://www.hupo.org/).

**The human genome**

There have been two main projects to sequence the human genome.

The [Human Genome Project (HGP)](https://www.genome.gov/) is the name for an international consortium of publicly funded projects to sequence the human genome and map every gene on every chromosome. The consortium includes the US Department of Energy, the US National Institutes of Health and the UK Wellcome Trust, alongside groups in Japan, France, Germany, China, and other countries. The main aims of the Human Genome Project are to:

- determine the sequence of the three billion or so nucleotides that constitute the human genome
- identify the 20,000 to 25,000 genes in the human genome
- develop tools for storing and analysing this information
- transfer some of the technologies involved to the private sector, to produce a biotechnology industry that can develop new medical applications
- examine the ethical, social and legal implications of the information obtained

The HGP uses the so-called **hierarchical shotgun sequencing** technique, in which the genome is divided into relatively large sections that are mapped onto the appropriate chromosomes before being sequenced.
The logo of the Human Genome Project

The other project was started later, in 1998, by the private industrial group Celera Genomics, led by Dr Craig Venter. Celera Genomics used the technique of **whole genome shotgun sequencing**, instead of dividing the genome into sections like the HGP.

In March 2000, the then US President Bill Clinton announced that human genes could not be patented. In June of the same year a draft sequence of the human genome was announced by Clinton and the then UK Prime Minister Tony Blair. The sequencing was essentially complete by May 2003, with a 'Gold Standard' version released in October 2004, although the full sequence of the last chromosome was published in the magazine Nature only in May 2006. The process is still not quite finished: for example, there are highly repetitive areas of DNA, especially around the centromeres and telomeres of the chromosomes, that are proving difficult to sequence.

Sequencing the human genome is just the first step in the process. The next stage is to identify the genes, and the proteins (or RNA molecules) for which they code, as well as elements of the genome that regulate the expression of genes, play a role in the replication of DNA and maintain the structure of the chromosomes. So far some 22,000 'gene loci' have been identified, including approximately 20,000 genes that code for proteins. Finding all the genes will not be easy, however. Relatively small genes are difficult to detect, some genes may overlap and some genes may code for a number of different products. The genome also contains 'pseudogenes', which are faulty copies of genes found elsewhere in the genome - stable, inherited, but not expressed in terms of protein formation.

The work of the **Encode Consortium** (Encyclopedia of DNA Elements) has indicated that approximately 3 per cent of the three billion base pairs in the human genome are associated with the 22,000 or so genes identified so far. The remaining 97 per cent, previously considered to be 'junk DNA' with little or no purpose, is now thought to play an important role in gene regulation. Some of this DNA is chemically active - producing RNA molecules that do not play a role in protein production but do appear to play a role in switching other genes on and off.

It is hoped that developments in our understanding of the human genome will improve our ability to diagnose, treat and prevent disease by improving our understanding of the molecular mechanisms involved, including the complex interactions between genetic factors and environmental factors. An insight into these molecular mechanisms will reveal new molecular targets - at the level of DNA and protein - for drugs, vaccines and genetic tests. Individual drug regimes can be designed in the light of an individual's genetic make-up and,
in time, **gene therapy** may lead to the possibility of the replacement or repair of defective genes.

This work is licensed under a [Creative Commons Licence](https://creativecommons.org/licenses/by-nc-sa/4.0/).