David Martindill

Anything but junk

Delving deep into the dark matter of the human genome

Key words
gene
genome
DNA
medicine

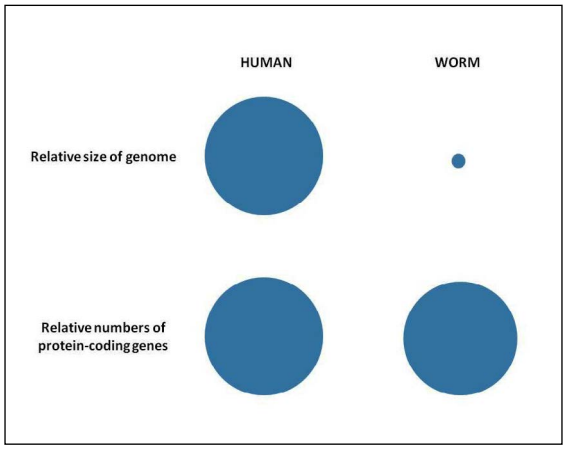
Most biology students know who discovered the structure of DNA. But few can recall Fred Sanger's contributions to genetics – and what about those of Francis Collins and John Sulston?

It is now more than a decade since the Human Genome Project concluded. Employing the technique of DNA sequencing developed by Sanger, an international consortium of scientists, including Collins and Sulston, determined the entire sequence of our genome, a code of 3.3 billion letters.

DNA sequencing machines at the Wellcome Trust Sanger Institute

This is much larger than that of many other species, for example the nematode worm, *C. elegans*, which has a genome around 3% the size of ours.

A tiny section of the human genome sequence. A staggering 3.3 billion letters long, it would take over 90 years, day and night, to read the complete sequence at a rate of one letter per second!



A comparison of genome size and gene number in humans and the nematode worm, C. elegans

The alphabet of molecular biology

DNA has a simplified alphabet of four letters, A, T, C and G. These are known as 'bases' and are read by the cell in groups of three. Each triplet encodes a specific amino acid, of which there are twenty different types, which the cell then joins together to form a protein. The function of a protein, encoded by a specific stretch of DNA called a gene, is entirely dependent on the sequence of its amino acids. Proteins control events during foetal development, and processes from respiration to reproduction in the adult.

Francis Crick claimed that, for cells to make proteins, the DNA code is first copied into a related molecule called RNA. This is like a 'photocopy' of a page (the gene) from an encyclopaedia (the genome), in a library (the nucleus). Structures in the cytoplasm called ribosomes convert this RNA copy into protein. RNA, claimed Crick, is a 'middle-man', a short-lived intermediate, between DNA and protein.

Given our complexity, scientists expected the human genome to contain a huge number of genes. It was therefore a shock when it was discovered that humans have just 20 000, roughly the same number as *C. elegans*. Until recently, there was no explanation for why so much of our genome, more than 98%, is non-coding. Scientists dismissed this as 'junk DNA'. This is mainly found in the dark regions of chromosomes, called heterochromatin, once they have been stained with dyes.

A human chromosome stained with a chemical called Giemsa (left) and a diagrammatic representation of the same chromosome (right). Junk DNA is predominantly found in the darker regions, called heterochromatin.

What we knew

Perhaps these figures should not have been as surprising. Even before our entire DNA sequence was revealed, sections of the genome were known not to encode proteins.

Lengths of junk DNA at the end of chromosomes (called telomeres) shorten as a cell divides, and this has been linked with the lifespan of a cell. Conditions characteristic of old age, including some lung diseases, are more common in people with undersized telomeres, while smoking, obesity and stress have all been found to accelerate their shortening. Beyond the telomeres, there appear to be regions of DNA within genes that do not encode amino acids, called introns, and sections of non-coding DNA surrounding genes that are copied by the cell into RNA.

For decades it has been known that mutations in these regions are responsible for a range of genetic disorders. These include Fragile X syndrome, myotonic dystrophy and some cases of the condition made prominent by the 'ice bucket challenge,' amyotrophic lateral sclerosis (ALS). On a more sinister note, remnants of viruses have been detected in the human genome, which slotted into our DNA during our evolutionary past. These 'genomic fossils' present potential dangers, not least an increased risk of cancer, if they decide to 'copy and paste' themselves from one place to another.

Fine tuning

Switching genes 'on' and 'off' should not be considered an issue of black versus white. The regulation of gene expression can occur in a whole host of shades of grey, and junk DNA is vital in this process.

Junk DNA contains all sorts of 'switches' to tell the cell where, when and for how long genes should be turned on to produce proteins. Some, called imprinting control elements, dictate which gene copy is turned on – the one from your mother or the one from your father. Others, called promoters, enable cells to regulate the extent of gene expression. They do this by acting as 'docking points' for proteins

called transcription factors. These in turn bind the cellular machinery that creates the RNA copy of the gene for protein synthesis.

However, due to mistakes in cell division, chromosomes can become cut and pasted with each other, sometimes leading to a promoter being placed near a gene it shouldn't be, with disastrous results. The best known example of this occurs in patients with a cancer called Burkitt's lymphoma, in which a very active promoter is placed next to a gene that controls cell division. Other junk DNA regions called enhancers also adjust gene expression and are very important in the control of correct development of the foetus. This is because gradients of chemicals called morphogens need to be established to ensure that the right structures develop in the right places. If enhancers are mutated, deformities can result.

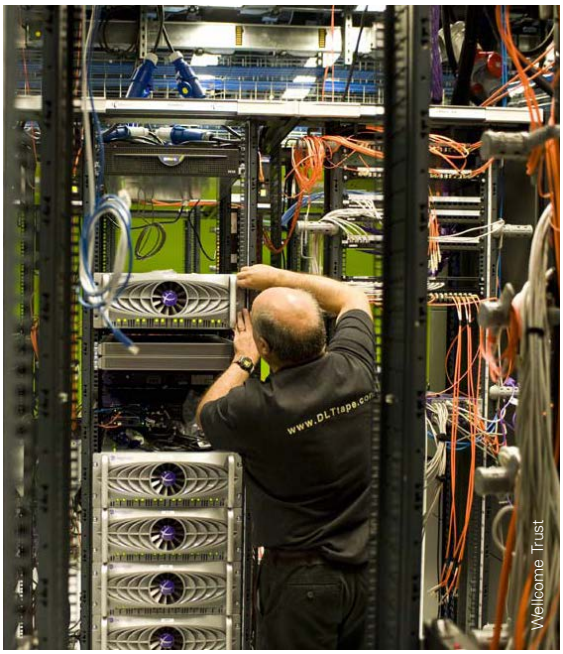
Mutation of one of the enhancers of the Sonic Hedgehog gene interferes with the gradient of this morphogen, resulting in deformities of the limbs. Look closely!

In DNA's shadow

The human genome does not just encode proteins as Crick thought. RNA molecules can have a role in their own right.

Some functional RNA molecules have been known for a long time. For example, ribosomes are made predominantly of RNA. However, the roles of others are only just emerging. Some are thought to be involved in maintaining populations of stem cells in the embryo, possibly by promoting the expression of some genes over others. Deciphering how RNA is involved in these pathways may enable us to increase the potency of cells, namely their ability to specialise into different types.

This has huge potential in the field of medicine which treats degenerative disorders such as diabetes and Parkinson's disease. Interestingly, many of these RNA molecules are often found exclusively in humans, and a high proportion of these are found in the brain. In experiments using rats, problems with production of these molecules during development have been shown to influence intellectual ability, impair muscle coordination and induce alcohol and drug dependency. Some pharmaceutical companies are already developing drugs that work by mimicking their function, with Alzheimer's disease being one such target.



The Wellcome Sanger Centre has 4 petabytes of computer storage to manage the vast amounts of data generated by genome sequencing.

A fruitful future

There are still many scientists who claim that we are paying too much attention to junk DNA. They say it might be like our appendix: important in our ancestors, but without much function in us. Indeed, there can be few explanations for a type of junk DNA called short tandem repeats (STRs), which consist of two- or three-base sequences (e.g. CT or CTG) repeated over and over again. However, our understanding of such sequences can still bring benefits. DNA fingerprinting relies on STRs, as the numbers of repeats differs between people. The coming decades promise opportunities for researchers to unlock further secrets in this sequence – we just have to look closely and use our imagination.

David Martindill teaches biology and has a research background in genetics.

Look here!

You can read more about this fascinating subject in Nessa Carey's new book, *Junk DNA: A Journey Through the Dark Matter of the Genome.*