Exploring genomes

All in sequence

Why is determining an organism's genome sequence important?

Most types of cell in an organism contain a complete copy of its genome. The organisation is quite complicated, but the simplest fact about any genome is that it is a collection of DNA sequences – long strings of the chemical 'letters' A, T, G and C (adenine, thymine, guanine and cytosine) in a particular order.

Learn to read an organism's genome sequence, and compare it with that of other organisms, and it can tell you lots of different things.

The human genome sequence contains a wealth of information about human biology, in both health and disease. Our DNA is a window on evolution and recent human history – including the migration of people around the world.

The genome sequences of other species have many other uses. The genomes of organisms used in farming, from rice and wheat to pigs and cattle, are being sequenced to help to breed improved strains. But the vast majority of the many thousands of genomes already completed are from bacteria. Some are species that cause diseases in people, as well as in agriculturally important animals or plants.

Others are important for maintaining health or have potential uses in the industrial production of biologically active chemicals and enzymes.

Genomic information is used to track harmful microbes such as those that cause infection in hospitals, as well as to aid the development of new drugs. New influenza strains have their genomes read quickly to understand how the virus spreads and to speed up vaccine production.

Knowledge of genome sequences also speeds up developments in biotechnology and is finding uses in tracking biodiversity and policing trade in protected species.

Genes, "junk" and non-coding DNA

The genome is far from just a collection of genes

Genes take up only a small part of the genome: protein-coding sequences account for only around 1-3 per cent of human DNA.

Biologists used to regard the other regions, which don't code for proteins, as 'junk'. Now, though, these regions are slowly yielding their secrets. Some have regulatory roles and determine the activities of the protein-coding sections. Others are probably parasitic bits and pieces that are along for the evolutionary ride.

It is not clear how much of the genome is transcribed into RNA. If, as some scientists suggest, it is as much as 90%, we must ask why. The non-coding regions contain thousands of long non-coding RNAs, the function of which remains largely unknown. However, a 2017 study by US researchers suggests some of these RNAs may play a role in cell growth and that many have functions that are specific to a single type of cell.

Scientists are still working to understand how much of the genome is functional – but there are different ways of defining 'functional'. For example, back in 2012, scientists working on the US-led Encyclopaedia of DNA Element (ENCODE) project published papers claiming that more than 80 per cent of the non-coding regions had a role in regulating the activity of genes. However, there is debate among scientists about how this project differentiated between activities that are biologically important and useful for humans and those that are not.

By contrast, a study published in 2014 found that just 8 per cent of the DNA in humans is functional. In this study, functional DNA was defined as DNA that has evolved unexpectedly slowly and is therefore thought to have an important role in the body. However, other scientists say that, for DNA, being functional is more than just being conserved evolutionarily, arguing that some non-conserved parts of the genome have a role in certain diseases.

Today, some scientists still argue that at least three quarters of the genome is non-functioning. A 2017 study came to this conclusion based on the assumption that if any more of it was functioning, then mutations in DNA that is important for our survival would build up to such high levels that women would need to give birth to multiple children – many of whom would die – in order for the human race to continue.

Understanding RNAs

Exploring the many roles of RNA

Ribonucleic acid (RNA) is similar to deoxyribonucleic acid (DNA). We usually think of RNAs as being single-stranded, although they may start out as double-stranded molecules like DNA before being split up by enzymes. RNA sequences are complementary to – rather than identical to – sequences in the DNA. The bases in an RNA sequence can bind to complementary RNA or DNA sequences. This seems to make RNA ideal for lots of the detailed regulation of gene activity.

As scientists have come to understand the many roles of RNA, they have started to give more specific names to the different types. There are miRNAs, siRNAs and piRNAs. There are A-RNAs and Y-RNAs. And there are some types of RNA that have only been found in non-human species so far, such as trans-acting siRNAs (ta-siRNAs), a subdivision of siRNAs that was discovered in the plant *Arabidopsis* and has since been recognised in rice and corn.

miRNAs (microRNAs) and siRNAs (short interfering RNAs) are probably the best understood, although there remains a lot to learn. They are both formed from short pieces of RNA, around 20 bases in length, and have regulatory roles. Both are involved in silencing processes that inhibit the expression of certain genes, but whereas the same miRNAs can regulate many different genes, siRNAs target specific gene locations.

miRNAs are made by genes that specifically encode miRNAs. They regulate other genes by inhibiting the translation of messenger RNAs that are destined to make proteins. They were discovered in 1993, in worms, where they regulate genes involved in the timing of developmental processes. We know that miRNAs can have important and widespread functions, because interfering with their processing in mice can cause abnormalities in the heart, kidneys and liver, and increase susceptibility to cancer and other diseases.

siRNAs are often regarded as having a protective role, which includes defending the genome against invading genetic materials from viruses. They are produced as strands that become incorporated into a “silencing complex” of proteins – it is this complex that binds to the mRNA. siRNA-based therapies are a rapidly expanding area of medicine: the molecules can be used to switch off genes involved in disease. In 2018, an siRNA-based drug was approved for use in clinics for the first time. Called Onpattro, it was approved for nerve damage in patients with a rare disease affecting the nervous system. However, there remain challenges in getting siRNAs into cells. Scientists are working on novel delivery approaches, such as biodegradable nanoparticle carriers.

REFERENCES

Origins and mechanisms of miRNAs and siRNAs

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2675692/

The role of noncoding RNAs in gene regulation

https://www.sciencedirect.com/science/article/pii/B9780128124338000095

Molecular mechanisms and biological functions of siRNA

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5542916/

Comparing genomes

How does the human genome compare with the genomes of other organisms?

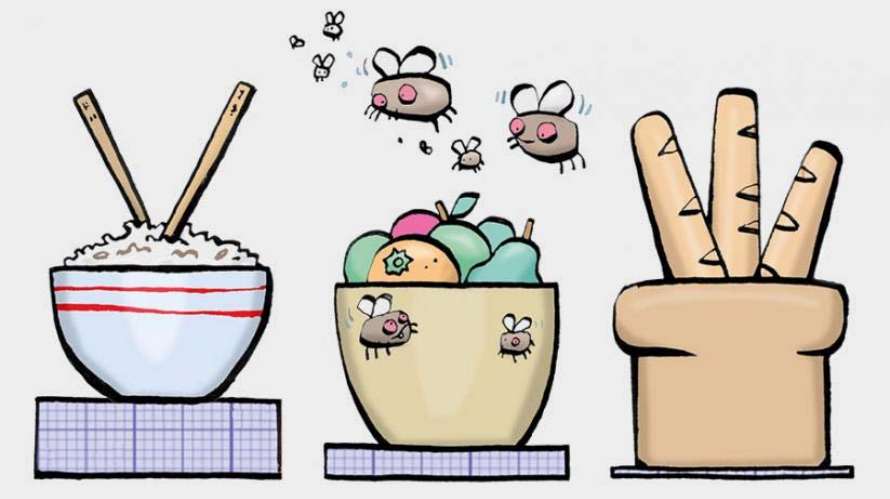


Illustration © Glen McBeth

Gene number and genome size by organism

Organism	Gene number (protein-coding)	Genome size (base pairs)
Loblolly pine	50,172	22.0bn
Human	20,454	3.6bn
Chimpanzee	23,534	3.4bn
Mouse	22,480	3.5bn
Rice (Japonica subspecies)	37,849	375m
Fruit fly	13,931	143m
Yeast (*Saccharomyces cerevisiae*)	6,600	12m
E. coli	5,494	5.4m
Influenza virus (A, Hong Kong)	14	13,500 (rounded)
Ebola (Zaire)	7	19,000 (rounded)

Now we have whole genomes from a range of organisms, comparing them is one way to investigate what makes each creature distinctive. In the simplest cases, this means counting genes and checking which ones are there. For example, the first non-human vertebrate animal to have its genome sequenced was the tiger pufferfish (*Fugu rubripes*) in 2001. Counting and checking shows that it has many genes in common with us. Animals without backbones, such as fruit flies and nematode worms, also share very many genes with humans, as do yeast – a 2015 study found that hundreds of yeast genes could be swapped for their human equivalents without harming the yeast. However, around a quarter of human genes have no equivalents in the fish, and we have almost ten times as much DNA as they do.

The chimpanzee genome was first published in draft in 2005. Initial comparisons with the human genome – made by simply lining the two genome sequences up together – suggested the chimp genome was 98.8 per cent identical to our own. This seemed to confirm our long-held belief that the chimp is our closest living

relative. However, in 2012 the publication of the genome sequence of the bonobo – which, like the chimpanzee, is an African ape – showed that bonobos are just as close a relative to us as chimpanzees are.

For very closely related organisms, like chimps and humans, we can align most of one organism's genome with that of the other. This reveals obvious differences and allows us to calculate a simple percentage similarity. Such a comparison is less meaningful when it involves two organisms that are less closely related.

But even for our closest relatives, the measure of 98.8 per cent similarity conceals important differences. In fact, despite the massive overlaps, there are still 17 million point changes in our genome compared to that of chimps and, in addition to differences in the precise sequence of genes, there are some differences in the overall set of genes. When it comes to our brains, there are differences in tens of genes. Comparing genome sequences also doesn't give us information on when particular genes are switched on and off during development – especially ones that affect the brain. These subtle differences in gene regulation and gene expression are likely to underlie the most important differences between closely related species.

REFERENCES

The bonobo genome compared with the chimpanzee and human genomes

https://www.nature.com/articles/nature11128

You And Yeast Have More In Common Than You Might Think

https://www.npr.org/sections/health-shots/2015/05/21/408322187/you-and-yeast-have-more-in-common-than-you-might-think?t=1564131814681&t=1569416543453

Chimps Can't Tell Us Much About Being Human

http://blogs.discovermagazine.com/crux/2018/07/09/chimps-cant-tell-us-much-about-being-human/#.XYtljrqcGFD

The human reference genome

Whose DNA made up the original human genome sequence?

The original of 'the' human genome was a reference sequence, compiled by analysing DNA from blood samples donated by a handful of anonymous volunteers. It stands as a comparison for genomes analysed later.

That sequence covers the nuclear DNA – the DNA contained in the cell's nucleus – which makes up over 99 per cent of the cell's total DNA. The mitochondrion, the tiny organelle that acts as the cell's energy generator, has its own separate genome, with 16,500 base pairs and just 37 genes, which was originally sequenced in 1981. Every cell has hundreds of copies of this DNA, and many different samples have been sequenced.

Since the Human Genome Project was completed, it's thought that over 200,00 individuals have had their whole genomes sequenced. There are also ongoing projects to sequence and share data from individual genomes, along with information about medical history and physical characteristics. In the USA, Professor George Church at Harvard Medical School started the Personal Genome Project in 2005 and was one of the first ten people on the project to make their genome sequences public (so-called 'open-source' publication). Participants must sign a consent form and fill in a survey when they donate their blood or saliva for sequencing. Today, hundreds of personal genomes have been published through the initiative, with national branches of the Personal Genome Project popping up in Canada, China, the UK and Austria.

REFERENCES

The Human Genome Project

https://www.genome.gov/human-genome-project

Understanding genomics

https://www.genomicsengland.co.uk/understanding-genomics/

Epigenetics

Another layer of genetic complexity

Understanding what the base code of DNA means is complex enough, but without an understanding of epigenetics it is meaningless. Epigenetics governs how the genome is played out in reality – it concerns chemical and physical modifications that change the activity of genes, without change the base code itself. It can even cause semi-permanent changes to our DNA that can be passed from one generation to the next.

Epigenetics modifications are indirect. Whilst the DNA sequences stay the same, there are changes to the set of chemical tags that are carried by the genes or the proteins bound to them. These tags, most often simple methyl (CH_3) groups, affect gene expression – such as by preventing genes being transcribed as proteins. Often, the gene affected is itself one that controls another genetic switch or switches. This means that there may be a cascade effect, where one small change leads to larger adjustments in gene expression.

The whole set of epigenetic tags is known as the epigenome. Non-coding RNAs that bind to DNA are considered to form part of the epigenome too, as are prion proteins like those that cause variant Creutzfeldt-Jakob disease or that may be involved in Alzheimer's Disease.

The epigenome is a bit like the personal settings that gradually customise the operating system on a computer. It is partly why twins grow to be different over the course of their lives. These changes are influenced by their different environments.

Most of the changes are 'reset' in sperm and egg. However, it is thought that some are passed on to the embryo, including changes associated with growth and development. Epigenetic changes passed on from the mother are more likely to be inherited than those from the father.

It is not clear how much influence our parents' experiences exert on us via the inherited parts of the epigenome. However, one environmental factor that has been better studied than others is the effect of the mother's diet. Malnutrition in pregnant women often causes growth problems in their babies, whilst obese women can give birth to babies with the early signs of diabetes or who grow to be obese as adulthoods themselves. Could the epigenome have a role to play? Studies in mice suggest the effects of a high-fat diet can be passed on through epigenetic changes to the genes for specific hormones involved in regulating appetite and sugar levels.

REFERENCES

The effects of a high-fat diet exposure *in utero* on the obesogenic and diabetogenic traits through epigenetic changes in adiponectin and lectin gene expression for multiple generations in female mice

https://www.ncbi.nlm.nih.gov/pubmed/25853666

Sequencing technologies

How have gene sequencing technologies developed, and what does this mean for researchers today?

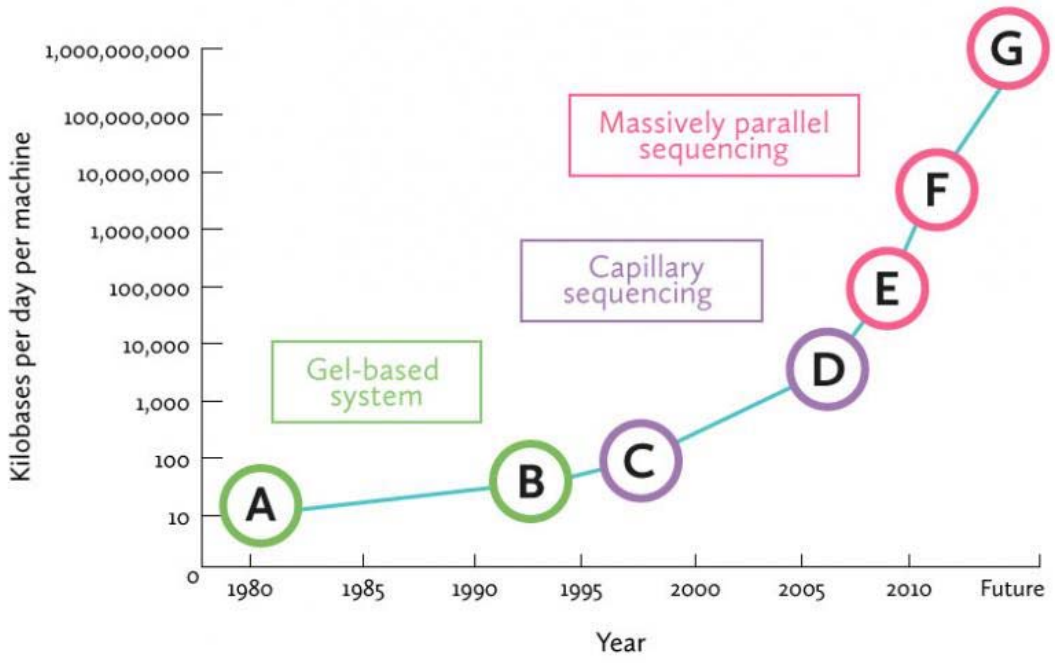
The first complete genome to be sequenced in 1977 came from a bacteriophage (phi X174) that infects *Escherichia coli*. It has just 11 genes and a little over 5,000 base pairs. Since then, biology has seen an explosion of sequence information.

The speed is a by-product of the efforts that went into the Human Genome Project. Ever-improving technologies have made DNA sequencing faster, more accurate and far cheaper. In the early 2000s, the cost of sequencing 1 megabase (1 million base pairs) of DNA was over US $5,000. By February 2019, the cost had fallen to a single cent.

After the Human Genome Project, the technology for "next generation" (massively parallel) DNA sequencing developed, meaning that hundreds of DNA sequences could be read at once in parallel reactions involving millions of fragments of DNA. The first commercial next generation sequencer was

released in 2005. Today, one person using a modern sequencing machine can generate more DNA sequence in a day than the entire human genome project generated in over a decade.

DNA sequences are stored in fast-growing computer databases. The data centre at the Wellcome Trust Sanger Institute outside Cambridge currently holds 55 petabytes (peta = 1 followed by 15 zeros) of storage, enough for about three million smart phones. Researchers need powerful software to search the DNA sequences. Bioinformatics is the art of harnessing computer power to make sense of mountains of biological data.



Improvements in the rate of sequence generation over a 30-year period.

A. Manual slab gel
B. Automated slab gel
C. First-generation capillary
D. Second-generation capillary sequencer
E. Microwell pyrosequencing
F. Short-read sequencers
G. Single molecule?

Credit: Graph reproduced with permission from Macmillan Publishers Ltd: Nature 458, 719-724 (2009).