In the fields of molecular biology and genetics, a genome is all the genetic information of an organism. It consists of nucleotide sequences of DNA (or RNA in RNA viruses). The nuclear genome includes protein-coding genes and non-coding genes, other functional regions of the genome such as regulatory sequences (see non-coding DNA), and often a substantial fraction of junk DNA with no evident function. Almost all eukaryotes have mitochondria and a small mitochondrial genome. Algae and plants also contain chloroplasts with a chloroplast genome.

The study of the genome is called genomics. The genomes of many organisms have been sequenced and various regions have been annotated. The Human Genome Project was started in October 1990, and then reported the sequence of the human genome in April 2003, although the initial "finished" sequence was missing 8% of the genome consisting mostly of repetitive sequences.

With advancements in technology that could handle sequencing of the many repetitive sequences found in human DNA that were not fully uncovered by the original Human Genome Project study, scientists reported the first end-to-end human genome sequence in March 2022.

The term genome was created in 1920 by Hans Winkler, professor of botany at the University of Hamburg, Germany. The website Oxford Dictionaries and the Online Etymology Dictionary suggest the name is a blend of the words gene and chromosome. However, see omics for a more thorough discussion. A few related -ome words already existed, such as biome and rhizome, forming a vocabulary into which genome fits systematically.

It's very difficult to come up with a precise definition of "genome." It usually refers to the DNA (or sometimes RNA) molecules that carry the genetic information in an organism but sometimes it is difficult to decide which molecules to include in the definition; for example, bacteria usually have one or two large DNA molecules (chromosomes) that contain all of the essential genetic material but they also contain smaller extrachromosomal plasmid molecules that carry important genetic information. The definition of 'genome' that's commonly used in the scientific literature is usually restricted to the large chromosomal DNA molecules in bacteria.

Nuclear genome

Eukaryotic genomes are even more difficult to define because almost all eukaryotic species contain nuclear chromosomes plus extra DNA molecules in the mitochondria. In addition, algae and plants have chloroplast DNA. Most textbooks make a distinction between the nuclear genome and the organelle (mitochondria and chloroplast) genomes so when they speak of, say, the human genome, they are only referring to the genetic material in the nucleus. This is the most common use of 'genome' in the scientific literature.

Ploidy

Most eukaryotes are diploid, meaning that there are two of each chromosome in the nucleus but the 'genome' refers to only one copy of each chromosome. Some eukaryotes have distinctive sex chromosomes, such as the X and Y chromosomes of mammals, so the technical definition of the genome must include both copies of the sex chromosomes. For example, the standard reference genome of humans consists of one copy of each of the 22 autosomes plus one X chromosome and one Y chromosome.

Sequencing and mapping

A genome sequence is the complete list of the nucleotides (A, C, G, and T for DNA genomes) that make up all the chromosomes of an individual or a species. Within a species, the vast majority of nucleotides are identical between individuals, but sequencing multiple individuals is necessary to understand the genetic diversity.

In 1976, Walter Fiers at the University of Ghent (Belgium) was the first to establish the complete nucleotide sequence of a viral RNA-genome (Bacteriophage MS2). The next year, Fred Sanger completed the first DNA-genome sequence: Phage Φ-X174, of 5386 base pairs. The first bacterial genome to be sequenced was that of Haemophilus influenzae, completed by a team at The Institute for Genomic Research in 1995. A few months later, the first eukaryotic genome was completed, with sequences of the 16 chromosomes of budding yeast Saccharomyces cerevisiae published as the result of a European-led effort begun in the mid-1980s. The first genome sequence for an archaeon, Methanococcus jannaschii, was completed in 1996, again by The Institute for Genomic Research.

The development of new technologies has made genome sequencing dramatically cheaper and easier, and the number of complete genome sequences is growing rapidly. The US National Institutes of Health maintains one of several comprehensive databases of genomic information. Among the thousands of completed genome sequencing projects include those for rice, a mouse, the plant Arabidopsis thaliana, the puffer fish, and the bacteria E. coli. In December 2013, scientists first sequenced the entire genome of a Neanderthal, an extinct species of humans. The genome was extracted from the toe bone of a 130,000-year-old Neanderthal found in a Siberian cave.

New sequencing technologies, such as massive parallel sequencing have also opened up the prospect of personal genome sequencing as a diagnostic tool, as pioneered by Manteia Predictive Medicine. A major step toward that goal was the completion in 2007 of the full genome of James D. Watson, one of the co-discoverers of the structure of DNA.

Whereas a genome sequence lists the order of every DNA base in a genome, a genome map identifies the landmarks. A genome map is less detailed than a genome sequence and aids in navigating around the genome. The Human Genome Project was organized to map and to sequence the human genome. A fundamental step in the project was the release of a detailed genomic map by Jean Weissenbach and his team at the Genoscope in Paris.

Reference genome sequences and maps continue to be updated, removing errors and clarifying regions of high allelic complexity. The decreasing cost of genomic mapping has permitted genealogical sites to offer it as a service, to the extent that one may submit one's genome to crowdsourced scientific endeavours such as DNA.LAND at the New York Genome Center, an example both of the economies of scale and of citizen science.

Viral genomes

Viral genomes can be composed of either RNA or DNA. The genomes of RNA viruses can be either single-stranded RNA or double-stranded RNA, and may contain one or more separate RNA molecules (segments: monopartit or multipartit genome). DNA viruses can have either single-stranded or double-stranded genomes. Most DNA virus genomes are composed of a single, linear molecule of DNA, but some are made up of a circular DNA molecule.

Prokaryotic genomes

Prokaryotes and eukaryotes have DNA genomes. Archaea and most bacteria have a single circular chromosome, however, some bacterial species have linear or multiple chromosomes. If the DNA is replicated faster than the bacterial cells divide, multiple copies of the chromosome can be present in a single cell, and if the cells divide faster than the DNA can be replicated, multiple replication of the chromosome is initiated before the division occurs, allowing daughter cells to inherit complete genomes and already partially replicated chromosomes. Most prokaryotes have very little repetitive DNA in their genomes. However, some symbiotic bacteria (e.g. Serratia symbiotica) have reduced genomes and a high fraction of pseudogenes: only ~40% of their DNA encodes proteins.

Some bacteria have auxiliary genetic material, also part of their genome, which is carried in plasmids. For this, the word genome should not be used as a synonym of chromosome.

Eukaryotic genomes are composed of one or more linear DNA chromosomes. The number of chromosomes varies widely from Jack jumper ants and an asexual nemotode, which each have only one pair, to a fern species that has 720 pairs. It is surprising the amount of DNA that eukaryotic genomes contain compared to other genomes. The amount is even more than what is necessary for DNA protein-coding and noncoding genes due to the fact that eukaryotic genomes show as much as 64,000-fold variation in their sizes. However, this special characteristic is caused by the presence of repetitive DNA, and transposable elements (TEs).

A typical human cell has two copies of each of 22 autosomes, one inherited from each parent, plus two sex chromosomes, making it diploid. Gametes, such as ova, sperm, spores, and pollen, are haploid, meaning they carry only one copy of each chromosome. In addition to the chromosomes in the nucleus, organelles such as the chloroplasts and mitochondria have their own DNA. Mitochondria are sometimes said to have their own genome often referred to as the "mitochondrial genome". The DNA found within the chloroplast may be referred to as the "plastome". Like the bacteria they originated from, mitochondria and chloroplasts have a circular chromosome.

Unlike prokaryotes where exon-intron organization of protein coding genes exists but is rather exceptional, eukaryotes generally have these features in their genes and their genomes contain variable amounts of repetitive DNA. In mammals and plants, the majority of the genome is composed of repetitive DNA.

DNA sequencing

High-throughput technology makes sequencing to assemble new genomes accessible to everyone. Sequence polymorphisms are typically discovered by comparing resequenced isolates to a reference, whereas analyses of coverage depth and mapping topology can provide details regarding structural variations such as chromosomal translocations and segmental duplications.

Coding sequences

DNA sequences that carry the instructions to make proteins are referred to as coding sequences. The proportion of the genome occupied by coding sequences varies widely. A larger genome does not necessarily contain more genes, and the proportion of non-repetitive DNA decreases along with increasing genome size in complex eukaryotes.

Noncoding sequences

Noncoding sequences include introns, sequences for non-coding RNAs, regulatory regions, and repetitive DNA. Noncoding sequences make up 98% of the human genome. There are two categories of repetitive DNA in the genome: tandem repeats and interspersed repeats.

Tandem repeats

Short, non-coding sequences that are repeated head-to-tail are called tandem repeats. Microsatellites consisting of 2–5 basepair repeats, while minisatellite repeats are 30–35 bp. Tandem repeats make up about 4% of the human genome and 9% of the fruit fly genome. Tandem repeats can be functional. For example, telomeres are composed of the tandem repeat TTAGGG in mammals, and they play an important role in protecting the ends of the chromosome.

In other cases, expansions in the number of tandem repeats in exons or introns can cause disease. For example, the human gene huntingtin (Htt) typically contains 6–29 tandem repeats of the nucleotides CAG (encoding a polyglutamine tract). An expansion to over 36 repeats results in Huntington's disease, a neurodegenerative disease. Twenty human disorders are known to result from similar tandem repeat expansions in various genes. The mechanism by which proteins with expanded polygulatamine tracts cause death of neurons is not fully understood. One possibility is that the proteins fail to fold properly and avoid degradation, instead accumulating in aggregates that also sequester important transcription factors, thereby altering gene expression.

Tandem repeats are usually caused by slippage during replication, unequal crossing-over and gene conversion.

Transposable elements

Transposable elements (TEs) are sequences of DNA with a defined structure that are able to change their location in the genome. TEs are categorized as either as a mechanism that replicates by copy-and-paste or as a mechanism that can be excised from the genome and inserted at a new location. In the human genome, there are three important classes of TEs that make up more than 45% of the human DNA; these classes are The long interspersed nuclear elements (LINEs), The interspersed nuclear elements (SINEs), and endogenous retroviruses. These elements have a big potential to modify the genetic control in a host organism.

The movement of TEs is a driving force of genome evolution in eukaryotes because their insertion can disrupt gene functions, homologous recombination between TEs can produce duplications, and TE can shuffle exons and regulatory sequences to new locations.

Retrotransposons

Retrotransposons are found mostly in eukaryotes but not found in prokaryotes. Retrotransposons form a large portion of the genomes of many eukaryotes. A retrotransposon is a transposable element that transposes through an RNA intermediate. Retrotransposons are composed of DNA, but are transcribed into RNA for transposition, then the RNA transcript is copied back to DNA formation with the help of a specific enzyme called reverse transcriptase. A retrotransposon that carries reverse transcriptase in its sequence can trigger its own transposition but retrotransposons that lack a reverse transcriptase must use reverse transcriptase synthesized by another retrotransposon. Retrotransposons can be transcribed into

RNA, which are then duplicated at another site into the genome. Retrotransposons can be divided into long terminal repeats (LTRs) and non-long terminal repeats (Non-LTRs).

Long terminal repeats (LTRs) are derived from ancient retroviral infections, so they encode proteins related to retroviral proteins including gag (structural proteins of the virus), pol (reverse transcriptase and integrase), pro (protease), and in some cases env (envelope) genes. These genes are flanked by long repeats at both 5' and 3' ends. It has been reported that LTRs consist of the largest fraction in most plant genome and might account for the huge variation in genome size.

Non-long terminal repeats (Non-LTRs) are classified as long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), and Penelope-like elements (PLEs). In Dictyostelium discoideum, there is another DIRS-like elements belong to Non-LTRs. Non-LTRs are widely spread in eukaryotic genomes.

Long interspersed elements (LINEs) encode genes for reverse transcriptase and endonuclease, making them autonomous transposable elements. The human genome has around 500,000 LINEs, taking around 17% of the genome.

Short interspersed elements (SINEs) are usually less than 500 base pairs and are non-autonomous, so they rely on the proteins encoded by LINEs for transposition. The Alu element is the most common SINE found in primates. It is about 350 base pairs and occupies about 11% of the human genome with around 1,500,000 copies.

DNA transposons

DNA transposons encode a transposase enzyme between inverted terminal repeats. When expressed, the transposase recognizes the terminal inverted repeats that flank the transposon and catalyzes its excision and reinsertion in a new site.[39] This cut-and-paste mechanism typically reinserts transposons near their original location (within 100 kb).[43] DNA transposons are found in bacteria and make up 3% of the human genome and 12% of the genome of the roundworm C. elegans.

Genome size

Genome size is the total number of the DNA base pairs in one copy of a haploid genome. Genome size varies widely across species. Invertebrates have small genomes, this is also correlated to a small number of transposable elements. Fish and Amphibians have intermediate-size genomes, and birds have relatively small genomes but it has been suggested that birds lost a substantial portion of their genomes during the phase of transition to flight. Before this loss, DNA methylation allows the adequate expansion of the genome.

In humans, the nuclear genome comprises approximately 3.1 billion nucleotides of DNA, divided into 24 linear molecules, the shortest 45 000 000 nucleotides in length and the longest 248 000 000 nucleotides, each contained in a different chromosome. There is no clear and consistent correlation between morphological complexity and genome size in either prokaryotes or lower eukaryotes. Genome size is largely a function of the expansion and contraction of repetitive DNA elements.

Since genomes are very complex, one research strategy is to reduce the number of genes in a genome to the bare minimum and still have the organism in question survive. There is experimental work being done on minimal genomes for single cell organisms as well as minimal genomes for multi-cellular organisms (see developmental biology). The work is both in vivo and in silico.

Comparison among genome sizes

There are many enormous differences in size in genomes, specially mentioned before in the multicellular eukaryotic genomes. Much of this is due to the differing abundances of transposable elements, which evolve by creating new copies of themselves in the chromosomes. Eukaryote genomes often contain many thousands of copies of these elements, most of which have acquired mutations that make them defective. Here is a table of some significant or representative genomes. See #See also for lists of sequenced genomes.

Genomic alterations

All the cells of an organism originate from a single cell, so they are expected to have identical genomes; however, in some cases, differences arise. Both the process of copying DNA during cell division and exposure to environmental mutagens can result in mutations in somatic cells. In some cases, such mutations lead to cancer because they cause cells to divide more quickly and invade surrounding tissues. In certain lymphocytes in the human immune system, V(D)J recombination generates different genomic sequences such that each cell produces a unique antibody or T cell receptors.

During meiosis, diploid cells divide twice to produce haploid germ cells. During this process, recombination results in a reshuffling of the genetic material from homologous chromosomes so each gamete has a unique genome.

Genome-wide reprogramming

Genome-wide reprogramming in mouse primordial germ cells involves epigenetic imprint erasure leading to totipotency. Reprogramming is facilitated by active DNA demethylation, a process that entails the DNA base excision repair pathway. This pathway is employed in the erasure of CpG methylation (5mC) in primordial germ cells. The erasure of 5mC occurs via its conversion to 5-hydroxymethylcytosine (5hmC) driven by high levels of the ten-eleven dioxygenase enzymes TET1 and TET2.

Genome evolution

Genomes are more than the sum of an organism's genes and have traits that may be measured and studied without reference to the details of any particular genes and their products. Researchers compare traits such as karyotype (chromosome number), genome size, gene order, codon usage bias, and GC-content to determine what mechanisms could have produced the great variety of genomes that exist today.

Duplications play a major role in shaping the genome. Duplication may range from extension of short tandem repeats, to duplication of a cluster of genes, and all the way to duplication of entire chromosomes or even entire genomes. Such duplications are probably fundamental to the creation of genetic novelty.

Horizontal gene transfer is invoked to explain how there is often an extreme similarity between small portions of the genomes of two organisms that are otherwise very distantly related. Horizontal gene transfer seems to be common among many microbes. Also, eukaryotic cells seem to have experienced a transfer of some genetic material from their chloroplast and mitochondrial genomes to their nuclear chromosomes. Recent empirical data suggest an important role of viruses and sub-viral RNA-networks to represent a main driving role to generate genetic novelty and natural genome editing.