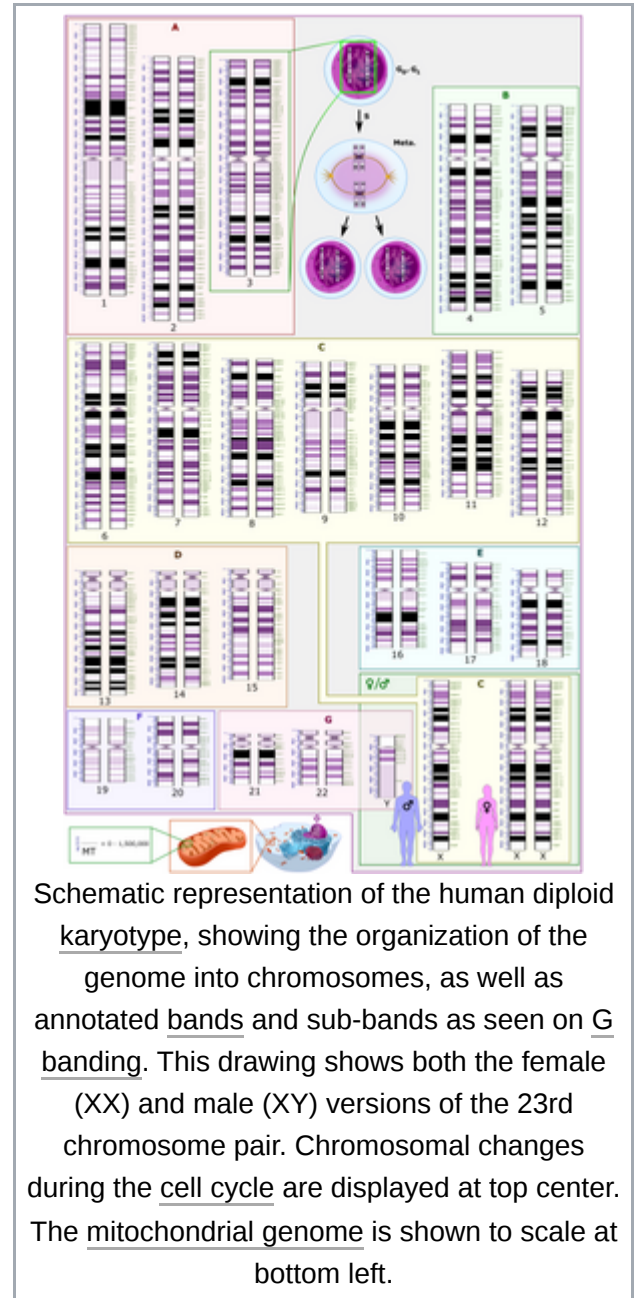# Human genome

The **human genome** is a complete set of nucleic acid sequences for humans, encoded as the DNA within each of the 24 distinct chromosomes in the cell nucleus. A small DNA molecule is found within individual mitochondria. These are usually treated separately as the nuclear genome and the mitochondrial genome.[1] Human genomes include both protein-coding DNA sequences and various types of DNA that does not encode proteins. The latter is a diverse category that includes DNA coding for non-translated RNA, such as that for ribosomal RNA, transfer RNA, ribozymes, small nuclear RNAs, and several types of regulatory RNAs. It also includes promoters and their associated gene-regulatory elements, DNA playing structural and replicatory roles, such as scaffolding regions, telomeres, centromeres, and origins of replication, plus large numbers of transposable elements, inserted viral DNA, non-functional pseudogenes and simple, highly repetitive sequences. Introns make up a large percentage of non-coding DNA. Some of this non-coding DNA is non-functional junk DNA, such as pseudogenes, but there is no firm consensus on the total amount of junk DNA.

Although the sequence of the human genome has been completely determined by DNA sequencing in 2022 (including methylome), it is not yet fully understood. Most, but not all, genes have been identified by a combination of high throughput experimental and bioinformatics approaches, yet much work still needs to be done to further elucidate the biological functions of their protein and RNA products.

Genomic information

Schematic representation of the human diploid karyotype, showing the organization of the genome into chromosomes, as well as annotated bands and sub-bands as seen on G banding. This drawing shows both the female (XX) and male (XY) versions of the 23rd chromosome pair. Chromosomal changes during the cell cycle are displayed at top center. The mitochondrial genome is shown to scale at bottom left.

Size of the human genome

In 2000, scientists reported the sequencing of 88% of human genome,[2] but as of 2020, at least 8% was still missing.[3] In 2021, scientists reported sequencing a complete, female genome (i.e., without the Y chromosome).[4][3] The human Y chromosome, consisting of 62,460,029 base pairs from a different cell line and found in all males, was sequenced completely in January 2022.[5]

The current version of the standard reference genome is called GRCh38.p14 (July 2023). It consists of 22 autosomes plus one copy of the X chromosome and one copy of the Y chromosome. It contains approximately 3.1 billion base pairs (3.1 Gb or 3.1 x 10^9 bp).[6] This represents the size of a composite genome based on data from multiple individuals but it is a good indication of the typical amount of DNA in a haploid set of chromosomes because the Y chromosome is quite small.[7] Most human cells are diploid so they contain twice as much DNA (~6.2 billion base pairs).

In 2023, a draft human pangenome reference was published.[8] It is based on 47 genomes from persons of varied ethnicity.[8] Plans are underway for an improved reference capturing still more biodiversity from a still wider sample.[8]

While there are significant differences among the genomes of human individuals (on the order of 0.1% due to single-nucleotide variants[9] and 0.6% when considering indels),[10] these are considerably smaller than the differences between humans and their closest living relatives, the bonobos and chimpanzees (~1.1% fixed single-nucleotide variants[11] and 4% when including indels).[12]

Molecular organization and gene content

The total length of the human reference genome does not represent the sequence of any specific individual, nor does it represent the sequence of all of the DNA found within a cell. The human reference genome only includes one copy of each of the paired, homologous autosomes plus one copy of each of the two sex chromosomes (X and Y). The total amount of DNA in this reference genome is 3.1 billion base pairs (3.1 Gb).[13]

Protein-coding genes

Protein-coding sequences represent the most widely studied and best understood component of the human genome. These sequences ultimately lead to the production of all human proteins, although several biological processes (e.g. DNA rearrangements and alternative pre-mRNA splicing) can lead to the production of many more unique proteins than the number of protein-coding genes.

The human reference genome contains somewhere between 19,000 and 20,000 protein-coding genes.[14][15] These genes contain an average of 10 introns and the average size of an intron is about 6 kb (6,000 bp).[16] This means that the average size of a protein-coding gene is about 62 kb and these genes take up about 40% of the genome.[17]

Exon sequences consist of coding DNA and untranslated regions (UTRs) at either end of the mature mRNA. The total amount of coding DNA is about 1-2% of the genome.[18][16]

Many people divide the genome into coding and non-coding DNA based on the idea that coding DNA is the most important functional component of the genome. About 98-99% of the human genome is non-coding DNA.

Non-coding genes

Noncoding RNA molecules play many essential roles in cells, especially in the many reactions of protein synthesis and RNA processing. Noncoding genes include those for tRNAs, ribosomal RNAs, microRNAs, snRNAs and long non-coding RNAs (lncRNAs).[19][20][21][22] The number of reported non-coding genes continues to rise slowly but the exact number in the human genome is yet to be determined. Many RNAs are thought to be non-functional.[23]

Many ncRNAs are critical elements in gene regulation and expression. Noncoding RNA also contributes to epigenetics, transcription, RNA splicing, and the translational machinery. The role of RNA in genetic regulation and disease offers a new potential level of unexplored genomic complexity.[24]

Pseudogenes

Pseudogenes are inactive copies of protein-coding genes, often generated by gene duplication, that have become nonfunctional through the accumulation of inactivating mutations. The number of pseudogenes in the human genome is on the order of 13,000,[25] and in some chromosomes is nearly the same as the number of functional protein-coding genes. Gene duplication is a major mechanism through which new genetic material is generated during molecular evolution.

For example, the olfactory receptor gene family is one of the best-documented examples of pseudogenes in the human genome. More than 60 percent of the genes in this family are non-functional pseudogenes in humans. By comparison, only 20 percent of genes in the mouse olfactory receptor gene family are pseudogenes. Research suggests that this is a species-specific characteristic, as the most closely related primates all have proportionally fewer pseudogenes. This genetic discovery helps to explain the less acute sense of smell in humans relative to other mammals.[26]

Regulatory DNA sequences

The human genome has many different regulatory sequences which are crucial to controlling gene expression. Conservative estimates indicate that these sequences make up 8% of the genome,[27] however extrapolations from the ENCODE project give that 20[28] or more[29] of the genome is gene regulatory sequence. Some types of non-coding DNA are genetic "switches" that do not encode proteins, but do regulate when and where genes are expressed (called enhancers).[30]

Regulatory sequences have been known since the late 1960s.[31] The first identification of regulatory sequences in the human genome relied on recombinant DNA technology.[32] Later with the advent of genomic sequencing, the identification of these sequences could be inferred by evolutionary conservation. The evolutionary branch between the primates and mouse, for example, occurred 70–90 million years ago.[33] So computer comparisons of gene sequences that identify conserved non-coding sequences will be an indication of their importance in duties such as gene regulation.[34]

Other genomes have been sequenced with the same intention of aiding conservation-guided methods, for exampled the pufferfish genome.[35] However, regulatory sequences disappear and re-evolve during evolution at a high rate.[36][37][38]

As of 2012, the efforts have shifted toward finding interactions between DNA and regulatory proteins by the technique ChIP-Seq, or gaps where the DNA is not packaged by histones (DNase hypersensitive sites), both of which tell where there are active regulatory sequences in the investigated cell type.[27]

Repetitive DNA sequences

Repetitive DNA sequences comprise approximately 50% of the human genome.[39]

About 8% of the human genome consists of tandem DNA arrays or tandem repeats, low complexity repeat sequences that have multiple adjacent copies (e.g. "CAGCAGCAG...").[40] The tandem sequences may be of variable lengths, from two nucleotides to tens of nucleotides. These sequences are highly variable, even among closely related individuals, and so are used for genealogical DNA testing and forensic DNA analysis.[41]

Repeated sequences of fewer than ten nucleotides (e.g. the dinucleotide repeat $(AC)_n$) are termed microsatellite sequences. Among the microsatellite sequences, trinucleotide repeats are of particular importance, as sometimes occur within coding regions of genes for proteins and may lead to genetic disorders. For example, Huntington's disease results from an expansion of the trinucleotide repeat $(CAG)_n$ within the *Huntingtin* gene on human chromosome 4. Telomeres (the ends of linear chromosomes) end with a microsatellite hexanucleotide repeat of the sequence $(TTAGGG)_n$.

Tandem repeats of longer sequences (arrays of repeated sequences 10–60 nucleotides long) are termed minisatellites.[42]

Transposable genetic elements, DNA sequences that can replicate and insert copies of themselves at other locations within a host genome, are an abundant component in the human genome. The most abundant transposon lineage, *Alu*, has about 50,000 active copies,[43] and can be inserted into intragenic and intergenic regions.[44] One other lineage, LINE-1, has about 100 active copies per genome (the number varies between people).[45] Together with non-functional relics of old transposons, they account for over half of total human DNA.[46] Sometimes called "jumping genes", transposons have played a major role in sculpting the human genome. Some of these sequences represent endogenous retroviruses, DNA copies of viral sequences that have become permanently integrated into the genome and are now passed on to succeeding generations. There are also a significant number of retroviruses in human DNA, at least 3 of which have been proven to possess an important function (i.e., HIV-like functional HERV-K; envelope genes of non-functional viruses HERV-W and HERV-FRD play a role in placenta formation by inducing cell-cell fusion).

Mobile elements within the human genome can be classified into LTR retrotransposons (8.3% of total genome), SINEs (13.1% of total genome) including Alu elements, LINEs (20.4% of total genome), SVAs (SINE-VNTR-Alu) and Class II DNA transposons (2.9% of total genome).

Junk DNA

There is no consensus on what constitutes a "functional" element in the genome since geneticists, evolutionary biologists, and molecular biologists employ different definitions and methods.[47][48] Due to the ambiguity in the terminology, different schools of thought have emerged.[49] In evolutionary definitions, "functional" DNA, whether it is coding or non-coding, contributes to the fitness of the organism, and therefore is maintained by negative evolutionary pressure whereas "non-functional" DNA

has no benefit to the organism and therefore is under neutral selective pressure. This type of DNA has been described as junk DNA.[50][51] In genetic definitions, "functional" DNA is related to how DNA segments manifest by phenotype and "nonfunctional" is related to loss-of-function effects on the organism.[47] In biochemical definitions, "functional" DNA relates to DNA sequences that specify molecular products (e.g. noncoding RNAs) and biochemical activities with mechanistic roles in gene or genome regulation (i.e. DNA sequences that impact cellular level activity such as cell type, condition, and molecular processes).[52][47] There is no consensus in the literature on the amount of functional DNA since, depending on how "function" is understood, ranges have been estimated from up to 90% of the human genome is likely nonfunctional DNA (junk DNA)[53] to up to 80% of the genome is likely functional.[54] It is also possible that junk DNA may acquire a function in the future and therefore may play a role in evolution,[55] but this is likely to occur only very rarely.[50] Finally DNA that is deliterious to the organism and is under negative selective pressure is called garbage DNA.[51]

Sequencing

The first human genome sequences were published in nearly complete draft form in February 2001 by the Human Genome Project[56] and Celera Corporation.[57] Completion of the Human Genome Project's sequencing effort was announced in 2004 with the publication of a draft genome sequence, leaving just 341 gaps in the sequence, representing highly repetitive and other DNA that could not be sequenced with the technology available at the time.[58] The human genome was the first of all vertebrates to be sequenced to such near-completion, and as of 2018, the diploid genomes of over a million individual humans had been determined using next-generation sequencing.[59]

These data are used worldwide in biomedical science, anthropology, forensics and other branches of science. Such genomic studies have led to advances in the diagnosis and treatment of diseases, and to new insights in many fields of biology, including human evolution.

By 2018, the total number of genes had been raised to at least 46,831,[60] plus another 2300 micro-RNA genes.[61] A 2018 population survey found another 300 million bases of human genome that was not in the reference sequence.[62] Prior to the acquisition of the full genome sequence, estimates of the number of human genes ranged from 50,000 to 140,000 (with occasional vagueness about whether these estimates included non-protein coding genes).[63] As genome sequence quality and the methods for identifying protein-coding genes improved,[58] the count of recognized protein-coding genes dropped to 19,000–20,000.[64]

In 2022, the Telomere-to-Telomere (T2T) consortium reported the complete sequence of a human female genome,[3] filling all the gaps in the X chromosome (2020) and the 22 autosomes (May 2021).[3][65] The previously unsequenced parts contain immune response genes that help to adapt to and survive infections, as well as genes that are important for predicting drug response.[66] The completed human genome sequence will also provide better understanding of human formation as an individual organism and how humans vary both between each other and other species.[66]

Although the 'completion' of the human genome project was announced in 2001,[2] there remained hundreds of gaps, with about 5–10% of the total sequence remaining undetermined. The missing genetic information was mostly in repetitive heterochromatic regions and near the centromeres and telomeres, but also some gene-encoding euchromatic regions.[67] There remained 160 euchromatic gaps in 2015 when

the sequences spanning another 50 formerly unsequenced regions were determined.[68] Only in 2020 was the first truly complete telomere-to-telomere sequence of a human chromosome determined, namely of the X chromosome.[69] The first complete telomere-to-telomere sequence of a human autosomal chromosome, chromosome 8, followed a year later.[70] The complete human genome (without Y chromosome) was published in 2021, while with Y chromosome in January 2022.[3][4][71]

In 2023, a draft human pangenome reference was published.[8] It is based on 47 genomes from persons of varied ethnicity.[8] Plans are underway for an improved reference capturing still more biodiversity from a still wider sample.[8]

Genomic variation in humans

Human reference genome

With the exception of identical twins, all humans show significant variation in genomic DNA sequences. The human reference genome (HRG) is used as a standard sequence reference.

There are several important points concerning the human reference genome:

- The HRG is a haploid sequence. Each chromosome is represented once.
- The HRG is a composite sequence, and does not correspond to any actual human individual.
- The HRG is periodically updated to correct errors, ambiguities, and unknown "gaps".
- The HRG in no way represents an "ideal" or "perfect" human individual. It is simply a standardized representation or model that is used for comparative purposes.

The Genome Reference Consortium is responsible for updating the HRG. Version 38 was released in December 2013.[72]

Measuring human genetic variation

Most studies of human genetic variation have focused on single-nucleotide polymorphisms (SNPs), which are substitutions in individual bases along a chromosome. Most analyses estimate that SNPs occur 1 in 1000 base pairs, on average, in the euchromatic human genome, although they do not occur at a uniform density. Thus follows the popular statement that "we are all, regardless of race, genetically 99.9% the same",[73] although this would be somewhat qualified by most geneticists. For example, a much larger fraction of the genome is now thought to be involved in copy number variation.[74] A large-scale collaborative effort to catalog SNP variations in the human genome is being undertaken by the International HapMap Project.

The genomic loci and length of certain types of small repetitive sequences are highly variable from person to person, which is the basis of DNA fingerprinting and DNA paternity testing technologies. The heterochromatic portions of the human genome, which total several hundred million base pairs, are also thought to be quite variable within the human population (they are so repetitive and so long that they cannot be accurately sequenced with current technology). These regions contain few genes, and it is unclear whether any significant phenotypic effect results from typical variation in repeats or heterochromatin.

Most gross genomic mutations in gamete germ cells probably result in inviable embryos; however, a number of human diseases are related to large-scale genomic abnormalities. Down syndrome, Turner Syndrome, and a number of other diseases result from nondisjunction of entire chromosomes. Cancer cells frequently have aneuploidy of chromosomes and chromosome arms, although a cause and effect relationship between aneuploidy and cancer has not been established.

Mapping human genomic variation

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.[75][76]

An example of a variation map is the HapMap being developed by the International HapMap Project. The HapMap is a haplotype map of the human genome, "which will describe the common patterns of human DNA sequence variation."[77] It catalogs the patterns of small-scale variations in the genome that involve single DNA letters, or bases.

Researchers published the first sequence-based map of large-scale structural variation across the human genome in the journal *Nature* in May 2008.[78][79] Large-scale structural variations are differences in the genome among people that range from a few thousand to a few million DNA bases; some are gains or losses of stretches of genome sequence and others appear as re-arrangements of stretches of sequence. These variations include differences in the number of copies individuals have of a particular gene, deletions, translocations and inversions.

Structural variation

Structural variation refers to genetic variants that affect larger segments of the human genome, as opposed to point mutations. Often, structural variants (SVs) are defined as variants of 50 base pairs (bp) or greater, such as deletions, duplications, insertions, inversions and other rearrangements. About 90% of structural variants are noncoding deletions but most individuals have more than a thousand such deletions; the size of deletions ranges from dozens of base pairs to tens of thousands of bp.[80] On average, individuals carry ~3 rare structural variants that alter coding regions, e.g. delete exons. About 2% of individuals carry ultra-rare megabase-scale structural variants, especially rearrangements. That is, millions of base pairs may be inverted within a chromosome; ultra-rare means that they are only found in individuals or their family members and thus have arisen very recently.[80]

SNP frequency across the human genome

Single-nucleotide polymorphisms (SNPs) do not occur homogeneously across the human genome. In fact, there is enormous diversity in SNP frequency between genes, reflecting different selective pressures on each gene as well as different mutation and recombination rates across the genome. However, studies on SNPs are biased towards coding regions, the data generated from them are unlikely to reflect the overall distribution of SNPs throughout the genome. Therefore, the SNP Consortium protocol was designed to identify SNPs with no bias towards coding regions and the Consortium's 100,000 SNPs generally reflect sequence diversity across the human chromosomes. The SNP Consortium aims to expand the number of SNPs identified across the genome to 300 000 by the end of the first quarter of 2001.[81]

Changes in **non-coding sequence** and synonymous changes in **coding sequence** are generally more common than non-synonymous changes, reflecting greater selective pressure reducing diversity at positions dictating amino acid identity. Transitional changes are more common than transversions, with CpG dinucleotides showing the highest mutation rate, presumably due to deamination.

TSC SNP distribution along the long arm of chromosome 22 (from https://web.archive.org/web/20130903043223/http://snp.cshl.org/). Each column represents a 1 Mb interval; the approximate cytogenetic position is given on the x-axis. Clear peaks and troughs of SNP density can be seen, possibly reflecting different rates of mutation, recombination and selection.

Personal genomes

A personal genome sequence is a (nearly) complete sequence of the chemical base pairs that make up the DNA of a single person. Because medical treatments have different effects on different people due to genetic variations such as single-nucleotide polymorphisms (SNPs), the analysis of personal genomes may lead to personalized medical treatment based on individual genotypes.[82]

The first personal genome sequence to be determined was that of Craig Venter in 2007. Personal genomes had not been sequenced in the public Human Genome Project to protect the identity of volunteers who provided DNA samples. That sequence was derived from the DNA of several volunteers from a diverse population.[83] However, early in the Venter-led Celera Genomics genome sequencing effort the decision was made to switch from sequencing a composite sample to using DNA from a single individual, later revealed to have been Venter himself. Thus the Celera human genome sequence released in 2000 was largely that of one man. Subsequent replacement of the early composite-derived data and determination of the diploid sequence, representing both sets of chromosomes, rather than a haploid sequence originally reported, allowed the release of the first personal genome.[84] In April 2008, that of James Watson was also completed. In 2009, Stephen Quake published his own genome sequence derived from a sequencer of his own design, the Heliscope.[85] A Stanford team led by Euan Ashley published a framework for the medical interpretation of human genomes implemented on Quake's genome and made whole genome-informed medical decisions for the first time.[86] That team further extended the approach to the West family, the first family sequenced as part of Illumina's Personal Genome Sequencing program.[87] Since then hundreds of personal genome sequences have been released,[88] including those of Desmond Tutu,[89][90] and of a Paleo-Eskimo.[91] In 2012, the whole genome sequences of two family trios among 1092 genomes was made public.[9] In November 2013, a Spanish family made four personal exome datasets (about 1% of the genome) publicly available under a Creative Commons public domain license.[92][93] The Personal Genome Project (started in 2005) is among the few to make both genome sequences and corresponding medical phenotypes publicly available.[94][95]

The sequencing of individual genomes further unveiled levels of genetic complexity that had not been appreciated before. Personal genomics helped reveal the significant level of diversity in the human genome attributed not only to SNPs but structural variations as well. However, the application of such knowledge to the treatment of disease and in the medical field is only in its very beginnings.[96] Exome
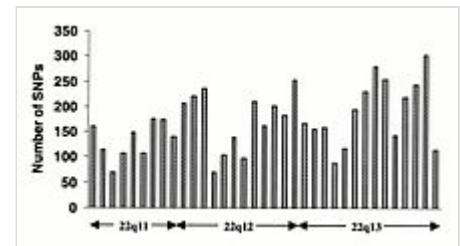
sequencing has become increasingly popular as a tool to aid in diagnosis of genetic disease because the exome contributes only 1% of the genomic sequence but accounts for roughly 85% of mutations that contribute significantly to disease.[97]
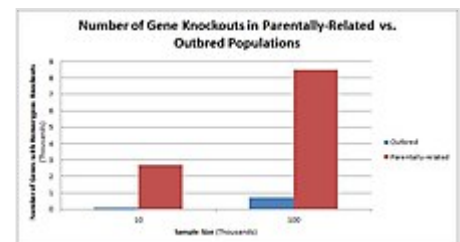
Human knockouts

In humans, gene knockouts naturally occur as heterozygous or homozygous loss-of-function gene knockouts. These knockouts are often difficult to distinguish, especially within heterogeneous genetic backgrounds. They are also difficult to find as they occur in low frequencies.
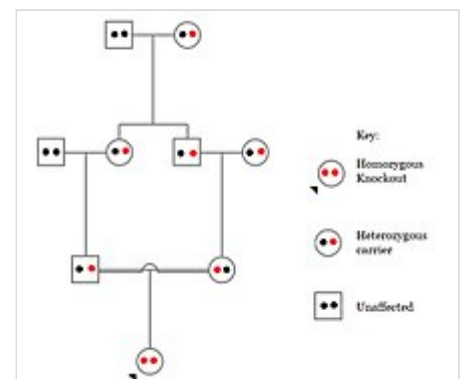
Populations with high rates of consanguinity, such as countries with high rates of first-cousin marriages, display the highest frequencies of homozygous gene knockouts. Such populations include Pakistan, Iceland, and Amish populations. These populations with a high level of parental-relatedness have been subjects of human knock out research which has helped to determine the function of specific genes in humans. By distinguishing specific knockouts, researchers are able to use phenotypic analyses of these individuals to help characterize the gene that has been knocked out.

Populations with a high level of parental-relatedness result in a larger number of homozygous gene knockouts as compared to outbred populations.[98]

Knockouts in specific genes can cause genetic diseases, potentially have beneficial effects, or even result in no phenotypic effect at all. However, determining a knockout's phenotypic effect and in humans can be challenging. Challenges to characterizing and clinically interpreting knockouts include difficulty calling of DNA variants, determining disruption of protein function (annotation), and considering the amount of influence mosaicism has on the phenotype.[98]

One major study that investigated human knockouts is the Pakistan Risk of Myocardial Infarction study. It was found that individuals possessing a heterozygous loss-of-function gene knockout for the APOC3 gene had lower triglycerides in the blood after consuming a high fat meal as compared to individuals without the mutation. However, individuals possessing homozygous loss-of-function gene knockouts of the APOC3 gene displayed the lowest level of triglycerides in the blood after the fat load test, as they produce no functional APOC3 protein.[99]

A pedigree displaying a first-cousin mating (carriers both carrying heterozygous knockouts mating as marked by double line) leading to offspring possessing a homozygous gene knockout

Human genetic disorders

Most aspects of human biology involve both genetic (inherited) and non-genetic (environmental) factors. Some inherited variation influences aspects of our biology that are not medical in nature (height, eye color, ability to taste or smell certain compounds, etc.). Moreover, some genetic disorders only cause disease in combination with the appropriate environmental factors (such as diet). With these caveats,

genetic disorders may be described as clinically defined diseases caused by genomic DNA sequence variation. In the most straightforward cases, the disorder can be associated with variation in a single gene. For example, cystic fibrosis is caused by mutations in the CFTR gene and is the most common recessive disorder in caucasian populations with over 1,300 different mutations known.[100]

Disease-causing mutations in specific genes are usually severe in terms of gene function and are rare, thus genetic disorders are similarly individually rare. However, since there are many genes that can vary to cause genetic disorders, in aggregate they constitute a significant component of known medical conditions, especially in pediatric medicine. Molecularly characterized genetic disorders are those for which the underlying causal gene has been identified. Currently there are approximately 2,200 such disorders annotated in the OMIM database.[100]

Studies of genetic disorders are often performed by means of family-based studies. In some instances, population based approaches are employed, particularly in the case of so-called founder populations such as those in Finland, French-Canada, Utah, Sardinia, etc. Diagnosis and treatment of genetic disorders are usually performed by a geneticist-physician trained in clinical/medical genetics. The results of the Human Genome Project are likely to provide increased availability of genetic testing for gene-related disorders, and eventually improved treatment. Parents can be screened for hereditary conditions and counselled on the consequences, the probability of inheritance, and how to avoid or ameliorate it in their offspring.

There are many different kinds of DNA sequence variation, ranging from complete extra or missing chromosomes down to single nucleotide changes. It is generally presumed that much naturally occurring genetic variation in human populations is phenotypically neutral, i.e., has little or no detectable effect on the physiology of the individual (although there may be fractional differences in fitness defined over evolutionary time frames). Genetic disorders can be caused by any or all known types of sequence variation. To molecularly characterize a new genetic disorder, it is necessary to establish a causal link between a particular genomic sequence variant and the clinical disease under investigation. Such studies constitute the realm of human molecular genetics.

With the advent of the Human Genome and International HapMap Project, it has become feasible to explore subtle genetic influences on many common disease conditions such as diabetes, asthma, migraine, schizophrenia, etc. Although some causal links have been made between genomic sequence variants in particular genes and some of these diseases, often with much publicity in the general media, these are usually not considered to be genetic disorders *per se* as their causes are complex, involving many different genetic and environmental factors. Thus there may be disagreement in particular cases whether a specific medical condition should be termed a genetic disorder.

Additional genetic disorders of mention are Kallman syndrome and Pfeiffer syndrome (gene FGFR1), Fuchs corneal dystrophy (gene TCF4), Hirschsprung's disease (genes RET and FECH), Bardet-Biedl syndrome 1 (genes CCDC28B and BBS1), Bardet-Biedl syndrome 10 (gene BBS10), and facioscapulohumeral muscular dystrophy type 2 (genes D4Z4 and SMCHD1).[101]

Genome sequencing is now able to narrow the genome down to specific locations to more accurately find mutations that will result in a genetic disorder. Copy number variants (CNVs) and single nucleotide variants (SNVs) are also able to be detected at the same time as genome sequencing with newer sequencing procedures available, called Next Generation Sequencing (NGS).[102] This only analyzes a small portion of the genome, around 1–2%. The results of this sequencing can be used for clinical diagnosis of a genetic condition, including Usher syndrome, retinal disease, hearing impairments,

diabetes, epilepsy, Leigh disease, hereditary cancers, neuromuscular diseases, primary immunodeficiencies, severe combined immunodeficiency (SCID), and diseases of the mitochondria.[103] NGS can also be used to identify carriers of diseases before conception. The diseases that can be detected in this sequencing include Tay-Sachs disease, Bloom syndrome, Gaucher disease, Canavan disease, familial dysautonomia, cystic fibrosis, spinal muscular atrophy, and fragile-X syndrome. The Next Genome Sequencing can be narrowed down to specifically look for diseases more prevalent in certain ethnic populations.[104]

Prevalence and associated gene/chromosome for some human genetic disorders

Disorder	Prevalence	Chromosome or gene involved
Chromosomal conditions		
Down syndrome	1:600	Chromosome 21
Klinefelter syndrome	1:500–1000 males	Additional X chromosome
Turner syndrome	1:2000 females	Loss of X chromosome
Sickle cell anemia	1 in 50 births in parts of Africa; rarer elsewhere	β-globin (on chromosome 11)
Bloom syndrome	1:48000 Ashkenazi Jews	BLM
Cancers		
Breast/Ovarian cancer (susceptibility)	~5% of cases of these cancer types	BRCA1, BRCA2
FAP (hereditary nonpolyposis coli)	1:3500	APC
Lynch syndrome	5–10% of all cases of bowel cancer	MLH1, MSH2, MSH6, PMS2
Fanconi anemia	1:130000 births	FANCC
Neurological conditions		
Huntington disease	1:20000	Huntingtin
Alzheimer disease - early onset	1:2500	PS1, PS2, APP
Tay-Sachs	1:3600 births in Ashkenazi Jews	HEXA gene (on chromosome 15)
Canavan disease	2.5% Eastern European Jewish ancestry	ASPA gene (on chromosome 17)
Familial dysautonomia	600 known cases worldwide since discovery	IKBKAP gene (on chromosome 9)
Fragile X syndrome	1.4:10000 in males, 0.9:10000 in females	FMR1 gene (on X chromosome)
Mucolipidosis type IV	1:90 to 1:100 in Ashkenazi Jews	MCOLN1
Other conditions		
Cystic fibrosis	1:2500	CFTR
Duchenne muscular dystrophy	1:3500 boys	Dystrophin
Becker muscular dystrophy	1.5–6:100000 males	DMD
Beta thalassemia	1:100000	HBB
Congenital adrenal hyperplasia	1:280 in Native Americans and Yupik Eskimos 1:15000 in American Caucasians	CYP21A2
Glycogen storage disease type I	1:100000 births in America	G6PC
Maple syrup urine disease	1:180000 in the U.S.	BCKDHA, BCKDHB, DBT, DLD

	1:176 in Mennonite/Amish communities 1:250000 in Austria	
Niemann–Pick disease, SMPD1-associated	1,200 cases worldwide	SMPD1
Usher syndrome	1:23000 in the U.S. 1:28000 in Norway 1:12500 in Germany	CDH23, CLRN1, DFNB31, GPR98, MYO7A, PCDH15, USH1C, USH1G, USH2A

Evolution

Comparative genomics studies of mammalian genomes suggest that approximately 5% of the human genome has been conserved by evolution since the divergence of extant lineages approximately 200 million years ago, containing the vast majority of genes.[105][106] The published chimpanzee genome differs from that of the human genome by 1.23% in direct sequence comparisons.[107] Around 20% of this figure is accounted for by variation within each species, leaving only ~1.06% consistent sequence divergence between humans and chimps at shared genes.[108] This nucleotide by nucleotide difference is dwarfed, however, by the portion of each genome that is not shared, including around

Hominin timeline
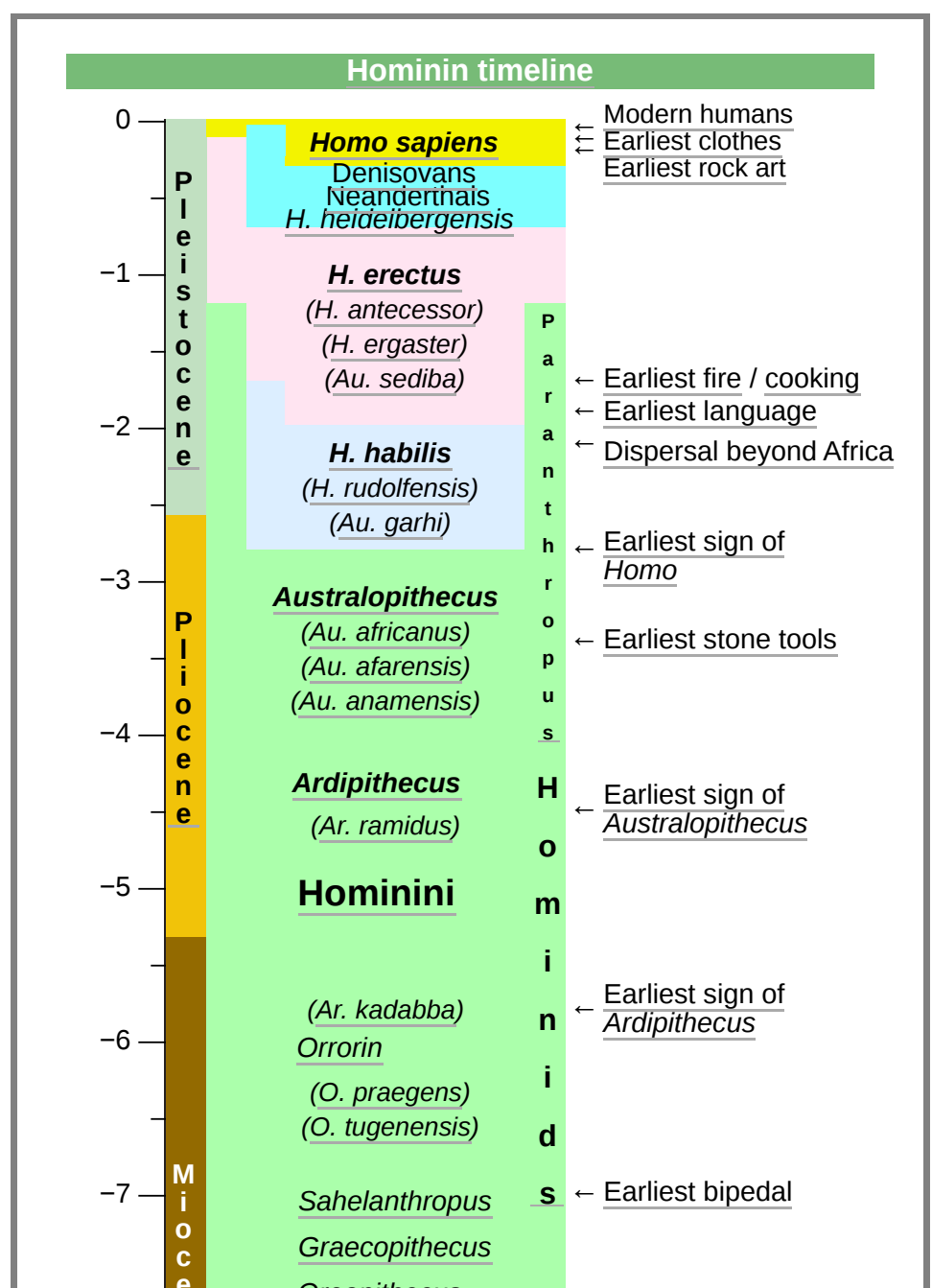
6% of functional genes that are unique to either humans or chimps.[109]

In other words, the considerable observable differences between humans and chimps may be due as much or more to genome level variation in the number, function and expression of genes rather

Oreopithecus
Chororapithecus
Sivapithecus
Ouranopithecus
(Ou. turkae)
(Ou. macedoniensis)
Samburupithecus
Nakalipithecus

← Chimpanzee split
← Gorilla split
← Earlier apes

−8

−9

−10

(million years ago)

than DNA sequence changes in shared genes. Indeed, even within humans, there has been found to be a previously unappreciated amount of copy number variation (CNV) which can make up as much as 5–15% of the human genome. In other words, between humans, there could be +/- 500,000,000 base pairs of DNA, some being active genes, others inactivated, or active at different levels. The full significance of this finding remains to be seen. On average, a typical human protein-coding gene differs from its chimpanzee ortholog by only two amino acid substitutions; nearly one third of human genes have exactly the same protein translation as their chimpanzee orthologs. A major difference between the two genomes is human chromosome 2, which is equivalent to a fusion product of chimpanzee chromosomes 12 and 13.[110] (later renamed to chromosomes 2A and 2B, respectively).

Humans have undergone an extraordinary loss of olfactory receptor genes during our recent evolution, which explains our relatively crude sense of smell compared to most other mammals. Evolutionary evidence suggests that the emergence of color vision in humans and several other primate species has diminished the need for the sense of smell.[111]

In September 2016, scientists reported that, based on human DNA genetic studies, all non-Africans in the world today can be traced to a single population that exited Africa between 50,000 and 80,000 years ago.[112]

Mitochondrial DNA

The human mitochondrial DNA is of tremendous interest to geneticists, since it undoubtedly plays a role in mitochondrial disease. It also sheds light on human evolution; for example, analysis of variation in the human mitochondrial genome has led to the postulation of a recent common ancestor for all humans on the maternal line of descent (see Mitochondrial Eve).

Due to the damage induced by the exposure to Reactive Oxygen Species mitochondrial DNA (mtDNA) has a more rapid rate of variation than nuclear DNA. This 20-fold higher mutation rate allows mtDNA to be used for more accurate tracing of maternal ancestry. Studies of mtDNA in populations have allowed ancient migration paths to be traced, such as the migration of Native Americans from Siberia[113] or Polynesians from southeastern Asia. It has also been used to show that there is no trace of Neanderthal DNA in the European gene mixture inherited through purely maternal lineage.[114] Due to the restrictive all or none manner of mtDNA inheritance, this result (no trace of Neanderthal mtDNA) would be likely unless there were a large percentage of Neanderthal ancestry, or there was strong positive selection for that mtDNA. For example, going back 5 generations, only 1 of a person's 32 ancestors contributed to that
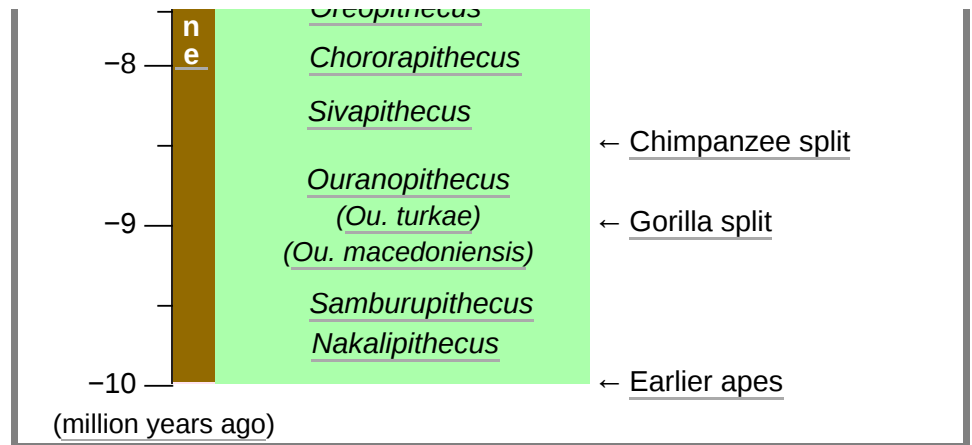
person's mtDNA, so if one of these 32 was pure Neanderthal an expected ~3% of that person's autosomal DNA would be of Neanderthal origin, yet they would have a ~97% chance of having no trace of Neanderthal mtDNA.

Epigenome

Epigenetics describes a variety of features of the human genome that transcend its primary DNA sequence, such as chromatin packaging, histone modifications and DNA methylation, and which are important in regulating gene expression, genome replication and other cellular processes. Epigenetic markers strengthen and weaken transcription of certain genes but do not affect the actual sequence of DNA nucleotides. DNA methylation is a major form of epigenetic control over gene expression and one of the most highly studied topics in epigenetics. During development, the human DNA methylation profile experiences dramatic changes. In early germ line cells, the genome has very low methylation levels. These low levels generally describe active genes. As development progresses, parental imprinting tags lead to increased methylation activity.[115][116]

Epigenetic patterns can be identified between tissues within an individual as well as between individuals themselves. Identical genes that have differences only in their epigenetic state are called **epialleles**. Epialleles can be placed into three categories: those directly determined by an individual's genotype, those influenced by genotype, and those entirely independent of genotype. The epigenome is also influenced significantly by environmental factors. Diet, toxins, and hormones impact the epigenetic state. Studies in dietary manipulation have demonstrated that methyl-deficient diets are associated with hypomethylation of the epigenome. Such studies establish epigenetics as an important interface between the environment and the genome.[117]

See also

- Human Genome Organisation
- Genome Reference Consortium
- Human Genome Project
- Genetics
- Genomics
- Genographic Project
- Genomic organization
- Karyotype
- Low copy repeats
- Non-coding DNA
- Whole genome sequencing
- Universal Declaration on the Human Genome and Human Rights

References

1. Brown TA (2002). *The Human Genome* (https://www.ncbi.nlm.nih.gov/books/NBK21134/) (2nd ed.). Oxford: Wiley-Liss.

2. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. (February 2001). "Initial sequencing and analysis of the human genome" (https://doi.org/10.1038%2F35057062). *Nature*. **409** (6822): 860–921. Bibcode:2001Natur.409..860L (https://ui.adsabs.harvard.edu/abs/2001Natur.409..860L). doi:10.1038/35057062 (https://doi.org/10.1038%2F35057062). hdl:2027.42/62798 (https://hdl.handle.net/2027.42%2F62798). PMID 11237011 (https://pubmed.ncbi.nlm.nih.gov/11237011).

3. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. (April 2022). "The complete sequence of a human genome" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9186530). *Science*. **376** (6588): 44–53. Bibcode:2022Sci...376...44N (https://ui.adsabs.harvard.edu/abs/2022Sci...376...44N). doi:10.1126/science.abj6987 (https://doi.org/10.1126%2Fscience.abj6987). PMC 9186530 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9186530). PMID 35357919 (https://pubmed.ncbi.nlm.nih.gov/35357919). S2CID 247854936 (https://api.semanticscholar.org/CorpusID:247854936).

4. "CHM13 T2T v1.1 – Genome – Assembly – NCBI" (https://www.ncbi.nlm.nih.gov/assembly/GCA_009914755.3). *www.ncbi.nlm.nih.gov*. Retrieved 26 July 2021.

5. Rhie A, Nurk S, Cechova M, Hoyt SJ, Taylor DJ, Altemose N, et al. (September 2023). "The complete sequence of a human Y chromosome" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10752217). *Nature*. **621** (7978): 344–354. Bibcode:2023Natur.621..344R (https://ui.adsabs.harvard.edu/abs/2023Natur.621..344R). doi:10.1038/s41586-023-06457-y (https://doi.org/10.1038%2Fs41586-023-06457-y). PMC 10752217 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10752217). PMID 37612512 (https://pubmed.ncbi.nlm.nih.gov/37612512). "Received 2 December 2022"

6. "Human assembly and gene annotation" (https://useast.ensembl.org/Homo_sapiens/Info/Annotation). *Ensembl*. 2022. Retrieved 28 February 2024.

7. Hallast P, Ebert P, Loftus M, Yilmaz F, Audano PA, Logsdon GA, et al. (September 2023). "Assembly of 43 human Y chromosomes reveals extensive complexity and variation" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10726138). *Nature*. **621** (7978): 355–364. Bibcode:2023Natur.621..355H (https://ui.adsabs.harvard.edu/abs/2023Natur.621..355H). doi:10.1038/s41586-023-06425-6 (https://doi.org/10.1038%2Fs41586-023-06425-6). ISSN 0028-0836 (https://search.worldcat.org/issn/0028-0836). PMC 10726138 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10726138). PMID 37612510 (https://pubmed.ncbi.nlm.nih.gov/37612510).

8. Liao WW, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, et al. (May 2023). "A draft human pangenome reference" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10172123). *Nature*. **617** (7960): 312–324. Bibcode:2023Natur.617..312L (https://ui.adsabs.harvard.edu/abs/2023Natur.617..312L). doi:10.1038/s41586-023-05896-x (https://doi.org/10.1038%2Fs41586-023-05896-x). PMC 10172123 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10172123). PMID 37165242 (https://pubmed.ncbi.nlm.nih.gov/37165242).

9. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. (November 2012). "An integrated map of genetic variation from 1,092 human genomes" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3498066). *Nature*. **491** (7422): 56–65. Bibcode:2012Natur.491...56T (https://ui.adsabs.harvard.edu/abs/2012Natur.491...56T). doi:10.1038/nature11632 (https://doi.org/10.1038%2Fnature11632). PMC 3498066 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3498066). PMID 23128226 (https://pubmed.ncbi.nlm.nih.gov/23128226).

10. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. (October 2015). "A global reference for human genetic variation" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4750478). *Nature*. **526** (7571): 68–74. Bibcode:2015Natur.526...68T (https://ui.adsabs.harvard.edu/abs/2015Natur.526...68T). doi:10.1038/nature15393 (https://doi.org/10.1038%2Fnature15393). PMC 4750478 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4750478). PMID 26432245 (https://pubmed.ncbi.nlm.nih.gov/26432245).

11. Chimpanzee Sequencing Analysis Consortium (September 2005). "Initial sequence of the chimpanzee genome and comparison with the human genome" (https://doi.org/10.1038%2F nature04072). *Nature*. **437** (7055): 69–87. Bibcode:2005Natur.437...69. (https://ui.adsabs.ha rvard.edu/abs/2005Natur.437...69.). doi:10.1038/nature04072 (https://doi.org/10.1038%2Fn ature04072). PMID 16136131 (https://pubmed.ncbi.nlm.nih.gov/16136131). S2CID 2638825 (https://api.semanticscholar.org/CorpusID:2638825).

12. Varki A, Altheide TK (December 2005). "Comparing the human and chimpanzee genomes: searching for needles in a haystack" (https://doi.org/10.1101%2Fgr.3737405). *Genome Research*. **15** (12): 1746–1758. doi:10.1101/gr.3737405 (https://doi.org/10.1101%2Fgr.3737 405). PMID 16339373 (https://pubmed.ncbi.nlm.nih.gov/16339373).

13. "Human genome assembly" (https://www.ensembl.org/Homo_sapiens/Info/Annotation). *Ensembl*. Retrieved 23 January 2024.

14. "Gene" (https://www.genome.gov/genetics-glossary/Gene#:~:text=And%20genes%20are%2 0the%20part,of%20the%20entire%20human%20genome.). *www.genome.gov*. Retrieved 7 January 2025.

15. Amaral P, Carbonell-Sala S, De La Vega FM, Faial T, Frankish A, Gingeras T, et al. (2023). "The status of the human gene catalogue" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1 0575709). *Nature*. **622** (7981): 41–47. arXiv:2303.13996 (https://arxiv.org/abs/2303.13996). Bibcode:2023Natur.622...41A (https://ui.adsabs.harvard.edu/abs/2023Natur.622...41A). doi:10.1038/s41586-023-06490-x (https://doi.org/10.1038%2Fs41586-023-06490-x). PMC 10575709 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10575709). PMID 37794265 (https://pubmed.ncbi.nlm.nih.gov/37794265).

16. Piovesan A, Antonaros F, Vitale L, Strippoli P, Pelleri MC, Caracausi M (2019). "Human protein-coding genes and gene feature statistics in 2019" (https://www.ncbi.nlm.nih.gov/pm c/articles/PMC6549324). *BMC Research Notes*. **12** (1): 315. doi:10.1186/s13104-019-4343-8 (https://doi.org/10.1186%2Fs13104-019-4343-8). PMC 6549324 (https://www.ncbi.nlm.nih. gov/pmc/articles/PMC6549324). PMID 31164174 (https://pubmed.ncbi.nlm.nih.gov/3116417 4).

17. Francis WR, Wörheide G (June 2017). "Similar Ratios of Introns to Intergenic Sequence across Animal Genomes" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5534336). *Genome Biology and Evolution*. **9** (6): 1582–1598. doi:10.1093/gbe/evx103 (https://doi.org/1 0.1093%2Fgbe%2Fevx103). PMC 5534336 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC 5534336). PMID 28633296 (https://pubmed.ncbi.nlm.nih.gov/28633296).

18. Hatje K, Mühlhausen S, Simm D, Killmar M (2019). "The Protein-Coding Human Genome: Annotating High-Hanging Fruits". *BioEssays*. **41** (11): 1900066. doi:10.1002/bies.201900066 (https://doi.org/10.1002%2Fbies.201900066). PMID 31544971 (https://pubmed.ncbi.nlm.nih.gov/31544971).

19. Pennisi E (September 2012). "Genomics. ENCODE project writes eulogy for junk DNA". *Science*. **337** (6099): 1159–1161. doi:10.1126/science.337.6099.1159 (https://doi.org/10.11 26%2Fscience.337.6099.1159). PMID 22955811 (https://pubmed.ncbi.nlm.nih.gov/2295581 1).

20. Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, et al. (March 2015). "The landscape of long noncoding RNAs in the human transcriptome" (https://www.ncbi.nlm.nih.g ov/pmc/articles/PMC4417758). *Nature Genetics*. **47** (3): 199–208. doi:10.1038/ng.3192 (http s://doi.org/10.1038%2Fng.3192). PMC 4417758 (https://www.ncbi.nlm.nih.gov/pmc/articles/ PMC4417758). PMID 25599403 (https://pubmed.ncbi.nlm.nih.gov/25599403).

21. Eddy SR (December 2001). "Non-coding RNA genes and the modern RNA world". *Nature Reviews Genetics*. **2** (12): 919–929. doi:10.1038/35103511 (https://doi.org/10.1038%2F351 03511). PMID 11733745 (https://pubmed.ncbi.nlm.nih.gov/11733745). S2CID 18347629 (htt ps://api.semanticscholar.org/CorpusID:18347629).

22. Managadze D, Lobkovsky AE, Wolf YI, Shabalina SA, Rogozin IB, Koonin EV (2013). "The vast, conserved mammalian lincRNome" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3585383). *PLOS Computational Biology*. **9** (2): e1002917. Bibcode:2013PLSCB...9E2917M (https://ui.adsabs.harvard.edu/abs/2013PLSCB...9E2917M). doi:10.1371/journal.pcbi.1002917 (https://doi.org/10.1371%2Fjournal.pcbi.1002917). PMC 3585383 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3585383). PMID 23468607 (https://pubmed.ncbi.nlm.nih.gov/23468607).

23. Palazzo AF, Lee ES (2015). "Non-coding RNA: what is functional and what is junk?" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4306305). *Frontiers in Genetics*. **6**: 2. doi:10.3389/fgene.2015.00002 (https://doi.org/10.3389%2Ffgene.2015.00002). PMC 4306305 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4306305). PMID 25674102 (https://pubmed.ncbi.nlm.nih.gov/25674102).

24. Mattick JS, Makunin IV (April 2006). "Non-coding RNA" (https://doi.org/10.1093%2Fhmg%2Fddl046). *Human Molecular Genetics*. **15** (Spec No 1): R17–29. doi:10.1093/hmg/ddl046 (https://doi.org/10.1093%2Fhmg%2Fddl046). PMID 16651366 (https://pubmed.ncbi.nlm.nih.gov/16651366).

25. Pei B, Sisu C, Frankish A, Howald C, Habegger L, Mu XJ, et al. (2012). "The GENCODE pseudogene resource" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3491395). *Genome Biology*. **13** (9): R51. doi:10.1186/gb-2012-13-9-r51 (https://doi.org/10.1186%2Fgb-2012-13-9-r51). PMC 3491395 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3491395). PMID 22951037 (https://pubmed.ncbi.nlm.nih.gov/22951037).

26. Gilad Y, Man O, Pääbo S, Lancet D (March 2003). "Human specific loss of olfactory receptor genes" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC152291). *Proceedings of the National Academy of Sciences of the United States of America*. **100** (6): 3324–3327. Bibcode:2003PNAS..100.3324G (https://ui.adsabs.harvard.edu/abs/2003PNAS..100.3324G). doi:10.1073/pnas.0535697100 (https://doi.org/10.1073%2Fpnas.0535697100). PMC 152291 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC152291). PMID 12612342 (https://pubmed.ncbi.nlm.nih.gov/12612342).

27. Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M (September 2012). "An integrated encyclopedia of DNA elements in the human genome" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3439153). *Nature*. **489** (7414): 57–74. Bibcode:2012Natur.489...57T (https://ui.adsabs.harvard.edu/abs/2012Natur.489...57T). doi:10.1038/nature11247 (https://doi.org/10.1038%2Fnature11247). PMC 3439153 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3439153). PMID 22955616 (https://pubmed.ncbi.nlm.nih.gov/22955616).

28. Birney E (5 September 2012). "ENCODE: My own thoughts" (http://genomeinformatician.blogspot.ca/2012/09/encode-my-own-thoughts.html). *Ewan's Blog: Bioinformatician at large*.

29. Stamatoyannopoulos JA (September 2012). "What does our genome encode?" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3431477). *Genome Research*. **22** (9): 1602–1611. doi:10.1101/gr.146506.112 (https://doi.org/10.1101%2Fgr.146506.112). PMC 3431477 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3431477). PMID 22955972 (https://pubmed.ncbi.nlm.nih.gov/22955972).

30. Carroll SB, Gompel N, Prudhomme B (May 2008). "Regulating Evolution". *Scientific American*. **298** (5): 60–67. Bibcode:2008SciAm.298e..60C (https://ui.adsabs.harvard.edu/abs/2008SciAm.298e..60C). doi:10.1038/scientificamerican0508-60 (https://doi.org/10.1038%2Fscientificamerican0508-60). PMID 18444326 (https://pubmed.ncbi.nlm.nih.gov/18444326).

31. Miller JH, Ippen K, Scaife JG, Beckwith JR (1968). "The promoter-operator region of the lac operon of Escherichia coli". *J. Mol. Biol.* **38** (3): 413–420. doi:10.1016/0022-2836(68)90395-1 (https://doi.org/10.1016%2F0022-2836%2868%2990395-1). PMID 4887877 (https://pubmed.ncbi.nlm.nih.gov/4887877).

32. Wright S, Rosenthal A, Flavell R, Grosveld F (1984). "DNA sequences required for regulated expression of beta-globin genes in murine erythroleukemia cells". *Cell*. **38** (1): 265–273. doi:10.1016/0092-8674(84)90548-8 (https://doi.org/10.1016%2F0092-8674%2884%29290548-8). PMID 6088069 (https://pubmed.ncbi.nlm.nih.gov/6088069). S2CID 34587386 (https://api.semanticscholar.org/CorpusID:34587386).

33. Nei M, Xu P, Glazko G (February 2001). "Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC30166). *Proceedings of the National Academy of Sciences of the United States of America*. **98** (5): 2497–2502. Bibcode:2001PNAS...98.2497N (https://ui.adsabs.harvard.edu/abs/2001PNAS...98.2497N). doi:10.1073/pnas.051611498 (https://doi.org/10.1073%2Fpnas.051611498). PMC 30166 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC30166). PMID 11226267 (https://pubmed.ncbi.nlm.nih.gov/11226267).

34. Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, et al. (April 2000). "Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons". *Science*. **288** (5463): 136–140. Bibcode:2000Sci...288..136L (https://ui.adsabs.harvard.edu/abs/2000Sci...288..136L). doi:10.1126/science.288.5463.136 (https://doi.org/10.1126%2Fscience.288.5463.136). PMID 10753117 (https://pubmed.ncbi.nlm.nih.gov/10753117). Summary (http://www.lbl.gov/Science-Articles/Archive/mouse-dna-model.html) Archived (https://web.archive.org/web/20091106091608/http://www.lbl.gov/Science-Articles/Archive/mouse-dna-model.html) 6 November 2009 at the Wayback Machine

35. Meunier M. "Genoscope and Whitehead announce a high sequence coverage of the Tetraodon nigroviridis genome" (https://web.archive.org/web/20061016085223/http://www.cns.fr/externe/English/Actualites/Presse/261001_1.html). Genoscope. Archived from the original (http://www.cns.fr/externe/English/Actualites/Presse/261001_1.html) on 16 October 2006. Retrieved 12 September 2006.

36. Romero IG, Ruvinsky I, Gilad Y (July 2012). "Comparative studies of gene expression and the evolution of gene regulation" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4034676). *Nature Reviews Genetics*. **13** (7): 505–516. doi:10.1038/nrg3229 (https://doi.org/10.1038%2Fnrg3229). PMC 4034676 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4034676). PMID 22705669 (https://pubmed.ncbi.nlm.nih.gov/22705669).

37. Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, et al. (May 2010). "Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3008766). *Science*. **328** (5981): 1036–1040. Bibcode:2010Sci...328.1036S (https://ui.adsabs.harvard.edu/abs/2010Sci...328.1036S). doi:10.1126/science.1186176 (https://doi.org/10.1126%2Fscience.1186176). PMC 3008766 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3008766). PMID 20378774 (https://pubmed.ncbi.nlm.nih.gov/20378774).

38. Wilson MD, Barbosa-Morais NL, Schmidt D, Conboy CM, Vanes L, Tybulewicz VL, et al. (October 2008). "Species-specific transcription in mice carrying human chromosome 21" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3717767). *Science*. **322** (5900): 434–438. Bibcode:2008Sci...322..434W (https://ui.adsabs.harvard.edu/abs/2008Sci...322..434W). doi:10.1126/science.1160930 (https://doi.org/10.1126%2Fscience.1160930). PMC 3717767 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3717767). PMID 18787134 (https://pubmed.ncbi.nlm.nih.gov/18787134).

39. Treangen TJ, Salzberg SL (January 2012). "Repetitive DNA and next-generation sequencing: computational challenges and solutions" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3324860). *Nature Reviews Genetics*. **13** (1): 36–46. doi:10.1038/nrg3117 (https://doi.org/10.1038%2Fnrg3117). PMC 3324860 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3324860). PMID 22124482 (https://pubmed.ncbi.nlm.nih.gov/22124482).

40. Duitama J, Zablotskaya A, Gemayel R, Jansen A, Belet S, Vermeesch JR, et al. (May 2014). "Large-scale analysis of tandem repeat variability in the human genome" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4027155). *Nucleic Acids Research*. **42** (9): 5728–5741. doi:10.1093/nar/gku212 (https://doi.org/10.1093%2Fnar%2Fgku212). PMC 4027155 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4027155). PMID 24682812 (https://pubmed.ncbi.nlm.nih.gov/24682812).

41. Pierce BA (2012). *Genetics : a conceptual approach* (4th ed.). New York: W.H. Freeman. pp. 538–540. ISBN 978-1-4292-3250-0.

42. "minisatellite, n. meanings, etymology and more | Oxford English Dictionary" (https://www.oed.com/dictionary/minisatellite_n?tab=meaning_and_use&tl=true). *www.oed.com*. Retrieved 8 October 2023.

43. Bennett EA, Keller H, Mills RE, Schmidt S, Moran JV, Weichenrieder O, et al. (December 2008). "Active Alu retrotransposons in the human genome" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2593586). *Genome Research*. **18** (12): 1875–1883. doi:10.1101/gr.081737.108 (https://doi.org/10.1101%2Fgr.081737.108). PMC 2593586 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2593586). PMID 18836035 (https://pubmed.ncbi.nlm.nih.gov/18836035).

44. Liang KH, Yeh CT (2013). "A gene expression restriction network mediated by sense and antisense Alu sequences located on protein-coding messenger RNAs" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3655826). *BMC Genomics*. **14**: 325. doi:10.1186/1471-2164-14-325 (https://doi.org/10.1186%2F1471-2164-14-325). PMC 3655826 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3655826). PMID 23663499 (https://pubmed.ncbi.nlm.nih.gov/23663499).

45. Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV, et al. (April 2003). "Hot L1s account for the bulk of retrotransposition in the human population" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC154336). *Proceedings of the National Academy of Sciences of the United States of America*. **100** (9): 5280–5285. Bibcode:2003PNAS..100.5280B (https://ui.adsabs.harvard.edu/abs/2003PNAS..100.5280B). doi:10.1073/pnas.0831042100 (https://doi.org/10.1073%2Fpnas.0831042100). PMC 154336 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC154336). PMID 12682288 (https://pubmed.ncbi.nlm.nih.gov/12682288).

46. Barton NH, Briggs DE, Eisen JA, Goldstein DB, Patel NH (2007). *Evolution*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press. ISBN 978-0-87969-684-9.

47. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, et al. (April 2014). "Defining functional DNA elements in the human genome" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4035993). *Proceedings of the National Academy of Sciences of the United States of America*. **111** (17): 6131–6138. Bibcode:2014PNAS..111.6131K (https://ui.adsabs.harvard.edu/abs/2014PNAS..111.6131K). doi:10.1073/pnas.1318948111 (https://doi.org/10.1073%2Fpnas.1318948111). PMC 4035993 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4035993). PMID 24753594 (https://pubmed.ncbi.nlm.nih.gov/24753594).

48. Linquist S, Doolittle WF, Palazzo AF (April 2020). "Getting clear about the F-word in genomics" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7153884). *PLOS Genetics*. **16** (4): e1008702. doi:10.1371/journal.pgen.1008702 (https://doi.org/10.1371%2Fjournal.pgen.1008702). PMC 7153884 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7153884). PMID 32236092 (https://pubmed.ncbi.nlm.nih.gov/32236092).

49. Doolittle WF (December 2018). "We simply cannot go on being so vague about 'function' " (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6299606). *Genome Biology*. **19** (1): 223. doi:10.1186/s13059-018-1600-4 (https://doi.org/10.1186%2Fs13059-018-1600-4). PMC 6299606 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6299606). PMID 30563541 (https://pubmed.ncbi.nlm.nih.gov/30563541).

50. Graur D (2017). "Rubbish DNA: the functionless fraction of the human genome.". *Evolution of the Human Genome I*. Evolutionary Studies. Tokyo: Springer. pp. 19–60. arXiv:1601.06047 (https://arxiv.org/abs/1601.06047). doi:10.1007/978-4-431-56603-8_2 (https://doi.org/10.1007%2F978-4-431-56603-8_2). ISBN 978-4-431-56603-8. S2CID 17826096 (https://api.semanticscholar.org/CorpusID:17826096).

51. Pena SD (2021). "An Overview of the Human Genome: Coding DNA and Non-Coding DNA" (https://books.google.com/books?id=cTYyEAAAQBAJ&dq=junk+DNA+controversy&pg=PA5). In Haddad LA (ed.). *Human Genome Structure, Function and Clinical Considerations*. Cham: Springer Nature. pp. 5–7. ISBN 978-3-03-073151-9.

52. Abascal F, Acosta R, Addleman NJ, Adrian J, et al. (30 July 2020). "Expanded Encyclopaedias of DNA elements in the Human and Mouse Genomes" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7410828). *Nature*. **583** (7818): 699–710. Bibcode:2020Natur.583..699E (https://ui.adsabs.harvard.edu/abs/2020Natur.583..699E). doi:10.1038/s41586-020-2493-4 (https://doi.org/10.1038%2Fs41586-020-2493-4). PMC 7410828 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7410828). PMID 32728249 (https://pubmed.ncbi.nlm.nih.gov/32728249). "Operationally, functional elements are defined as discrete, linearly ordered sequence features that specify molecular products (for example, protein-coding genes or noncoding RNAs) or biochemical activities with mechanistic roles in gene or genome regulation (for example, transcriptional promoters or enhancers)."

53. Graur D (July 2017). "An Upper Limit on the Functional Fraction of the Human Genome" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5570035). *Genome Biology and Evolution*. **9** (7): 1880–1885. doi:10.1093/gbe/evx121 (https://doi.org/10.1093%2Fgbe%2Fevx121). PMC 5570035 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5570035). PMID 28854598 (https://pubmed.ncbi.nlm.nih.gov/28854598).Lay summary in: Le Page M (17 July 2017). "At least 75 per cent of our DNA really is useless junk after all" (https://www.newscientist.com/article/2140926-at-least-75-per-cent-of-our-dna-really-is-useless-junk-after-all/). *NewScientist*.

54. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. (The ENCODE Project Consortium) (September 2012). "An integrated encyclopedia of DNA elements in the human genome" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3439153). *Nature*. **489** (7414): 57–74. Bibcode:2012Natur.489...57T (https://ui.adsabs.harvard.edu/abs/2012Natur.489...57T). doi:10.1038/nature11247 (https://doi.org/10.1038%2Fnature11247). PMC 3439153 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3439153). PMID 22955616 (https://pubmed.ncbi.nlm.nih.gov/22955616). "These data enabled us to assign biochemical functions for 80% of the genome, in particular outside of the well-studied protein-coding regions.".

55. Andolfatto P (October 2005). "Adaptive evolution of non-coding DNA in Drosophila". *Nature*. **437** (7062): 1149–52. Bibcode:2005Natur.437.1149A (https://ui.adsabs.harvard.edu/abs/2005Natur.437.1149A). doi:10.1038/nature04107 (https://doi.org/10.1038%2Fnature04107). PMID 16237443 (https://pubmed.ncbi.nlm.nih.gov/16237443). S2CID 191219 (https://api.semanticscholar.org/CorpusID:191219). Lay summary in: "UCSD Study Shows 'Junk' DNA Has Evolutionary Importance" (https://www.sciencedaily.com/releases/2005/10/051020090946.htm). *ScienceDaily*. Rockville, MD. 20 October 2005.

56. "International Human Genome Sequencing Consortium Publishes Sequence and Analysis of the Human Genome" (https://www.genome.gov/10002192). *National Human Genome Research Institute*. National Institutes of Health, U.S. Department of Health and Human Resources. 12 February 2001.

57. Pennisi E (February 2001). "The human genome". *Science*. **291** (5507): 1177–1180. doi:10.1126/science.291.5507.1177 (https://doi.org/10.1126%2Fscience.291.5507.1177). PMID 11233420 (https://pubmed.ncbi.nlm.nih.gov/11233420). S2CID 38355565 (https://api.semanticscholar.org/CorpusID:38355565).

58. International Human Genome Sequencing Consortium (October 2004). "Finishing the euchromatic sequence of the human genome" (https://doi.org/10.1038%2Fnature03001). *Nature*. **431** (7011): 931–945. Bibcode:2004Natur.431..931H (https://ui.adsabs.harvard.edu/abs/2004Natur.431..931H). doi:10.1038/nature03001 (https://doi.org/10.1038%2Fnature03001). PMID 15496913 (https://pubmed.ncbi.nlm.nih.gov/15496913).

59. Molteni M (19 November 2018). "Now You Can Sequence Your Whole Genome For Just $200" (https://www.wired.com/story/whole-genome-sequencing-cost-200-dollars/). *Wired*.

60. Saey TH (17 September 2018). "A recount of human genes ups the number to at least 46,831" (https://www.sciencenews.org/article/recount-human-genes-ups-number-least-46831). *Science News*.

61. Alles J, Fehlmann T, Fischer U, Backes C, Galata V, Minet M, et al. (April 2019). "An estimate of the total number of true human miRNAs" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6468295). *Nucleic Acids Research*. **47** (7): 3353–3364. doi:10.1093/nar/gkz097 (https://doi.org/10.1093%2Fnar%2Fgkz097). PMC 6468295 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6468295). PMID 30820533 (https://pubmed.ncbi.nlm.nih.gov/30820533).

62. Zhang S (28 November 2018). "300 Million Letters of DNA Are Missing From the Human Genome". *The Atlantic*.

63. Wade N (23 September 1999). "Number of Human Genes Is Put at 140,000, a Significant Gain" (https://archive.nytimes.com/www.nytimes.com/library/national/science/092399sci-human-genome.html). *The New York Times*.

64. Ezkurdia I, Juan D, Rodriguez JM, Frankish A, Diekhans M, Harrow J, et al. (November 2014). "Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4204768). *Human Molecular Genetics*. **23** (22): 5866–5878. doi:10.1093/hmg/ddu309 (https://doi.org/10.1093%2Fhmg%2Fddu309). PMC 4204768 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4204768). PMID 24939910 (https://pubmed.ncbi.nlm.nih.gov/24939910).

65. Wrighton K (February 2021). "Filling in the gaps telomere to telomere" (https://www.nature.com/articles/d42859-020-00117-1). *Nature Milestones: Genomic Sequencing*: S21.

66. "Scientists sequence the complete human genome for the first time" (https://edition.cnn.com/2022/03/31/health/first-complete-human-genome-sequence/index.html). *CNN*. 31 March 2022. Retrieved 1 April 2022.

67. Zhang S (28 November 2018). "300 Million Letters of DNA Are Missing From the Human Genome" (https://www.theatlantic.com/science/archive/2018/11/human-genome-300-million-missing-letters-dna/576481/). *The Atlantic*. Retrieved 16 August 2019.

68. Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, et al. (January 2015). "Resolving the complexity of the human genome using single-molecule sequencing" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4317254). *Nature*. **517** (7536): 608–611. Bibcode:2015Natur.517..608C (https://ui.adsabs.harvard.edu/abs/2015Natur.517..608C). doi:10.1038/nature13907 (https://doi.org/10.1038%2Fnature13907). PMC 4317254 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4317254). PMID 25383537 (https://pubmed.ncbi.nlm.nih.gov/25383537).

69. Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, et al. (September 2020). "Telomere-to-telomere assembly of a complete human X chromosome" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7484160). *Nature*. **585** (7823): 79–84. Bibcode:2020Natur.585...79M (https://ui.adsabs.harvard.edu/abs/2020Natur.585...79M). doi:10.1038/s41586-020-2547-7 (https://doi.org/10.1038%2Fs41586-020-2547-7). PMC 7484160 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7484160). PMID 32663838 (https://pubmed.ncbi.nlm.nih.gov/32663838).

70. Logsdon GA, Vollger MR, Hsieh P, Mao Y, Liskovykh MA, Koren S, et al. (May 2021). "The structure, function and evolution of a complete human chromosome 8" (https://www.ncbi.nl m.nih.gov/pmc/articles/PMC8099727). *Nature*. **593** (7857): 101–107. Bibcode:2021Natur.593..101L (https://ui.adsabs.harvard.edu/abs/2021Natur.593..101L). doi:10.1038/s41586-021-03420-7 (https://doi.org/10.1038%2Fs41586-021-03420-7). PMC 8099727 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8099727). PMID 33828295 (https://pubmed.ncbi.nlm.nih.gov/33828295).

71. "Genome List – Genome – NCBI" (https://www.ncbi.nlm.nih.gov/genome/browse/#!/eukaryot es/51/). *www.ncbi.nlm.nih.gov*. Retrieved 26 July 2021.

72. NCBI. "GRCh38 – hg38 – Genome – Assembly" (https://www.ncbi.nlm.nih.gov/assembly/GC F_000001405.26/). *ncbi.nlm.nih.gov*. Retrieved 15 March 2019.

73. "from Bill Clinton's 2000 State of the Union address" (https://web.archive.org/web/20170221 025546/https://clinton4.nara.gov/WH/SOTU00/sotu-text.html). Archived from the original (htt ps://clinton4.nara.gov/WH/SOTU00/sotu-text.html) on 21 February 2017. Retrieved 14 June 2007.

74. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. (November 2006). "Global variation in copy number in the human genome" (https://www.ncbi.nlm.nih.gov/pmc/ articles/PMC2669898). *Nature*. **444** (7118): 444–454. Bibcode:2006Natur.444..444R (http s://ui.adsabs.harvard.edu/abs/2006Natur.444..444R). doi:10.1038/nature05329 (https://doi.o rg/10.1038%2Fnature05329). PMC 2669898 (https://www.ncbi.nlm.nih.gov/pmc/articles/PM C2669898). PMID 17122850 (https://pubmed.ncbi.nlm.nih.gov/17122850).

75. "What's a Genome?" (http://www.genomenewsnetwork.org/resources/whats_a_genome/Ch p3_1.shtml). Genomenewsnetwork.org. 15 January 2003. Retrieved 31 May 2009.

76. "Fact Sheet: Genome Mapping: A Guide to the Genetic Highway We Call the Human Genome" (https://web.archive.org/web/20100719235548/http://www.ncbi.nlm.nih.gov/About/ primer/mapping.html). *National Center for Biotechnology Information*. U.S. National Library of Medicine, National Institutes of Health. 29 March 2004. Archived from the original (https:// www.ncbi.nlm.nih.gov/About/primer/mapping.html) on 19 July 2010. Retrieved 31 May 2009.

77. "About the Project" (https://web.archive.org/web/20080515164605/http://www.hapmap.org/a bouthapmap.html). *International HapMap Project*. Archived from the original (http://www.hap map.org/abouthapmap.html) on 15 May 2008. Retrieved 31 May 2009.

78. "2008 Release: Researchers Produce First Sequence Map of Large-Scale Structural Variation in the Human Genome" (http://www.genome.gov/27026113). genome.gov. Retrieved 31 May 2009.

79. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, et al. (May 2008). "Mapping and sequencing of structural variation from eight human genomes" (https://www.n cbi.nlm.nih.gov/pmc/articles/PMC2424287). *Nature*. **453** (7191): 56–64. Bibcode:2008Natur.453...56K (https://ui.adsabs.harvard.edu/abs/2008Natur.453...56K). doi:10.1038/nature06862 (https://doi.org/10.1038%2Fnature06862). PMC 2424287 (https:// www.ncbi.nlm.nih.gov/pmc/articles/PMC2424287). PMID 18451855 (https://pubmed.ncbi.nl m.nih.gov/18451855).

80. Abel HJ, Larson DE, Regier AA, Chiang C, Das I, Kanchi KL, et al. (July 2020). "Mapping and characterization of structural variation in 17,795 human genomes" (https://www.ncbi.nl m.nih.gov/pmc/articles/PMC7547914). *Nature*. **583** (7814): 83–89. Bibcode:2020Natur.583...83A (https://ui.adsabs.harvard.edu/abs/2020Natur.583...83A). doi:10.1038/s41586-020-2371-0 (https://doi.org/10.1038%2Fs41586-020-2371-0). PMC 7547914 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7547914). PMID 32460305 (https://pubmed.ncbi.nlm.nih.gov/32460305).

81. Gray IC, Campbell DA, Spurr NK (2000). "Single nucleotide polymorphisms as tools in human genetics" (https://doi.org/10.1093%2Fhmg%2F9.16.2403). *Human Molecular Genetics*. **9** (16): 2403–2408. doi:10.1093/hmg/9.16.2403 (https://doi.org/10.1093%2Fhm g%2F9.16.2403). PMID 11005795 (https://pubmed.ncbi.nlm.nih.gov/11005795).

82. Lai E (June 2001). "Application of SNP technologies in medicine: lessons learned and future challenges" (https://doi.org/10.1101%2Fgr.192301). *Genome Research*. **11** (6): 927–929. doi:10.1101/gr.192301 (https://doi.org/10.1101%2Fgr.192301). PMID 11381021 (https://pubmed.ncbi.nlm.nih.gov/11381021).

83. "Human Genome Project Completion: Frequently Asked Questions" (http://www.genome.gov/11006943). genome.gov. Retrieved 31 May 2009.

84. Singer E (4 September 2007). "Craig Venter's Genome" (https://web.archive.org/web/20110607104347/http://www.technologyreview.com/biomedicine/19328/?a=f). *MIT Technology Review*. Archived from the original (http://www.technologyreview.com/biomedicine/19328/?a=f) on 7 June 2011. Retrieved 25 May 2010.

85. Pushkarev D, Neff NF, Quake SR (September 2009). "Single-molecule sequencing of an individual human genome" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4117198). *Nature Biotechnology*. **27** (9): 847–850. doi:10.1038/nbt.1561 (https://doi.org/10.1038%2Fnbt.1561). PMC 4117198 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4117198). PMID 19668243 (https://pubmed.ncbi.nlm.nih.gov/19668243).

86. Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, Dewey FE, et al. (May 2010). "Clinical assessment incorporating a personal genome" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2937184). *Lancet*. **375** (9725): 1525–1535. doi:10.1016/S0140-6736(10)60452-7 (https://doi.org/10.1016%2FS0140-6736%2810%2960452-7). PMC 2937184 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2937184). PMID 20435227 (https://pubmed.ncbi.nlm.nih.gov/20435227).

87. Dewey FE, Chen R, Cordero SP, Ormond KE, Caleshu C, Karczewski KJ, et al. (September 2011). "Phased whole-genome genetic risk in a family quartet using a major allele reference sequence" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3174201). *PLOS Genetics*. **7** (9): e1002280. doi:10.1371/journal.pgen.1002280 (https://doi.org/10.1371%2Fjournal.pgen.1002280). PMC 3174201 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3174201). PMID 21935354 (https://pubmed.ncbi.nlm.nih.gov/21935354).

88. "Complete Genomics Adds 29 High-Coverage, Complete Human Genome Sequencing Datasets to Its Public Genomic Repository" (http://globenewswire.com/news-release/2011/04/06/443810/218032/en/Complete-Genomics-Adds-29-High-Coverage-Complete-Human-Genome-Sequencing-Datasets-to-Its-Public-Genomic-Repository.html?print=1) (Press release).

89. Sample I (17 February 2010). "Desmond Tutu's genome sequenced as part of genetic diversity study" (https://www.theguardian.com/science/2010/feb/17/desmond-tutu-genome-genetic-diversity). *The Guardian*.

90. Schuster SC, Miller W, Ratan A, Tomsho LP, Giardine B, Kasson LR, et al. (February 2010). "Complete Khoisan and Bantu genomes from southern Africa" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3890430). *Nature*. **463** (7283): 943–947. Bibcode:2010Natur.463..943S (https://ui.adsabs.harvard.edu/abs/2010Natur.463..943S). doi:10.1038/nature08795 (https://doi.org/10.1038%2Fnature08795). PMC 3890430 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3890430). PMID 20164927 (https://pubmed.ncbi.nlm.nih.gov/20164927).

91. Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, et al. (February 2010). "Ancient human genome sequence of an extinct Palaeo-Eskimo" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3951495). *Nature*. **463** (7282): 757–762. Bibcode:2010Natur.463..757R (https://ui.adsabs.harvard.edu/abs/2010Natur.463..757R). doi:10.1038/nature08835 (https://doi.org/10.1038%2Fnature08835). PMC 3951495 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3951495). PMID 20148029 (https://pubmed.ncbi.nlm.nih.gov/20148029).

92. Corpas M, Cariaso M, Coletta A, Weiss D, Harrison AP, Moran F, et al. (12 November 2013). "A Complete Public Domain Family Genomics Dataset". bioRxiv 10.1101/000216 (https://doi.org/10.1101%2F000216).

93. Corpas M (June 2013). "Crowdsourcing the corpasome" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3706263). *Source Code for Biology and Medicine*. **8** (1): 13. doi:10.1186/1751-0473-8-13 (https://doi.org/10.1186%2F1751-0473-8-13). PMC 3706263 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3706263). PMID 23799911 (https://pubmed.ncbi.nlm.nih.gov/23799911).

94. Mao Q, Ciotlos S, Zhang RY, Ball MP, Chin R, Carnevali P, et al. (October 2016). "The whole genome sequences and experimentally phased haplotypes of over 100 personal genomes" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5057367). *GigaScience*. **5** (1): 42. doi:10.1186/s13742-016-0148-z (https://doi.org/10.1186%2Fs13742-016-0148-z). PMC 5057367 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5057367). PMID 27724973 (https://pubmed.ncbi.nlm.nih.gov/27724973).

95. Cai B, Li B, Kiga N, Thusberg J, Bergquist T, Chen YC, et al. (September 2017). "Matching phenotypes to whole genomes: Lessons learned from four iterations of the personal genome project community challenges" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5645203). *Human Mutation*. **38** (9): 1266–1276. doi:10.1002/humu.23265 (https://doi.org/10.1002%2Fhumu.23265). PMC 5645203 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5645203). PMID 28544481 (https://pubmed.ncbi.nlm.nih.gov/28544481).

96. Gonzaga-Jauregui C, Lupski JR, Gibbs RA (2012). "Human genome sequencing in health and disease" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3656720). *Annual Review of Medicine*. **63**: 35–61. doi:10.1146/annurev-med-051010-162644 (https://doi.org/10.1146%2Fannurev-med-051010-162644). PMC 3656720 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3656720). PMID 22248320 (https://pubmed.ncbi.nlm.nih.gov/22248320).

97. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, et al. (November 2009). "Genetic diagnosis by whole exome capture and massively parallel DNA sequencing" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2768590). *Proceedings of the National Academy of Sciences of the United States of America*. **106** (45): 19096–19101. Bibcode:2009PNAS..10619096C (https://ui.adsabs.harvard.edu/abs/2009PNAS..10619096C). doi:10.1073/pnas.0910672106 (https://doi.org/10.1073%2Fpnas.0910672106). PMC 2768590 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2768590). PMID 19861545 (https://pubmed.ncbi.nlm.nih.gov/19861545).

98. Narasimhan VM, Xue Y, Tyler-Smith C (April 2016). "Human Knockout Carriers: Dead, Diseased, Healthy, or Improved?" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4826344). *Trends in Molecular Medicine*. **22** (4): 341–351. doi:10.1016/j.molmed.2016.02.006 (https://doi.org/10.1016%2Fj.molmed.2016.02.006). PMC 4826344 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4826344). PMID 26988438 (https://pubmed.ncbi.nlm.nih.gov/26988438).

99. Saleheen D, Natarajan P, Armean IM, Zhao W, Rasheed A, Khetarpal SA, et al. (April 2017). "Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5600291). *Nature*. **544** (7649): 235–239. Bibcode:2017Natur.544..235S (https://ui.adsabs.harvard.edu/abs/2017Natur.544..235S). doi:10.1038/nature22034 (https://doi.org/10.1038%2Fnature22034). PMC 5600291 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5600291). PMID 28406212 (https://pubmed.ncbi.nlm.nih.gov/28406212).

100. Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA (January 2002). "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC99152). *Nucleic Acids Research*. **30** (1): 52–55. doi:10.1093/nar/30.1.52 (https://doi.org/10.1093%2Fnar%2F30.1.52). PMC 99152 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC99152). PMID 11752252 (https://pubmed.ncbi.nlm.nih.gov/11752252).

101. Katsanis N (November 2016). "The continuum of causality in human genetic disorders" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5114767). *Genome Biology*. **17** (1): 233. doi:10.1186/s13059-016-1107-9 (https://doi.org/10.1186%2Fs13059-016-1107-9). PMC 5114767 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5114767). PMID 27855690 (https://pubmed.ncbi.nlm.nih.gov/27855690).

102. Alekseyev YO, Fazeli R, Yang S, Basran R, Maher T, Miller NS, et al. (2018). "A Next-Generation Sequencing Primer-How Does It Work and What Can It Do?" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5944141). *Academic Pathology*. **5**: 2374289518766521. doi:10.1177/2374289518766521 (https://doi.org/10.1177%2F2374289518766521). PMC 5944141 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5944141). PMID 29761157 (https://pubmed.ncbi.nlm.nih.gov/29761157).

103. Wong JC (2017). "Overview of the Clinical Utility of Next Generation Sequencing in Molecular Diagnoses of Human Genetic Disorders". In Wong LJ (ed.). *Next Generation Sequencing Based Clinical Molecular Diagnosis of Human Genetic Disorders*. Cham: Springer International Publishing. pp. 1–11. doi:10.1007/978-3-319-56418-0_1 (https://doi.org/10.1007%2F978-3-319-56418-0_1). ISBN 978-3-319-56416-6.

104. Fedick A, Zhang J (2017). "Next Generation of Carrier Screening". In Wong LJ (ed.). *Next Generation Sequencing Based Clinical Molecular Diagnosis of Human Genetic Disorders*. Cham: Springer International Publishing. pp. 339–354. doi:10.1007/978-3-319-56418-0_16 (https://doi.org/10.1007%2F978-3-319-56418-0_16). ISBN 978-3-319-56416-6.

105. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, et al. (December 2002). "Initial sequencing and comparative analysis of the mouse genome" (https://doi.org/10.1038%2Fnature01262). *Nature*. **420** (6915): 520–562. Bibcode:2002Natur.420..520W (https://ui.adsabs.harvard.edu/abs/2002Natur.420..520W). doi:10.1038/nature01262 (https://doi.org/10.1038%2Fnature01262). PMID 12466850 (https://pubmed.ncbi.nlm.nih.gov/12466850). "the proportion of small (50–100 bp) segments in the mammalian genome that is under (purifying) selection can be estimated to be about 5%. This proportion is much higher than can be explained by protein-coding sequences alone, implying that the genome contains many additional features (such as untranslated regions, regulatory elements, non-protein-coding genes, and chromosomal structural elements) under selection for biological function."

106. Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, et al. (June 2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2212820). *Nature*. **447** (7146): 799–816. Bibcode:2007Natur.447..799B (https://ui.adsabs.harvard.edu/abs/2007Natur.447..799B). doi:10.1038/nature05874 (https://doi.org/10.1038%2Fnature05874). PMC 2212820 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2212820). PMID 17571346 (https://pubmed.ncbi.nlm.nih.gov/17571346).

107. The Chimpanzee Sequencing Analysis Consortium (September 2005). "Initial sequence of the chimpanzee genome and comparison with the human genome" (https://doi.org/10.1038%2Fnature04072). *Nature*. **437** (7055): 69–87. Bibcode:2005Natur.437...69. (https://ui.adsabs.harvard.edu/abs/2005Natur.437...69.). doi:10.1038/nature04072 (https://doi.org/10.1038%2Fnature04072). PMID 16136131 (https://pubmed.ncbi.nlm.nih.gov/16136131). "We calculate the genome-wide nucleotide divergence between human and chimpanzee to be 1.23%, confirming recent results from more limited studies."

108. The Chimpanzee Sequencing Analysis Consortium (September 2005). "Initial sequence of the chimpanzee genome and comparison with the human genome" (https://doi.org/10.1038%2Fnature04072). *Nature*. **437** (7055): 69–87. Bibcode:2005Natur.437...69. (https://ui.adsabs.harvard.edu/abs/2005Natur.437...69.). doi:10.1038/nature04072 (https://doi.org/10.1038%2Fnature04072). PMID 16136131 (https://pubmed.ncbi.nlm.nih.gov/16136131). "we estimate that polymorphism accounts for 14–22% of the observed divergence rate and thus that the fixed divergence is ~1.06% or less"

109. Demuth JP, De Bie T, Stajich JE, Cristianini N, Hahn MW (2006). "The evolution of mammalian gene families" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1762380). *PLOS ONE*. **1** (1): e85. Bibcode:2006PLoSO...1...85D (https://ui.adsabs.harvard.edu/abs/2006PLoSO...1...85D). doi:10.1371/journal.pone.0000085 (https://doi.org/10.1371%2Fjournal.pone.0000085). PMC 1762380 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1762380). PMID 17183716 (https://pubmed.ncbi.nlm.nih.gov/17183716). "Our results imply that humans and chimpanzees differ by at least 6% (1,418 of 22,000 genes) in their complement of genes, which stands in stark contrast to the oft-cited 1.5% difference between orthologous nucleotide sequences"

110. The Chimpanzee Sequencing Analysis Consortium (September 2005). "Initial sequence of the chimpanzee genome and comparison with the human genome" (https://doi.org/10.1038%2Fnature04072). *Nature*. **437** (7055): 69–87. Bibcode:2005Natur.437...69. (https://ui.adsabs.harvard.edu/abs/2005Natur.437...69.). doi:10.1038/nature04072 (https://doi.org/10.1038%2Fnature04072). PMID 16136131 (https://pubmed.ncbi.nlm.nih.gov/16136131). "Human chromosome 2 resulted from a fusion of two ancestral chromosomes that remained separate in the chimpanzee lineage"
Olson MV, Varki A (January 2003). "Sequencing the chimpanzee genome: insights into human evolution and disease". *Nature Reviews Genetics*. **4** (1): 20–28. doi:10.1038/nrg981 (https://doi.org/10.1038%2Fnrg981). PMID 12509750 (https://pubmed.ncbi.nlm.nih.gov/12509750). S2CID 205486561 (https://api.semanticscholar.org/CorpusID:205486561). "Large-scale sequencing of the chimpanzee genome is now imminent."

111. Gilad Y, Wiebe V, Przeworski M, Lancet D, Pääbo S (January 2004). "Loss of olfactory receptor genes coincides with the acquisition of full trichromatic vision in primates" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC314465). *PLOS Biology*. **2** (1): E5. doi:10.1371/journal.pbio.0020005 (https://doi.org/10.1371%2Fjournal.pbio.0020005). PMC 314465 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC314465). PMID 14737185 (https://pubmed.ncbi.nlm.nih.gov/14737185). "Our findings suggest that the deterioration of the olfactory repertoire occurred concomitant with the acquisition of full trichromatic color vision in primates."

112. Zimmer C (21 September 2016). "How We Got Here: DNA Points to a Single Migration From Africa" (https://www.nytimes.com/2016/09/22/science/ancient-dna-human-history.html). *The New York Times*. Retrieved 22 September 2016.

113. Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E (January 2017). "Tracing the peopling of the world through genomics" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5772775). *Nature*. **541** (7637): 302–310. Bibcode:2017Natur.541..302N (https://ui.adsabs.harvard.edu/abs/2017Natur.541..302N). doi:10.1038/nature21347 (https://doi.org/10.1038%2Fnature21347). PMC 5772775 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5772775). PMID 28102248 (https://pubmed.ncbi.nlm.nih.gov/28102248).

114. Sykes B (9 October 2003). "Mitochondrial DNA and human history" (https://web.archive.org/web/20150907140051/http://genome.wellcome.ac.uk/doc_WTD020876.html). The Human Genome. Archived from the original (http://genome.wellcome.ac.uk/doc_WTD020876.html) on 7 September 2015. Retrieved 19 September 2006.

115. Misteli T (February 2007). "Beyond the sequence: cellular organization of genome function" (https://doi.org/10.1016%2Fj.cell.2007.01.028). *Cell*. **128** (4): 787–800. doi:10.1016/j.cell.2007.01.028 (https://doi.org/10.1016%2Fj.cell.2007.01.028). PMID 17320514 (https://pubmed.ncbi.nlm.nih.gov/17320514). S2CID 9064584 (https://api.semanticscholar.org/CorpusID:9064584).

116. Bernstein BE, Meissner A, Lander ES (February 2007). "The mammalian epigenome" (https://doi.org/10.1016%2Fj.cell.2007.01.033). *Cell*. **128** (4): 669–681. doi:10.1016/j.cell.2007.01.033 (https://doi.org/10.1016%2Fj.cell.2007.01.033). PMID 17320505 (https://pubmed.ncbi.nlm.nih.gov/17320505). S2CID 2722988 (https://api.semanticscholar.org/CorpusID:2722988).

117. Scheen AJ, Junien C (May–June 2012). "[Epigenetics, interface between environment and genes: role in complex diseases]". *Revue Médicale de Liège*. **67** (5–6): 250–257. PMID 22891475 (https://pubmed.ncbi.nlm.nih.gov/22891475).

External links

- Annotated (version 110) genome viewer of T2T-CHM13 v2.0 (https://www.ncbi.nlm.nih.gov/genome/gdv/browser/genome/?id=GCF_009914755.1)
- Complete human genome T2T-CHM13 v2.0 (no gaps) (https://www.ncbi.nlm.nih.gov/assembly/GCA_009914755.4)
- Ensembl (https://www.ensembl.org/Homo_sapiens/Location/Genome?db=core;g=ENSG00000139618;r=13:32315086-32400268) The Ensembl Genome Browser Project
- National Library of Medicine Genome Data Viewer (GDV) (https://www.ncbi.nlm.nih.gov/genome/gdv/?org=homo-sapiens&group=catarrhini)
- UCSC Genome Browser using T2T-CHM13 v2.0 (https://genome.ucsc.edu/h/GCA_009914755.4)
- Uniprot: per chromosome gene list (https://www.uniprot.org/proteomes/UP000005640)
- Human Genome Project (https://web.archive.org/web/20130102065343/http://www.ornl.gov/sci/techresources/Human_Genome/project/info.shtml)
- The National Human Genome Research Institute (https://www.genome.gov/)
- The National Office of Public Health Genomics (https://www.cdc.gov/genomics/default.htm)
- Simple Human Genome viewer (https://grrreggg.github.io/gene/)