Human Genome

From Wikipedia, the free encyclopedia

*For a non-technical introduction to the topic, see*[*Introduction to genetics*](https://en.wikipedia.org/wiki/Introduction_to_genetics)*.*

|  |  |
| --- | --- |
| **Genomic information** | |
| Graphical representation of the idealized human diploid [karyotype](https://en.wikipedia.org/wiki/Karyotype), showing the organization of the genome into chromosomes. This drawing shows both the female (XX) and male (XY) versions of the 23rd chromosome pair. Chromosomes are shown aligned at their [centromeres](https://en.wikipedia.org/wiki/Centromeres). The mitochondrial DNA is not shown. | |
| [**NCBI**](https://en.wikipedia.org/wiki/National_Center_for_Biotechnology_Information)**genome ID** | [51](http://www.ncbi.nlm.nih.gov/genome/51) |
| [**Ploidy**](https://en.wikipedia.org/wiki/Ploidy) | diploid |
| [**Genome size**](https://en.wikipedia.org/wiki/Genome_size) | 3,234.83 Mb (Mega-base pairs) per haploid genome  6,469.66 Mb total (diploid). |
| **Number of** [**chromosomes**](https://en.wikipedia.org/wiki/Chromosome) | 23 pairs |

The **human genome** is the complete set of [nucleic acid sequence](https://en.wikipedia.org/wiki/Nucleic_acid_sequence) for [humans](https://en.wikipedia.org/wiki/Humans) ([*Homo sapiens*](https://en.wikipedia.org/wiki/Homo_sapiens)), encoded as [DNA](https://en.wikipedia.org/wiki/DNA) within the 23 [chromosome](https://en.wikipedia.org/wiki/Chromosome) pairs in [cell nuclei](https://en.wikipedia.org/wiki/Cell_nuclei) and in a small DNA molecule found within individual [mitochondria](https://en.wikipedia.org/wiki/Mitochondrial_DNA). Human [genomes](https://en.wikipedia.org/wiki/Genome) include both protein-coding DNA genes and DNA. Haploid human genomes, which are contained in [germ cells](https://en.wikipedia.org/wiki/Germ_cells) (the [egg](https://en.wikipedia.org/wiki/Egg) and [sperm](https://en.wikipedia.org/wiki/Sperm) [gamete](https://en.wikipedia.org/wiki/Gamete) cells created in the [meiosis](https://en.wikipedia.org/wiki/Meiosis) phase of [sexual reproduction](https://en.wikipedia.org/wiki/Sexual_reproduction) before fertilization creates a [zygote](https://en.wikipedia.org/wiki/Zygote)) consist of three billion [DNA](https://en.wikipedia.org/wiki/DNA) [base pairs](https://en.wikipedia.org/wiki/Base_pair), while [diploid](https://en.wikipedia.org/wiki/Diploid) genomes (found in [somatic cells](https://en.wikipedia.org/wiki/Somatic_cells)) have twice the DNA content. While there are significant differences among the genomes of human individuals (on the order of 0.1%), these are considerably smaller than the differences between humans and their closest living relatives, the [chimpanzees](https://en.wikipedia.org/wiki/Chimpanzee) (approximately 4%) and [bonobos](https://en.wikipedia.org/wiki/Bonobo).

The [Human Genome Project](https://en.wikipedia.org/wiki/Human_Genome_Project) produced the first complete sequences of individual human genomes, with the first draft sequence and initial analysis being published on February 12, 2001. The human genome was the first of all vertebrates to be completely sequenced. As of 2012, thousands of human genomes have been completely sequenced, and many more have been mapped at lower levels of resolution. The resulting data are used worldwide in [biomedical science](https://en.wikipedia.org/wiki/Biomedical_science), [anthropology](https://en.wikipedia.org/wiki/Biological_anthropology), [forensics](https://en.wikipedia.org/wiki/Forensic_DNA) and other branches of science. There is a widely held expectation that genomic studies will lead to advances in the diagnosis and treatment of diseases, and to new insights in many fields of biology, including [human evolution](https://en.wikipedia.org/wiki/Human_evolution).

Although the sequence of the human genome has been (almost) completely determined by DNA sequencing, it is not yet fully understood. Most (though probably not all) genes have been identified by a combination of high throughput experimental and [bioinformatics](https://en.wikipedia.org/wiki/Bioinformatics) approaches, yet much work still needs to be done to further elucidate the biological functions of their protein and [RNA](https://en.wikipedia.org/wiki/RNA) products. Recent results suggest that most of the vast quantities of noncoding DNA within the genome have associated biochemical activities, including [regulation of gene expression](https://en.wikipedia.org/wiki/Regulation_of_gene_expression), organization of [chromosome architecture](https://en.wikipedia.org/wiki/Chromatin), and signals controlling [epigenetic inheritance](https://en.wikipedia.org/wiki/Epigenetic_inheritance).

There are an estimated 19,000-20,000 human protein-coding genes. The estimate of the number of human genes has been repeatedly revised down from initial predictions of 100,000 or more as genome sequence quality and gene finding methods have improved, and could continue to drop further. [Protein](https://en.wikipedia.org/wiki/Protein)-coding sequences account for only a very small fraction of the genome (approximately 1.5%), and the rest is associated with [non-coding RNA](https://en.wikipedia.org/wiki/Non-coding_RNA) molecules, [regulatory DNA sequences](https://en.wikipedia.org/wiki/Regulatory_sequences), [LINEs](https://en.wikipedia.org/wiki/Retrotransposon#LINEs), [SINEs](https://en.wikipedia.org/wiki/Retrotransposon#SINEs), [introns](https://en.wikipedia.org/wiki/Intron), and sequences for which as yet [no function](https://en.wikipedia.org/wiki/Noncoding_DNA) has been determined.

In June 2016, scientists formally announced [HGP-Write](https://en.wikipedia.org/wiki/Human_Genome_Project_-_Write), a plan to synthesize the human genome.

Molecular organization and gene content

*See also:*[*Lists of human genes by chromosome*](https://en.wikipedia.org/wiki/Lists_of_human_genes_by_chromosome)

The total length of the human genome is over 3 billion base pairs. The genome is organized into 22 paired chromosomes, plus the X chromosome (one in males, two in females) and, in males only, one Y chromosome. These are all large linear DNA molecules contained within the cell nucleus. The genome also includes the [mitochondrial DNA](https://en.wikipedia.org/wiki/Mitochondrial_DNA), a comparatively small circular molecule present in each [mitochondrion](https://en.wikipedia.org/wiki/Mitochondrion). Basic information about these molecules and their gene content, based on a [reference genome](https://en.wikipedia.org/wiki/Reference_genome) that does not represent the sequence of any specific individual, are provided in the following table. (Data source: [Ensemble genome browser release 87](http://useast.ensembl.org/Homo_sapiens/Location/Genome?r=Y:1-1000), December 2016 for most values; [Ensemble genome browser release 68](http://jul2012.archive.ensembl.org/Homo_sapiens/Location/Chromosome?r=1:1-1000000), July 2012 for miRNA, rRNA, snRNA, snoRNA.)

| **Chromosome** | **Length (mm)** | **Base pairs** | **Variations** | **Protein-coding genes** | **Pseudogenes** | **Total long ncRNA** | **Total small ncRNA** | **miRNA** | **rRNA** | **snRNA** | **snoRNA** | **Misc. ncRNA** | **Links** | **Centromere position (Mbp)** | **Cumulative (%)** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| [1](https://en.wikipedia.org/wiki/Chromosome_1_(human)) | 85 | 248956422 | 12151146 | 2058 | 1220 | 1200 | 496 | 134 | 66 | 221 | 145 | 192 | [EBI](http://useast.ensembl.org/Homo_sapiens/Location/Chromosome?r=1) | 125 | 7.9 |
| [2](https://en.wikipedia.org/wiki/Chromosome_2_(human)) | 83 | 242193529 | 12945965 | 1309 | 1023 | 1037 | 375 | 115 | 40 | 161 | 117 | 176 | [EBI](http://useast.ensembl.org/Homo_sapiens/Location/Chromosome?r=2) | 93.3 | 16.2 |
| [3](https://en.wikipedia.org/wiki/Chromosome_3_(human)) | 67 | 198295559 | 10638715 | 1078 | 763 | 711 | 298 | 99 | 29 | 138 | 87 | 134 | [EBI](http://useast.ensembl.org/Homo_sapiens/Location/Chromosome?r=3) | 91 | 23 |
| [4](https://en.wikipedia.org/wiki/Chromosome_4_(human)) | 65 | 190214555 | 10165685 | 752 | 727 | 657 | 228 | 92 | 24 | 120 | 56 | 104 | [EBI](http://useast.ensembl.org/Homo_sapiens/Location/Chromosome?r=4) | 50.4 | 29.6 |
| [5](https://en.wikipedia.org/wiki/Chromosome_5_(human)) | 62 | 181538259 | 9519995 | 876 | 721 | 844 | 235 | 83 | 25 | 106 | 61 | 119 | [EBI](http://useast.ensembl.org/Homo_sapiens/Location/Chromosome?r=5) | 48.4 | 35.8 |
| [6](https://en.wikipedia.org/wiki/Chromosome_6_(human)) | 58 | 170805979 | 9130476 | 1048 | 801 | 639 | 234 | 81 | 26 | 111 | 73 | 105 | [EBI](http://useast.ensembl.org/Homo_sapiens/Location/Chromosome?r=6) | 61 | 41.6 |
| [7](https://en.wikipedia.org/wiki/Chromosome_7_(human)) | 54 | 159345973 | 8613298 | 989 | 885 | 605 | 208 | 90 | 24 | 90 | 76 | 143 | [EBI](http://useast.ensembl.org/Homo_sapiens/Location/Chromosome?r=7) | 59.9 | 47.1 |
| [8](https://en.wikipedia.org/wiki/Chromosome_8_(human)) | 50 | 145138636 | 8221520 | 677 | 613 | 735 | 214 | 80 | 28 | 86 | 52 | 82 | [EBI](http://useast.ensembl.org/Homo_sapiens/Location/Chromosome?r=8) | 45.6 | 52 |
| [9](https://en.wikipedia.org/wiki/Chromosome_9_(human)) | 48 | 138394717 | 6590811 | 786 | 661 | 491 | 190 | 69 | 19 | 66 | 51 | 96 | [EBI](http://useast.ensembl.org/Homo_sapiens/Location/Chromosome?r=9) | 49 | 56.3 |
| [10](https://en.wikipedia.org/wiki/Chromosome_10_(human)) | 46 | 133797422 | 7223944 | 733 | 568 | 579 | 204 | 64 | 32 | 87 | 56 | 89 | [EBI](http://useast.ensembl.org/Homo_sapiens/Location/Chromosome?r=10) | 40.2 | 60.9 |
| [11](https://en.wikipedia.org/wiki/Chromosome_11_(human)) | 46 | 135086622 | 7535370 | 1298 | 821 | 710 | 233 | 63 | 24 | 74 | 76 | 97 | [EBI](http://useast.ensembl.org/Homo_sapiens/Location/Chromosome?r=11) | 53.7 | 65.4 |
| [12](https://en.wikipedia.org/wiki/Chromosome_12_(human)) | 45 | 133275309 | 7228129 | 1034 | 617 | 848 | 227 | 72 | 27 | 106 | 62 | 115 | [EBI](http://useast.ensembl.org/Homo_sapiens/Location/Chromosome?r=12) | 35.8 | 70 |
| [13](https://en.wikipedia.org/wiki/Chromosome_13_(human)) | 39 | 114364328 | 5082574 | 327 | 372 | 397 | 104 | 42 | 16 | 45 | 34 | 75 | [EBI](http://useast.ensembl.org/Homo_sapiens/Location/Chromosome?r=13) | 17.9 | 73.4 |
| [14](https://en.wikipedia.org/wiki/Chromosome_14_(human)) | 36 | 107043718 | 4865950 | 830 | 523 | 533 | 239 | 92 | 10 | 65 | 97 | 79 | [EBI](http://useast.ensembl.org/Homo_sapiens/Location/Chromosome?r=14) | 17.6 | 76.4 |
| [15](https://en.wikipedia.org/wiki/Chromosome_15_(human)) | 35 | 101991189 | 4515076 | 613 | 510 | 639 | 250 | 78 | 13 | 63 | 136 | 93 | [EBI](http://useast.ensembl.org/Homo_sapiens/Location/Chromosome?r=15) | 19 | 79.3 |
| [16](https://en.wikipedia.org/wiki/Chromosome_16_(human)) | 31 | 90338345 | 5101702 | 873 | 465 | 799 | 187 | 52 | 32 | 53 | 58 | 51 | [EBI](http://useast.ensembl.org/Homo_sapiens/Location/Chromosome?r=16) | 36.6 | 82 |
| [17](https://en.wikipedia.org/wiki/Chromosome_17_(human)) | 28 | 83257441 | 4614972 | 1197 | 531 | 834 | 235 | 61 | 15 | 80 | 71 | 99 | [EBI](http://useast.ensembl.org/Homo_sapiens/Location/Chromosome?r=17) | 24 | 84.8 |
| [18](https://en.wikipedia.org/wiki/Chromosome_18_(human)) | 27 | 80373285 | 4035966 | 270 | 247 | 453 | 109 | 32 | 13 | 51 | 36 | 41 | [EBI](http://useast.ensembl.org/Homo_sapiens/Location/Chromosome?r=18) | 17.2 | 87.4 |
| [19](https://en.wikipedia.org/wiki/Chromosome_19_(human)) | 20 | 58617616 | 3858269 | 1472 | 512 | 628 | 179 | 110 | 13 | 29 | 31 | 61 | [EBI](http://useast.ensembl.org/Homo_sapiens/Location/Chromosome?r=19) | 26.5 | 89.3 |
| [20](https://en.wikipedia.org/wiki/Chromosome_20_(human)) | 21 | 64444167 | 3439621 | 544 | 249 | 384 | 131 | 57 | 15 | 46 | 37 | 68 | [EBI](http://useast.ensembl.org/Homo_sapiens/Location/Chromosome?r=20) | 27.5 | 91.4 |
| [21](https://en.wikipedia.org/wiki/Chromosome_21_(human)) | 16 | 46709983 | 2049697 | 234 | 185 | 305 | 71 | 16 | 5 | 21 | 19 | 24 | [EBI](http://useast.ensembl.org/Homo_sapiens/Location/Chromosome?r=21) | 13.2 | 92.6 |
| [22](https://en.wikipedia.org/wiki/Chromosome_22_(human)) | 17 | 50818468 | 2135311 | 488 | 324 | 357 | 78 | 31 | 5 | 23 | 23 | 62 | [EBI](http://useast.ensembl.org/Homo_sapiens/Location/Chromosome?r=22) | 14.7 | 93.8 |
| [X](https://en.wikipedia.org/wiki/X_chromosome) | 53 | 156040895 | 5753881 | 842 | 874 | 271 | 258 | 128 | 22 | 85 | 64 | 100 | [EBI](http://useast.ensembl.org/Homo_sapiens/Location/Chromosome?r=X) | 60.6 | 99.1 |
| [Y](https://en.wikipedia.org/wiki/Y_chromosome) | 20 | 57227415 | 211643 | 71 | 388 | 71 | 30 | 15 | 7 | 17 | 3 | 8 | [EBI](http://useast.ensembl.org/Homo_sapiens/Location/Chromosome?r=Y) | 12.5 | 100 |
| [mtDNA](https://en.wikipedia.org/wiki/Mitochondrial_DNA) | 0.0054 | 16569 | 929 | 13 | 0 | 0 | 24 | 0 | 2 | 0 | 0 | 0 | [EBI](http://useast.ensembl.org/Homo_sapiens/Location/Chromosome?r=MT) | N/A | 100 |
| total |  | 3088286401 | 155630645 | 20412 | 14600 | 14727 | 5037 | 1756 | 532 | 1944 | 1521 | 2213 |  |  |  |

**Table 1** (above) summarizes the physical organization and gene content of the human [reference genome](https://en.wikipedia.org/wiki/Reference_genome), with links to the original analysis, as published in the [Ensemble](https://en.wikipedia.org/wiki/Ensembl) database at the [European Bioinformatics Institute](https://en.wikipedia.org/wiki/European_Bioinformatics_Institute) (EBI) and [Welcome Trust Sanger Institute](https://en.wikipedia.org/wiki/Wellcome_Trust_Sanger_Institute). Chromosome lengths were estimated by multiplying the number of base pairs by 0.34 nanometers, the distance between base pairs in the [DNA double helix](https://en.wikipedia.org/wiki/DNA#Properties). The number of proteins is based on the number of initial [precursor mRNA](https://en.wikipedia.org/wiki/Precursor_mRNA) transcripts, and does not include products of [alternative pre-mRNA splicing](https://en.wikipedia.org/wiki/Alternative_splicing), or modifications to protein structure that occur after [translation](https://en.wikipedia.org/wiki/Translation_(biology)).

[Variations](https://en.wikipedia.org/wiki/Human_genetic_variation) are unique DNA sequence differences that have been identified in the individual human genome sequences analyzed by Ensemble as of December, 2016. The number of identified variations is expected to increase as further [personal genomes](https://en.wikipedia.org/wiki/Personal_genomics) are sequenced and analyzed. In addition to the gene content shown in this table, a large number of non-expressed functional sequences have been identified throughout the human genome (see below). Links open windows to the reference chromosome sequences in the EBI genome browser.

Small [non-coding RNAs](https://en.wikipedia.org/wiki/Non-coding_RNA) are RNAs of as many as 200 bases that do not have protein-coding potential. These include: [microRNAs](https://en.wikipedia.org/wiki/MicroRNA), or miRNAs (post-transcriptional regulators of gene expression), [small nuclear RNAs](https://en.wikipedia.org/wiki/Small_nuclear_RNA), or snRNAs (the RNA components of [spliceosomes](https://en.wikipedia.org/wiki/Spliceosome)), and [small nucleolar RNAs](https://en.wikipedia.org/wiki/Small_nucleolar_RNA), or snoRNA (involved in guiding chemical modifications to other RNA molecules). [Long non-coding RNAs](https://en.wikipedia.org/wiki/Long_non-coding_RNA) are RNA molecules longer than 200 bases that do not have protein-coding potential. These include: [ribosomal RNAs](https://en.wikipedia.org/wiki/Ribosomal_RNA), or rRNAs (the RNA components of [ribosomes](https://en.wikipedia.org/wiki/Ribosome)), and a variety of other long RNAs that are involved in [regulation of gene expression](https://en.wikipedia.org/wiki/Regulation_of_gene_expression), [epigenetic](https://en.wikipedia.org/wiki/Epigenetic) modifications of DNA nucleotides and [histone](https://en.wikipedia.org/wiki/Histone) proteins, and regulation of the activity of protein-coding genes. Small discrepancies between total-small-ncRNA numbers and the numbers of specific types of small ncNRAs result from the former values being sourced from Ensemble release 87 and the latter from Ensemble release 68.

**Completeness of the human genome sequence**

Although the human genome has been completely sequenced for all practical purposes, there are still hundreds of gaps in the sequence. A recent study noted more than 160 [euchromatic](https://en.wikipedia.org/wiki/Euchromatin) gaps of which 50 gaps were closed. However, there are still numerous gaps in the [heterochromatic](https://en.wikipedia.org/wiki/Heterochromatin) parts of the genome which is much harder to [sequence](https://en.wikipedia.org/wiki/DNA_sequencing) due to numerous [repeats](https://en.wikipedia.org/wiki/Repetitive_DNA) and other intractable sequence features.

Coding vs. noncoding DNA

The content of the human genome is commonly divided into coding and noncoding DNA sequences. [Coding DNA](https://en.wikipedia.org/wiki/Coding_region) is defined as those sequences that can be transcribed into [mRNA](https://en.wikipedia.org/wiki/MRNA) and [translated](https://en.wikipedia.org/wiki/Translation_(biology)) into proteins during the human life cycle; these sequences occupy only a small fraction of the genome (<2%). [Noncoding DNA](https://en.wikipedia.org/wiki/Noncoding_DNA) is made up of all of those sequences (ca. 98% of the genome) that are not used to encode proteins.

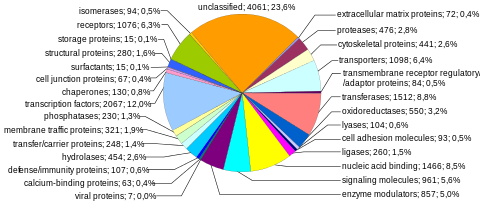
Some noncoding DNA contains genes for RNA molecules with important biological functions ([noncoding RNA](https://en.wikipedia.org/wiki/Noncoding_RNA), for example [ribosomal RNA](https://en.wikipedia.org/wiki/Ribosomal_RNA) and [transfer RNA](https://en.wikipedia.org/wiki/Transfer_RNA)). The exploration of the function and evolutionary origin of noncoding DNA is an important goal of contemporary genome research, including the [ENCODE](https://en.wikipedia.org/wiki/ENCODE) (Encyclopedia of DNA Elements) project, which aims to survey the entire human genome, using a variety of experimental tools whose results are indicative of molecular activity.

Because non-coding DNA greatly outnumbers coding DNA, the concept of the sequenced genome has become a more focused analytical concept than the classical concept of the DNA-coding gene.

Mutation rate

Mutation rate of human genome is a very important factor in calculating evolutionary time points. Researchers calculated the number of genetic variations between human and apes. Dividing that number by age of fossil of most recent common ancestor of humans and ape, researchers calculated the mutation rate. Recent studies using next generation sequencing technologies concluded a slow mutation rate which doesn't add up with human migration pattern time points and suggesting a new evolutionary time scale. 100,000 year old human fossils found in Israel have served to compound this new found uncertainty of the human migration timeline.

Coding sequences (protein-coding genes)



Human genes categorized by function of the transcribed proteins, given both as number of encoding genes and percentage of all 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. rearrangements and [alternative pre-mRNA splicing](https://en.wikipedia.org/wiki/Alternative_splicing)) can lead to the production of many more unique proteins than the number of protein-coding genes.

The complete modular protein-coding capacity of the genome is contained within the [exome](https://en.wikipedia.org/wiki/Exome), and consists of DNA sequences encoded by [exons](https://en.wikipedia.org/wiki/Exon) that can be translated into proteins. Because of its biological importance, and the fact that it constitutes less than 2% of the genome, sequencing of the exome was the first major milepost of the Human Genome Project.

**Number of protein-coding genes**. About 20,000 human proteins have been annotated in databases such as [Uniprot](https://en.wikipedia.org/wiki/Uniprot). Historically, estimates for the number of protein genes have varied widely, ranging up to 2,000,000 in the late 1960s, but several researchers pointed out in the early 1970s that the estimated [mutational load](https://en.wikipedia.org/wiki/Genetic_load) from deleterious mutations placed an upper limit of approximately 40,000 for the total number of functional loci (this includes protein-coding and functional non-coding genes).

The number of human protein-coding genes is not significantly larger than that of many less complex organisms, such as the roundworm and the [fruit fly](https://en.wikipedia.org/wiki/Drosophila_melanogaster). This difference may result from the extensive use of [alternative pre-mRNA splicing](https://en.wikipedia.org/wiki/Alternative_splicing) in humans, which provides the ability to build a very large number of modular proteins through the selective incorporation of exons.

**Protein-coding capacity per chromosome**. Protein-coding genes are distributed unevenly across the chromosomes, ranging from a few dozen to more than 2000, with an especially high [gene density](https://en.wikipedia.org/wiki/Gene_density) within chromosomes 19, 11, and 1 (Table 1). Each chromosome contains various gene-rich and gene-poor regions, which may be correlated with [chromosome bands](https://en.wikipedia.org/wiki/Cytogenetics) and [GC-content](https://en.wikipedia.org/wiki/GC-content). The significance of these nonrandom patterns of gene density is not well understood.

**Size of protein-coding genes**. The size of protein-coding genes within the human genome shows enormous variability (Table 2). For example, the gene for [histone H1a](https://en.wikipedia.org/wiki/Histone_H1) (HIST1HIA) is relatively small and simple, lacking introns and encoding mRNA sequences of 781 nt and a 215 amino acid protein (648 nt [open reading frame](https://en.wikipedia.org/wiki/Open_reading_frame)). [Dystrophin](https://en.wikipedia.org/wiki/Dystrophin) (DMD) is the largest protein-coding gene in the human reference genome, spanning a total of 2.2 MB, while [Titin](https://en.wikipedia.org/wiki/Titin) (TTN) has the longest coding sequence (114,414 bp), the largest number of [exons](https://en.wikipedia.org/wiki/Exons) (363), and the longest single exon (17,106 bp). Over the whole genome, the median size of an exon is 122 bp (mean = 145 bp), the median number of exons is 7 (mean = 8.8), and the median coding sequence encodes 367 amino acids (mean = 447 amino acids; Table 21 in ).

| **Protein** | **Chrom** | **Gene** | **Length** | **Exons** | **Exon length** | **Intron length** | **Alt splicing** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| [Breast cancer type 2 susceptibility protein](https://en.wikipedia.org/wiki/BRCA2) | 13 | [BRCA2](http://omim.org/entry/600185) | 83,736 | 27 | 11,386 | 72,350 | yes |
| [Cystic fibrosis transmembrane conductance regulator](https://en.wikipedia.org/wiki/Cystic_fibrosis_transmembrane_conductance_regulator) | 7 | [CFTR](http://omim.org/entry/602421) | 202,881 | 27 | 4,440 | 198,441 | yes |
| [Cytochrome b](https://en.wikipedia.org/wiki/Cytochrome_b) | MT | [MTCYB](http://omim.org/entry/516020) | 1,140 | 1 | 1,140 | 0 | no |
| [Dystrophin](https://en.wikipedia.org/wiki/Dystrophin) | X | [DMD](http://omim.org/entry/300377) | 2,220,381 | 79 | 10,500 | 2,209,881 | yes |
| [Glyceraldehyde-3-phosphate dehydrogenase](https://en.wikipedia.org/wiki/Glyceraldehyde-3-phosphate_dehydrogenase) | 12 | [GAPDH](http://omim.org/entry/138400) | 4,444 | 9 | 1,425 | 3,019 | yes |
| [Hemoglobin beta subunit](https://en.wikipedia.org/wiki/Hemoglobin) | 11 | [HBB](http://omim.org/entry/141900?search=hbb%20gene&highlight=gene%20hbb) | 1,605 | 3 | 626 | 979 | no |
| [Histone H1A](https://en.wikipedia.org/wiki/Histone_H1) | 6 | [HIST1H1A](http://omim.org/entry/142709?search=histone%20h1&highlight=h1%20histone) | 781 | 1 | 781 | 0 | no |
| [Titin](https://en.wikipedia.org/wiki/Titin) | 2 | [TTN](http://omim.org/entry/188840?search=ttn&highlight=ttn) | 281,434 | 364 | 104,301 | 177,133 | yes |

**Table 2. Examples of human protein-coding genes.** Chrom, chromosome. Alt splicing, alternative pre-mRNA splicing. (Data source: [Ensemble genome browser](http://useast.ensembl.org/Homo_sapiens/Location/Genome?r=6:133017695-133161157) release 68, July 2012)

Noncoding DNA (ncDNA)

*Main article:*[*Noncoding DNA*](https://en.wikipedia.org/wiki/Noncoding_DNA)

Noncoding DNA is defined as all of the DNA sequences within a genome that are not found within protein-coding exons, and so are never represented within the amino acid sequence of expressed proteins. By this definition, more than 98% of the human genomes is composed of ncDNA.

Numerous classes of noncoding DNA have been identified, including genes for noncoding RNA (e.g. tRNA and rRNA), pseudogenes, introns, untranslated regions of mRNA, regulatory DNA sequences, repetitive DNA sequences, and sequences related to mobile genetic elements.

Numerous sequences that are included within genes are also defined as noncoding DNA. These include genes for noncoding RNA (e.g. tRNA, rRNA), and untranslated components of protein-coding genes (e.g. introns, and 5' and 3' untranslated regions of mRNA).

Protein-coding sequences (specifically, coding [exons](https://en.wikipedia.org/wiki/Exon)) constitute less than 1.5% of the human genome. In addition, about 26% of the human genome is [introns](https://en.wikipedia.org/wiki/Introns). Aside from genes (exons and introns) and known regulatory sequences (8–20%), the human genome contains regions of noncoding DNA. The exact amount of noncoding DNA that plays a role in cell physiology has been hotly debated. Recent analysis by the ENCODE project indicates that 80% of the entire human genome is either transcribed, binds to regulatory proteins, or is associated with some other biochemical activity.

It however remains controversial whether all of this biochemical activity contributes to cell physiology, or whether a substantial portion of this is the result transcriptional and biochemical noise, which must be actively filtered out by the organism. Excluding protein-coding sequences, introns, and regulatory regions, much of the non-coding DNA is composed of: Many DNA sequences that do not play a role in expression have important biological functions. [Comparative genomics](https://en.wikipedia.org/wiki/Comparative_genomics) studies indicate that about 5% of the genome contains sequences of noncoding DNA that are highly [conserved](https://en.wikipedia.org/wiki/Conserved_sequence), sometimes on time-scales representing hundreds of millions of years, implying that these noncoding regions are under strong [evolutionary](https://en.wikipedia.org/wiki/Evolution) pressure and [positive selection](https://en.wikipedia.org/wiki/Natural_selection).

Many of these sequences regulate the structure of chromosomes by limiting the regions of [heterochromatin](https://en.wikipedia.org/wiki/Heterochromatin) formation and regulating structural features of the chromosomes, such as the [telomeres](https://en.wikipedia.org/wiki/Telomeres) and [centromeres](https://en.wikipedia.org/wiki/Centromeres). Other noncoding regions serve as [origins of DNA replication](https://en.wikipedia.org/wiki/DNA_replication#Origins). Finally several regions are transcribed into functional noncoding RNA that regulate the expression of protein-coding genes (for example), mRNA translation and stability (see [miRNA](https://en.wikipedia.org/wiki/MiRNA)), chromatin structure (including [histone](https://en.wikipedia.org/wiki/Histone) modifications, for example), DNA methylation (for example), DNA recombination (for example), and cross-regulate other noncoding RNAs (for example). It is also likely that many transcribed noncoding regions do not serve any role and that this transcription is the product of non-specific [RNA Polymerase](https://en.wikipedia.org/wiki/RNA_Polymerase) activity.

**Pseudogenes**

*Main article:*[*Pseudogene*](https://en.wikipedia.org/wiki/Pseudogene)

Pseudogenes are inactive copies of protein-coding genes, often generated by [gene duplication](https://en.wikipedia.org/wiki/Gene_duplication), that have become nonfunctional through the accumulation of inactivating mutations. **Table 1** shows that the number of pseudogenes in the human genome is on the order of 13,000, 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](https://en.wikipedia.org/wiki/Molecular_evolution).

For example, the [olfactory receptor](https://en.wikipedia.org/wiki/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.

**Genes for noncoding RNA (ncRNA)**

*Main article:*[*Noncoding RNA*](https://en.wikipedia.org/wiki/Noncoding_RNA)

Noncoding RNA molecules play many essential roles in cells, especially in the many reactions of [protein synthesis](https://en.wikipedia.org/wiki/Translation_(biology)) and [RNA processing](https://en.wikipedia.org/wiki/Post-transcriptional_modification). Noncoding RNA include [tRNA](https://en.wikipedia.org/wiki/TRNA), [ribosomal](https://en.wikipedia.org/wiki/Ribosome) RNA, [microRNA](https://en.wikipedia.org/wiki/MicroRNA), [snRNA](https://en.wikipedia.org/wiki/SnRNA) and other non-coding RNA genes including about 60,000 [long non coding RNAs](https://en.wikipedia.org/wiki/LncRNA) (lncRNAs). It should be noted that while the number of reported lncRNA genes continues to rise and the exact number in the human genome is yet to be defined, many of them are argued to be non-functional.

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.

**Introns and untranslated regions of mRNA**

In addition to the ncRNA molecules that are encoded by discrete genes, the initial transcripts of protein coding genes usually contain extensive noncoding sequences, in the form of [introns](https://en.wikipedia.org/wiki/Introns), [5'-untranslated regions](https://en.wikipedia.org/wiki/Five_prime_untranslated_region) (5'-UTR), and [3'-untranslated regions](https://en.wikipedia.org/wiki/Three_prime_untranslated_region) (3'-UTR). Within most protein-coding genes of the human genome, the length of intron sequences is 10- to 100-times the length of exon sequences (Table 2).

**Regulatory DNA sequences**

The human genome has many different [regulatory sequences](https://en.wikipedia.org/wiki/Regulatory_regions) which are crucial to controlling [gene expression](https://en.wikipedia.org/wiki/Gene_expression). Conservative estimates indicate that these sequences make up 8% of the genome, however extrapolations from the [ENCODE](https://en.wikipedia.org/wiki/ENCODE) project give that 20-40% 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](https://en.wikipedia.org/wiki/Enhancer_(genetics))).

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

Other genomes have been sequenced with the same intention of aiding conservation-guided methods, for example the [pufferfish](https://en.wikipedia.org/wiki/Pufferfish) genome. However, regulatory sequences disappear and re-evolve during evolution at a high rate.

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

**Repetitive DNA sequences**

[Repetitive DNA sequences](https://en.wikipedia.org/wiki/Repetitive_DNA) comprise approximately 50% of the human genome.

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..."). 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](https://en.wikipedia.org/wiki/Genealogical_DNA_testing) and [forensic DNA analysis](https://en.wikipedia.org/wiki/Forensic_DNA).

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](https://en.wikipedia.org/wiki/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](https://en.wikipedia.org/wiki/Minisatellite).

**Mobile genetic elements (transposons) and their relics**

[Transposable genetic elements](https://en.wikipedia.org/wiki/Transposable_element), 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, and can be inserted into intragenic and intergenic regions. One other lineage, LINE-1, has about 100 active copies per genome (the number varies between people). Together with non-functional relics of old transposons, they account for over half of total human DNA. Sometimes called "jumping genes", transposons have played a major role in sculpting the human genome. Some of these sequences represent [endogenous retroviruses](https://en.wikipedia.org/wiki/Endogenous_retroviruses), DNA copies of viral sequences that have become permanently integrated into the genome and are now passed on to succeeding generations.

Mobile elements within the human genome can be classified into [LTR retrotransposons](https://en.wikipedia.org/wiki/Retrotransposon#LTR_retrotransposons) (8.3% of total genome), [SINEs](https://en.wikipedia.org/wiki/Short_interspersed_nuclear_element) (13.1% of total genome) including [Alu elements](https://en.wikipedia.org/wiki/Alu_elements), [LINEs](https://en.wikipedia.org/wiki/Long_interspersed_nuclear_element) (20.4% of total genome), SVAs and [Class II DNA transposons](https://en.wikipedia.org/wiki/Transposable_element#Classification) (2.9% of total genome).

Genomic variation in humans

*Main articles:*[*Human genetic variation*](https://en.wikipedia.org/wiki/Human_genetic_variation)*and*[*Human genetic clustering*](https://en.wikipedia.org/wiki/Human_genetic_clustering)

**Human reference genome**

With the exception of identical twins, all humans show significant variation in genomic DNA sequences. The human [reference genome](https://en.wikipedia.org/wiki/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 and ambiguities.
* 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.

**Measuring human genetic variation**

Most studies of human genetic variation have focused on [single-nucleotide polymorphisms](https://en.wikipedia.org/wiki/Single-nucleotide_polymorphism) (SNPs), which are substitutions in individual bases along a chromosome. Most analysts estimate that SNPs occur 1 in 1000 base pairs, on average, in the [euchromatic](https://en.wikipedia.org/wiki/Euchromatin) human genome, although they do not occur at a uniform density. Thus follows the popular statement that "we are all, regardless of [race](https://en.wikipedia.org/wiki/Race_(classification_of_human_beings)), genetically 99.9% the same",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](https://en.wikipedia.org/wiki/Copy_number_variation). A large-scale collaborative effort to catalog SNP variations in the human genome is being undertaken by the [International HapMap Project](https://en.wikipedia.org/wiki/International_HapMap_Project).

The genomic loci and length of certain types of small [repetitive sequences](https://en.wikipedia.org/wiki/Repeated_sequence_(DNA)) are highly variable from person to person, which is the basis of [DNA fingerprinting](https://en.wikipedia.org/wiki/DNA_fingerprinting) and [DNA paternity testing](https://en.wikipedia.org/wiki/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](https://en.wikipedia.org/wiki/Phenotype) effect results from typical variation in repeats or heterochromatin.

Most gross genomic mutations in [gamete](https://en.wikipedia.org/wiki/Gamete) germ cells probably result in inviable embryos; however, a number of human diseases are related to large-scale genomic abnormalities. [Down syndrome](https://en.wikipedia.org/wiki/Down_syndrome), [Turner Syndrome](https://en.wikipedia.org/wiki/Turner_Syndrome), and a number of other diseases result from [nondisjunction](https://en.wikipedia.org/wiki/Nondisjunction) of entire chromosomes [Cancer](https://en.wikipedia.org/wiki/Cancer) cells frequently have [aneuploidy](https://en.wikipedia.org/wiki/Aneuploidy) of chromosomes and chromosome arms, although a [cause and effect](https://en.wikipedia.org/wiki/Causality) relationship between aneuploidy and cancer has not been established.

**Mapping human genomic variation**

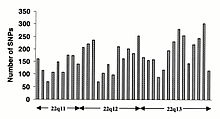
Whereas a genome sequence lists the order of every DNA base in a genome, a genome map identifies the landmarks. A genome map is less detailed than a genome sequence and aids in navigating around the genome.

An example of a variation map is the HapMap being developed by the [International HapMap Project](https://en.wikipedia.org/wiki/International_HapMap_Project). The HapMap is a [haplotype](https://en.wikipedia.org/wiki/Haplotype) map of the human genome, "which will describe the common patterns of human DNA sequence variation." 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*](https://en.wikipedia.org/wiki/Nature_(journal)) in May 2008. 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](https://en.wikipedia.org/wiki/Copy_number_variation) individuals have of a particular gene, deletions, translocations and inversions.

**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](https://en.wikipedia.org/wiki/Single-nucleotide_polymorphism) 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](https://en.wikipedia.org/wiki/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](https://en.wikipedia.org/wiki/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.



TSC SNP distribution along the long arm of chromosome 22 (from<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.

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.

**Personal genomes**

*See also:*[*Personal genomics*](https://en.wikipedia.org/wiki/Personal_genomics)

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

The first personal genome sequence to be determined was that of [Craig Venter](https://en.wikipedia.org/wiki/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. However, early in the Venter-led [Celera Genomics](https://en.wikipedia.org/wiki/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](https://en.wikipedia.org/wiki/Chromosomes), rather than a haploid sequence originally reported, allowed the release of the first personal genome. In April 2008, that of [James Watson](https://en.wikipedia.org/wiki/James_Watson) was also completed. Since then hundreds of personal genome sequences have been released, including those of [Desmond Tutu](https://en.wikipedia.org/wiki/Desmond_Tutu), and of a [Paleo-Eskimo](https://en.wikipedia.org/wiki/Paleo-Eskimo). In November 2013, a Spanish family made their personal genomics data publicly available under a [Creative Commons public domain license](https://en.wikipedia.org/wiki/CC0#Zero_.2F_Public_domain). The work was led by Manuel Corpas and the data obtained by [direct-to-consumer genetic testing](https://en.wikipedia.org/wiki/Genetic_testing#Direct-to-Consumer_genetic_testing) with [23andMe](https://en.wikipedia.org/wiki/23andMe) and the [Beijing Genomics Institute](https://en.wikipedia.org/wiki/Beijing_Genomics_Institute)). This is believed to be the first such public genomics dataset for a whole family.

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. [Exome sequencing](https://en.wikipedia.org/wiki/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.

Human genetic disorders

*For more details on this topic, see*[*Genetic disorder*](https://en.wikipedia.org/wiki/Genetic_disorder)*.*

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](https://en.wikipedia.org/wiki/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.

Disease-causing mutations in specific genes are usually severe in terms of gene function, and are fortunately 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.

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](https://en.wikipedia.org/wiki/Geneticist)-physician trained in clinical/medical genetics. The results of the Human Genome Project are likely to provide increased availability of [genetic testing](https://en.wikipedia.org/wiki/Genetic_testing) for gene-related disorders, and eventually improved treatment. Parents can be screened for hereditary conditions and [counselled](https://en.wikipedia.org/wiki/Genetic_counseling) on the consequences, the probability it will be inherited, and how to avoid or ameliorate it in their offspring.

As noted above, 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](https://en.wikipedia.org/wiki/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. The categorized table below provides the prevalence as well as the genes or chromosomes associated with some human genetic disorders.

|  |  |  |
| --- | --- | --- |
| **Disorder** | **Prevalence** | **Chromosome or gene involved** |
| **Chromosomal conditions** | | |
| [Down syndrome](https://en.wikipedia.org/wiki/Down_syndrome) | 1:600 | Chromosome 21 |
| [Klinefelter syndrome](https://en.wikipedia.org/wiki/Klinefelter_syndrome) | 1:500–1000 males | Additional X chromosome |
| [Turner syndrome](https://en.wikipedia.org/wiki/Turner_syndrome) | 1:2000 females | Loss of X chromosome |
| [Sickle cell anemia](https://en.wikipedia.org/wiki/Sickle_cell_anemia) | 1 in 50 births in parts of Africa; rarer elsewhere | [β-globin (on chromosome 11)](https://en.wikipedia.org/wiki/HBB) |
| **Cancers** | | |
| [Breast](https://en.wikipedia.org/wiki/Breast_cancer)/[Ovarian cancer](https://en.wikipedia.org/wiki/Ovarian_cancer) (susceptibility) | ~5% of cases of these cancer types | BRCA1, BRCA2 |
| [FAP](https://en.wikipedia.org/wiki/Familial_adenomatous_polyposis) (hereditary nonpolyposis coli) | 1:3500 | [APC](https://en.wikipedia.org/wiki/Adenomatous_polyposis_coli) |
| [Lynch syndrome](https://en.wikipedia.org/wiki/Lynch_syndrome) | 5–10% of all cases of bowel cancer | MLH1, MSH2, MSH6, PMS2 |
| **Neurological conditions** | | |
| [Huntington disease](https://en.wikipedia.org/wiki/Huntington_disease) | 1:20000 | Huntingtin |
| [Alzheimer disease](https://en.wikipedia.org/wiki/Alzheimer_disease) ‐ early onset | 1:2500 | PS1, PS2, APP |
| **Other conditions** | | |
| [Cystic fibrosis](https://en.wikipedia.org/wiki/Cystic_fibrosis) | 1:2500 | CFTR |
| [Duchenne muscular dystrophy](https://en.wikipedia.org/wiki/Duchenne_muscular_dystrophy) | 1:3500 boys | Dystrophin |

Evolution

[**Human timeline**](https://en.wikipedia.org/wiki/Timeline_of_human_evolution)

-10 — 9 — 8 — 7 — 6 — 5 — 4 —3 — 2 — 1 — 0

[**Human-like  
apes**](https://en.wikipedia.org/wiki/Timeline_of_human_evolution#Hominidae)

[*Nakalipithecus*](https://en.wikipedia.org/wiki/Nakalipithecus)

[*Ouranopithecus*](https://en.wikipedia.org/wiki/Ouranopithecus)

[*Sahelanthropus*](https://en.wikipedia.org/wiki/Sahelanthropus)

[*Orrorin*](https://en.wikipedia.org/wiki/Orrorin)

[*Ardipithecus*](https://en.wikipedia.org/wiki/Ardipithecus)

[*Australopithecus*](https://en.wikipedia.org/wiki/Australopithecus)

[***Homo habilis***](https://en.wikipedia.org/wiki/Homo_habilis)

[***Homo erectus***](https://en.wikipedia.org/wiki/Homo_erectus)

[***Neanderthal***](https://en.wikipedia.org/wiki/Neanderthal)

[***Homo sapiens***](https://en.wikipedia.org/wiki/Anatomically_modern_human)

[Earlier apes](https://en.wikipedia.org/wiki/Timeline_of_human_evolution#Hominidae)←

[Possibly bipedal](https://en.wikipedia.org/wiki/Sahelanthropus#Fossils)←

[Earliest bipedal](https://en.wikipedia.org/wiki/Human_skeletal_changes_due_to_bipedalism)←

[Earliest stone tools](https://en.wikipedia.org/wiki/Stone_Age#Beginning_of_the_Stone_Age)←

[Earliest exit from Africa](https://en.wikipedia.org/wiki/Out_of_Africa_I)←

[Earliest fire use](https://en.wikipedia.org/wiki/Control_of_fire_by_early_humans#Lower_Paleolithic_Evidence)←

[Earliest cooking](https://en.wikipedia.org/wiki/Cooking#History)←

[Earliest clothes](https://en.wikipedia.org/wiki/Clothing#Origin_of_clothing)←

[Modern humans](https://en.wikipedia.org/wiki/Behavioral_modernity)← [Pleistocene](https://en.wikipedia.org/wiki/Pleistocene)← [Pliocene](https://en.wikipedia.org/wiki/Pliocene)← [Miocene](https://en.wikipedia.org/wiki/Miocene)← [Hominids](https://en.wikipedia.org/wiki/Hominidae)←

[Axis scale](https://en.wikipedia.org/wiki/Cartesian_coordinate_system#Two_dimensions): [millions of years](https://en.wikipedia.org/wiki/Geologic_time_scale).  
Also see: [*Life timeline*](https://en.wikipedia.org/wiki/Template:Life_timeline) and [*Nature timeline*](https://en.wikipedia.org/wiki/Template:Nature_timeline)

*See also:*[*Human evolution*](https://en.wikipedia.org/wiki/Human_evolution)*and*[*Chimpanzee Genome Project*](https://en.wikipedia.org/wiki/Chimpanzee_Genome_Project)

[Comparative genomics](https://en.wikipedia.org/wiki/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. The published [chimpanzee](https://en.wikipedia.org/wiki/Chimpanzee) genome differs from that of the human genome by 1.23% in direct sequence comparisons. 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. This nucleotide by nucleotide difference is dwarfed, however, by the portion of each genome that is not shared, including around 6% of functional genes that are unique to either humans or chimps.

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 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](https://en.wikipedia.org/wiki/Ortholog) by only two [amino acid](https://en.wikipedia.org/wiki/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](https://en.wikipedia.org/wiki/Chromosome_2_(human)), which is equivalent to a fusion product of chimpanzee chromosomes 12 and 13. (later renamed to chromosomes 2A and 2B, respectively).

Humans have undergone an extraordinary loss of [olfactory receptor](https://en.wikipedia.org/wiki/Olfactory_receptor) genes during our recent evolution, which explains our relatively crude sense of [smell](https://en.wikipedia.org/wiki/Olfaction) compared to most other mammals. Evolutionary evidence suggests that the emergence of [color vision](https://en.wikipedia.org/wiki/Color_vision) in humans and several other [primate](https://en.wikipedia.org/wiki/Primate) species has diminished the need for the sense of smell.

In September 2016, scientists reported that, based on human DNA genetic studies, all [non-Africans in the world today](https://en.wikipedia.org/wiki/Behavioral_modernity) can be traced to a [single population](https://en.wikipedia.org/wiki/Anatomically_modern_human) that [exited Africa](https://en.wikipedia.org/wiki/Recent_African_origin_of_modern_humans) between 50,000 and 80,000 years ago.

Mitochondrial DNA

The human [mitochondrial DNA](https://en.wikipedia.org/wiki/Mitochondrial_DNA) is of tremendous interest to geneticists, since it undoubtedly plays a role in [mitochondrial disease](https://en.wikipedia.org/wiki/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](https://en.wikipedia.org/wiki/Mitochondrial_Eve)).

Due to the lack of a system for checking for copying errors, 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](https://en.wikipedia.org/wiki/Indigenous_peoples_of_the_Americas) from [Siberia](https://en.wikipedia.org/wiki/Siberia) or [Polynesians](https://en.wikipedia.org/wiki/Polynesia) from southeastern [Asia](https://en.wikipedia.org/wiki/Asia). It has also been used to show that there is no trace of [Neanderthal](https://en.wikipedia.org/wiki/Neanderthal) DNA in the European gene mixture inherited through purely maternal lineage. 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 your 32 ancestors contributed to your mtDNA, so if one of these 32 was pure Neanderthal you would expect that ~3% of your autosomal DNA would be of Neanderthal origin, yet you would have a ~97% chance to have no trace of Neanderthal mtDNA).

Epigenome

*See also:*[*Epigenetics*](https://en.wikipedia.org/wiki/Epigenetics)

Epigenetics describes a variety of features of the human genome that transcend its primary DNA sequence, such as [chromatin](https://en.wikipedia.org/wiki/Chromatin) packaging, [histone](https://en.wikipedia.org/wiki/Histone) modifications and [DNA methylation](https://en.wikipedia.org/wiki/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.

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.

See also

* [Genetics](https://en.wikipedia.org/wiki/Genetics)
* [Genomics](https://en.wikipedia.org/wiki/Genomics)
* [Genographic Project](https://en.wikipedia.org/wiki/Genographic_Project)
* [Genomic organization](https://en.wikipedia.org/wiki/Genomic_organization)
* [Low copy repeats](https://en.wikipedia.org/wiki/Low_copy_repeats)
* [Noncoding DNA](https://en.wikipedia.org/wiki/Noncoding_DNA)
* [Whole genome sequencing](https://en.wikipedia.org/wiki/Whole_genome_sequencing)
* [Universal Declaration on the Human Genome and Human Rights](https://en.wikipedia.org/wiki/Universal_Declaration_on_the_Human_Genome_and_Human_Rights)

This page was last modified on 8 April 2017, at 17:36.