

# HUMAN GENOME

Dec 29 2024  
Phuc Loi Luu, PhD  
Email: [Luu.p.loi@googlemail.com](mailto:Luu.p.loi@googlemail.com)  
Zalo: 0901802182

# Content

1. Human Genome Build
2. Introduction to GENCODE
3. Introduction to UCSC Genome Browser

# Human Genome Build

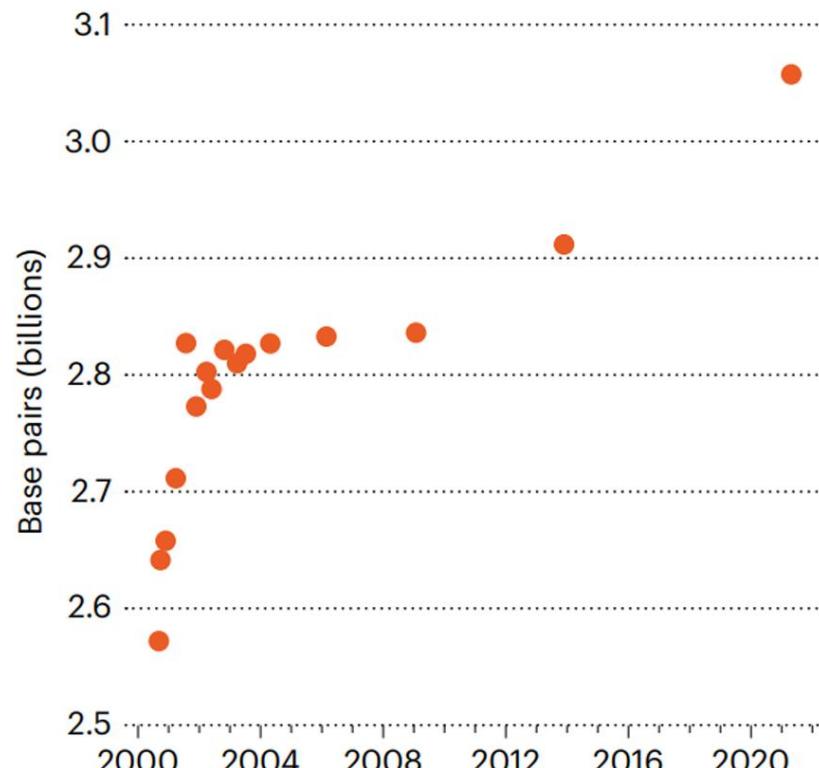
Human	hs1	Jan. 2022	T2T Consortium CHM13v2.0	Available
	hg38	Dec. 2013	Genome Reference Consortium GRCh38	Available
	hg19	Feb. 2009	Genome Reference Consortium GRCh37	Available
	hg18	Mar. 2006	NCBI Build 36.1	Available
	hg17	May 2004	NCBI Build 35	Available
	hg16	Jul. 2003	NCBI Build 34	Available
	hg15	Apr. 2003	NCBI Build 33	Archived
	hg13	Nov. 2002	NCBI Build 31	Archived
	hg12	Jun. 2002	NCBI Build 30	Archived
	hg11	Apr. 2002	NCBI Build 29	Archived (data only)
	hg10	Dec. 2001	NCBI Build 28	Archived (data only)
	hg8	Aug. 2001	UCSC-assembled	Archived (data only)
	hg7	Apr. 2001	UCSC-assembled	Archived (data only)
	hg6	Dec. 2000	UCSC-assembled	Archived (data only)
	hg5	Oct. 2000	UCSC-assembled	Archived (data only)
	hg4	Sep. 2000	UCSC-assembled	Archived (data only)
	hg3	Jul. 2000	UCSC-assembled	Archived (data only)
	hg2	Jun. 2000	UCSC-assembled	Archived (data only)
	hg1	May 2000	UCSC-assembled	Archived (data only)

## A COMPLETE HUMAN GENOME IS CLOSE: HOW THE GAPS WERE FILLED

Researchers added 200 million DNA base pairs and 115 genes – but they've yet to finish the Y chromosome.

## COMPLETING THE HUMAN GENOME

Researchers have been filling in incompletely sequenced parts of the human reference genome for 20 years, and have now almost finished it, with 3.05 billion DNA base pairs.



0.3% of sequence might still have errors. Includes X but not Y chromosome. Count excludes mitochondrial DNA.

SOURCE: ADAM PHILLIPY

# Fasta Format

## Single Sequence Fasta Format

```
>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken  
MADQLTEEQIAEFKEAFSLFDKDGDTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTID  
FPEFLTMMARKMKDTDSEEEIREAFRVFDKGNGYISAAELRHVMTNLGEKLTDEEVDEMIREA  
DIDGDGQVNYEEFVQMMTAK*
```

## Multiple Sequence Fasta Format

```
>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken  
MADQLTEEQIAEFKEAFSLFDKDGDTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTID  
FPEFLTMMARKMKDTDSEEEIREAFRVFDKGNGYISAAELRHVMTNLGEKLTDEEVDEMIREA  
DIDGDGQVNYEEFVQMMTAK*
```

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]  
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPIGTNLV  
EWIWGGFSVDKATLNRRFAFHFILEPFTMVALAGVHLTFHETGSNNPLGLTSSDKIPFHPYYTIKDFLG  
LILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL  
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTWIGSQVEYPYTIIGQMASILYFSIILAFLPIAGX  
IENY
```

# Where to download the reference genomes

All the files in this directory are freely available for public use.			
Name	Last modified	Size	Description
<a href="#">Parent_Directory</a>	-	-	
<a href="#">analysisSet/</a>	2020-03-13 17:39	-	
<a href="#">chromAgp.tar.gz</a>	2009-03-20 09:02	538K	
<a href="#">chromFa.tar.gz</a>	2009-03-20 09:21	905M	
<a href="#">chromFaMasked.tar.gz</a>	2009-03-20 09:30	477M	
<a href="#">chromOut.tar.gz</a>	2009-03-20 09:03	163M	
<a href="#">chromTrf.tar.gz</a>	2009-03-20 09:30	7.6M	
<a href="#">est.fa.gz</a>	2019-10-14 15:08	1.5G	
<a href="#">est.fa.gz.md5</a>	2019-10-14 15:08	44	
<a href="#">genes/</a>	2024-07-31 12:33	-	
<a href="#">hg19.2bit</a>	2009-03-08 15:29	778M	
<a href="#">hg19.agp.gz</a>	2009-05-06 15:22	532K	
<a href="#">hg19.chrom.sizes</a>	2009-03-08 14:56	1.9K	
<a href="#">hg19.chromAlias.bb</a>	2023-02-23 12:59	38K	
<a href="#">hg19.chromAlias.txt</a>	2023-02-22 11:47	4.8K	
<a href="#">hg19.fa.align.gz</a>	2009-03-08 22:08	2.2G	
<a href="#">hg19.fa.gz</a>	2018-08-21 12:56	905M	
<a href="#">hg19.fa.masked.gz</a>	2018-09-12 10:33	477M	
<a href="#">hg19.fa.out.gz</a>	2009-03-08 21:55	163M	
<a href="#">hg19.gc5Base.wib</a>	2019-01-17 14:49	571M	
<a href="#">hg19.gc5Base.wig.gz</a>	2019-01-17 14:49	11M	
<a href="#">hg19.gc5Base.wigVarStep.gz</a>	2018-09-28 15:21	1.5G	
<a href="#">hg19.trf.bed.gz</a>	2009-03-08 15:00	7.6M	
<a href="#">initial/</a>	2024-02-29 14:20	-	
<a href="#">latest/</a>	2020-03-25 13:33	-	
<a href="#">md5sum.txt</a>	2019-01-17 15:55	967	
<a href="#">mrna.fa.gz</a>	2019-10-14 14:50	370M	
<a href="#">mrna.fa.gz.md5</a>	2019-10-14 14:50	45	
<a href="#">p13_plusMT/</a>	2024-07-23 16:43	-	
<a href="#">refMrna.fa.gz</a>	2019-10-14 15:08	80M	
<a href="#">refMrna.fa.gz.md5</a>	2019-10-14 15:08	48	
<a href="#">upstream1000.fa.gz</a>	2019-10-14 15:09	9.7M	
<a href="#">upstream1000.fa.gz.md5</a>	2019-10-14 15:09	53	
<a href="#">upstream2000.fa.gz</a>	2019-10-14 15:10	18M	
<a href="#">upstream2000.fa.gz.md5</a>	2019-10-14 15:10	53	
<a href="#">upstream5000.fa.gz</a>	2019-10-14 15:10	47M	
<a href="#">upstream5000.fa.gz.md5</a>	2019-10-14 15:10	53	
<a href="#">xenoMrna.fa.gz</a>	2019-10-14 15:00	6.4G	
<a href="#">xenoMrna.fa.gz.md5</a>	2019-10-14 15:00	49	
<a href="#">xenoRefMrna.fa.gz</a>	2019-10-14 15:08	250M	
<a href="#">xenoRefMrna.fa.gz.md5</a>	2019-10-14 15:08	52	

<https://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/>

restricted permission concerning the use, copying, or distribution  
the information contained in GenBank.

Name	Last modified	Size	Description
<a href="#">Parent_Directory</a>	-	-	
<a href="#">analysisSet/</a>	2023-01-06 17:06	-	
<a href="#">est.fa.gz</a>	2019-10-14 13:54	1.5G	
<a href="#">est.fa.gz.md5</a>	2019-10-14 13:54	44	
<a href="#">genes/</a>	2024-12-23 12:50	-	
<a href="#">hg38.2bit</a>	2015-04-30 16:16	797M	
<a href="#">hg38.agp.gz</a>	2014-01-15 20:55	842K	
<a href="#">hg38.chrom.sizes</a>	2013-12-24 21:06	11K	
<a href="#">hg38.chromAlias.bb</a>	2024-04-08 20:13	243K	
<a href="#">hg38.chromAlias.txt</a>	2021-10-06 13:44	27K	
<a href="#">hg38.chromFa.tar.gz</a>	2014-01-23 17:18	938M	
<a href="#">hg38.chromFaMasked.tar.gz</a>	2014-01-23 17:18	487M	
<a href="#">hg38.fa.align.gz</a>	2014-01-08 23:43	2.4G	
<a href="#">hg38.fa.gz</a>	2014-01-15 21:14	938M	
<a href="#">hg38.fa.masked.gz</a>	2014-01-15 21:24	487M	
<a href="#">hg38.fa.out.gz</a>	2014-01-15 20:56	172M	
<a href="#">hg38.gc5Base.bw</a>	2013-12-24 21:28	1.6G	
<a href="#">hg38.gc5Base.wib</a>	2019-01-17 14:50	591M	
<a href="#">hg38.gc5Base.wig.gz</a>	2019-01-17 14:50	11M	
<a href="#">hg38.gc5Base.wigVarStep.gz</a>	2013-12-24 21:14	1.5G	
<a href="#">hg38.trf.bed.gz</a>	2014-01-15 20:56	7.9M	
<a href="#">initial/</a>	2024-04-15 18:59	-	
<a href="#">latest/</a>	2023-02-22 16:20	-	
<a href="#">md5sum.txt</a>	2023-02-23 14:28	720	
<a href="#">mrna.fa.gz</a>	2019-10-14 13:37	370M	
<a href="#">mrna.fa.gz.md5</a>	2019-10-14 13:37	45	
<a href="#">p11/</a>	2024-01-25 20:54	-	
<a href="#">p12/</a>	2024-01-25 20:54	-	
<a href="#">p13/</a>	2024-01-25 20:54	-	
<a href="#">p14/</a>	2023-01-25 00:44	-	
<a href="#">refMrna.fa.gz</a>	2019-10-14 13:55	80M	
<a href="#">refMrna.fa.gz.md5</a>	2019-10-14 13:55	48	
<a href="#">upstream1000.fa.gz</a>	2019-10-14 14:46	10M	
<a href="#">upstream1000.fa.gz.md5</a>	2019-10-14 14:46	53	
<a href="#">upstream2000.fa.gz</a>	2019-10-14 14:46	20M	
<a href="#">upstream2000.fa.gz.md5</a>	2019-10-14 14:46	53	
<a href="#">upstream5000.fa.gz</a>	2019-10-14 14:47	50M	
<a href="#">upstream5000.fa.gz.md5</a>	2019-10-14 14:47	53	
<a href="#">xenoMrna.fa.gz</a>	2019-10-14 13:46	6.4G	
<a href="#">xenoMrna.fa.gz.md5</a>	2019-10-14 13:47	49	
<a href="#">xenoRefMrna.fa.gz</a>	2019-10-14 13:55	250M	
<a href="#">xenoRefMrna.fa.gz.md5</a>	2019-10-14 13:55	52	

<https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/>

analysis on CHM13.

Files:

hs1.fa.gz - "Soft-masked" assembly CHM13 sequence in one file.  
Repeats from RepeatMasker and Tandem Repeats Finder (with period of 12 or less) are shown in lower case; non-repeating sequence is shown in upper case.

hs1.2bit - contains the complete human CHM13 genome sequence in the 2bit file format. Repeats from RepeatMasker and Tandem Repeats Finder (with period of 12 or less) are shown in lower case; non-repeating sequence is shown in upper case. The utility program, twoBitToFa (available from the kent src tree), can be used to extract .fa file(s) from this file. A pre-compiled version of the command line tool can be found at:  
[http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86\\_64/](http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/)

See also:  
<http://genome.ucsc.edu/admin/git.html>  
<http://genome.ucsc.edu/admin/jk-install.html>

hs1.repeatMasker.out - RepeatMasker.out file. RepeatMasker was run with the -s (sensitive) setting.

hs1.repeatMasker.version.txt - version of repeatmasker that was used.

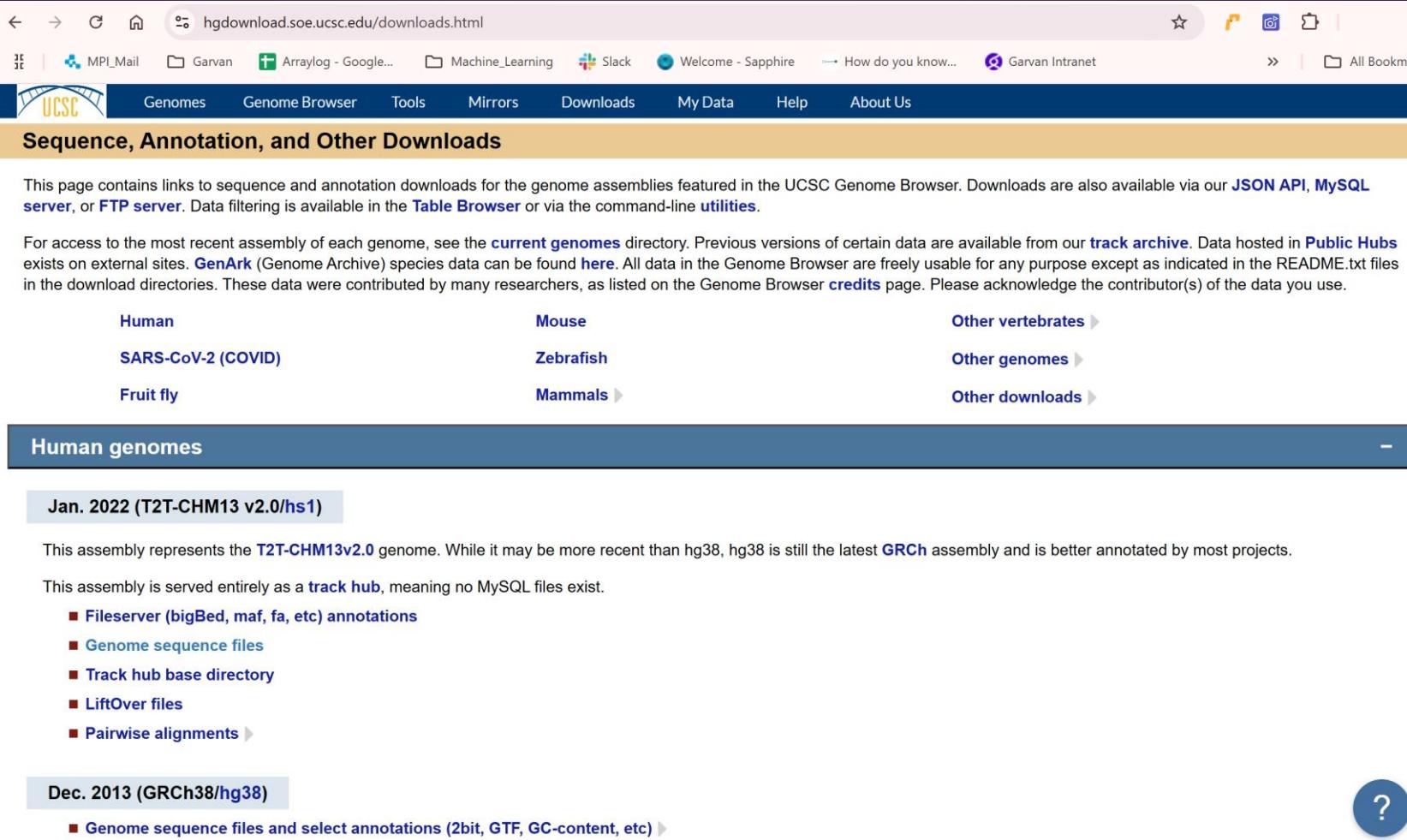
genes/hs1.ncbiRefSeq.gp.gz - gene annotations made by NCBI RefSeq in UCSC genePred format.

genes/hs1.ncbiRefSeq.gtf.gz - gene annotations made by NCBI RefSeq in GFF/GTF format.

Name	Last modified	Size	Description
<a href="#">Parent_Directory</a>	-	-	
<a href="#">GCA_009914755.4_assembly_report.txt</a>	2022-08-11 03:26	3.4K	
<a href="#">THIS_IS_GENOME_ASSEMBLY_T2T-CHM13v2.0</a>	2022-07-18 12:06	0	
<a href="#">genes/</a>	2023-08-30 15:16	-	
<a href="#">hs1.2bit</a>	2022-07-16 14:27	775M	
<a href="#">hs1.2bit.bpt</a>	2022-07-16 14:50	3.3K	
<a href="#">hs1.chrom.sizes.txt</a>	2022-06-22 21:16	375	
<a href="#">hs1.chromAlias.bb</a>	2022-07-18 11:38	43K	
<a href="#">hs1.chromAlias.txt</a>	2022-07-18 11:38	866	
<a href="#">hs1.fa.gz</a>	2022-07-18 12:08	930M	
<a href="#">hs1.repeatMasker.out.gz</a>	2022-07-18 11:48	173M	
<a href="#">hs1.repeatMasker.version.txt</a>	2022-02-04 10:35	922	
<a href="#">hs1.trans.efidx</a>	2022-07-16 14:33	4.6G	
<a href="#">hs1.untrans.efidx</a>	2022-07-16 14:32	2.16	
<a href="#">md5sum.txt</a>	2022-09-15 11:21	658	

<https://hgdownload.soe.ucsc.edu/goldenPath/hs1/bigZips/>

# Where to download the reference genomes



The screenshot shows a web browser window displaying the UCSC Genome Browser Downloads page at [hgdownload.soe.ucsc.edu/downloads.html](https://hgdownload.soe.ucsc.edu/downloads.html). The page has a dark blue header with the UCSC logo and navigation links for Genomes, Genome Browser, Tools, Mirrors, Downloads, My Data, Help, and About Us. Below the header is a yellow banner titled "Sequence, Annotation, and Other Downloads". A message on the page states: "This page contains links to sequence and annotation downloads for the genome assemblies featured in the UCSC Genome Browser. Downloads are also available via our [JSON API](#), [MySQL server](#), or [FTP server](#). Data filtering is available in the [Table Browser](#) or via the command-line [utilities](#)". Another message below says: "For access to the most recent assembly of each genome, see the [current genomes](#) directory. Previous versions of certain data are available from our [track archive](#). Data hosted in [Public Hubs](#) exists on external sites. [GenArk](#) (Genome Archive) species data can be found [here](#). All data in the Genome Browser are freely usable for any purpose except as indicated in the README.txt files in the download directories. These data were contributed by many researchers, as listed on the Genome Browser [credits](#) page. Please acknowledge the contributor(s) of the data you use." The main content area is divided into sections for Human, Mouse, Other vertebrates, SARS-CoV-2 (COVID), Zebrafish, Other genomes, Fruit fly, Mammals, and Other downloads. A "Human genomes" section is expanded, showing the "Jan. 2022 (T2T-CHM13 v2.0/hs1)" assembly. It describes the assembly as representing the T2T-CHM13v2.0 genome and being served as a track hub. It lists annotations available: Fileserver (bigBed, maf, fa, etc) annotations, Genome sequence files, Track hub base directory, LiftOver files, and Pairwise alignments. A "Dec. 2013 (GRCh38/hg38)" section is also partially visible. A question mark icon is in the bottom right corner.

hgdownload.soe.ucsc.edu/downloads.html

Genomes Genome Browser Tools Mirrors Downloads My Data Help About Us

Sequence, Annotation, and Other Downloads

This page contains links to sequence and annotation downloads for the genome assemblies featured in the UCSC Genome Browser. Downloads are also available via our [JSON API](#), [MySQL server](#), or [FTP server](#). Data filtering is available in the [Table Browser](#) or via the command-line [utilities](#).

For access to the most recent assembly of each genome, see the [current genomes](#) directory. Previous versions of certain data are available from our [track archive](#). Data hosted in [Public Hubs](#) exists on external sites. [GenArk](#) (Genome Archive) species data can be found [here](#). All data in the Genome Browser are freely usable for any purpose except as indicated in the README.txt files in the download directories. These data were contributed by many researchers, as listed on the Genome Browser [credits](#) page. Please acknowledge the contributor(s) of the data you use.

Human      Mouse      Other vertebrates ▶

SARS-CoV-2 (COVID)      Zebrafish      Other genomes ▶

Fruit fly      Mammals ▶      Other downloads ▶

Human genomes

Jan. 2022 (T2T-CHM13 v2.0/hs1)

This assembly represents the [T2T-CHM13v2.0](#) genome. While it may be more recent than hg38, hg38 is still the latest [GRCh](#) assembly and is better annotated by most projects.

This assembly is served entirely as a [track hub](#), meaning no MySQL files exist.

- [Fileserver \(bigBed, maf, fa, etc\) annotations](#)
- [Genome sequence files](#)
- [Track hub base directory](#)
- [LiftOver files](#)
- [Pairwise alignments](#) ▶

Dec. 2013 (GRCh38/hg38)

- [Genome sequence files and select annotations \(2bit, GTF, GC-content, etc\)](#) ▶

# Where to download the reference genomes

National Library of Medicine  
National Center for Biotechnology Information

Search NCBI ...

Log in

NCBI Datasets Taxonomy Genome Gene Command-line tools Documentation

## Genome assembly GRCh37.p13

Download datasets URL FTP

⚠ See latest version: GCF\_000001405.40

		Actions
NCBI RefSeq assembly	GCF_000001405.25 (replaced)	⋮
Submitted GenBank assembly	GCA_000001405.14 (replaced)	⋮
Taxon	Homo sapiens (human)	
Synonym	hg19	

Additional genomes  
Browse all Homo sapiens genomes (1754)

BioProject  
PRJNA31257  
The Human Genome Project, currently maintained by the Genome Reference Consortium (GRC)

Publications  
Showing 5 of 400  
Genome Biol 2008  
Finishing the finished human chromosome

[https://www.ncbi.nlm.nih.gov/datasets/genome/GCF\\_000001405.25/](https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.25/)

Index of /genomes/all/GCF/000/001/405/GCF\_000001405.25\_GRCh37.p13

Name	Last modified	Size
Parent Directory	2024-09-07 15:39	-
Annotation_comparison/	2024-09-07 15:39	-
GCF_000001405.25_GRCh37.p13_assembly_structure/	2013-09-16 15:36	-
GRCh37_segs_for_alignment_pipelines/	2024-09-07 15:39	-
RefSeq_transcripts_alignments/	2024-09-07 15:38	109K
GCF_000001405.25_RS_2024_09_annotation_report.xml	2024-09-07 15:38	17K
GCF_000001405.25_GRCh37.p13_assembly_regions.txt	2024-09-07 15:38	28K
GCF_000001405.25_GRCh37.p13_assembly_stats.txt	2024-09-07 15:38	52K
GCF_000001405.25_GRCh37.p13_cds_from_genomic.fna.gz	2024-09-07 15:38	19M
GCF_000001405.25_GRCh37.p13_fcs_report.txt	2024-11-21 14:25	460
GCF_000001405.25_GRCh37.p13_feature_count.txt	2024-09-07 15:38	9.0K
GCF_000001405.25_GRCh37.p13_feature_table.txt.gz	2024-09-07 15:38	5.3M
GCF_000001405.25_GRCh37.p13_genomic.fna.gz	2024-09-07 15:39	900M
GCF_000001405.25_GRCh37.p13_genomic.gbff.gz	2024-09-07 15:39	1.2G
GCF_000001405.25_GRCh37.p13_genomic.gff.gz	2024-09-07 15:39	46M
GCF_000001405.25_GRCh37.p13_genomic.gtf.gz	2024-09-07 15:39	25M
GCF_000001405.25_GRCh37.p13_genomic_gaps.txt.gz	2024-09-07 15:39	6.1K
GCF_000001405.25_GRCh37.p13_protein.faa.gz	2024-09-07 15:39	15M
GCF_000001405.25_GRCh37.p13_protein.gpff.gz	2024-09-07 15:39	125M
GCF_000001405.25_GRCh37.p13_pseudo_without_product.fna.gz	2024-09-07 15:39	6.7M
GCF_000001405.25_GRCh37.p13_rm.out.gz	2024-09-07 15:39	179M
GCF_000001405.25_GRCh37.p13_rm.run	2024-09-07 15:39	874
GCF_000001405.25_GRCh37.p13_rna.fna.gz	2024-09-07 15:39	62M
GCF_000001405.25_GRCh37.p13_rna.gbff.gz	2024-09-07 15:39	290M
GCF_000001405.25_GRCh37.p13_rna_from_genomic.fna.gz	2024-09-07 15:39	51M
GCF_000001405.25_GRCh37.p13_translated_cds.faa.gz	2024-09-07 15:39	13M
README.txt	2024-08-27 13:56	55K
README_GCF_000001405.25_RS_2024_09	2024-09-07 15:39	1.2K
README_patch_release.txt	2024-09-07 15:39	1.9K
annotation_hashes.txt	2024-09-18 08:49	411
assembly_status.txt	2024-12-29 01:21	16
md5checksums.txt	2024-09-18 08:49	93K
uncompressed_checksums.txt	2024-09-18 08:51	12K

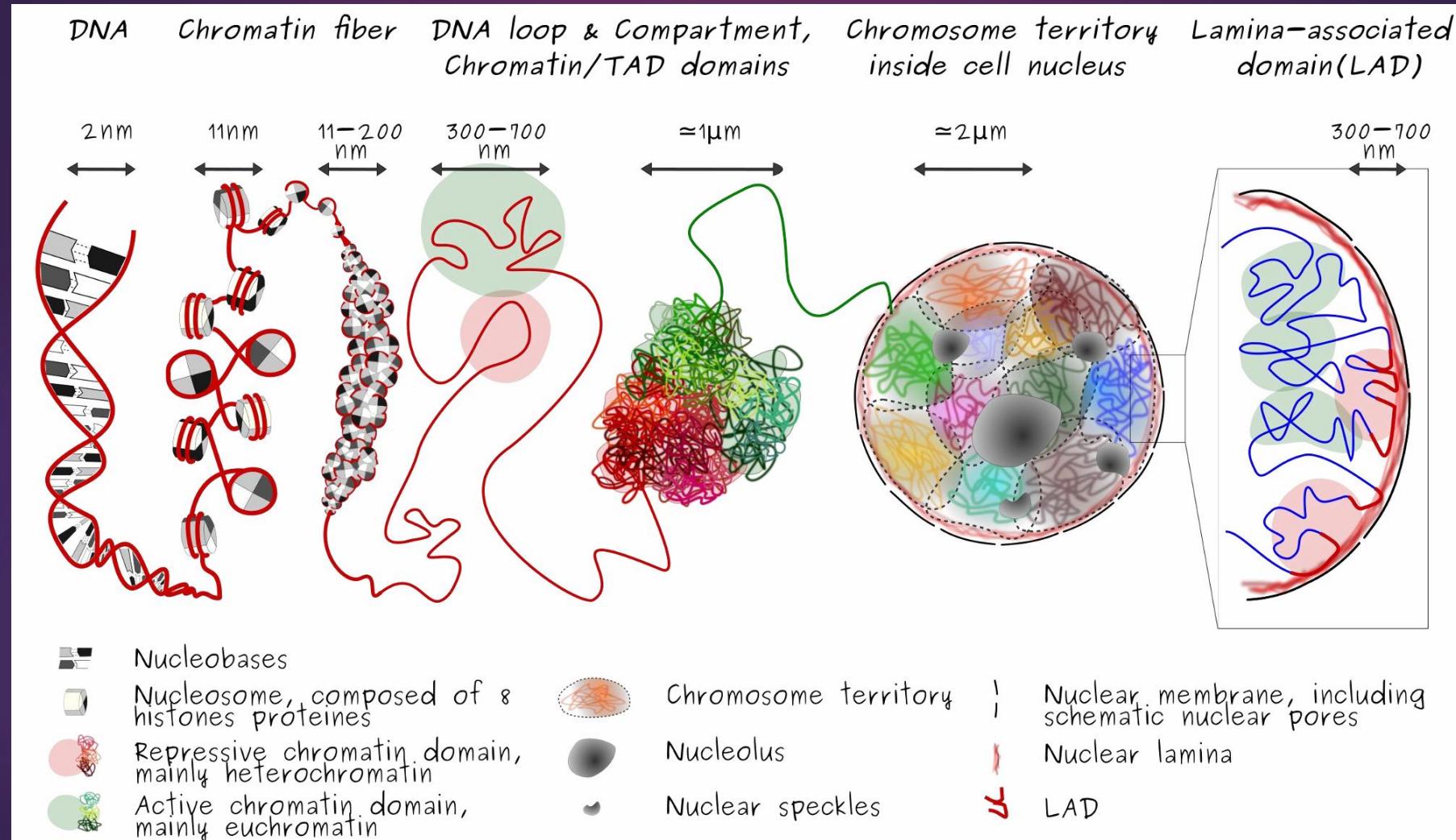
[https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/405/GCF\\_000001405.25\\_GRCh37.p13/](https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/405/GCF_000001405.25_GRCh37.p13/)

# Where to download the reference genomes

The screenshot shows a web browser window with the URL [gencodegenes.org/human/](https://www.gencodegenes.org/human/). The page displays a table of Fasta files available for download. A note at the top states: "• This dataset does **not** form part of the main annotation file".

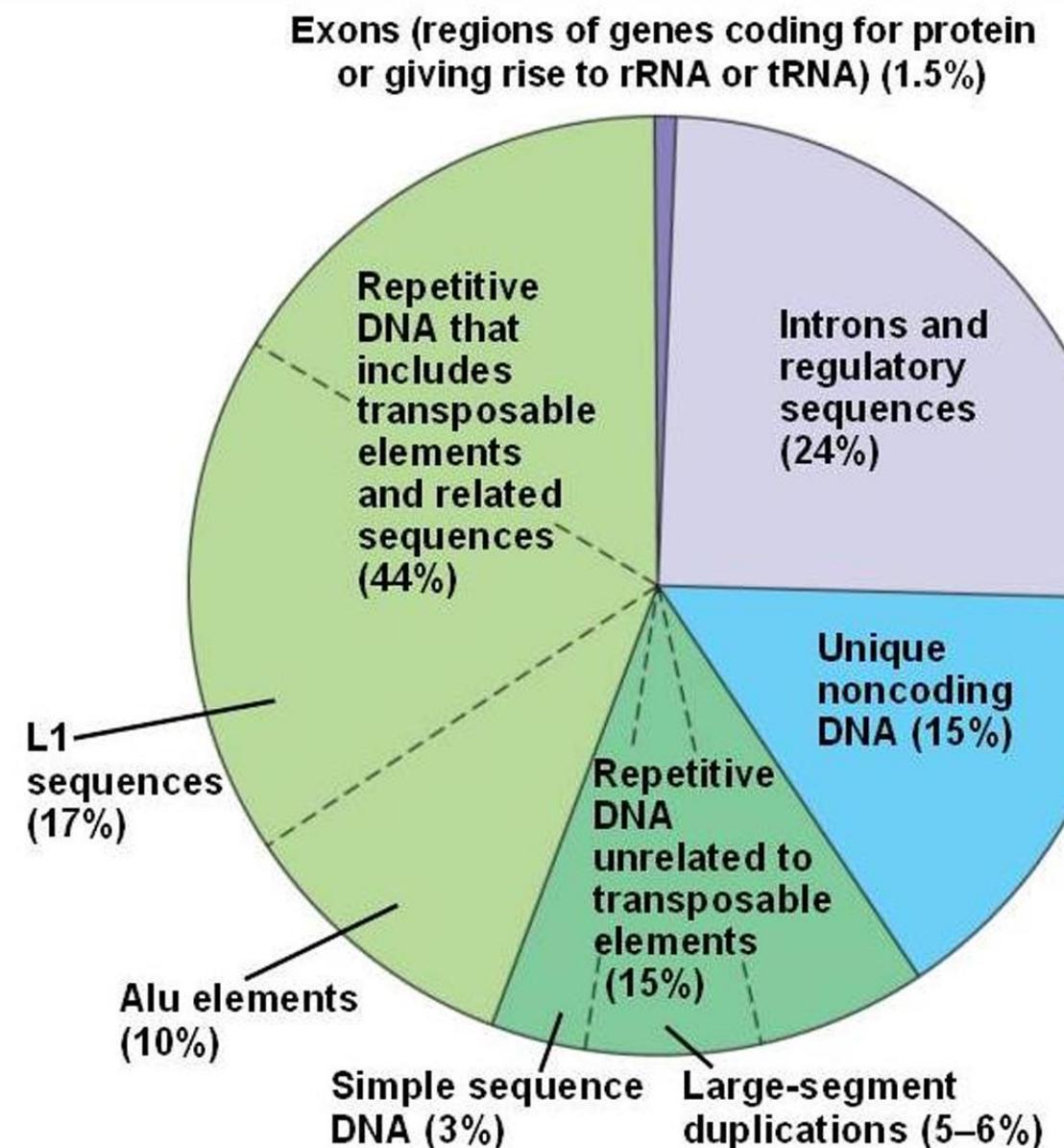
Content	Regions	Description	Download
Transcript sequences	CHR	<ul style="list-style-type: none"><li>Nucleotide sequences of all transcripts on the reference chromosomes</li></ul>	<a href="#">Fasta</a>
Protein-coding transcript sequences	CHR	<ul style="list-style-type: none"><li>Nucleotide sequences of coding transcripts on the reference chromosomes</li><li>Transcript biotypes: protein_coding, nonsense-mediated_decay, non_stop_decay, IG_*_gene, TR_*_gene, polymorphic_pseudogene, protein_coding_LoF</li></ul>	<a href="#">Fasta</a>
Protein-coding transcript translation sequences	CHR	<ul style="list-style-type: none"><li>Amino acid sequences of coding transcript translations on the reference chromosomes</li><li>Transcript biotypes: protein_coding, nonsense-mediated_decay, non_stop_decay, IG_*_gene, TR_*_gene, polymorphic_pseudogene, protein_coding_LoF</li></ul>	<a href="#">Fasta</a>
Long non-coding RNA transcript sequences	CHR	<ul style="list-style-type: none"><li>Nucleotide sequences of long non-coding RNA transcripts on the reference chromosomes</li></ul>	<a href="#">Fasta</a>
Genome sequence (GRCh38.p14)	ALL	<ul style="list-style-type: none"><li>Nucleotide sequence of the GRCh38.p14 genome assembly version on all regions, including reference chromosomes, scaffolds, assembly patches and haplotypes</li><li>The sequence region names are the same as in the GTF/GFF3 files</li></ul>	<a href="#">Fasta</a>
Genome sequence, primary assembly (GRCh38)	PRI	<ul style="list-style-type: none"><li>Nucleotide sequence of the GRCh38 primary genome assembly (chromosomes and scaffolds)</li><li>The sequence region names are the same as in the GTF/GFF3 files</li></ul>	<a href="#">Fasta</a>

# Human Genome Organization

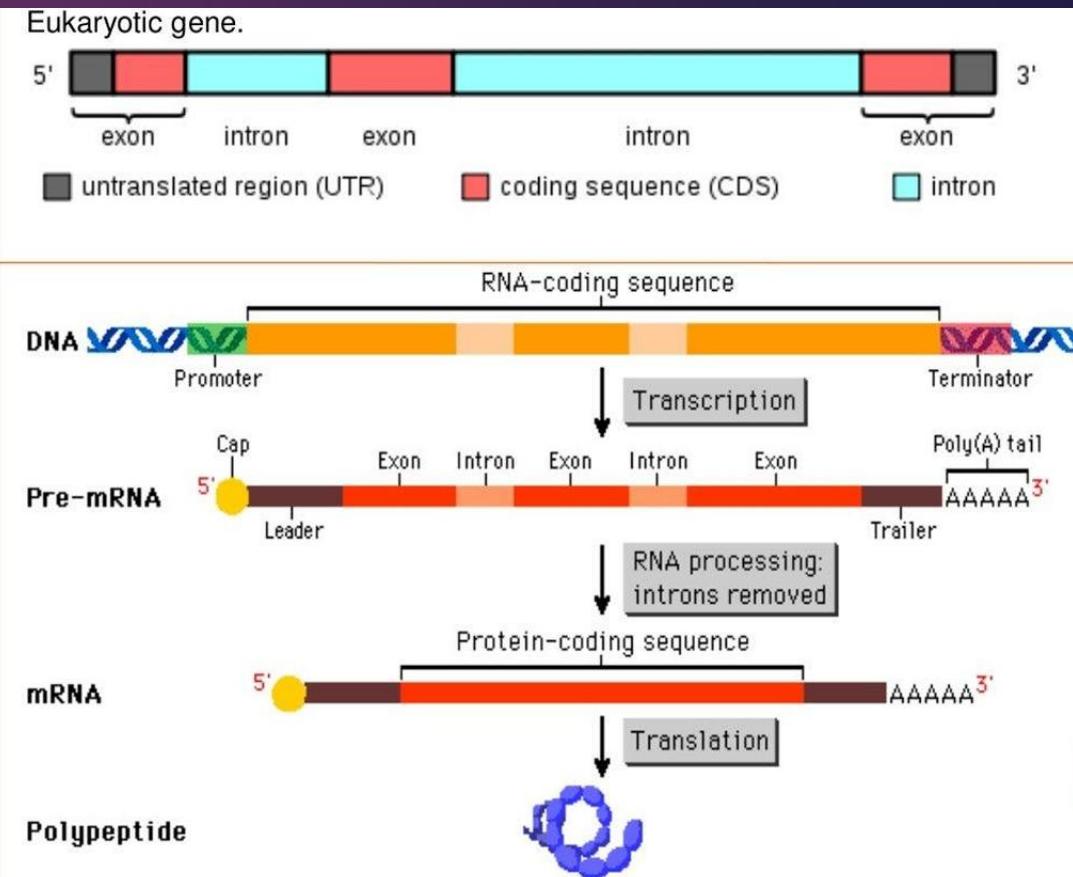
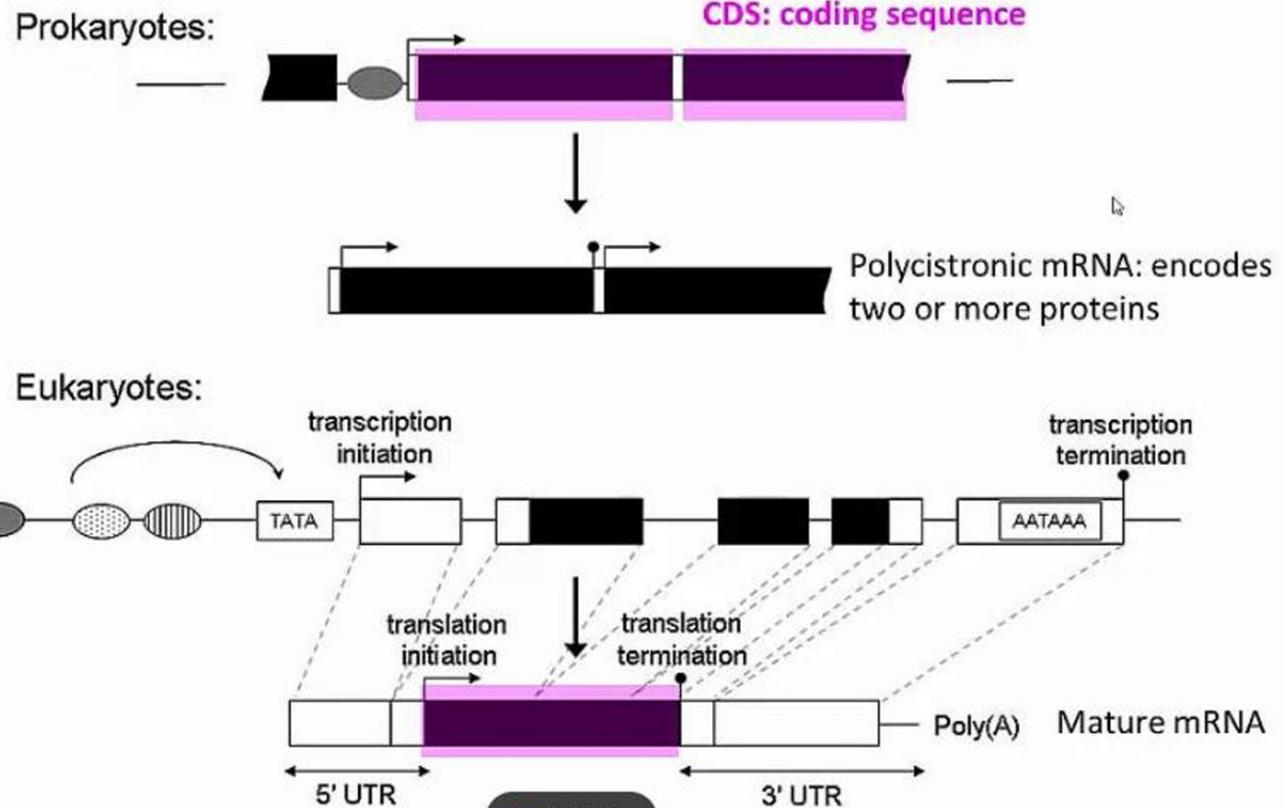


# Genome components

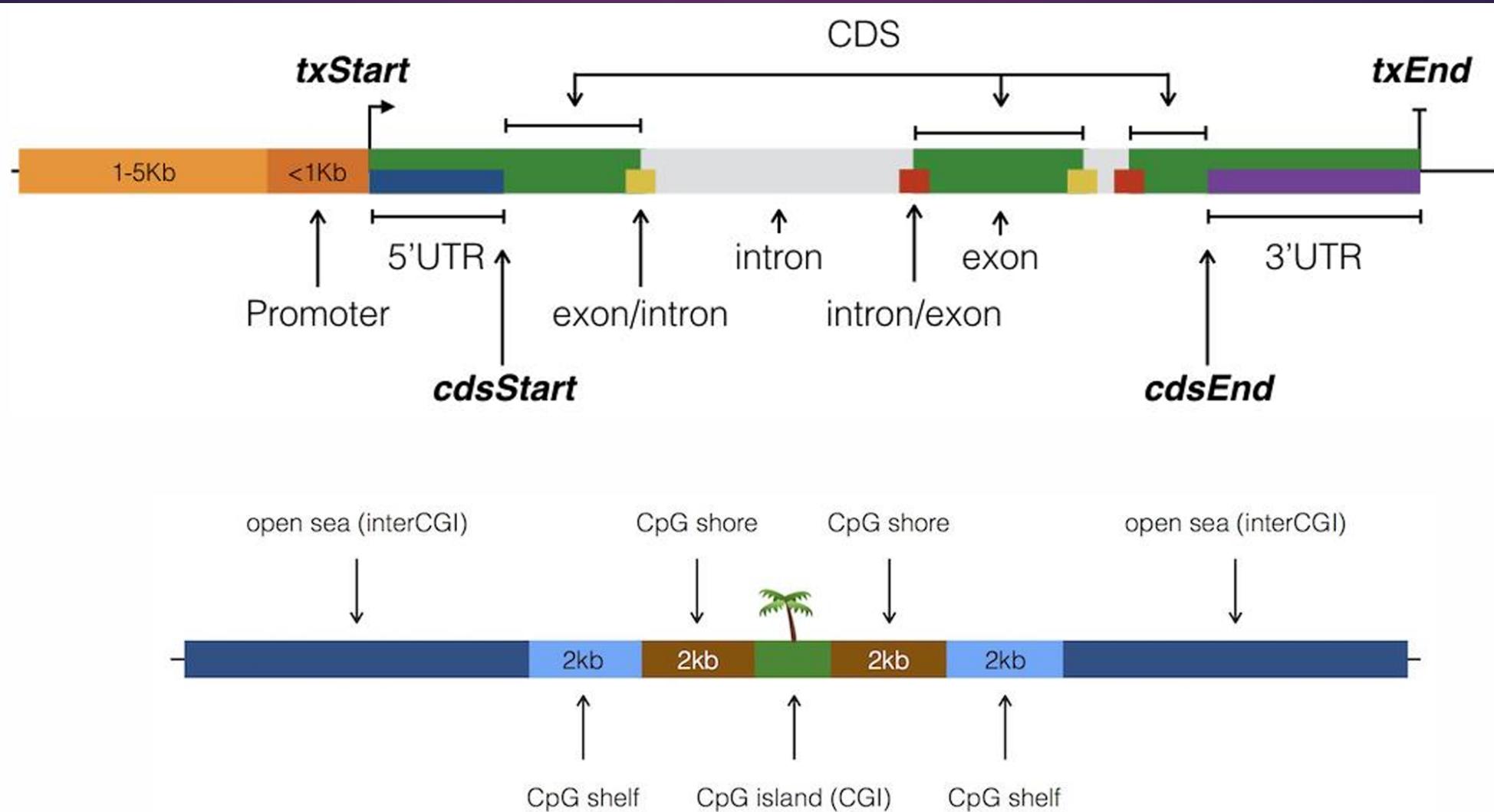
- The human genome contains less than 2% coding exons within genes.
- The remaining DNA consist of:
  - introns and regulatory sequences such as enhancers
  - unique noncoding DNA contains many pseudogenes (genes that have accumulated mutations and became nonfunctional)
  - repetitive DNA are sequences that are repeated many times; much of these are
  - transposable elements that can "jump" between chromosomes, leading to transpositions.



# Gene structure



# Gene structure



# Understanding Gene Annotation through GENCODE

(<https://www.gencodegenes.org/>)

The screenshot shows the GENCODE website homepage. At the top, there's a navigation bar with links for Human, Mouse, How to access data, FAQ, Documentation, and About us. Below the navigation bar, there are two main sections: "HUMAN" and "MOUSE". The HUMAN section features a composite image of a person's face split vertically, with the left side showing darker skin and the right side showing lighter skin. It also includes the text "GENCODE 47 (October 2024)". The MOUSE section features a photograph of a grey mouse. It also includes the text "GENCODE M36 (October 2024)". Below these sections, a large text block states the project's goal: "The goal of the GENCODE project is to identify and classify all gene features in the human and mouse genomes with high accuracy based on biological evidence, and to release these annotations for the benefit of biomedical research and genome interpretation." At the bottom, another text block mentions the expansion of IncRNA annotations: "The GENCODE human and mouse lncRNA annotations are significantly expanding as we integrate models from our [Capture Long-read Sequencing project](#)".

**HUMAN**  
GENCODE 47 (October 2024)



**MOUSE**  
GENCODE M36 (October 2024)



The goal of the GENCODE project is to identify and classify all gene features in the human and mouse genomes with high accuracy based on biological evidence, and to release these annotations for the benefit of biomedical research and genome interpretation.

The GENCODE human and mouse lncRNA annotations are significantly expanding as we integrate models from our [Capture Long-read Sequencing project](#).

# Understanding Gene Annotation through GENCODE

The screenshot shows the GENCODE Human release 47 (GRCh38.p14) page. At the top, there's a navigation bar with links for Human, Mouse, How to access data, FAQ, Documentation, and About us. Below this is a header with the GENCODE logo and a "Human" section. The main content area features a "Release 47 (GRCh38.p14)" section with three bullet points: "Statistics of this release", "More information about this assembly (including patches, scaffolds and haplotypes)", and "Go to GRCh37 version of this release". To the right, there's a sidebar titled "More about GENCODE Human" with links for Current human data, Release history, Statistics, Data format, and FTP site. Below this is a "GTF / GFF3 files" section with a table. The table has four columns: Content, Regions, Description, and Download. It lists four rows:

Content	Regions	Description	Download
Comprehensive gene annotation	CHR	• It contains the comprehensive gene annotation on the reference chromosomes only	<a href="#">GTF</a> <a href="#">GFF3</a>
Comprehensive gene annotation	ALL	• It contains the comprehensive gene annotation on the reference chromosomes, scaffolds, assembly patches and alternate loci (haplotypes)	<a href="#">GTF</a> <a href="#">GFF3</a>
Comprehensive gene annotation	PRI	• It contains the comprehensive gene annotation on the primary assembly (chromosomes and scaffolds) sequence regions	<a href="#">GTF</a> <a href="#">GFF3</a>
Basic gene annotation	CHR	• It contains the basic gene annotation on the reference chromosomes only • This is a <b>subset</b> of the corresponding comprehensive annotation, including only those transcripts tagged as 'basic' in every gene	<a href="#">GTF</a> <a href="#">GFF3</a>

<https://www.gencodegenes.org/human/>

# GTF (General Feature Format): 9 columns of data

## gencode.v41.annotation.gtf

Annotation file from GENCODE

```
##description: evidence-based annotation of the human genome (GRCh38), version 41 (Ensembl 107)
##provider: GENCODE
##contact: gencode-help@ebi.ac.uk
##format: gtf
##date: 2022-05-12

chr14 HAVANA gene 100725705 100738224 . + . gene_id "ENSG00000185559.16"; gene_type "protein_coding"; gene_name "DLK1"; level 2; hgnc_id "HGNC:2907"; havana_gene "OTTHUMG00000171600.5";
chr14 HAVANA transcript 100725705 100732070 . + . gene_id "ENSG00000185559.16"; transcript_id "ENST00000392848.9"; gene_type "protein_coding"; gene_name "DLK1"; transcript_type "protein_coding"; transcript_name "DLK1-203"; level 2; protein_id "ENSP00000376589.5"; transcript_support_level "4"; hgnc_id "HGNC:2907"; tag "alternative_5_UTR"; tag "mRNA_end_NF"; tag "cds_end_NF"; havana_gene "OTTHUMG00000171600.5"; havana_transcript "OTTHUMT00000414388.2";
chr14 HAVANA exon 100725705 100725833 . + . gene_id "ENSG00000185559.16"; transcript_id "ENST00000392848.9"; gene_type "protein_coding"; gene_name "DLK1"; transcript_type "protein_coding"; transcript_name "DLK1-203"; exon_number 1; exon_id "ENSE00002444405.1"; level 2; protein_id "ENSP00000376589.5"; transcript_support_level "4"; hgnc_id "HGNC:2907"; tag "alternative_5_UTR"; tag "mRNA_end_NF"; tag "cds_end_NF"; havana_gene "OTTHUMG00000171600.5"; havana_transcript "OTTHUMT00000414388.2";
chr14 HAVANA exon 100726534 100726595 . + . gene_id "ENSG00000185559.16"; transcript_id "ENST00000392848.9"; gene_type "protein_coding"; gene_name "DLK1"; transcript_type "protein_coding"; transcript_name "DLK1-203"; exon_number 2; exon_id "ENSE00002444247.1"; level 2; protein_id "ENSP00000376589.5"; transcript_support_level "4"; hgnc_id "HGNC:2907"; tag "alternative_5_UTR"; tag "mRNA_end_NF"; tag "cds_end_NF"; havana_gene "OTTHUMG00000171600.5"; havana_transcript "OTTHUMT00000414388.2";
chr14 HAVANA exon 100726996 100727135 . + . gene_id "ENSG00000185559.16"; transcript_id "ENST00000392848.9"; gene_type "protein_coding"; gene_name "DLK1"; transcript_type "protein_coding"; transcript_name "DLK1-203"; exon_number 3; exon_id "ENSE00002444248.1"; level 2; protein_id "ENSP00000376589.5"; transcript_support_level "4"; hgnc_id "HGNC:2907"; tag "alternative_5_UTR"; tag "mRNA_end_NF"; tag "cds_end_NF"; havana_gene "OTTHUMG00000171600.5"; havana_transcript "OTTHUMT00000414388.2";
```



# Understanding Gene Annotation through GENCODE



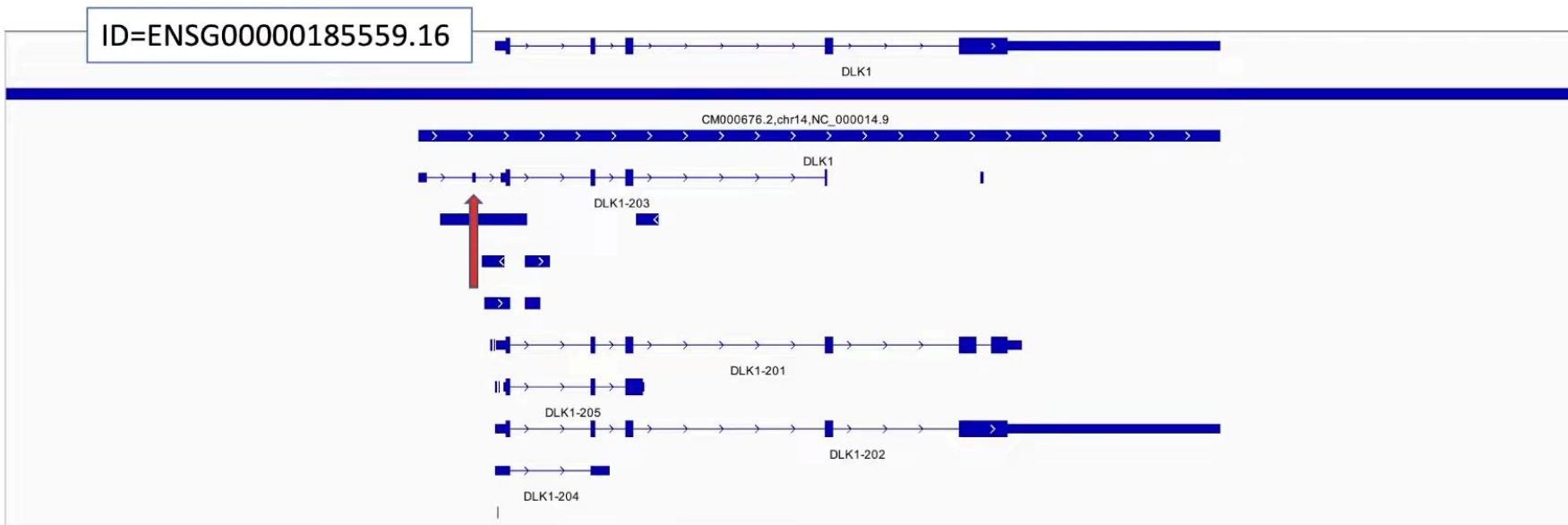
# Understanding Gene Annotation through GENCODE



# Understanding Gene Annotation through GENCODE



# Understanding Gene Annotation through GENCODE



```
chr14 HAVANA gene 100725705 100738224 . + . ID=ENSG00000185559.16;gene_id=ENSG00000185559.16;gene_type=protein_coding;gene_name=DLK1;level=2;hgnc_id=HGNC:2907;havana_gene=OTTHUMG00000171600.5
chr14 HAVANA transcript 100725705 100732070 . + . ID=ENST00000392848.9;Parent=ENSG00000185559.16;gene_id=ENSG0000185559.16;transcript_id=ENST00000392848.9;gene_type=protein_coding;gene_name=DLK1;transcript_type=protein_coding;transcript_name=DLK1-203;level=2;protein_id=ENSP00000376589.5;transcript_support_level=4;hgnc_id=HGNC:2907;tag=alternative_5_UTR,mRNA_end_NF,cds_end_NF;havana_gene=OTTHUMG00000171600.5;havana_transcript=OTTHUMT00000414388.2
chr14 HAVANA exon 100725705 100725833 . + . ID=exon:ENST00000392848.9:1;Parent=ENST00000392848.9;gene_id=ENSG00000185559.16;transcript_id=ENST00000392848.9;gene_type=protein_coding;gene_name=DLK1;transcript_type=protein_coding;transcript_name=DLK1-203;exon_number=1;exon_id=ENSE00002444405.1;level=2;protein_id=ENSP00000376589.5;transcript_support_level=4;hgnc_id=HGNC:2907;tag=alternative_5_UTR,mRNA_end_NF,cds_end_NF;havana_gene=OTTHUMG00000171600.5;havana_transcript=OTTHUMT00000414388.2
chr14 HAVANA exon 100726534 100726595 . + . ID=exon:ENST00000392848.9:2;Parent=ENST00000392848.9;gene_id=ENSG00000185559.16;transcript_id=ENST00000392848.9;gene_type=protein_coding;gene_name=DLK1;transcript_type=protein_coding;transcript_name=DLK1-203;exon_number=2;exon_id=ENSE00002444247.1;level=2;protein_id=ENSP00000376589.5;transcript_support_level=4;hgnc_id=HGNC:2907;tag=alternative_5_UTR,mRNA_end_NF,cds_end_NF;havana_gene=OTTHUMG00000171600.5;havana_transcript=OTTHUMT00000414388.2
```



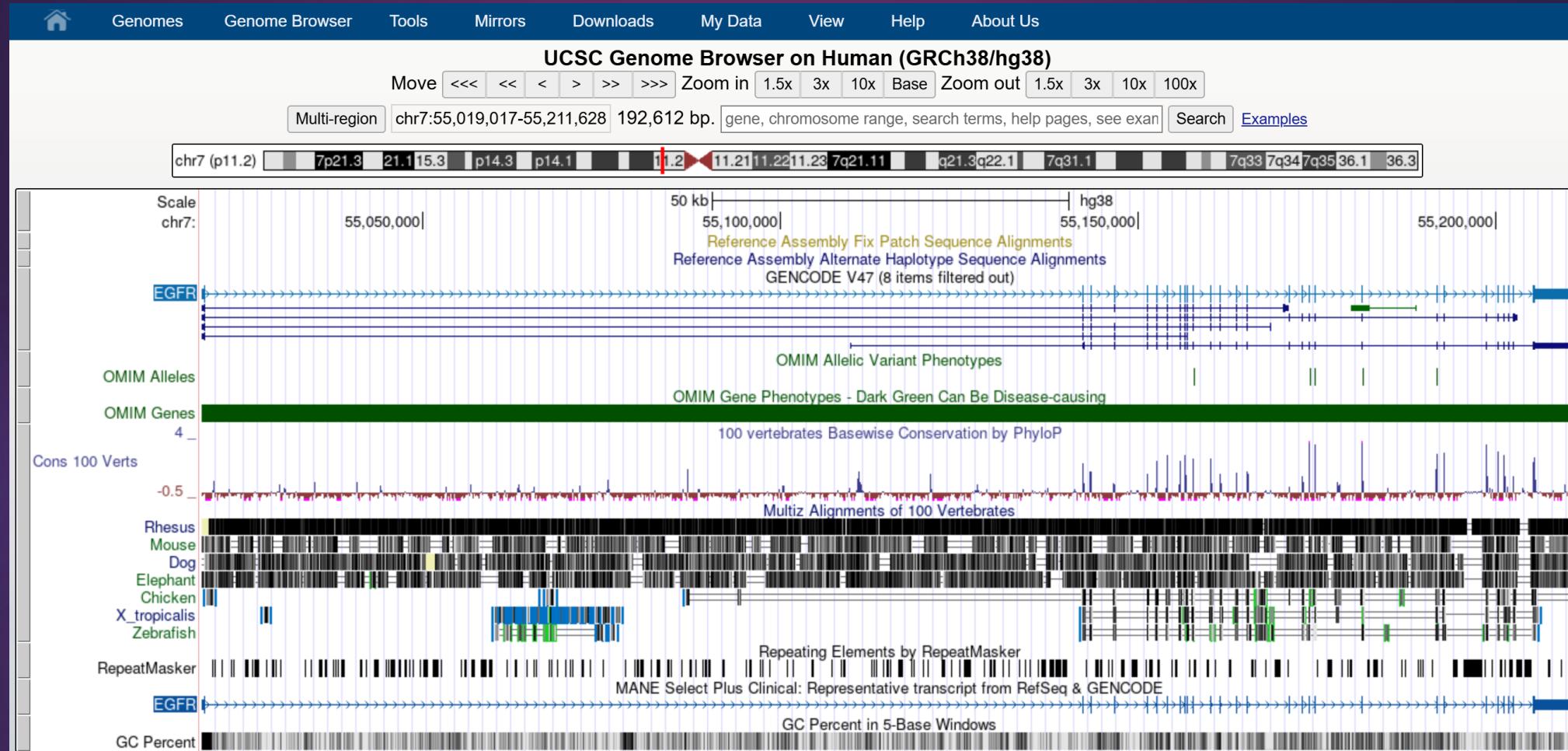
# Introduction to UCSC Genome Browser



The screenshot shows the UCSC Genome Browser homepage. At the top, there is a banner featuring the University of California Santa Cruz Genomics Institute logo and the UCSC logo. Below the banner, a navigation bar includes links for Genomes, Genome Browser, Tools, Mirrors, Downloads, My Data, Projects, Help, and About Us. The main content area displays a genomic track for the SAT2 gene, showing various tracks like ATBP2, TPS3, and WRAP53. A red box highlights a "Try our new clinical tutorial!" button with a magnifying glass icon. Below this, there is a search bar with the placeholder "Search genes, data, help docs and more..." and a "Search" button. The page is divided into sections: "Tools" on the left and "News" on the right. The "Tools" section lists links for Genome Browser, BLAT, In-Silico PCR, Table Browser, LiftOver, REST API, Variant Annotation Integrator, and More tools... Each link is preceded by a small icon. The "News" section lists recent news items with dates: Nov. 8, 2024 - New GENCODE gene tracks: Human V47 (hg19/hg38) - Mouse M36; Nov. 4, 2024 - GIAB Problematic Regions tracks for human (hg38 and hs1); Oct. 23, 2024 - New GENCODE gene tracks: V47 (hg38) - VM36 (mm39); Oct. 9, 2024 - CADD v1.7 and ClinGen CSpec for hg19 and hg38; Oct. 1, 2024 - New clinical tutorial; Sep. 30, 2024 - New gnomAD v4.1 tracks for hg38. At the bottom, there are links for CSHL Genome Informatics 2024 and American Society of Human Genetics (ASHG), along with a "Meetings and Workshops: Come see us in person!" section and a question mark icon.

<https://genome.ucsc.edu/index.html>

# Introduction to UCSC Genome Browser



# Getting Help with the UCSC Genome Browser

## Online training and tutorials

Our video tutorials address some common questions we've gathered from our [mailing list](#). Along the way we try to show you interesting features of the Browser you may not have found on your own.

Visit our [YouTube channel](#) or use the links below.

### Video tutorials

- [Browser Basics, Part One: Getting around in the Browser.](#) [[transcript](#)]
- [Browser Basics, Part Two: Configuring the Browser.](#) [[transcript](#)]
- [Browser Basics, Part Three: Configuration + DNA navigation.](#) [[transcript](#)]
- [Making Links, Part One: Understanding the URL.](#) [[transcript](#)]
- [Making Links, Part Two: Jump into genes.](#) [[transcript](#)]
- [Making Links, Part Three: Composites, custom tracks, spreadsheets.](#) [[transcript](#)]
- [Coronavirus Basics: Coronavirus Browser SARS-CoV-2.](#) [[transcript](#)]
- [Saving and sharing sessions in the Browser.](#) [[transcript](#)]
- [Controlling visibility of data tracks in the Browser.](#) [[transcript](#)]
- [Using the isPCR tool \(isPCR\) in the UCSC Genome Browser.](#) [[transcript](#)]
- [dbSNP resources in the UCSC Genome Browser database.](#) [[transcript](#)]
- [Using the UCSC Genome Browser Data Integrator.](#) [[transcript](#)]

### Video tutorials

- [Finding a list of genes in a region.](#) [[transcript](#)]
- [Finding exon numbers.](#) [[transcript](#)]
- [Finding all SNPs in a gene.](#) [[transcript](#)]
- [Finding SNPs upstream from a gene.](#) [[transcript](#)]
- [Find which tables belong to a data track.](#) [[transcript](#)]
- [Identifying codon numbers in a gene.](#) [[transcript](#)]
- [Obtaining exon coordinates and sequences.](#) [[transcript](#)]
- [Multi-Region View: Exon-only display mode.](#) [[transcript](#)]
- [Multi-Region View: Alternate haplotypes.](#) [[transcript](#)]
- [Multi-Region View: Discontinuous regions.](#) [[transcript](#)]
- [How-to: Genome Browser in the Cloud.](#)
- [How-to: Genome Browser Gateway.](#)



# Xin chân thành cảm ơn!

LUU PHUC LOI, PHD

ZALO: 0901802182

LUU.P.LOI@GOOGLEMAIL.COM