HUMAN GENOME

Dec 29 2024

Phuc Loi Luu, PhD
Email: 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)

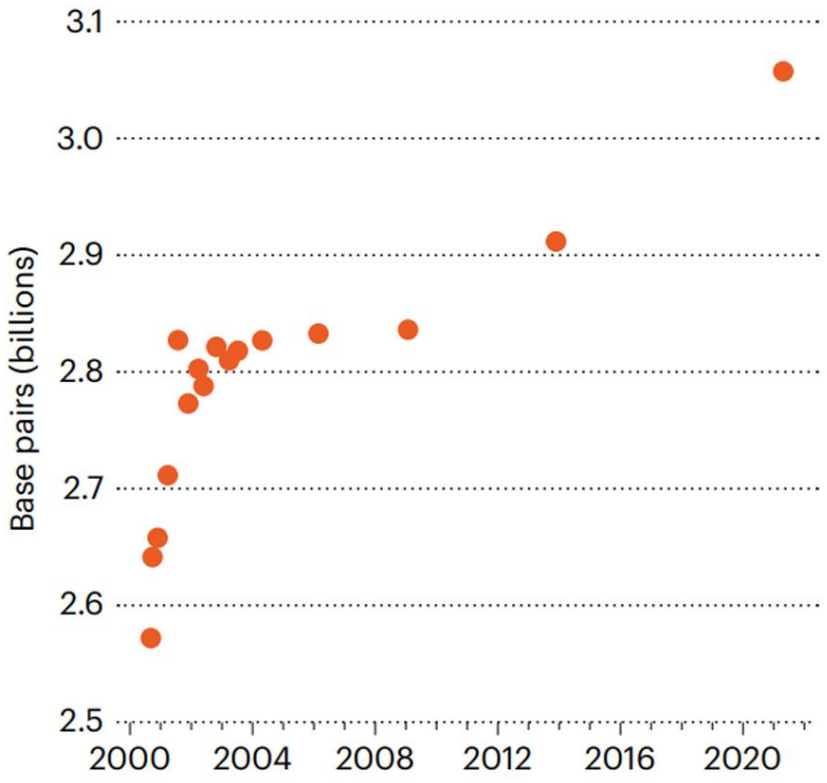
https://genome.ucsc.edu/FAQ/FAQreleases.html

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 PHILLIPPY

Fasta Format

Single Sequence Fasta Format

```
>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken
MADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTID
FPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREA
DIDGDGQVNYEEFVQMMTAK*
```

Multiple Sequence Fasta Format

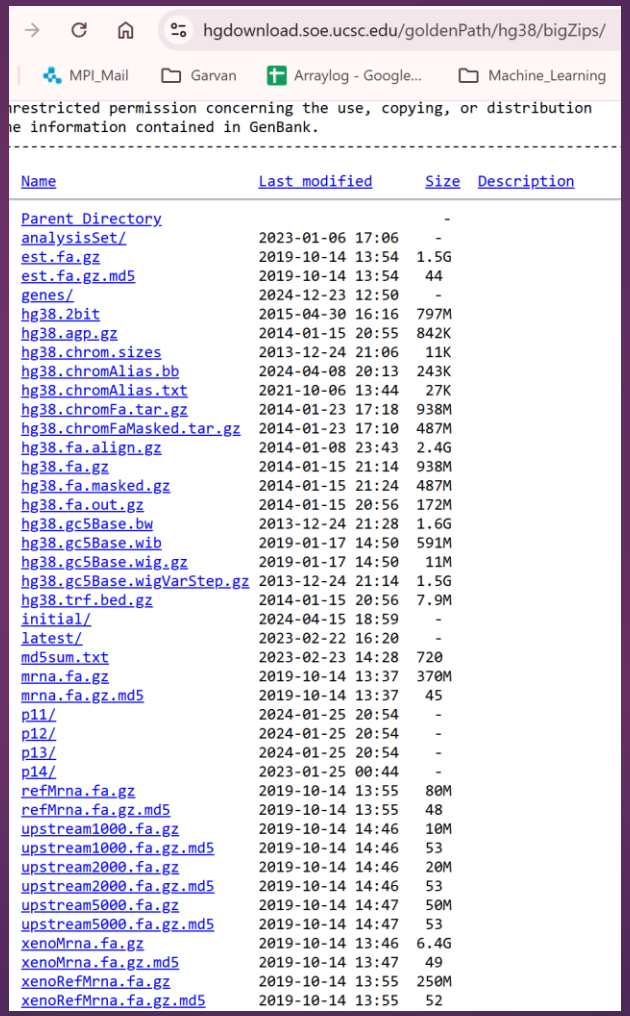
```
>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken
MADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTID
FPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREA
DIDGDGQVNYEEFVQMMTAK*

>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

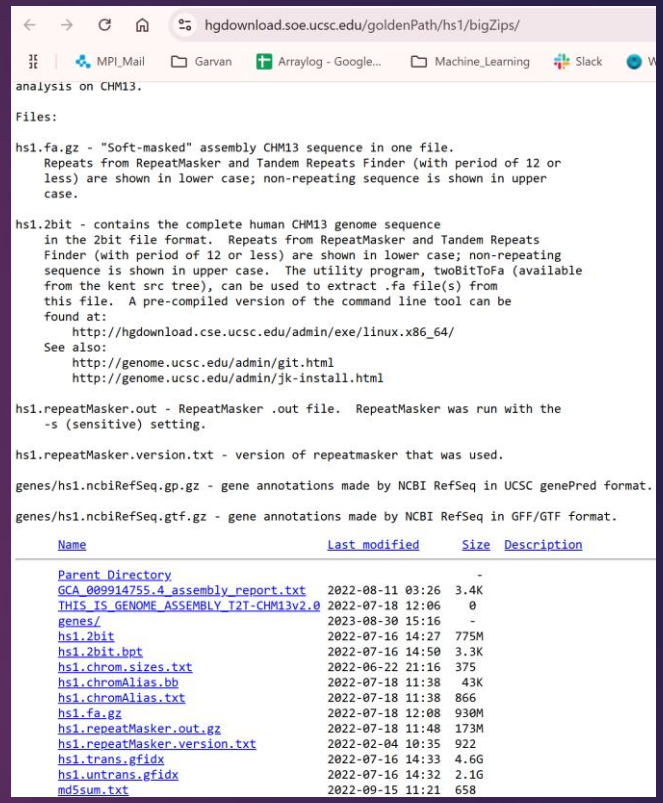
https://en.wikipedia.org/wiki/FASTA_format

Where to download the reference genomes

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

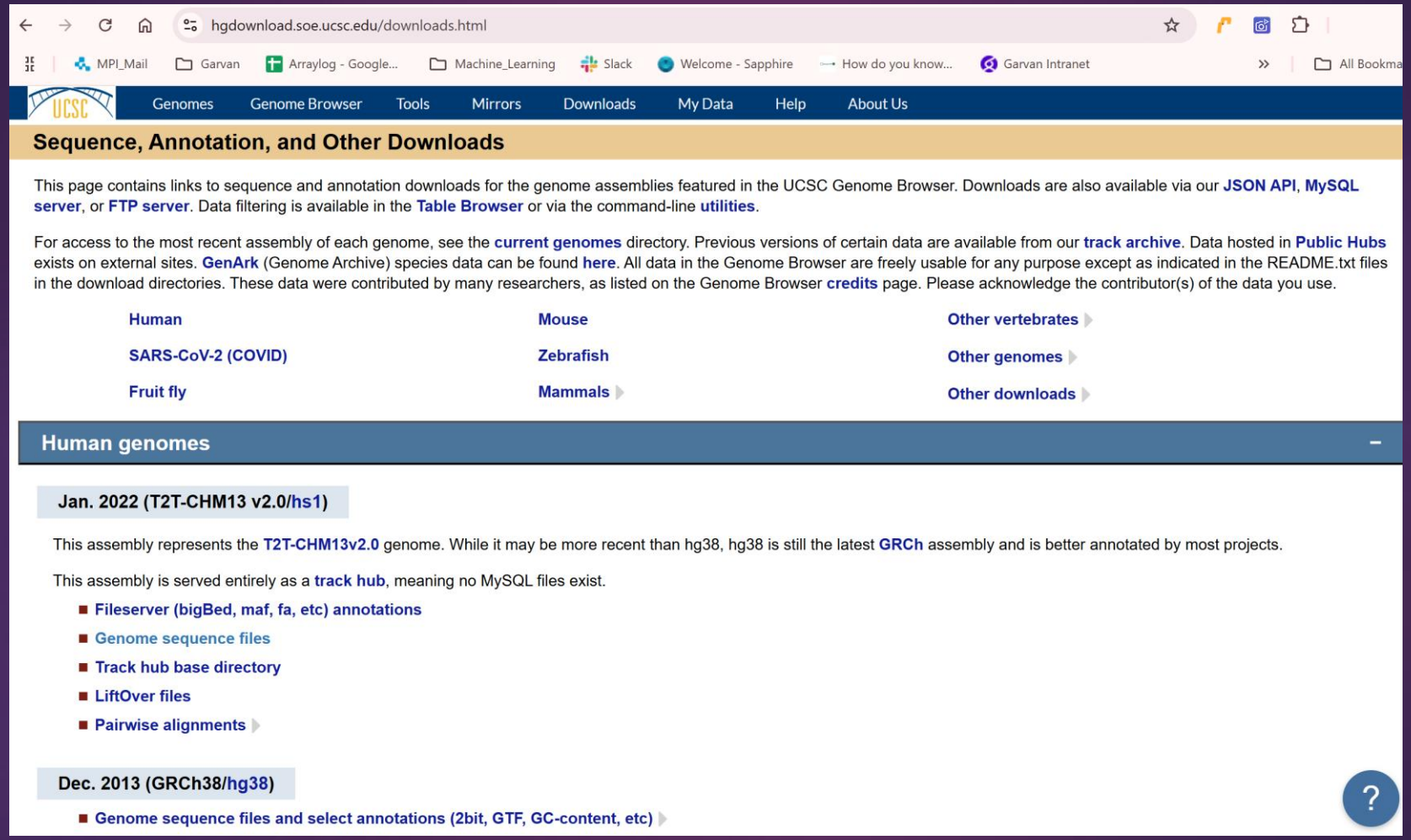


https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/



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

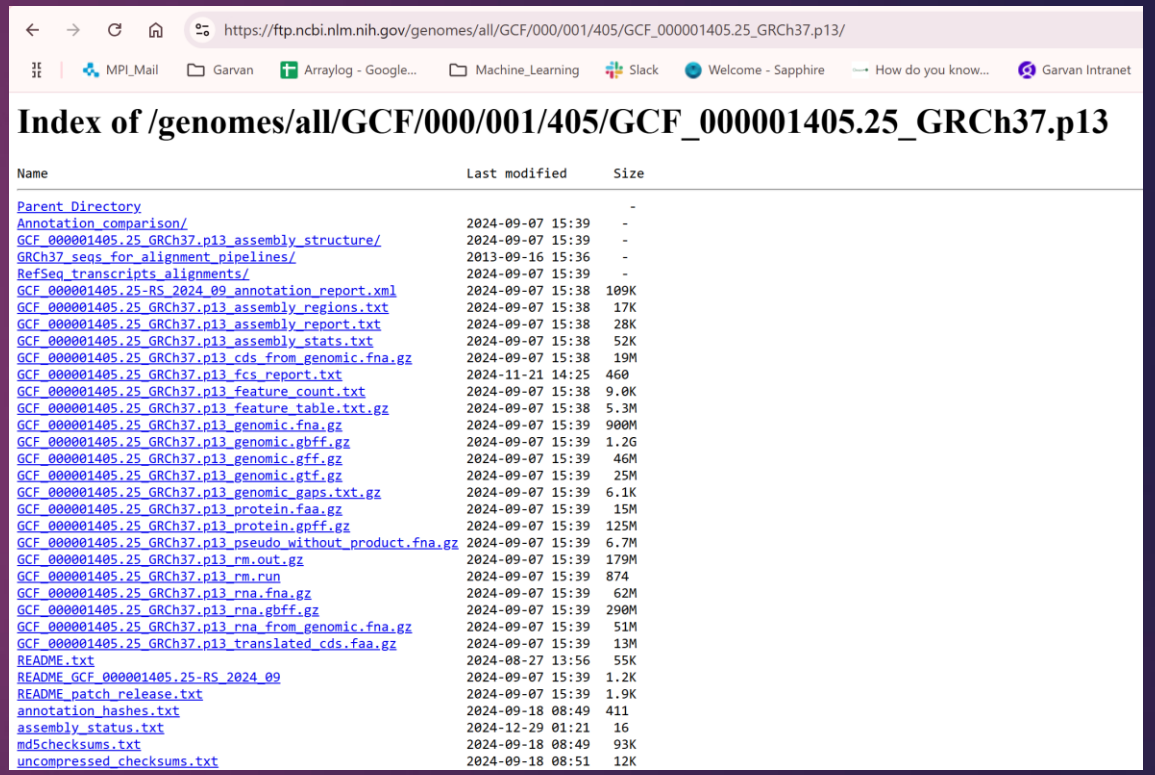
Where to download the reference genomes



https://hgdownload.soe.ucsc.edu/downloads.html

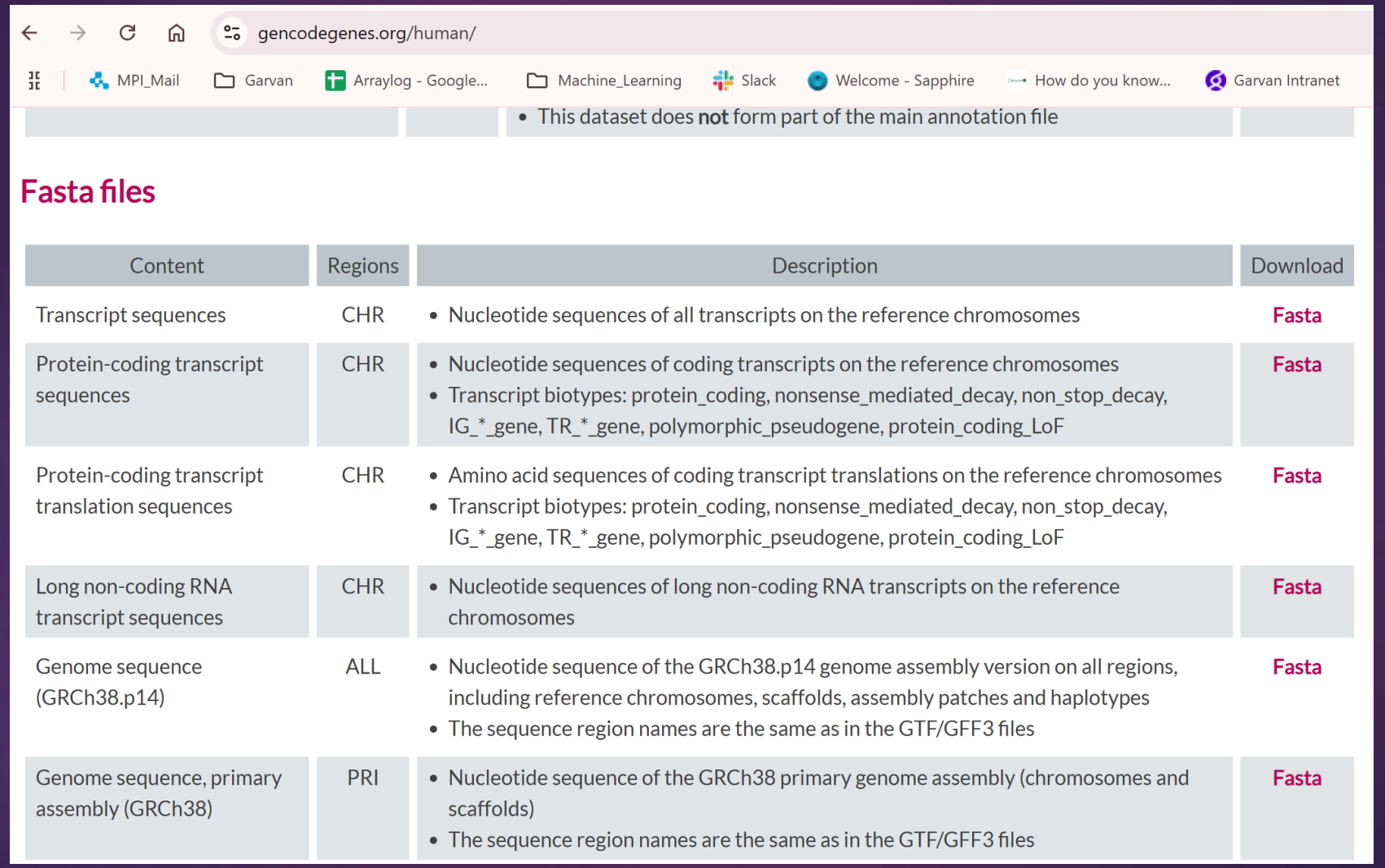
Where to download the reference genomes

https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.25/



https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/405/GCF_000001405.25_GRCh37.p13/

Where to download the reference genomes



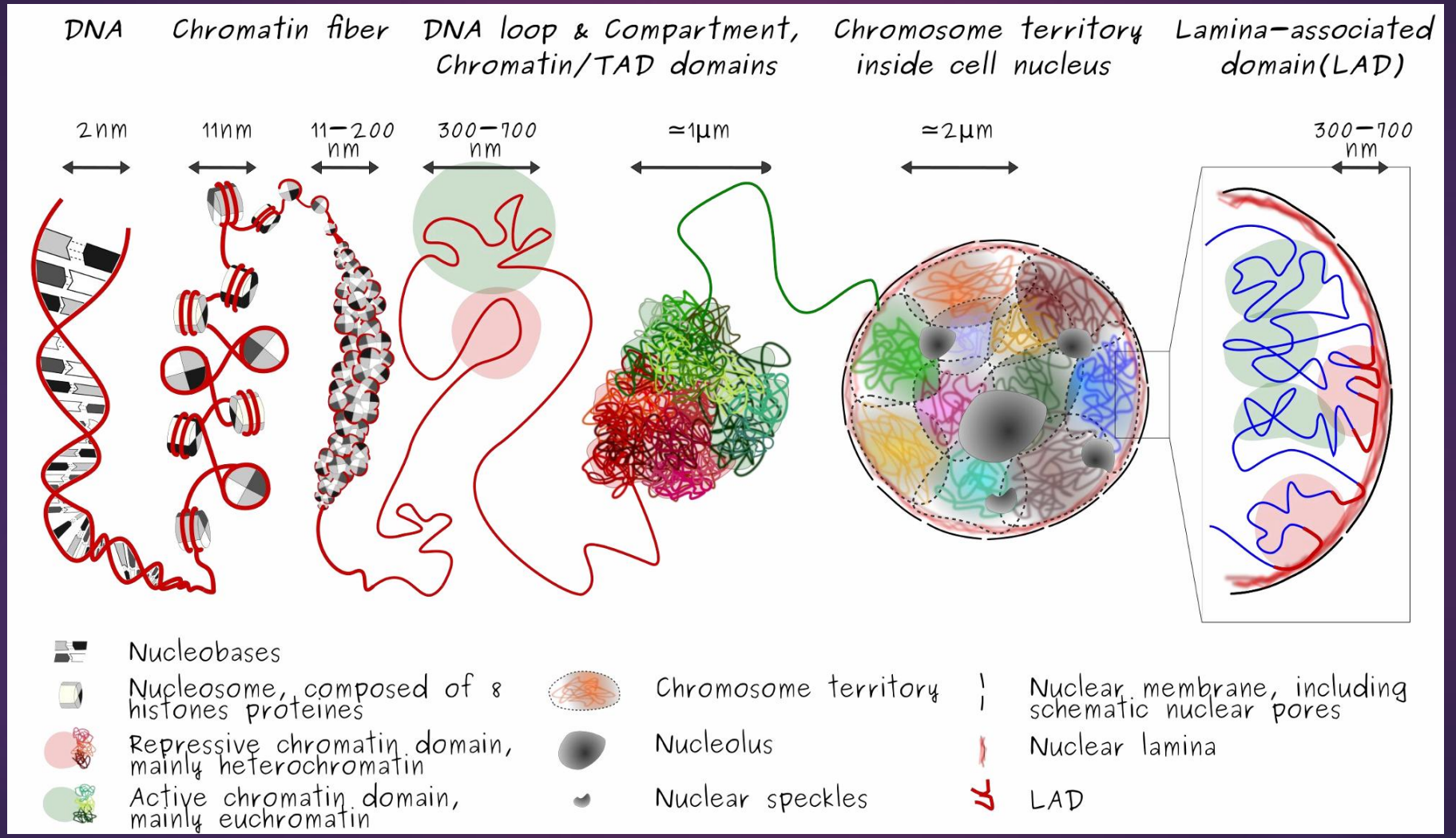
- This dataset does **not** form part of the main annotation file

Fasta files

Content	Regions	Description	Download
Transcript sequences	CHR	• Nucleotide sequences of all transcripts on the reference chromosomes	Fasta
Protein-coding transcript sequences	CHR	• Nucleotide sequences of coding transcripts on the reference chromosomes • Transcript biotypes: protein_coding, nonsense_mediated_decay, non_stop_decay, IG_*_gene, TR_*_gene, polymorphic_pseudogene, protein_coding_LoF	Fasta
Protein-coding transcript translation sequences	CHR	• Amino acid sequences of coding transcript translations on the reference chromosomes • Transcript biotypes: protein_coding, nonsense_mediated_decay, non_stop_decay, IG_*_gene, TR_*_gene, polymorphic_pseudogene, protein_coding_LoF	Fasta
Long non-coding RNA transcript sequences	CHR	• Nucleotide sequences of long non-coding RNA transcripts on the reference chromosomes	Fasta
Genome sequence (GRCh38.p14)	ALL	• Nucleotide sequence of the GRCh38.p14 genome assembly version on all regions, including reference chromosomes, scaffolds, assembly patches and haplotypes • The sequence region names are the same as in the GTF/GFF3 files	Fasta
Genome sequence, primary assembly (GRCh38)	PRI	• Nucleotide sequence of the GRCh38 primary genome assembly (chromosomes and scaffolds) • The sequence region names are the same as in the GTF/GFF3 files	Fasta

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

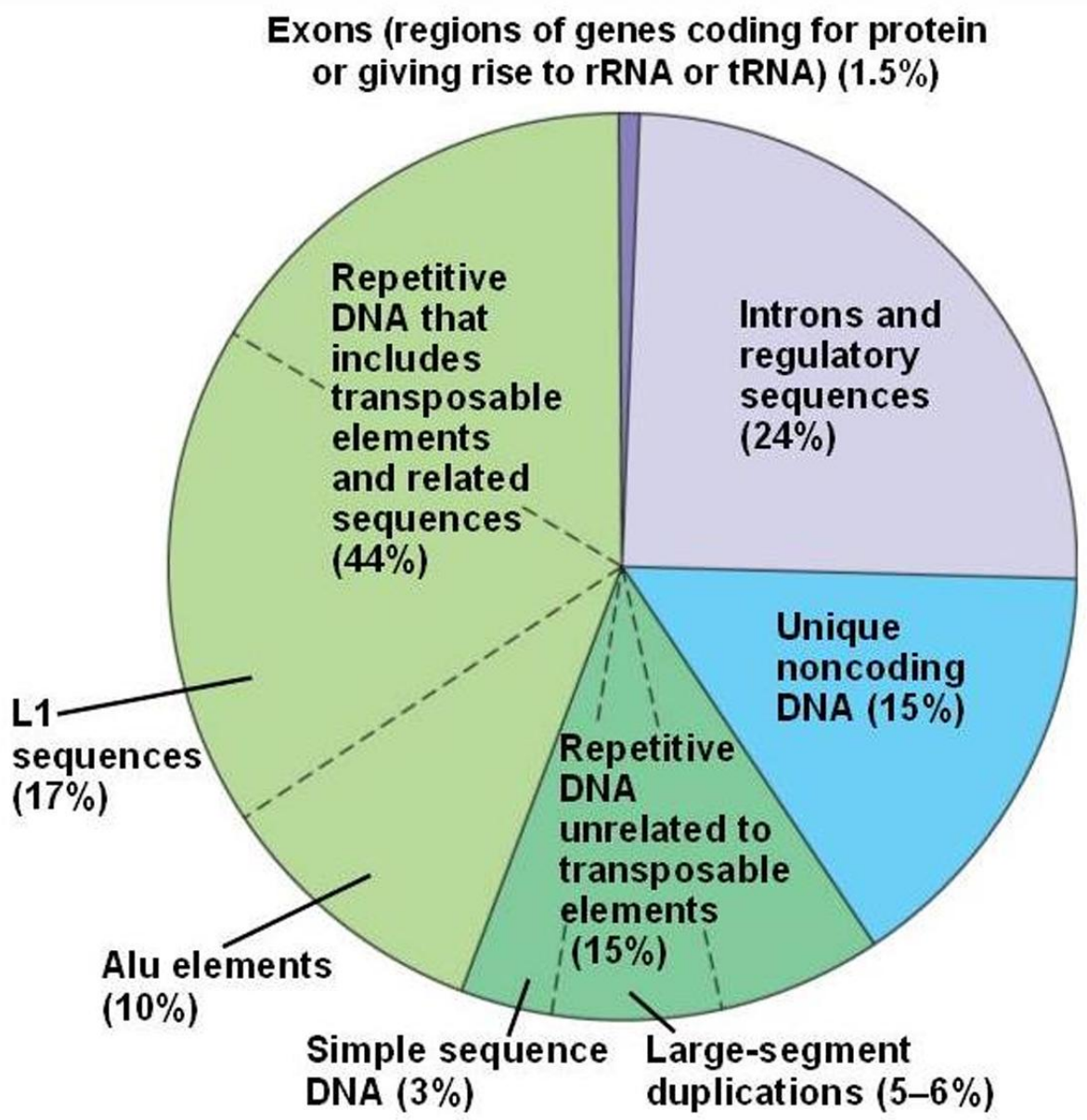
Human Genome Organization



index_3DOrgGenome_SquareShape.png (2523×1422)

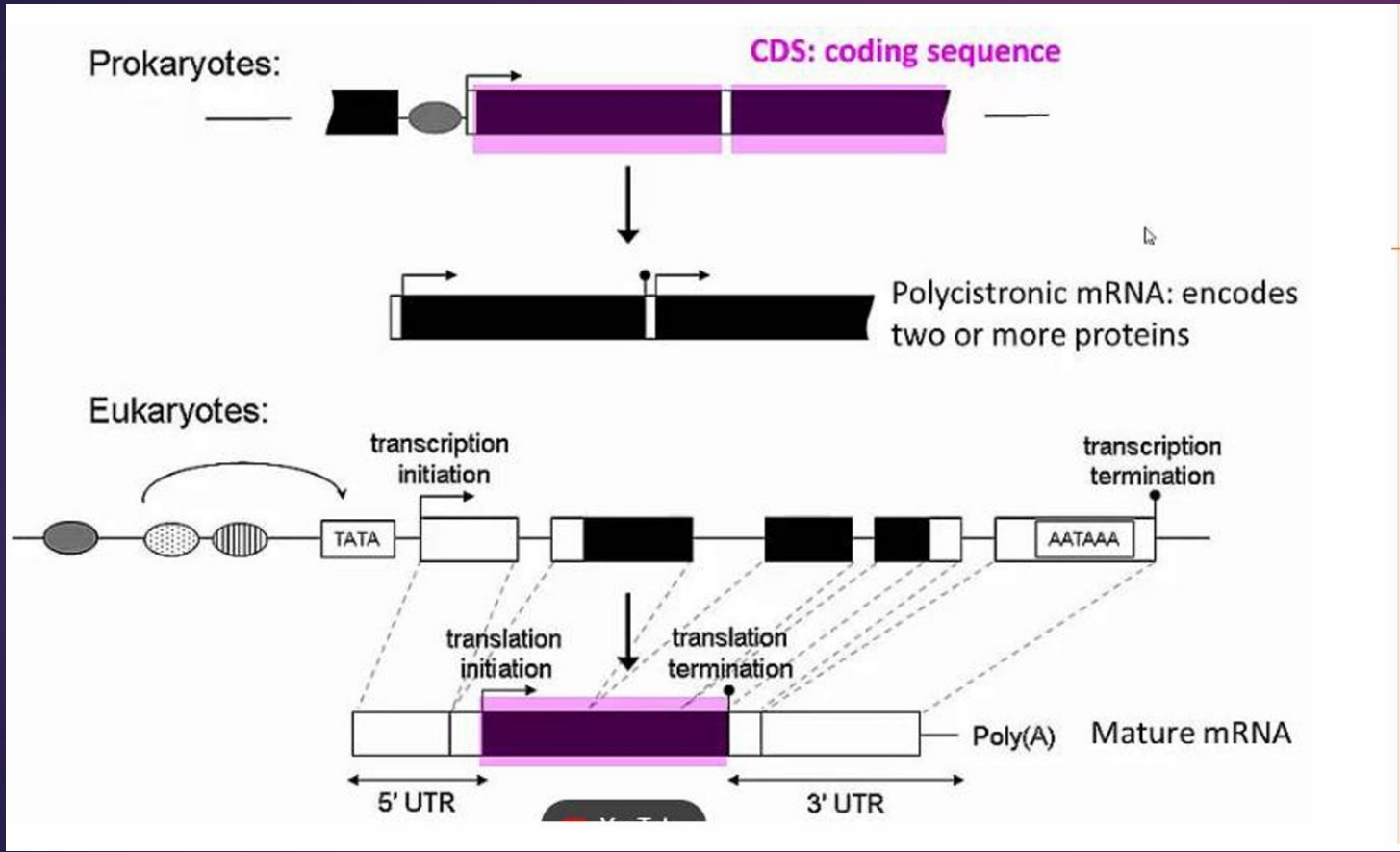
Genome components

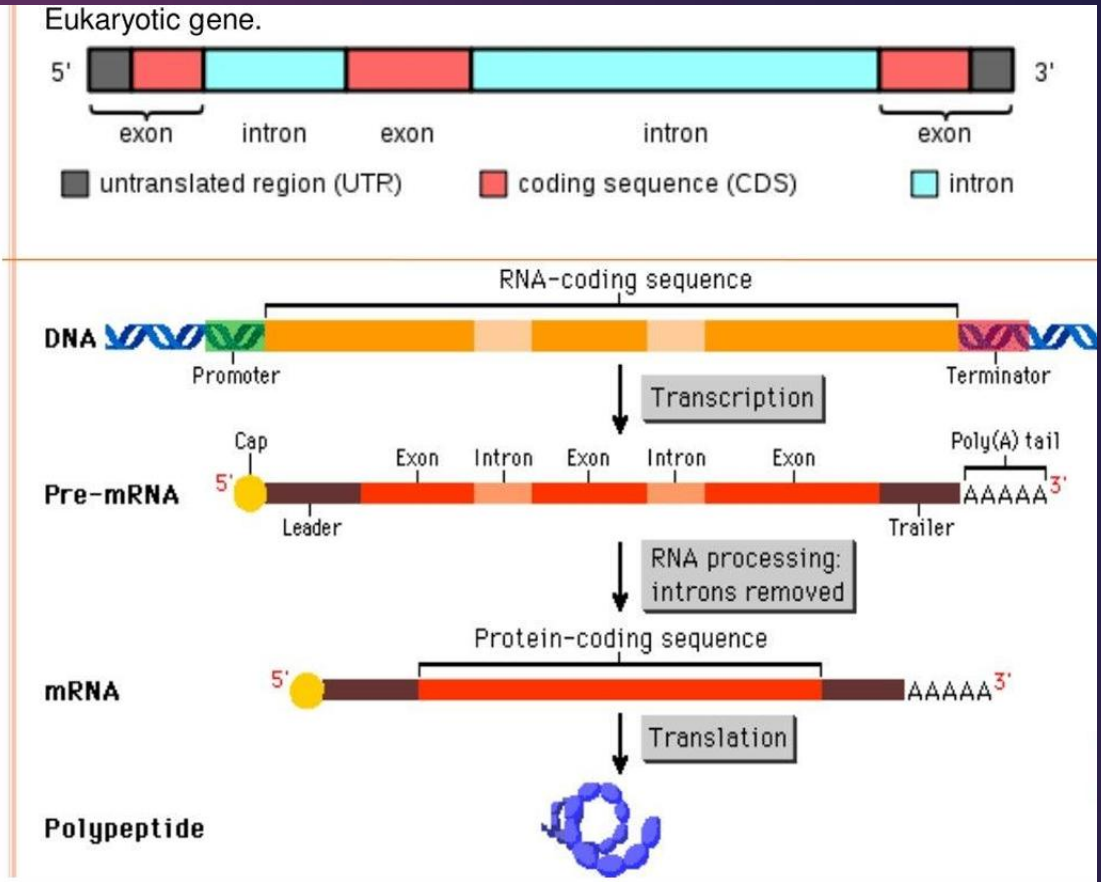
- The human genome contains less than 2% coding exons within genes.
- The remaining DNA consist of:
  - introns and regulatory sequences such as enhancers
  - unique noncoding DNA contains many pseudogenes (genes that have accumulated mutations and became nonfunctional)
  - repetitive DNA are sequences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



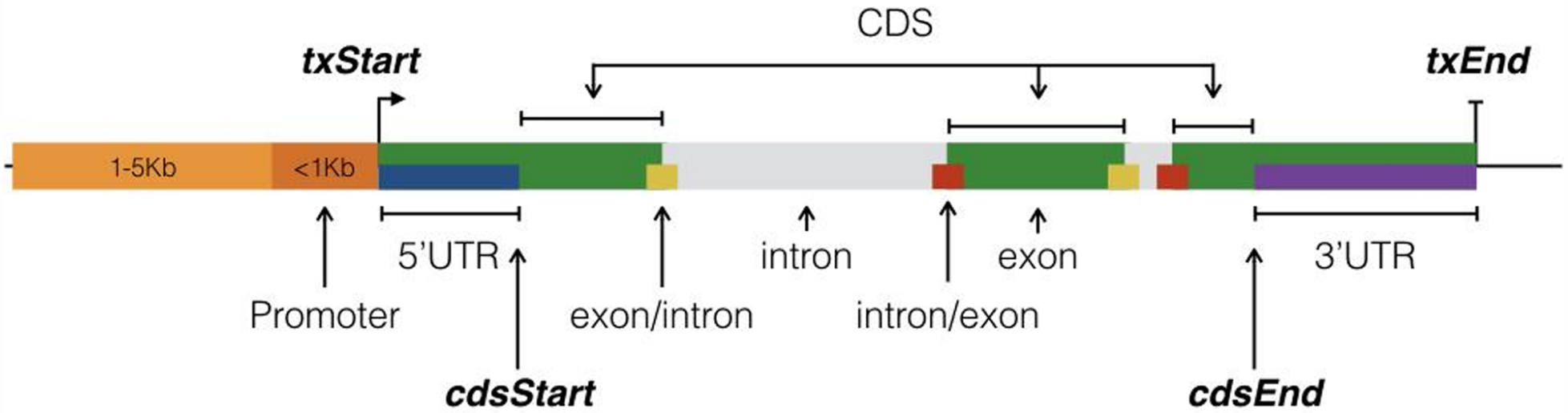
human_genome.html 21_07HumanDNATypes-L.jpg (nicerweb.com)

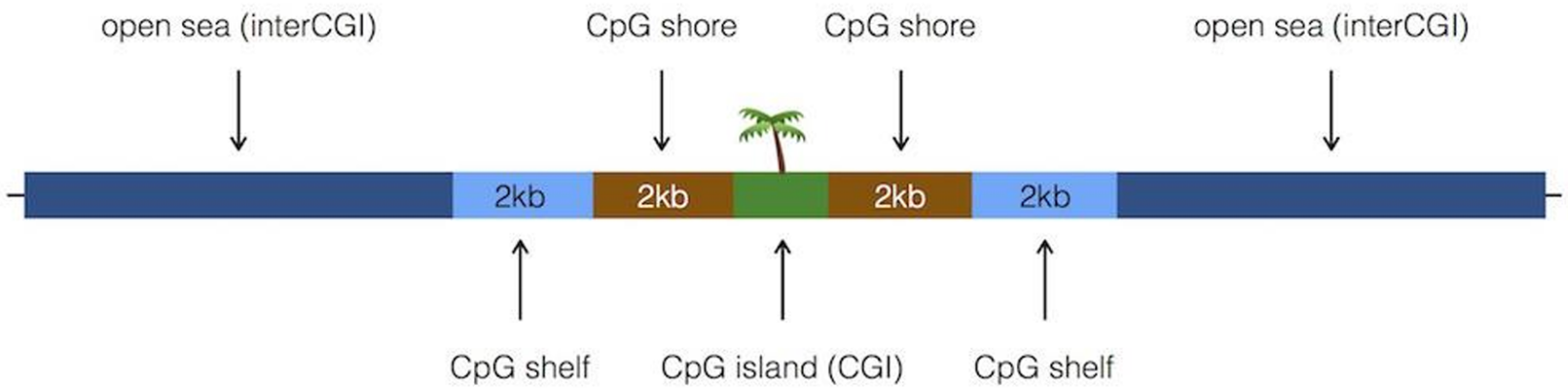
Gene structure





Gene structure





https://www.bioconductor.org/packages/devel/bioc/vignettes/annotatr/inst/doc/annotatr-vignette.html

Understanding Gene Annotation through GENCODE (https://www.gencodegenes.org/)

GENCODE

Human | Mouse | How to access data | FAQ | Documentation | About us

HUMAN

GENCODE 47 (October 2024)

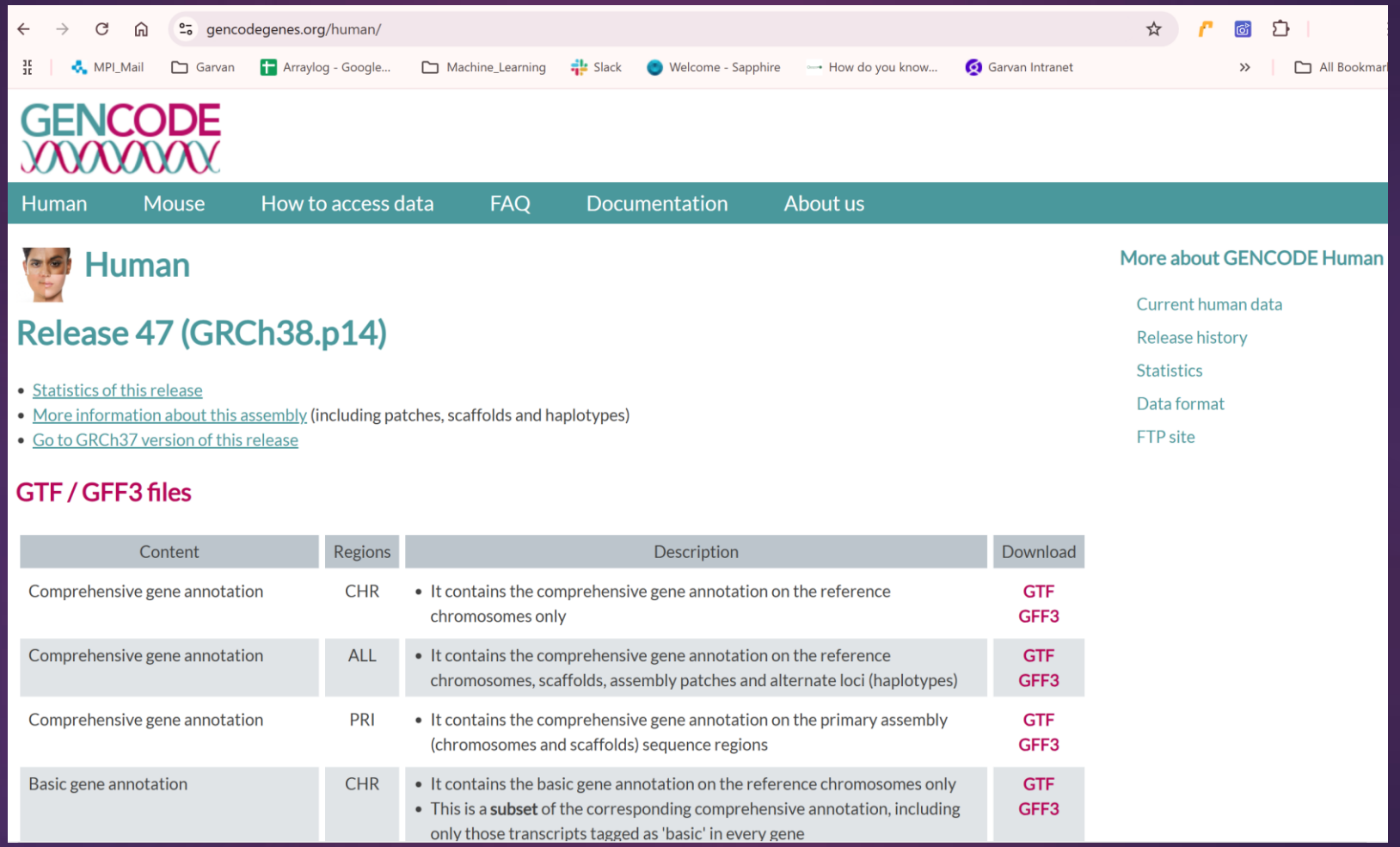
MOUSE

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



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 "D
LK1"; 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"; lev
el 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"; lev
el 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 "ENST0000039[illegible].9"; gene_
type "protein_coding"; gene_name "DLK1"; transcript_type "protein_coding"; transcript_name "DLK1-203"; exon_number 3; exon_id "ENSE0000[illegible]8.1"; lev
el 2; protein_id "ENSP00000376589.5"; transcript_support_level "4"; hgnc_id "HGNC:2907"; tag "alternative_5_UTR"; tag "mRNA_end_NF"; t[illegible]end_NF";
 havana_gene "OTTHUMG00000171600.5"; havana_transcript "OTTHUMT00000414388.2";
```

https://www.youtube.com/watch?v=2Y7zxh1BiFs

Understanding Gene Annotation through GENCODE

https://www.youtube.com/watch?v=2Y7zxh1BiFs

Understanding Gene Annotation through GENCODE

https://www.youtube.com/watch?v=2Y7zxh1BiFs

Understanding Gene Annotation through GENCODE

https://www.youtube.com/watch?v=2Y7zxh1BiFs

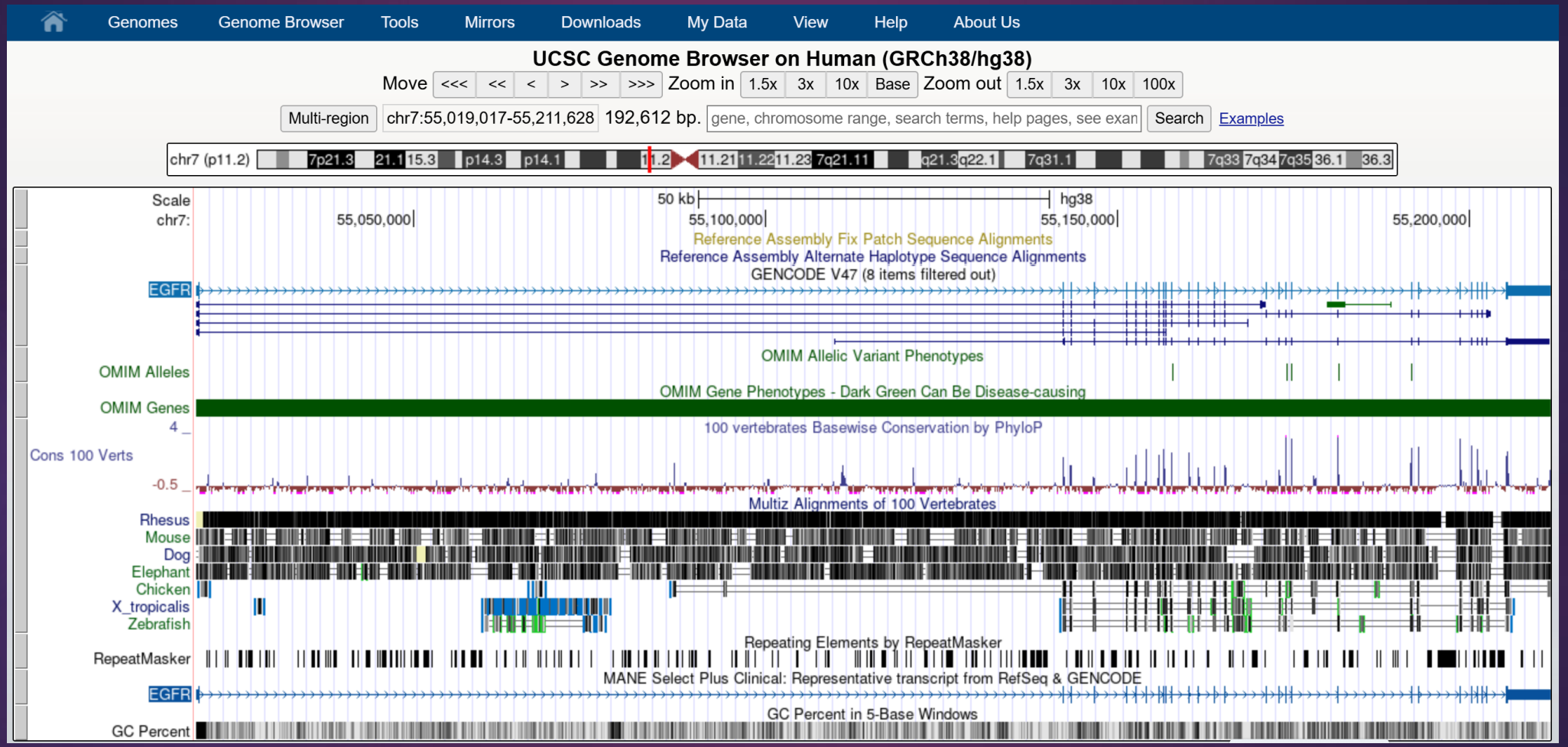
Understanding Gene Annotation through GENCODE

https://www.youtube.com/watch?v=2Y7zxh1BiFs

Introduction to UCSC Genome Browser

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

Introduction to UCSC Genome Browser



https://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr7%3A55019017%2D55211628&hgsid=2410221991_gQpMUpHGCVZMeYat2VhYyvvaQEGY

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.

https://genome.ucsc.edu/training/

Xin chân thành cảm ơn!

LUU PHUC LOI, PHD

ZALO: 0901802182

LUU.P.LOI@GOOGLEMAIL.COM