

Cancer Genomics & Bioinformatics

Day 1: Introduction to biological databases

Aristeidis G. Telonis, PhD

Research Assistant Professor

Department of Biochemistry & Molecular Biology

Sylvester Comprehensive Cancer Center

University of Miami Miller school of Medicine

<https://github.com/ceccarellilab/CancerBioinformaticsCourse>



Learning objectives

Databases:

- How is biological information organized in databases?
- What are some commonly used biological databases?
- How can I access genomic data?

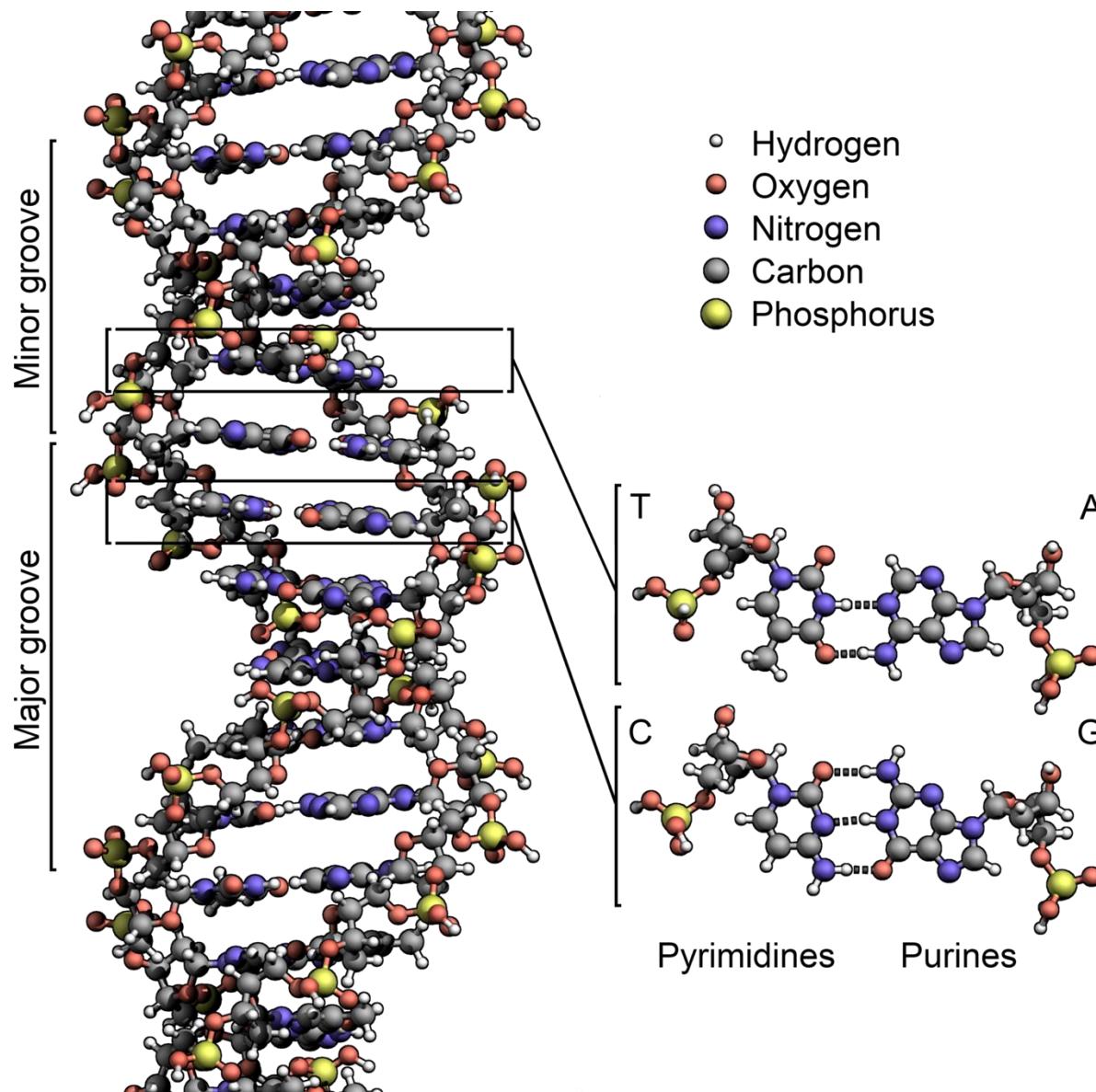
How do I use linux to answer questions like:

- Which chromosome has the most coding genes?
- Which coding gene has the most coding splice variants?
- Which coding gene has the most non-coding variants?

@SylvesterCancer



Genome organization: from biology to computer bytes



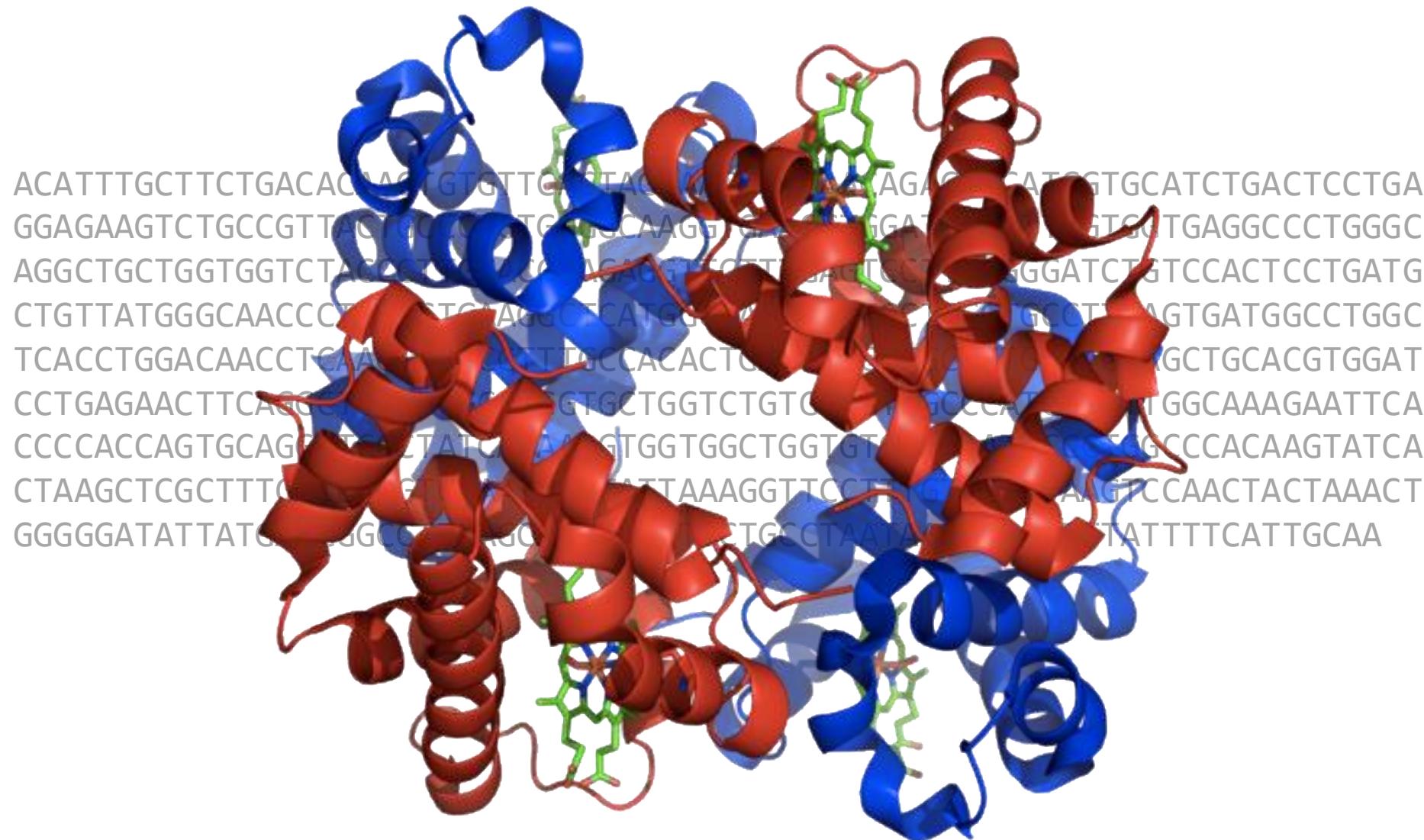
By Zephyris - Own work, CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=15027555>

DNA is a sequence of four letters

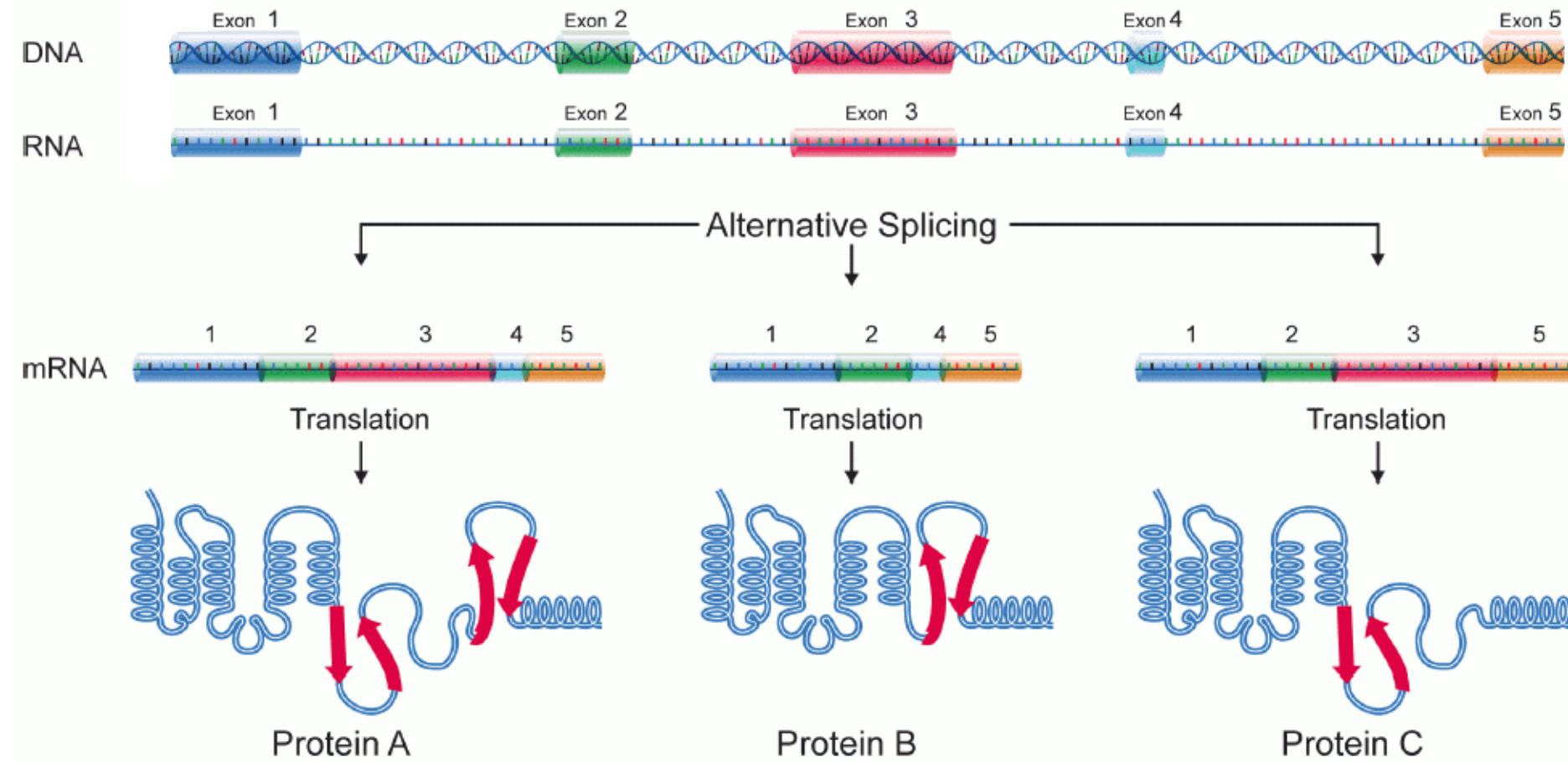
ACATTGCTTCTGACACAACGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCATCTGACTCCTGA
GGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGC
AGGCTGCTGGTGGTCTACCCTGGACCCAGAGGTTCTTGAGTCCTTGGGATCTGTCCACTCCTGATG
CTGTTATGGGCAACCCTAACGGTGAAGGCTCATGGCAAGAAAGTGCCTGGTGCCTTAGTGATGGCCTGGC
TCACCTGGACAACCTCAAGGGCACCTTGCCACACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGAT
CCTGAGAACCTCAGGCTCCTGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTGGCAAAGAATTCA
CCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATCA
CTAAGCTCGCTTCTTGCTGCCAATTCTATTAAAGGTTCCCTTGTCCCTAACGTCCAACTAACACTAAACT
GGGGGATATTATGAAGGGCCTGAGCATCTGGATTCTGCCTAACAAAAAACATTATTCATTGCAA

+ **6.4 billion letters**

DNA is a sequence of four letters



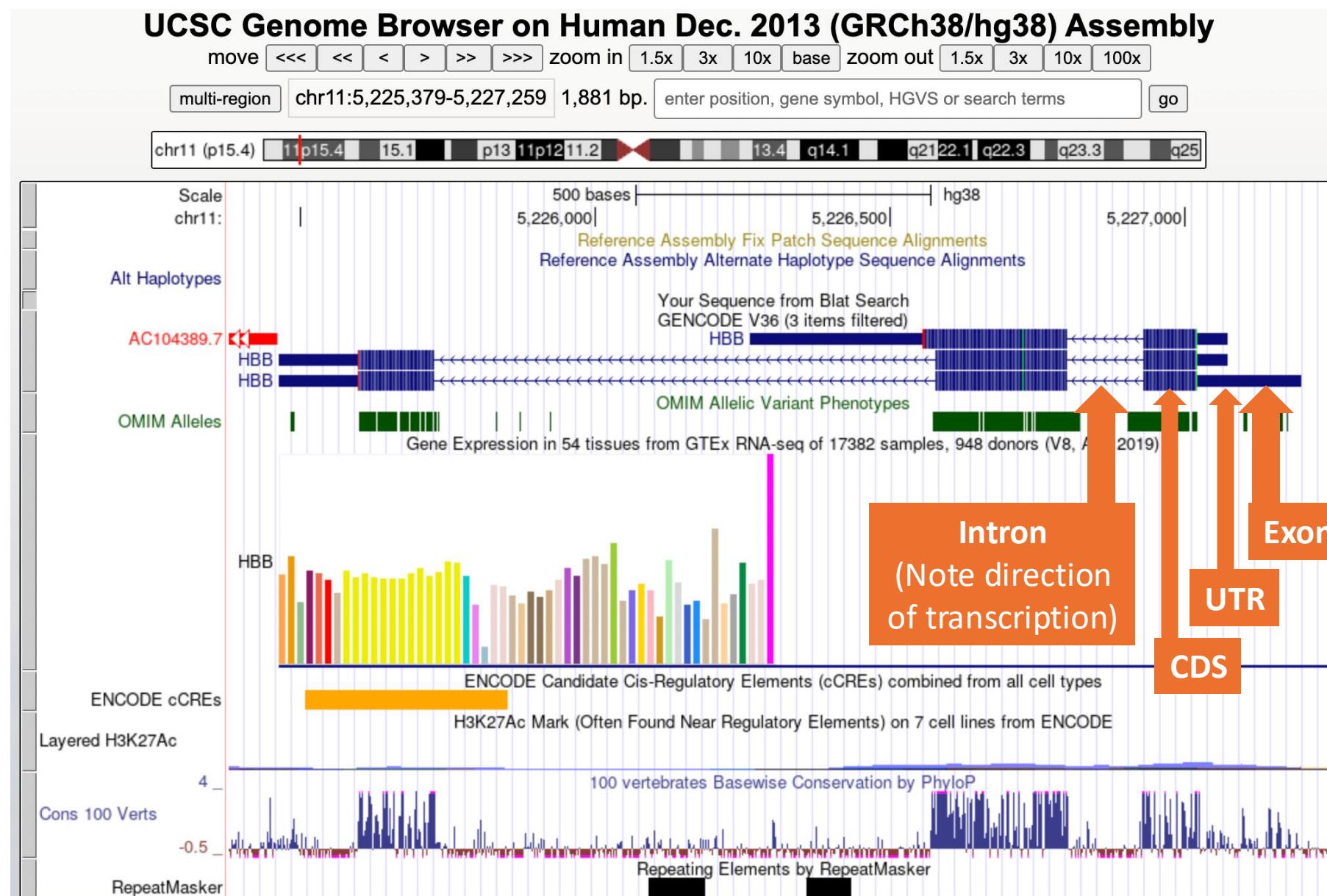
We need to account for the complexity of molecular biology



By National Human Genome Research Institute
http://www.genome.gov/Images/EdKit/bio2j_large.gif, Public Domain
<https://commons.wikimedia.org/w/index.php?curid=2132737>

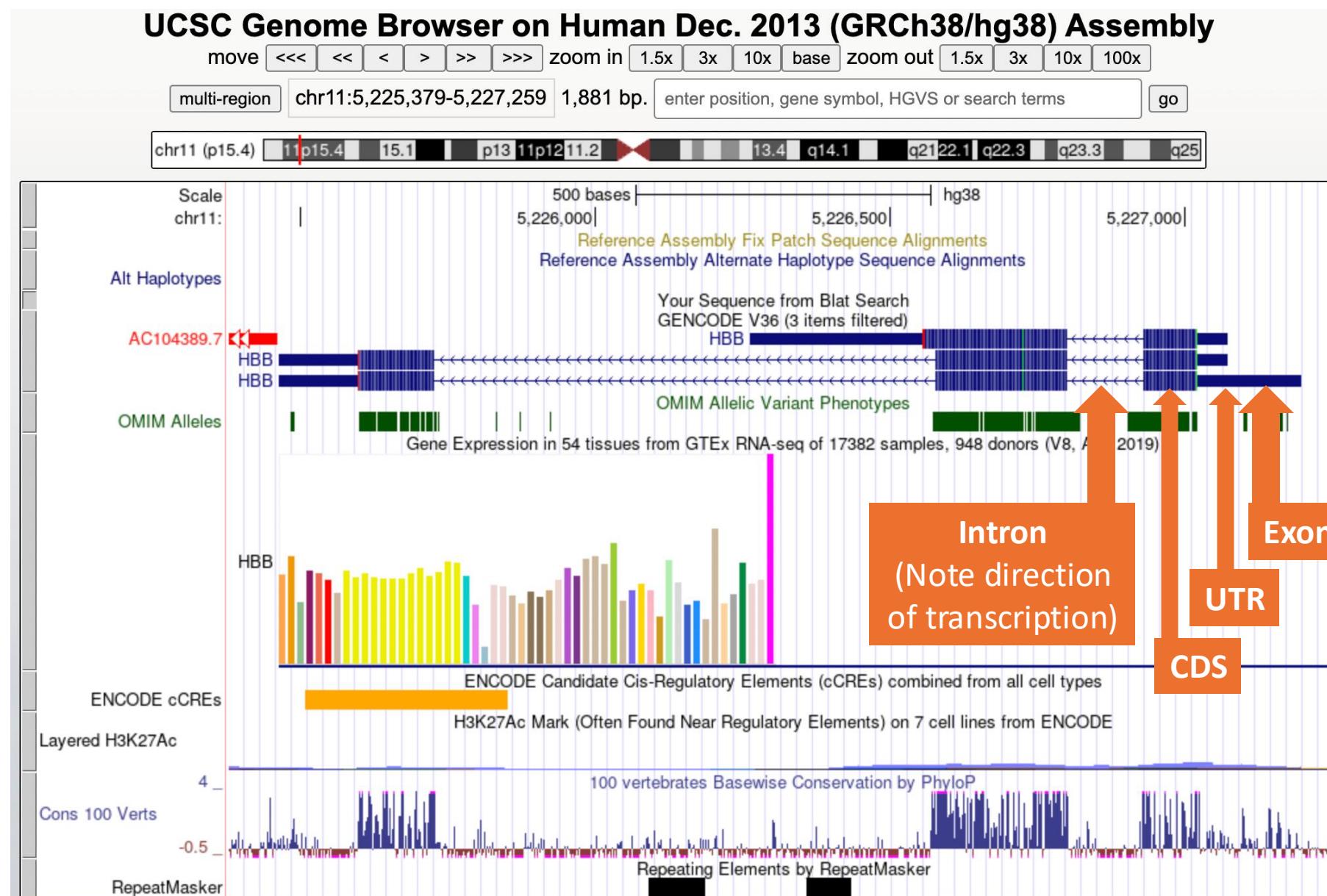
The above mRNAs are protein-coding transcripts. There are a lot of transcripts that are not translated into proteins and are therefore non-coding

The UCSC Genome Browser: view of HBB (Hemoglobin beta)



http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr11%3A5225379%2D5227259&hgsid=1106581317_2aa2ojxb5YEYzbxFYrRt0jEidpa8

The UCSC Genome Browser: view of HBB (Hemoglobin beta)



http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr11%3A5225379%2D5227259&hgsid=1106581317_2aa2ojxb5YEYzbxFYrRt0jEidpa8

We know a LOT of things

- Feel free to explore the databases:

UCSC

http://genome.ucsc.edu/cgi-bin/hgGene?hg_gene=ENST00000335295.4&hg_chrom=chr11&hg_start=5225463&hg_end=5227071&hg_type=knownGene&db=hg38

ENSEMBL

http://useast.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000244734;r=11:5225464-5229395

NCBI

https://www.ncbi.nlm.nih.gov/gene?cmd=retrieve&dopt=default&rn=1&list_uids=3043

We know the genome sequence

We know the protein coding genes (location, size, exons, introns)

We know the non-coding genes

→ http://useast.ensembl.org/Homo_sapiens/Info/Annotation

We know a LOT of things, we can access a LOT of information

Gene counts (Primary assembly)

Coding genes	19,871 (excl 661 readthrough)
Non coding genes	42,126
Small non coding genes	4,866
Long non coding genes	35,044 (excl 301 readthrough)
Misc non coding genes	2,216
Pseudogenes	15,198 (excl 1 readthrough)
Gene transcripts	387,954

http://www.ensembl.org/Homo_sapiens/Info/Annotation

An official website of the United States government [Here's how you know](#)

National Library of Medicine
National Center for Biotechnology Information

Gene Advanced Log in

Full Report Hide sidebar >>

HBB hemoglobin subunit beta [*Homo sapiens* (human)]

Gene ID: 3043, updated on 3-May-2025

Summary

Official Symbol HBB provided by HGNC
Official Full Name hemoglobin subunit beta provided by HGNC
Primary source HGNC:HGNC:4827
See related Ensembl:ENSG00000244734 MIM:141900; AllianceGenome:HGNC:4827
Gene type protein coding
RefSeq status REVIEWED
Organism *Homo sapiens*
Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Homo
Also known as ECYT6; CD113t-C; beta-globin
Summary The alpha (HBA) and beta (HBB) loci determine the structure of the 2 types of polypeptide chains that form the hemoglobin tetramer. The beta-globin gene locus consists of two alpha chains and two beta chains. Mutant beta globin causes sickle cell disease. Reduced amounts of detectable beta globin causes beta-plus-thalassemia. Gamma-G -- gamma-A -- delta -- beta--3'. [provided by RefSeq, Jul 2008]
Orthologs [all](#)
NEW Try the new [Gene table](#)
Try the new [Transcript table](#)

Genomic context

Location: 11p15.4
Exon count: 3

<https://www.ncbi.nlm.nih.gov/gene/3043>

Table of contents Go to the HIV-1, Human Interaction Database

Pathways from PubChem

14 items

Transport of small molecules
Data Source: Reactome External ID: R-HSA-382551
Taxonomy Name: *Homo sapiens* (human)

O₂/CO₂ exchange in erythrocytes
Data Source: Reactome External ID: R-HSA-1480926
Taxonomy Name: *Homo sapiens* (human)

Erythrocytes take up carbon dioxide and release oxygen
Data Source: Reactome External ID: R-HSA-1237044
Taxonomy Name: *Homo sapiens* (human)

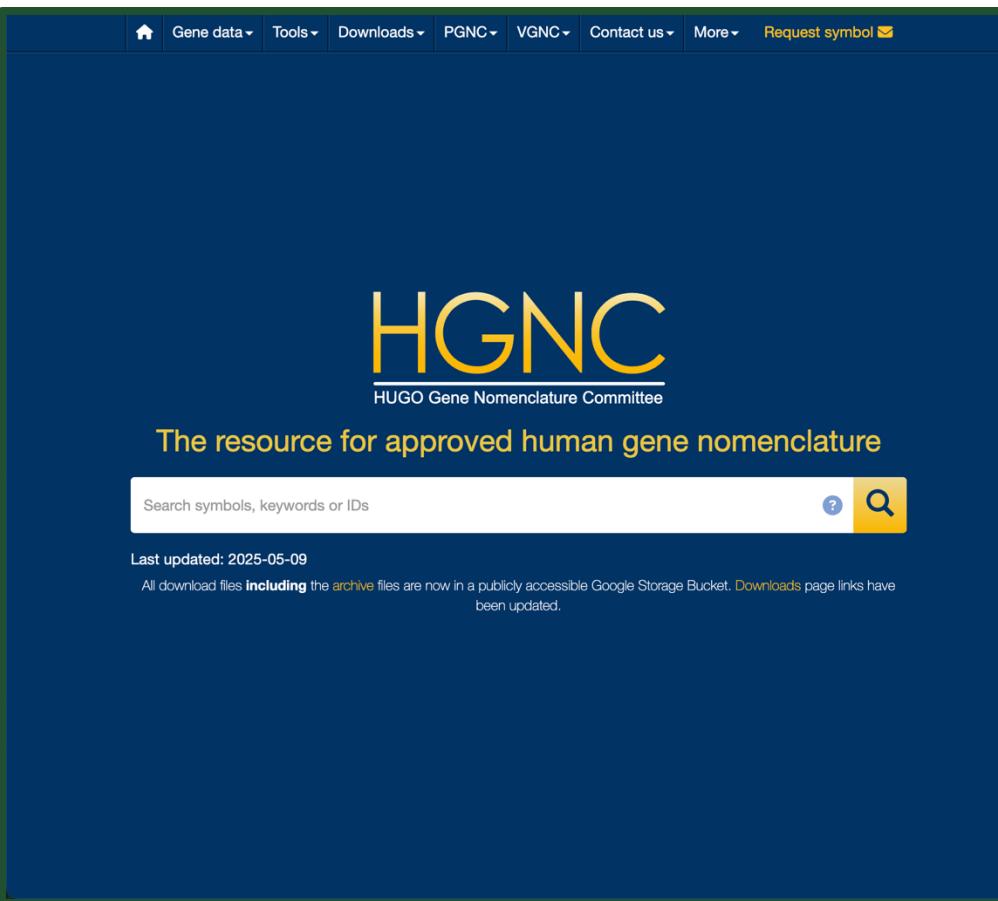
Erythrocytes take up oxygen and release carbon dioxide
Data Source: Reactome External ID: R-HSA-1247673
Taxonomy Name: *Homo sapiens* (human)

Vesicle-mediated transport
Data Source: Reactome External ID: R-HSA-5653656
Taxonomy Name: *Homo sapiens* (human)

Binding and Uptake of Ligands by Scavenger Receptors
Data Source: Reactome External ID: R-HSA-2173782
Taxonomy Name: *Homo sapiens* (human)

Scavenging of heme from plasma
Data Source: Reactome External ID: R-HSA-2168880
Taxonomy Name: *Homo sapiens* (human)

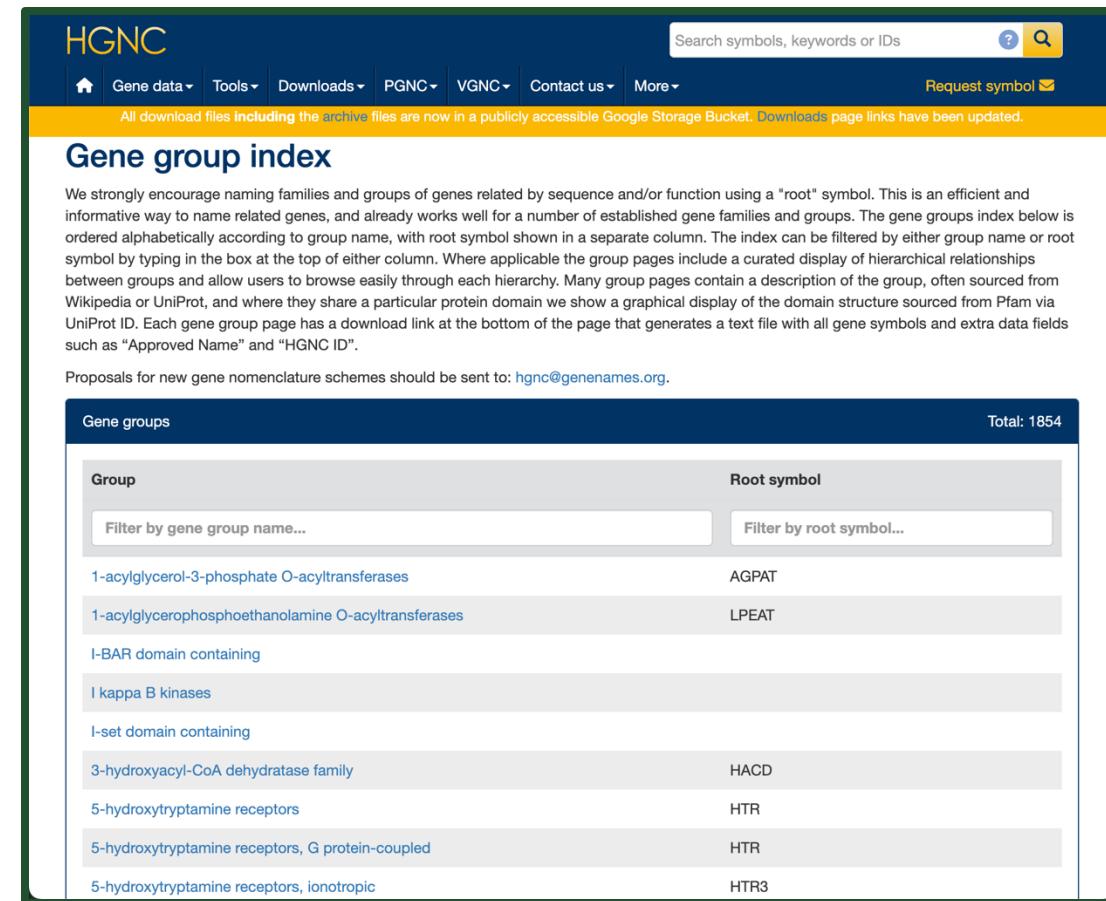
Hemostasis
Data Source: Reactome External ID: R-HSA-109582
Taxonomy Name: *Homo sapiens* (human)



The HGNC homepage features a large yellow "HGNC" logo at the top left. Below it is the text "HUGO Gene Nomenclature Committee". A sub-headline reads "The resource for approved human gene nomenclature". A search bar at the top right contains the placeholder "Search symbols, keywords or IDs" and includes a magnifying glass icon. At the bottom left, a "Last updated: 2025-05-09" message is displayed, followed by a note about download files being available in a Google Storage Bucket.

<https://www.genenames.org/>

Gene data > Gene group reports



The "Gene group index" page has a header with the HGNC logo and navigation links. A banner at the top states: "All download files including the archive files are now in a publicly accessible Google Storage Bucket. Downloads page links have been updated." The main content area is titled "Gene group index" and contains a detailed description of gene naming conventions. It includes a table titled "Gene groups" with two columns: "Group" and "Root symbol". The table lists several gene groups with their corresponding root symbols.

Group	Root symbol
1-acylglycerol-3-phosphate O-acyltransferases	AGPAT
1-acylglycerophosphoethanolamine O-acyltransferases	LPEAT
I-BAR domain containing	
I kappa B kinases	
I-set domain containing	
3-hydroxyacyl-CoA dehydratase family	HACD
5-hydroxytryptamine receptors	HTR
5-hydroxytryptamine receptors, G protein-coupled	HTR
5-hydroxytryptamine receptors, ionotropic	HTR3

Gene groups

Total: 8

Group	Root symbol
acetylase	AADAC
Arylacetamide deacetylase family	AADAC
Histone deacetylase superfamily	HDAC, SIRT
Histone deacetylases, class IIA	HDAC
Histone deacetylases, class IIB	HDAC
Histone deacetylases, class I	HDAC
Histone deacetylases, class II	HDAC
Histone deacetylases, class IV	HDAC
SIN3 histone deacetylase complex subunits	

Filter by root symbol...

10 25 50 100

Gene group: Histone deacetylases, class I (HDAC) [?](#)

A subgroup of [?](#): "Histone deacetylase superfamily (HDAC, SIRT)"

Gene group hierarchy map [?](#)

Rearrange mode

```

graph LR
    HDAC_Superfamily[Histone deacetylase superfamily] --- HDAC_Class_I[Histone deacetylases, class I]
  
```

The mapped domains of Q13547, encoded by the HDAC1 gene, an example gene within the group. [Source: Pfam & UniProt]

Gene: HDAC1, UniProt: Q13547

482 amino acids

Genes contained within the group: 4 [?](#) [Download](#)

HGNC ID (gene)	Approved symbol	Approved name	Previous symbols	Aliases	Chromosome
HGNC:4852	HDAC1	histone deacetylase 1	RPD3L1	HD1, GON-10, KDAC1	1p35.2-p35.1
HGNC:4853	HDAC2	histone deacetylase 2	RPD3, YAF1, KDAC2		6q21
HGNC:4854	HDAC3	histone deacetylase 3	RPD3, HD3, RPD3-2, KDAC3		5q31.3
HGNC:13315	HDAC8	histone deacetylase 8	HDACL1, WTS, MRXS6	RPD3, KDAC8	Xq13.1

Downloads [?](#)

Genes within the current group only: [TXT](#) [JSON](#)

[back to index ...](#)



Human Mouse How to access data FAQ Documentation About us



Statistics about the GENCODE Release 48

The statistics derive from the [gtf file](#) that contains only the annotation of the main chromosomes.

For details about the calculation of these statistics please see the [README_stats.txt file](#).

General stats

Total No of Genes	78686	Total No of Transcripts	385669
Protein-coding genes	19435	Protein-coding transcripts	89843
- readthrough genes (not included)	661	- full length protein-coding	65024
Long non-coding RNA genes	35901	- partial length protein-coding	24819
Small non-coding RNA genes	7563	Nonsense mediated decay transcripts	21902
Pseudogenes	14695	Long non-coding RNA loci transcripts	191076
- processed pseudogenes	10643		
- unprocessed pseudogenes	3549		
- unitary pseudogenes	266		
Immunoglobulin/T-cell receptor gene segments		Total No of distinct translations	65814
- protein coding segments	412	Genes that have more than one distinct translations	13648
- pseudogenes	237		

More about GENCODE Human



Human Mouse How to access data FAQ Documentation About us

The GENCODE Project: Encyclopædia of genes and gene variants

Background

The National Human Genome Research Institute (NHGRI) launched a public research consortium named [ENCODE](#), the Encyclopedia Of DNA Elements, in September 2003, to carry out a project to identify all functional elements in the human genome sequence. After a successful pilot phase on 1% of the genome, the scale-up to the entire genome is now underway. The Wellcome Sanger Institute was [awarded a grant](#) to carry out a scale-up of the GENCODE project for integrated annotation of gene features.

Having been involved in successfully delivering the definitive annotation of functional elements in the human genome, the GENCODE group were [awarded a second grant](#) in 2013 in order to continue their human genome annotation work and expand GENCODE to include annotation of the mouse genome. A [third grant](#) was awarded in 2017 for the continued improvement of the annotation of the human and mouse genomes, and a fourth grant followed in 2021. Details of the grants awarded can be found [here](#).

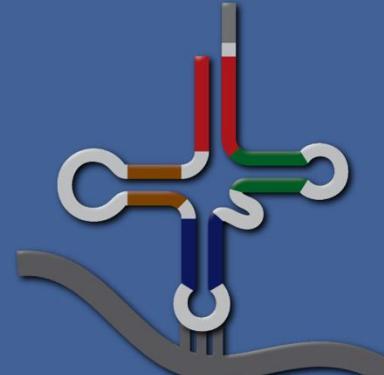
GENCODE gene annotation is used by a large number of international consortia.

Home

GtRNAdb

tRNAscan-SE analysis of complete genomes

The genomic tRNA database contains tRNA gene predictions made by **tRNAscan-SE** on complete or nearly complete genomes. Unless otherwise noted, all annotation is automated, and has not been inspected for agreement with published literature.



Data Release 22 (Sept 2024)

To access the previous releases of GtRNAdb or check out the change log, please click [here](#)

Links to Most Viewed Genomes

<i>Homo sapiens</i> (GRCh37/hg19)	<i>Homo sapiens</i> (GRCh38/hg38)
<i>Mus musculus</i> (GRCm38/mm10)	<i>Mus musculus</i> (GRCm39/mm39)
<i>Saccharomyces cerevisiae</i> S288c	<i>Schizosaccharomyces pombe</i> 972h-
<i>Caenorhabditis elegans</i>	<i>Drosophila melanogaster</i> (BDGP Rel.)

Additional tRNA Tools

- [tDRnamer](#) - standardized naming for tRNA-derived RNAs
- [tRAX](#) - tRNA and tRNA-derived RNA sequencing data analysis

miRBase

Home Search Browse Help Downloads 

miRBase: the microRNA database
the archive for microRNA sequences and annotations

[More about miRBase →](#)

 By identifier or keyword

 By genomic location

 Browse

 By tissue expression

 By sequence

 Downloads

MANCHESTER 1824
The University of Manchester

Comments or questions? Email mirbase@manchester.ac.uk

<https://gtrnadb.ucsc.edu/>

<https://www.mirbase.org/>

Find your protein

UniProtKB

Examples: Insulin, APP, Human, P05067, organism_id:9606

UniProt is the world's leading high-quality, comprehensive and freely available protein knowledgebase.

Cite


<https://www.uniprot.org/>

UniProtKB BLAST Align Peptide search ID mapping SPARQL Advanced | List Search Help

Q12888 · TP53B_HUMAN

Function	Protein ⁱ	TP53-binding protein 1	Amino acids	1972 (go to sequence)
Names & Taxonomy	Gene ⁱ	TP53BP1	Protein existence ⁱ	Evidence at protein level
Subcellular Location	Status ⁱ	UniProtKB reviewed (Swiss-Prot)	Annotation score ⁱ	5/5
Disease & Variants	Organism ⁱ	Homo sapiens (Human)		
PTM/Processing				
Expression				
Interaction				
Structure				
Family & Domains				
Sequence & Isoforms				
Similar Proteins				

Entry Variant viewer 2,324 Feature viewer Genomic coordinates Publications External links History

Tools Download Add Add a publication Entry feedback

Functionⁱ

Double-strand break (DSB) repair protein involved in response to DNA damage, telomere dynamics and class-switch recombination (CSR) during antibody genesis (PubMed:12364621, PubMed:17190600, PubMed:21144835, PubMed:22553214, PubMed:23333306, PubMed:27153538, PubMed:28241136, PubMed:31135337, PubMed:37696958). Plays a key role in the repair of double-strand DNA breaks (DSBs) in response to DNA damage by promoting non-homologous end joining (NHEJ)-mediated repair of DSBs and specifically counteracting the function of the homologous recombination (HR) repair protein BRCA1 (PubMed:22553214, PubMed:23333306, PubMed:23727112, PubMed:27153538, PubMed:31135337). In response to DSBs, phosphorylation by ATM promotes interaction with RIF1 and dissociation from NUDT16L1/TIRR, leading to recruitment to DSBs sites (PubMed:28241136). Recruited to DSBs sites by recognizing and binding histone H2A monoubiquitinated at 'Lys-15' (H2AK15Ub) and histone H4

Feedback Help

PATHWAYS | KEGG as a reference database for metabolic, signaling and organismal processes

KEGG

Databases

Tools

Auto annotation

Kanehisa Lab



KEGG PATHWAY Database

Wiring diagrams of molecular interaction networks

KEGG2 PATHWAY BRITE MODULE KO GENES

Select prefix

Enter keywords

Pathway Maps

KEGG PATHWAY is a collection of manually drawn pathway reaction and relation networks for:

1. Metabolism
Global/overview Carbohydrate Energy Lipid Nucleic acid
Cofactor/vitamin Terpenoid/PK Other secondary metabolites
2. Genetic Information Processing
3. Environmental Information Processing
4. Cellular Processes
5. Organismal Systems
6. Human Diseases
7. Drug Development

The pathway map viewer linked from this page is a part of

Pathway Identifiers

Each pathway map is identified by the combination of 2-4 letters and numbers starting with the following:
map manually drawn reference pathway

Annotation
ko reference pathway highlighting KO identifiers

ec reference metabolic pathway highlighting EC numbers

rn reference metabolic pathway highlighting reaction numbers

<org> organism-specific pathway generated by conversion

vg viruses pathway generated by converting KOs to viruses

vx viruses extended pathway generated by including

and the numbers starting with the following:

011 global map (lines linked to KOs)

012 overview map (lines linked to KOs)

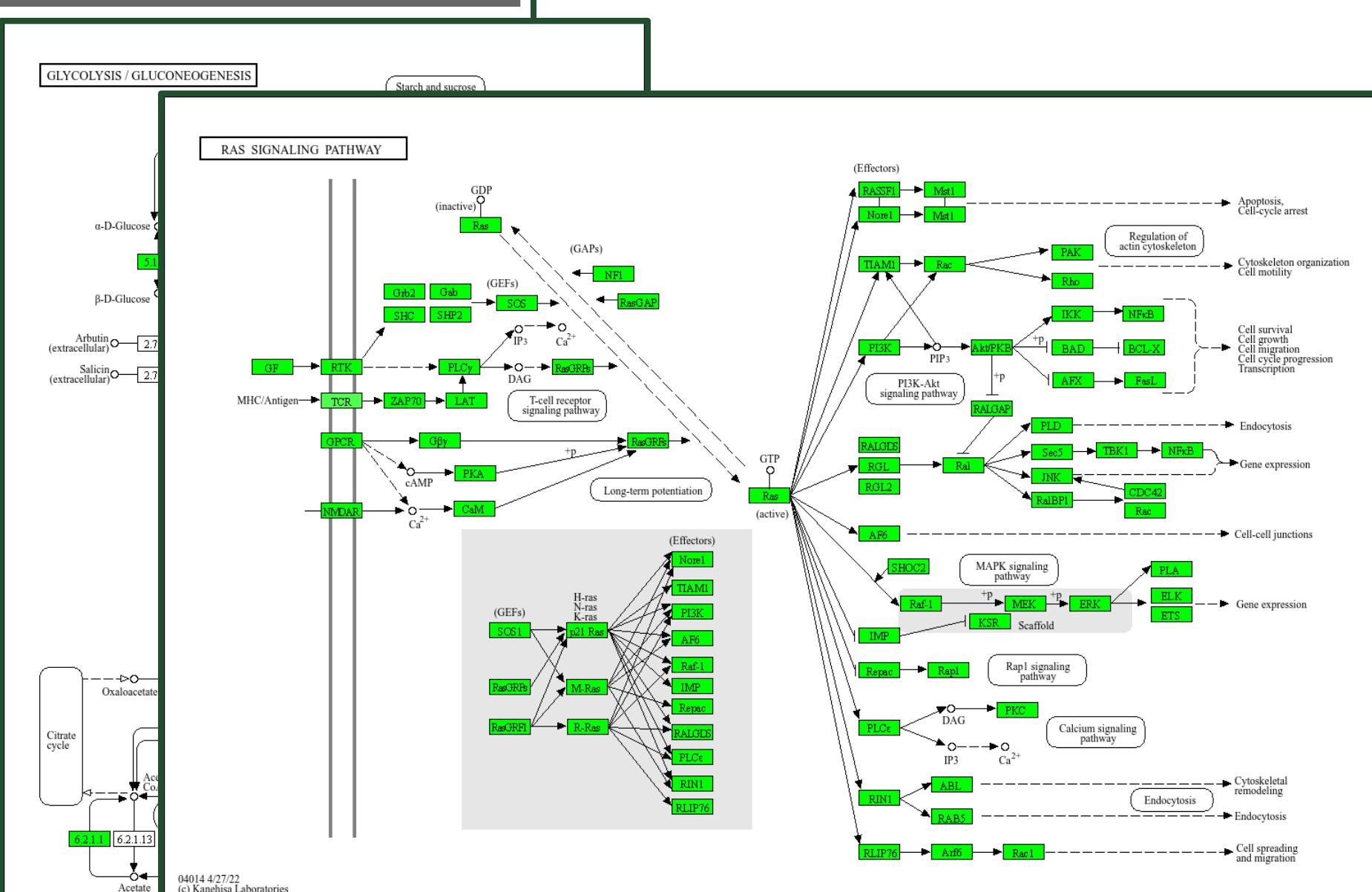
013 higher (multi-organism) level overview map (lines linked to KOs)

010 chemical structure map (no KO expansion)

07 drug structure map (no KO expansion)

other regular map (boxes linked to KOs)

are used for different types of maps.



PATHWAYS | The Molecular Signature Database has a rich collection across multiple biological conditions



GSEA Home Downloads Molecular Signatures Database Documentation Contact Team

MSigDB Home

Human Collections

- About
- Browse
- Search
- Investigate
- Gene Families

Mouse Collections

- About
- Browse
- Search
- Investigate

Help



Overview

The Molecular Signatures Database (MSigDB) is a resource of tens of thousands of annotated gene sets for use with GSEA software, divided into **Human** and **Mouse** collections. From this web site, you can

- Examine a gene set and its annotations. See, for example, the [HALLMARK_APOPTOSIS](#) human gene set page.
- Browse gene sets by name or collection.
- Search for gene sets by keyword.
- Investigate gene sets:
 - Compute overlaps between your gene set and gene sets in MSigDB.
 - Categorize members of a gene set by gene families.
 - View the expression profile of a gene set in a provided public expression compendia.
 - Investigate the gene set in the online **biological network repository** [NDEx](#)
- Download gene sets.

License Terms

GSEA and MSigDB are available for use under [these license terms](#).

Please [register](#) to download the GSEA software and the MSigDB gene sets, and to use our web tools. After registering, you can log in at any time using your email address. Registration is free. Its only purpose is to help us track usage for reports to our funding agencies.

Current Version

UC San Diego

BROAD
INSTITUTE

<https://www.gsea-msigdb.org/gsea/msigdb/>

Human Collections

- H** hallmark gene sets are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.
- C1** positional gene sets corresponding to human chromosome cytogenetic bands.
- C2** curated gene sets from online pathway databases, publications in PubMed, and knowledge of domain experts.
- C3** regulatory target gene sets based on gene target predictions for microRNA seed sequences and predicted transcription factor binding sites.
- C4** computational gene sets defined by mining large collections of cancer-oriented expression data.

Mouse Collections

- MH** mouse-ortholog hallmark gene sets are versions of gene sets in the MSigDB Hallmarks collection mapped to their mouse orthologs.

- C2** (curated gene sets, 7411 gene sets)
- CP** (chemical and genetic perturbations, 3494 gene sets)
- CP** (canonical pathways, 3917 gene sets)
 - CP:BIOCARTA** (BioCarta gene sets, 292 gene sets)
 - CP:KEGG_MEDICUS** (KEGG Medicus gene sets, 658 gene sets)
 - CP:PID** (PID gene sets, 196 gene sets)
 - CP:REACTOME** (Reactome gene sets, 1736 gene sets)
 - CP:WIKIPATHWAYS** (WikiPathways gene sets, 830 gene sets)
 - CP:KEGG_LEGACY** (KEGG Legacy gene sets, 186 gene sets)

BECKER_TAMOXIFEN_RESISTANCE_DN	IIIZUKA_LIVER_CANCER_PROGRESSION_L0	PYEON_CANCER_HEAD_AND_NECK_VS_CERV
BECKER_TAMOXIFEN_RESISTANCE_UP	L1_DN	CAL_DN
BEGUM_TARGETS_OF_PAX3_FOXO1_FUSION_AND_PAX3	IIIZUKA_LIVER_CANCER_PROGRESSION_L0	PYEON_CANCER_HEAD_AND_NECK_VS_CERV
BEGUM_TARGETS_OF_PAX3_FOXO1_FUSION_UP	L1_UP	CAL_UP
BEGUM_TARGETS_OF_PAX3_FOXO1_FUSION_UP	IIIZUKA_LIVER_CANCER_PROGRESSION_L1	PYEON HPV_POSITIVE_TUMORS_DN
BEIER_GLIOMA_STEM_CELL_DN	G1_DN	PYEON HPV_POSITIVE_TUMORS_UP
BEIER_GLIOMA_STEM_CELL_UP	IIKEDA_MIR133_TARGETS_DN	QI_HYPoxIA
BENITEZ_GBM_PROTEASOME_INHIBITION_RESPONSE	IIKEDA_MIR133_TARGETS_UP	QI_HYPoxIA_TARGETS_OF_HIF1A_AND_FOX
BENNETT_SYSTEMIC_LUPUS_ERYTHEMATOSUS	IIKEDA_MIR1 TARGETS_DN	A2
BENPORATH_CYCLING_GENES	IIKEDA_MIR1 TARGETS_UP	QI_PLASMACYTOMA_DN
BENPORATH_EED_TARGETS	IIKEDA_MIR133 TARGETS_DN	QI_PLASMACYTOMA_UP
BENPORATH_ES_1	IIKEDA_MIR133 TARGETS_UP	QUINTENS_EMBRYONIC_BRAIN_RESPONSE_TO_O
BENPORATH_ES_2	IIKEDA_MIR30 TARGETS_DN	RADAeva_RESPONSE_TO_IFNA1_DN
BENPORATH_ES_CORE_NINE	IIKEDA_MIR30 TARGETS_UP	RADAeva_RESPONSE_TO_IFNA1_UP
BENPORATH_ES_CORE_NINE_CORRELATED	IM_SREBF1A TARGETS	RADMACHER_AMI_PROGNOSIS
BENPORATH_ES_WITH_H3K27ME3	INAMURA_LUNG_CANCER_SCC_DN	RAFFEL_VEGFA_TARGETS_DN
BENPORATH_MYC_MAX_TARGETS	INAMURA_LUNG_CANCER_SCC_SUBTYPES_UP	RAGHAVACHARI_PLATELET_SPECIFIC_GENE
BENPORATH_MYC_TARGETS_WITH_EBOX	INAKAWA_STING_SIGNALING	RAHMAN_TP53_TARGETS_PHOSPHORYLATED
BENPORATH_NANOG_TARGETS	ISSHAKA_MLL2 TARGETS	RAMALHO_STEMNESS_DN
BENPORATH_NOS_TARGETS	ITO_PTIG1 TARGETS_DN	RAMALHO_STEMNESS_UP
BENPORATH_OCT4_TARGETS	ITO_PTIG1 TARGETS_UP	RAMASWAMY_METASTASIS_DN
BENPORATH_PRC2_TARGETS	IVANOV_MUTATED_IN_COLON_CANCER	RAMASWAMY_METASTASIS_UP
BENPORATH_PROLIFERATION	IVANOV_HEMATOPOIESIS_EARLY_PROGENITOR	RAMJAUN_APOPTOSIS_BY_TGFBI_VIA_MAPK
BENPORATH_SOX2_TARGETS	IVANOV_HEMATOPOIESIS_INTERMEDIATE_PROGENITOR	1_UP
BENPORATH_SUZ12_TARGETS	IVANOV_HEMATOPOIESIS_LATE_PROGENITOR	RAMJAUN_APOPTOSIS_BY_TGFBI_VIA_SMAD
BERENJENO_ROCK_SIGNALING_NOT_VIA_RHOA_ER_UP	IVANOV_HEMATOPOIESIS_MATURE_CELL	4_UP
BERENJENO_ROCK_SIGNALING_NOT_VIA_RHOA_ER_UP	IVANOV_HEMATOPOIESIS_STEM_CELL	RAMPON_ENRICHED_LEARNING_ENVIRONMEN
BERENJENO_TRANSFORMED_BY_RHOA_DN	IVANOV_HEMATOPOIESIS_STEM_CELL_AND	T_EARLY_DN
BERENJENO_TRANSFORMED_BY_RHOA_FOREVER	IVANOV_HEMATOPOIESIS_STEM_CELL_LON	RAMPON_ENRICHED_LEARNING_ENVIRONMEN
SIBLY_DN	IVANOV_HEMATOPOIESIS_STEM_CELL_G_TERM	T_EARLY_UP
BERENJENO_TRANSFORMED_BY_RHOA_REVER	IVANOV_HEMATOPOIESIS_STEM_CELL_SHO	RAMPON_ENRICHED_LEARNING_ENVIRONMEN
SIBLY_UP	IVANOV_HEMATOPOIESIS_STEM_CELL_RT_TERM	T_LATE_UP
BERENJENO_TRANSFORMED_BY_RHOA_UP	IVANOVSKA_MIR106B TARGETS	RANKIN_ANIOGENIC_TARGETS_OF_VHL_HI
BERGER_MBD2_TARGETS	IWANAGA_CARCINOGENESIS_BY_KRAS_DN	F2A_DN
BERNARD_PPAPDC1B TARGETS_DN	RAO_BOUND_BY_SALL4_ISOFORM_A	RANKIN_ANIOGENIC_TARGETS_OF_VHL_HI
		F2A_UP
		RAO_BOUND_BY_SALL4
		RAO_BOUND_BY_SALL4_ISOFORM_A
		RAO_BOUND_BY_SALL4_ISOFORM_B

Journals have dedicated issues to databases



Volume 53, Issue D1

6 January 2025

[Cover image](#)

EISSN 1362-4962

EDITORIAL

MAJOR MULTI-DATABASE RESOURCES

NUCLEIC ACID SEQUENCE, STRUCTURE AND REGULATION

PROTEIN SEQUENCE AND STRUCTURE, MOTIFS AND DOMAINS

METABOLIC AND SIGNALLING PATHWAYS, ENZYMES

VIRUSES, BACTERIA, PROTOZOA AND FUNGI

HUMAN GENOME, MODEL ORGANISMS, COMPARATIVE GENOMICS

GENOMIC VARIATION, DISEASES AND DRUGS

PLANTS

OTHER DATABASES

CORRECTIONS

JOURNAL ARTICLE

CancerSCEM 2.0: an updated data resource of single-cell expression map across various human cancers

Jingyao Zeng, Zhi Nie, Yunfei Shang, Jialin Mai, Yadong Zhang, Yuntian Yang, Chenle Xu, Jing Zhao, Zhuojing Fan, Jingfa Xiao [Author Notes](#)

Nucleic Acids Research, Volume 53, Issue D1, 6 January 2025, Pages D1278–D1286, <https://doi.org/10.1093/nar/gkae954>

Published: 26 October 2024 [Article history ▾](#)

JOURNAL ARTICLE

canSAR 2024—an update to the public drug discovery knowledgebase

Phillip W Gingrich, Rezvan Chitsazi, Ansuman Biswas, Chunjie Jiang, Li Zhao, Joseph E Tym, Kevin M Brammer, Jun Li, Zhigang Shu, David S Maxwell, Jeffrey A Tacy, Ioan L Mica, Michael Darkoh, Patrizio di Micco, Kaitlyn P Russell, Paul Workman, Bissan Al-Lazikani

Nucleic Acids Research, Volume 53, Issue D1, 6 January 2025, Pages D1287–D1294, <https://doi.org/10.1093/nar/gkae1050>

Published: 13 November 2024 [Article history ▾](#)

JOURNAL ARTICLE

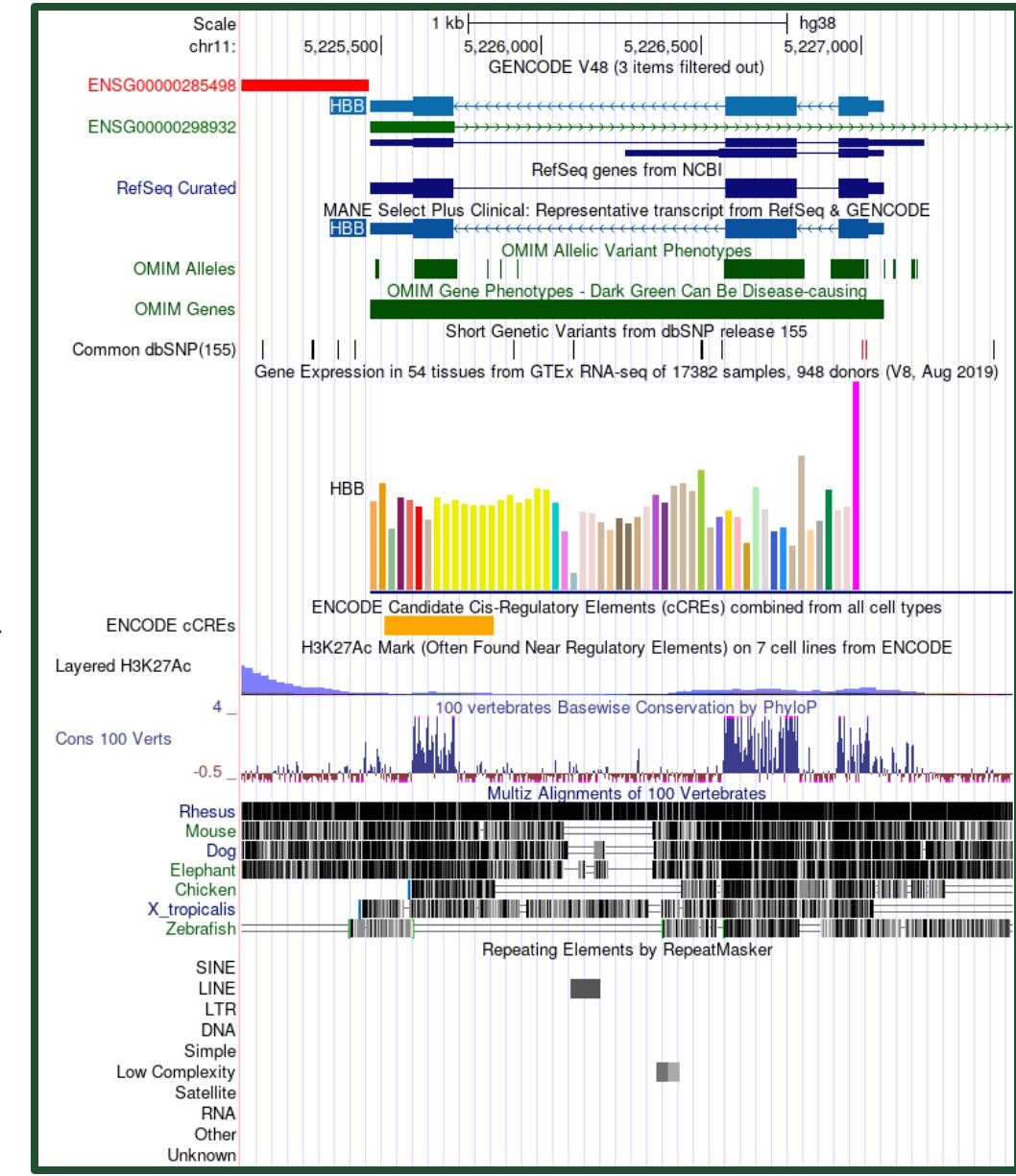
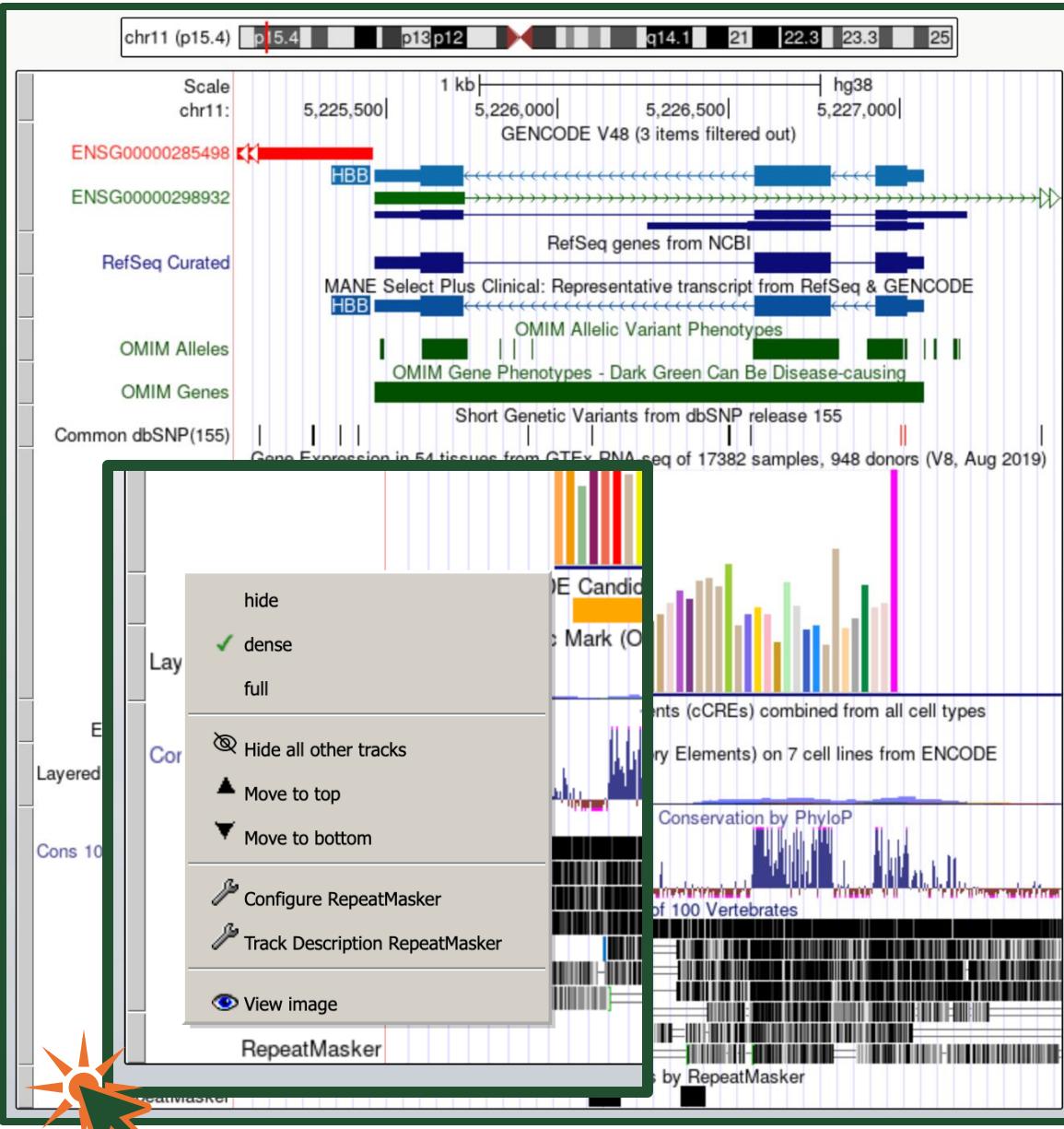
VISTA Enhancer browser: an updated database of tissue-specific developmental enhancers

Michael Kosicki, Fotis A Baltoumas, Guy Kelman, Joshua Boverhof, Yeongshnn Ong, Laura E Cook, Diane E Dickel, Georgios A Pavlopoulos, Len A Pennacchio Axel Visel

Nucleic Acids Research, Volume 53, Issue D1, 6 January 2025, Pages D324–D330, <https://doi.org/10.1093/nar/gkae940>

Published: 29 October 2024 [Article history ▾](#)

The UCSC Genome Browser contains a rich collection of data tracks



The UCSC Genome Browser contains a rich collection of data tracks

The screenshot displays the UCSC Genome Browser interface, highlighting the COSMIC Track Settings page and the main genome browser controls.

COSMIC Track Settings (Left Panel):

- Header: Genomes, Genome Browser, Tools, Mirrors, Downloads, My Data, Projects, Help, About.
- Panel Title: COSMIC Track Settings.
- Section: Catalogue of Somatic Mutations in Cancer V101 ([All Phenotypes, Variants, and Literature tracks](#))
- Display mode: hide [Duplicate track](#)
- Display data as a density graph:
- Data schema/format description and download
- Source data version: COSMIC v101
- Assembly: Human Dec. 2013 (GRCh38/hg38)
- Data last updated at UCSC: 2025-01-23 23:45:21
- Updated Note: Updated January 31, 2025
- Description:

[COSMIC](#), the "Catalogue Of Somatic Mutations In Cancer," is an online database of somatic mutations found in human cancer. Focused exclusively on non-inherited acquired mutations, COSMIC combines information from a range of sources, curating the described relationships between cancer phenotypes and gene (and genomic) mutations. These data are then made available in a number of ways including here in the UCSC genome browser, on the COSMIC website with custom analytical tools, or via the [COSMIC sftp server](#). Publications using COSMIC as a data source may cite our reference below.
- Repeating Elements by RepeatMasker visualization.

Main Genome Browser Controls (Right Panel):

- Header: Refresh, Updated, NCBI LocusLink, GCRS.
- Section: Genes and Gene Predictions.
- Category: CRISPR Targets (hide), Gencode Archive (hide), Gencode Versions (hide), Gencode Mapped (hide).
- Category: Non-coding RNA (hide), Other RefSeq (hide), Pfam in Gencode (hide), Prediction Archive (hide).
- Category: UniProt (hide).
- Section: Phenotypes, Variants, and Literature.
- Category: CADD 1.7 (hide), Cancer Gene Expr (hide), ClinGen (hide), ClinGen CNVs (hide).
- Category: COSMIC Regions (hide), COVID Data (hide), DECIPHER (hide), Development Delay (hide).
- Category: GWAS Catalog (hide), HGMD_public (hide), LOVD Variants (hide), MITOMAP (No data-chr11).
- Category: SNPedia (hide), Splicing Impact (hide), TCGA Pan-Cancer (hide), UniProt Variants (hide), Variants in Papers (hide).
- Section: Variation.
- Category: dbSNP Archive (hide), dbVar Common Struct Var (hide), DGV Struct (hide), Genome In a Bottle (hide), gnomAD Variants (hide).
- Section: Human Pangenome - HPRC.
- Category: Multiple Alignment (hide), Pairwise Alignments (hide), Rearrangement (hide), Short Variants (hide).

cBioPortal includes cancer-relevant genomic information

<https://www.cancer.org/cancer/cancer-type/ovarian-cancer/ovarian-cancer-questions-and-answers/what-is-ovarian-cancer.html>

- Mutation
- Amplification
- Deep Deletion
- Multiple Alterations
- Structural Variant

cBioPortal includes cancer-relevant genomic information

cBioPortal FOR CANCER GENOMICS

Data Sets Web API Tutorials/Webinars FAQ News Visualize Your Data About cBioPortal Installations Donate [Login](#)

TCGA PanCancer Atlas Studies
User-defined Patient List (10967 samples / 10953 patients) - TP53

Queried gene is altered in • 3939 (36%) of queried patients
• 3943 (36%) of queried samples

OncoPrint Cancer Types Summary Plots Mutations Structural Variants **Beta!** Comparison/Survival CN Segments Pathways Download [Cancer Type Summary Help](#)

TP53

Category Type
 Cancer Study
 Cancer Type
 Cancer Type Detailed

Filter Categories
Min. # Total Patients 10
0 1084
Min. % Altered Patients 0% 92%
Sort X-Axis By

Occurrence
Count By Samples Patients
Y-Axis Value

Show Genomic Alteration Types

Alteration Frequency

Cancer Type	Structural variant data	Mutation data	CNA data
Uveal Melanoma (TCGA, PanCancer Atlas)	~88%	~1%	~1%
Pancreatic Adenocarcinoma (TCGA, PanCancer Atlas)	~85%	~1%	~1%
Head and Neck Squamous Cell Carcinoma (TCGA, PanCancer Atlas)	~82%	~1%	~1%
Esophageal Adenocarcinoma (TCGA, PanCancer Atlas)	~70%	~1%	~1%
Ovarian Serous Cystadenocarcinoma (TCGA, PanCancer Atlas)	~65%	~1%	~1%
Lung Adenocarcinoma (TCGA, PanCancer Atlas)	~58%	~1%	~1%
Bladder Urothelial Carcinoma (TCGA, PanCancer Atlas)	~55%	~1%	~1%
Stomach Adenocarcinoma (TCGA, PanCancer Atlas)	~52%	~1%	~1%
Sarcoma (TCGA, PanCancer Atlas)	~48%	~1%	~1%
Brain Lower Grade Glioma (TCGA, PanCancer Atlas)	~45%	~1%	~1%
Breast Invasive Carcinoma (TCGA, PanCancer Atlas)	~35%	~1%	~1%
Uterine Corpus Endometrial Carcinoma (TCGA, PanCancer Atlas)	~32%	~1%	~1%
Kidney Renal Papillary Cell Carcinoma (TCGA, PanCancer Atlas)	~30%	~1%	~1%
Glioblastoma Multiforme (TCGA, PanCancer Atlas)	~25%	~1%	~1%
Adrenocortical Carcinoma (TCGA, PanCancer Atlas)	~20%	~1%	~1%
Skin Cutaneous Melanoma (TCGA, PanCancer Atlas)	~18%	~1%	~1%
Mesothelioma (TCGA, PanCancer Atlas)	~15%	~1%	~1%
Prostate Adenocarcinoma (TCGA, PanCancer Atlas)	~12%	~1%	~1%
Cholangiocarcinoma (TCGA, PanCancer Atlas)	~10%	~1%	~1%
Diffuse Large B-Cell Lymphoma (TCGA, PanCancer Atlas)	~8%	~1%	~1%
Cervical鳞状上皮细胞癌 (TCGA, PanCancer Atlas)	~6%	~1%	~1%
Acute Myeloid Leukemia (TCGA, PanCancer Atlas)	~5%	~1%	~1%
Thymoma (TCGA, PanCancer Atlas)	~3%	~1%	~1%
Kidney Renal Clear Cell Carcinoma (TCGA, PanCancer Atlas)	~2%	~1%	~1%
Phaeochromocytoma and Paraganglioma (TCGA, PanCancer Atlas)	~1%	~1%	~1%
Testicular Germ Cell Tumors (TCGA, PanCancer Atlas)	~1%	~1%	~1%
Thyroid Carcinoma (TCGA, PanCancer Atlas)	~1%	~1%	~1%

[Download](#)

cBioPortal includes cancer-relevant genomic information

cBioPortal FOR CANCER GENOMICS

Data Sets Web API Tutorials/Webinars FAQ News Visualize Your Data About cBioPortal Installations ❤️ Donate Login

TCGA PanCancer Atlas Studies
User-defined Patient List (10967 samples / 10953 patients) - TP53

Queried gene is altered in • 3939 (36%) of queried patients
• 3943 (36%) of queried samples

Modify Query Comparison/Survival CN Segments Pathways Download Comparison/Survival Help

OncoPrint Cancer Types Summary Plots Mutations Structural Variants Beta! Comparison/Survival CN Segments Pathways Download Comparison/Survival Help

Groups: (drag to reorder) Altered group (3943/3939) Unaltered group (7017/7011) Unprofiled group (7) TP53 (3943/3939) Select all | Deselect all

Overlap Survival Clinical Genomic Alterations mRNA Protein Arm-level CNA Genetic Ancestry Methylation Phosphosite Quantification

① Interpret all results with caution, as they can be confounded by many different variables that are not controlled for in these analyses. Consider consulting a statistician.

② The log-rank test is used to test the null hypothesis that there is no difference between the groups in the probability of an event at any time point. Hazard ratios are derived from the log-rank test.

The survival data on patients from different cohorts may have been defined by different criteria.

Min. # Patients per Group: 3 Columns ▾ Q

Survival Type	Number of Patients	p-Value	q-Value
Overall	10803	< 10 ⁻¹⁰	< 10 ⁻¹⁰
Disease Free	5383	< 10 ⁻¹⁰	< 10 ⁻¹⁰
Disease-specific	10258	< 10 ⁻¹⁰	< 10 ⁻¹⁰
Progression Free	10613	< 10 ⁻¹⁰	< 10 ⁻¹⁰

Showing 1-4 of 4

Overall

Calculate hazard ratios Altered group Add landmarks Add landmark values X-Axis Max: 370 Months Survival Logrank Test P-Value: 0.00

Probability of Overall Survival

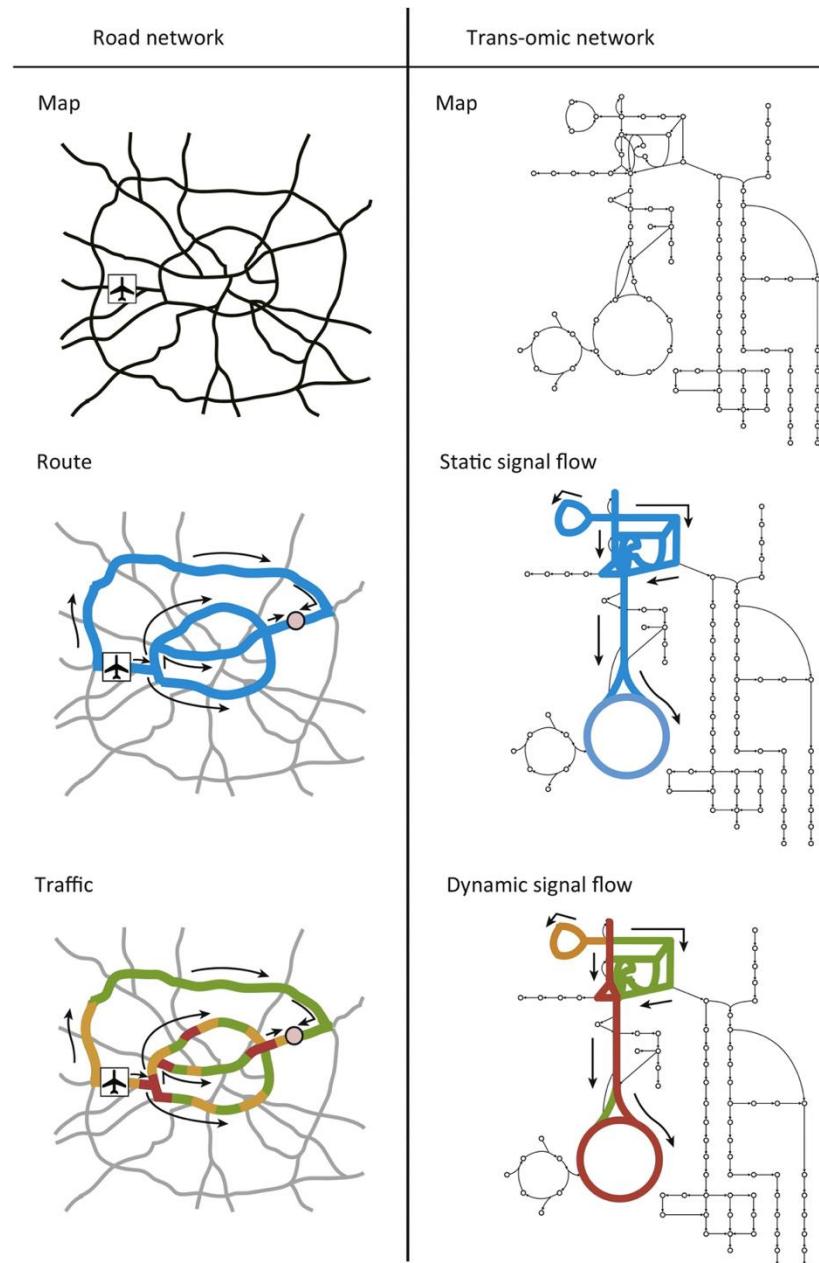
Overall Survival (Months)

Number at risk (n)

Time (months)	Altered group (n)	Unaltered group (n)
0	3890	6913
20	2069	4034
40	980	2127
60	539	1219
80	287	710
100	164	393
120	84	229
140	45	130
160	23	86
180	14	53
200	12	41
220	9	25
240	5	13
260	3	9
280	3	6
300	1	5
320	0	4
340	0	4
360	0	2

Survival plot summary

	Number of Cases, Total	Number of Events	Median Months Overall (95% CI)
Altered group	3890	1480	55.20 (51.16 - 60.66)
Unaltered group	6913	2033	95.34 (89.16 - 105.04)



Genome

What reactions are able to happen under any physiological conditions

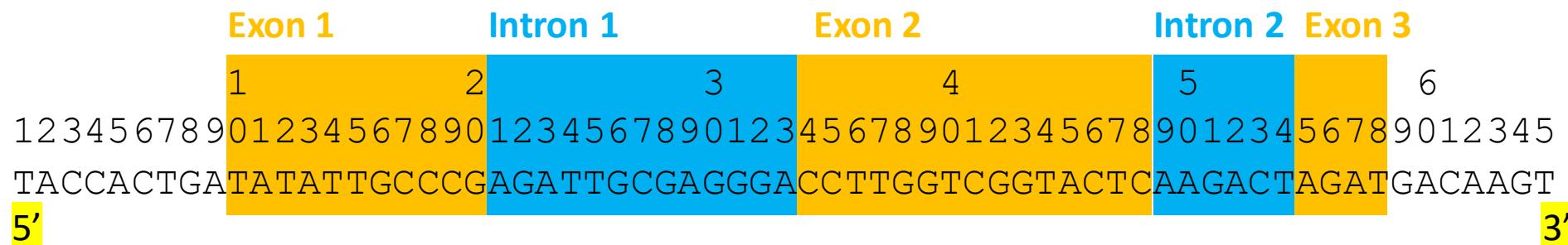
Epigenome, Transcriptome

What reactions seem to be active under specific physiological conditions

Proteome, Metabolome

What reactions are active under specific physiological conditions

Genome organization: from biology to computer bytes



Gene = exon(s) + intron(s)

22 autosomal chromosomes: 1, 2,.. 22
1 pair of sex chromosomes: X, Y
1 mitochondrial chromosome: M or MT

Example: “gene001” starts at position 10 and ends at position 58.

It has 3 exons, at positions: 10-20, 34-48, 55-58

It has 2 introns, at positions: 21-33, 49-54

Let's assume, this gene has three transcript variants:

“transcript001”: Exon1 + Exon2 + Exon3

“transcript002”: Exon2 + Exon3 (different transcription start site)

“transcript003”: Exon1 + Exon3 (alternative splicing)

Genomic coordinates:

Gene001: 10-58

Transcript001: 10-58

Transcript002: 34-58

Transcript003: 10-58

In BED format (chromosome, start, end, annotation):

chr1 10 58 gene001

chr1 10 58 transcript001|gene001

chr1 34 58 transcript002|gene001

chr1 10 58 transcript003|gene001

<https://genome.ucsc.edu/FAQ/FAQformat.html#format1>

Working with BioMart

- Go to www.ensembl.org
- Click on “BioMart” at the top of the page
- Choose “Ensembl genes 114” as the database
- Choose “Human genes (GRCh38.p14)” as the dataset.

From this page you can access all human genes.

Clicking on “Attributes” you can choose what information you want to see.

Clicking on “Filtering” you can filter the information.

We will use BioMart, a tool that allows easy access of genomic data.

More info: <http://useast.ensembl.org/info/data/biomart/index.html>

-- Our goal --

Get a file with the following information (attributes):

- genomic location of the transcript
- transcript ID and type
- Gene ID, name and type

Show only the genes (filtering):

- on primary chromosomes

Working with BioMart

The screenshot shows the Ensembl BioMart interface. On the left, there's a sidebar with sections for 'Dataset' (Human genes (GRCh38.p14)), 'Filters' (listing chromosomes 1-22, MT, X, Y), 'Attributes' (listing various gene-related attributes), and another 'Dataset' section with '[None Selected]'. The main area has tabs for 'New', 'Count', and 'Results' (which is selected). At the top right are links for 'Login/Register', a search bar with placeholder 'Search all species...', and icons for 'URL', 'XML', and 'Help'. Below the search bar are buttons for 'Compressed file (.gz)' (selected), 'TSV', 'Unique results only' (unchecked), and a 'Go' button. A large orange arrow points from step 1 to the 'Filters' section. Another orange arrow points from step 2 to the 'Attributes' section. An orange box labeled '3' points to the 'Results' tab. An orange box labeled '4' points to the 'Go' button. An orange box labeled '5' points to the 'Compressed file (.gz)' dropdown.

Chromosome/scaffold name	Transcript start (bp)	Transcript end (bp)	Transcript stable ID	Transcript type	Gene stable ID	Gene name	Gene type
1	3069168	3434342	ENST00000511072	protein_coding	ENSG00000142611	PRDM16	protein_coding
1	3069183	3186591	ENST00000607632	retained_intron	ENSG00000142611	PRDM16	protein_coding
1	3069197	3435421	ENST00000378391	protein_coding	ENSG00000142611	PRDM16	protein_coding
1	3069211	3434342	ENST00000514189	protein_coding	ENSG00000142611	PRDM16	protein_coding
1	3069203	3438621	ENST00000270722	protein_coding	ENSG00000142611	PRDM16	protein_coding
1	3237931	3433925	ENST00000512462	protein_coding_CDS_not_defined	ENSG00000142611	PRDM16	protein_coding
1	3244132	3386918	ENST00000463591	protein_coding	ENSG00000142611	PRDM16	protein_coding
1	3396491	3434293	ENST00000509860	protein_coding	ENSG00000142611	PRDM16	protein_coding
1	3424919	3434095	ENST00000378389	protein_coding_CDS_not_defined	ENSG00000142611	PRDM16	protein_coding
1	3425216	3426072	ENST00000606170	retained_intron	ENSG00000142611	PRDM16	protein_coding

1. Click “Filters” and select these chromosomes
(you will find them under “REGION”)
2. Click “Attributes” and select these ones in **this specific order**
(you will find them under “GENE”)
3. Click “Results”
4. Select “Compressed file (.gz)”
5. Click “Go” and save the file in the directory you will be working (e.g. make a folder “workshop” on your desktop)

Questions we can answer using the file from biomart:

- Which chromosome has the highest number of protein coding genes?
- Which protein-coding gene has the most coding splice variants?
- Which protein-coding gene has the most non-coding splice variants?

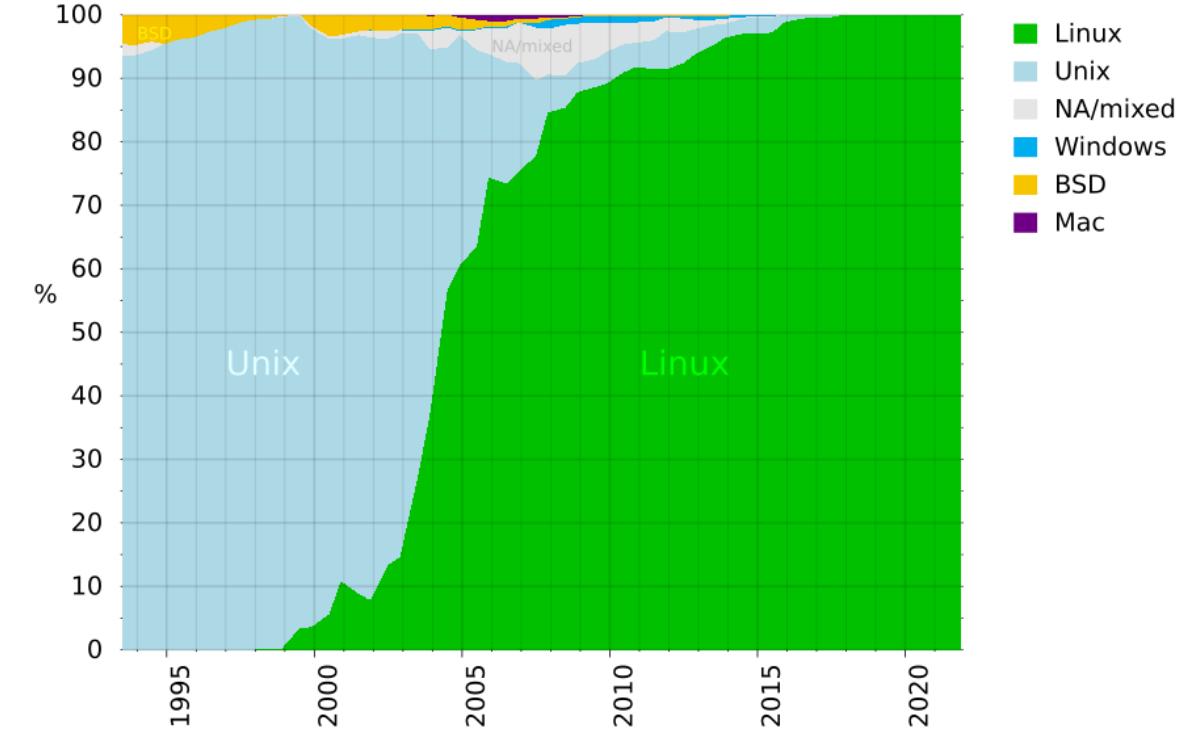
Which tool can we use?

Linux / bash

Operating systems used on top 500 supercomputers

By Benedikt.Seidl - Eigenes Werk (own work) / source top500.org, Public Domain

https://en.m.wikipedia.org/wiki/File:Operating_systems_used_on_top_500_supercomputers.svg



Working with linux / bash

Unix is a command-line programming language.

Instead of clicking, you type commands. For example, when you want to make a font bold, then you click on the “B” button. This click is translated to the command “make the selected text bold”. In unix you would need to type this.

Open the terminal, type “bash” and hit “return” (or “enter”). This will get you into the bash environment (the default may be zsh, which uses slightly different commands and syntax).

Type “pwd” and hit return. This will tell you in which directory you currently are.

If you made the folder “workshop” on your desktop, you can move to it with the “cd” command.

This is what I got when running the above (the orange ones are what I typed):

```
Last login: Sun Jun 01 14:30:17 on ttys001

The default interactive shell is now zsh.
To update your account to use zsh, please run `chsh -s /bin/zsh`.
For more details, please visit https://support.apple.com/kb/HT208050.
UH-GDG7LR6RT7:~ axt5207$ pwd
/Users/axt5207
UH-GDG7LR6RT7:~ axt5207$ cd Desktop/workshop/
UH-GDG7LR6RT7:workshop axt5207$ pwd
/Users/axt5207/Desktop/workshop
UH-GDG7LR6RT7:workshop axt5207$ ls
mart_export.txt.gz
UH-GDG7LR6RT7:workshop axt5207$
```

To see the manual of a specific command,e.g. of pwd , type “man pwd” hit return. Scroll up/down with arrows, hit “q” to exit the manual page. Read the manual pages of the commands you use!

An online tutorial: <https://linuxcommand.org/>

We should be ready to
start with **exercise 1**

**The most commonly used
unix commands**

cat	mv
cd	paste
colrm	pwd
cut	rm
cp	sed
exit	sort
ftp	scp
grep	ssh
head	sftp
join	tail
less	uniq
ls	xterm
man	wc
more	wget



BLAST: the “google” of the genome

 U.S. National Library of Medicine
National Center for Biotechnology Information

Log in

BLAST®

Home Recent Results Saved Strategies Help

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

NEWS

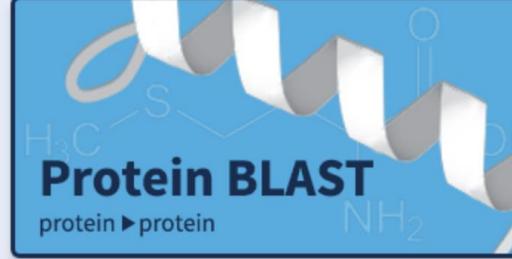
A new feature was added to Primer-BLAST.
We now offer the ability for user to run primer-blast from NCBI assembly page..
Tue, 23 Feb 2021 12:00:00 EST [More BLAST news...](#)

Web BLAST


Nucleotide BLAST
nucleotide ▶ nucleotide


blastx
translated nucleotide ▶ protein


tblastn
protein ▶ translated nucleotide


Protein BLAST
protein ▶ protein

BLAST Genomes

Enter organism common name, scientific name, or tax id

Search

Human Mouse Rat Microbes

<https://blast.ncbi.nlm.nih.gov/>

If you didn't trust me that the sequence was HBB

Descriptions Graphic Summary Alignments Taxonomy

Sequences producing significant alignments Download New Select columns Show 100 ?

select all 100 sequences selected

GenBank Graphics Distance tree of results New MSA Viewer

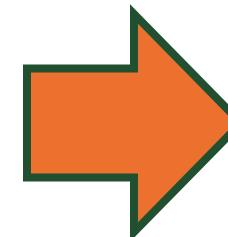
	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	Homo sapiens hemoglobin subunit beta (HBB), mRNA	Homo sapiens	1160	1160	100%	0.0	100.00%	628	NM_000518.5
<input checked="" type="checkbox"/>	PREDICTED: Pan troglodytes hemoglobin subunit beta (HBB), mRNA	Pan troglodytes	1155	1155	100%	0.0	99.84%	748	XM_508242.4
<input checked="" type="checkbox"/>	PREDICTED: Pan paniscus hemoglobin subunit beta (LOC100976465), mRNA	Pan paniscus	1149	1149	100%	0.0	99.68%	646	XM_003819029.3
<input checked="" type="checkbox"/>	Homo sapiens hemoglobin, beta, mRNA (cDNA clone MGC:14540 IMAGE:4292125), complete cds	Homo sapiens	1142	1142	99%	0.0	99.52%	658	BC007075.1
<input checked="" type="checkbox"/>	Human messenger RNA for beta-globin	Homo sapiens	1140	1140	99%	0.0	99.52%	626	V00497.1
<input checked="" type="checkbox"/>	PREDICTED: Gorilla gorilla gorilla hemoglobin subunit beta (LOC101126932), mRNA	Gorilla gorilla go...	1133	1133	100%	0.0	99.20%	753	XM_019036164.1
<input checked="" type="checkbox"/>	Homo sapiens hemoglobin beta mRNA, complete cds	Homo sapiens	1133	1133	99%	0.0	99.36%	647	AY509193.1
<input checked="" type="checkbox"/>	PREDICTED: Pongo abelii hemoglobin subunit beta (HBB), mRNA	Pongo abelii	1098	1098	99%	0.0	98.25%	627	XM_002822127.4
<input checked="" type="checkbox"/>	PREDICTED: Nomascus leucogenys hemoglobin subunit beta (HBB), mRNA	Nomascus leuc...	1088	1088	100%	0.0	97.93%	753	XM_004090649.3
<input checked="" type="checkbox"/>	PREDICTED: Hylobates moloch hemoglobin subunit beta (HBB), mRNA	Hylobates moloch	1077	1077	100%	0.0	97.61%	753	XM_032166808.1
<input checked="" type="checkbox"/>	Homo sapiens hemoglobin beta chain variant (HBB) mRNA, HBB-Dothan allele, complete cds	Homo sapiens	1064	1064	93%	0.0	99.32%	587	EU694432.1
<input checked="" type="checkbox"/>	Homo sapiens hemoglobin beta subunit variant (HBB) mRNA, complete cds	Homo sapiens	1064	1064	92%	0.0	99.83%	579	AF181989.1
<input checked="" type="checkbox"/>	Homo sapiens beta globin chain variant (HBB) mRNA, complete cds	Homo sapiens	1057	1057	92%	0.0	99.48%	581	AF349114.1
<input checked="" type="checkbox"/>	PREDICTED: Trachypithecus francoisi hemoglobin subunit beta (LOC117081249), mRNA	Trachypithecus f...	1038	1038	100%	0.0	96.51%	749	XM_033207068.1
<input checked="" type="checkbox"/>	PREDICTED: Colobus angolensis palliatus hemoglobin, beta (HBB), mRNA	Colobus angole...	1038	1038	100%	0.0	96.50%	753	XM_011963601.1
<input checked="" type="checkbox"/>	PREDICTED: Rhinopithecus roxellana hemoglobin subunit beta (LOC104662339), mRNA	Rhinopithecus r...	1033	1033	100%	0.0	96.35%	749	XM_010363344.2
<input checked="" type="checkbox"/>	PREDICTED: Rhinopithecus bieti hemoglobin subunit beta (LOC108521664), mRNA	Rhinopithecus b...	1033	1033	100%	0.0	96.35%	749	XM_017861621.1
<input checked="" type="checkbox"/>	Homo sapiens hemoglobin beta chain (HBB) mRNA, complete cds	Homo sapiens	1027	1027	89%	0.0	99.64%	562	AF117710.1
<input checked="" type="checkbox"/>	Human sickle beta-hemoglobin mRNA	Homo sapiens	1024	1024	91%	0.0	98.44%	576	M25113.1
<input checked="" type="checkbox"/>	PREDICTED: Mandrillus leucophaeus hemoglobin subunit beta (LOC105535916), transcript variant X2, mRNA	Mandrillus leuco...	1022	1022	100%	0.0	96.02%	753	XM_0119751
<input checked="" type="checkbox"/>	PREDICTED: Piliocolobus tephrosceles hemoglobin subunit beta (HRR1) mRNA	Piliocolobus teph...	1016	1016	100%	n n	95.87%	749	XM_0232008

Fe

What you need to run blast, or any alignment algorithm (including STAR)

Query file

```
>MyQuery
ACATTTGCTTCTGACACAACACTGTGTTCACTAG
CAACCTCAAACAGACACCAGTGGTCATCTGAC
TCCTGAGGAGAAGTCTGCCGTTACTGCCCTGT
GGGGCAAGGTGAACGTGGATGAAGTTGGTGGT
GAGGCCTGGCAGGCTGCTGGTGGTCTACCC
TTGGACCCAGAGGTTCTTGAGTCCTTGGGG
ATCTGTCCACTCCTGATGCTGTTATGGGCAAC
CCTAAGGTGAAGGCTCATGGCAAGAAAGTGCT
CGGTGCCTTAGTGATGGCTGGCTCACCTGG
ACAACCTCAAGGGCACCTTGCCACACTGAGT
GAGCTGCACTGTGACAAGCTGCACTGGATCC
TGAGAACATTCAAGGCTCCTGGCAACGTGCTGG
TCTGTGTGCTGGCCATCACTTGGCAAAGAA
TTCACCCCCACCAAGTGCAGGCTGCCTATCAGAA
AGTGGTGGCTGGTGTGGCTAATGCCCTGGCCC
ACAAGTATCACTAAGCTCGCTTCTGCTGTC
CAATTCTATTAAAGGTTCTTGTCCCTAA
GTCCAACACTAAACTGGGGATATTATGAAG
GGCCTTGAGCATCTGGATTCTGCCTAATAAAA
AACATTTATTTCATTGCAA
```



Database

```
>chr1
ACAACCTGTGTTCACTAGCAAACATTGCTTCTGAC.....
>chr2
ACATTTGCTTCTGACACAACACTGTGTTCACTAGCAA.....
.....
>chrX
TCTGACACAACACTGTGTTCACTAGCAAACATTGCT.....
```

Fasta files (.fa) are used to store sequences in a simple format: Lines starting with ">" contain the name, description or any other information about the sequence and are followed by the sequence (in one or multiple lines) until the next ">" (if any).

Table Browser: querying annotation tables

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute UCSC Genome Browser Gateway

Genomes Genome Browser Tools Mirrors Downloads My Data Projects Help About Us

Browse>Select Species

POPULAR SPECIES Human Mouse Rat Zebrafish Fruitfly

Search through thousands of genome browsers Enter species, common name or assembly ID

Unable to find a genome? Send us a request.

UCSC SPECIES TREE AND CONNECTED ASSEMBLIES

UCSC Genome Browser assembly ID: hg38
Sequencing/Assembly provider ID: Genome Reference Consortium Human GRCh38.p14 (GCA_000001405.29)
Assembly date: Dec. 2013 initial release; June 2022 patch release 14
Assembly accession: GCA_000001405.29
NCBI Genome ID: 51 (Homo sapiens (human))
NCBI Assembly ID: GCF_000001405.40 (GRCh38.p14, GCA_000001405.29)
BioProject ID: PRJNA31257

(Graphic courtesy of CBSE)

<http://genome.ucsc.edu/goldenPath/help/hgTablesHelp.html#Introduction>

How to access the COSMIC mutations from the Table Browser

Table Browser

Use this tool to retrieve and export data from the Genome Browser annotation track database. You can limit retrieval based on data attributes and intersect or merge with data from another track, or retrieve DNA sequence covered by a track. [More...](#)

Select dataset [i](#)

Clade: Mammal Genome: Human Assembly: Dec. 2013 (GRCh38/hg38)

Group: Phenotypes, Variants, and Literature Track: COSMIC

Table: cosmicMuts Data format description

Define region of interest [i](#)

Region: Genome Position chr17:1-83,257,441 Lookup Define regions

Identifiers (names/accessions): Paste list Upload list

Optional: Subset, combine, compare with another track [i](#)

Filter: Create

Intersection: Create

Retrieve and display data [i](#)

Output format: BED - browser extensible data Send output to Galaxy GREAT

Output filename: cosmicMuts_chr17.bed.gz (leave blank to keep output in browser)

File type returned: Plain text Gzip compressed

Get output Summary/statistics

How COSMIC entries are stored

Genomes Genome Browser Tools Mirrors Downloads My Data Projects

Schema for COSMIC - Catalogue of Somatic Mutations in Cancer V101

Database: hg38 **Primary Table:** cosmicMuts **Data last updated:** 2025-01-23
Big Bed File Download: <https://hgdownload.soe.ucsc.edu/gbdb/hg38/cosmic/cosmic.bb>
Item Count: 14,005,634

The data is stored in the binary [BigBed](#) format. Our `bigBedToBed` tool accepts a file or the URL above as the input and converts it to text.

To download this table in different text formats or to intersect or correlate it with other tables, use the [Table Browser](#).

Format description: COSMIC data

field	example	description
chrom	chr1	Reference sequence chromosome or scaffold
chromStart	166070060	Start position in chrom
chromEnd	166070061	End position in chrom
name	COSV100597382	Genomic Mutation ID
score	0	Not used
strand	-	Which DNA strand contains the observed alleles
refAllele	G	Sequence of reference allele
altAllele	A	Sequence of alternate allele
cosmicLegIden	COSN30539494	Cosmic legacy mutation identifier

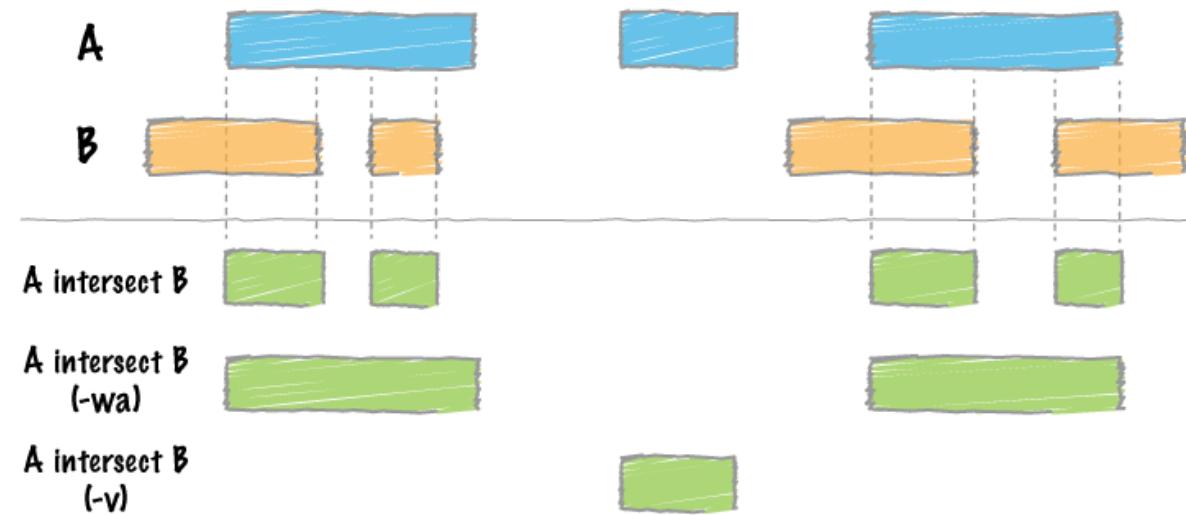
Sample Rows

chrom	chromStart	chromEnd	name	score	strand	refAllele	altAllele	cosmicLegIden
chr1	166070060	166070061	COSV100597382	0	-	G	A	COSN30539494
chr1	166070088	166070089	COSV100597761	0	-	C	T	COSN30454149
chr1	166070107	166070108	COSV100597504	0	-	T	G	COSN31525649
chr1	166070116	166070117	COSV57983747	0	-	G	T	COSN13486892
chr1	166070123	166070124	COSV100597351	0	-	A	G	COSN30504526
chr1	166070136	166070137	COSV100597301	0	-	C	T	COSN30460455
chr1	166070147	166070148	COSV57975849	0	-	C	T	COSN8636682
chr1	166070159	166070160	COSV100597775	0	-	A	C	COSN30113831
chr1	166070166	166070167	COSV100597396	0	-	G	T	COSN30505482
chr1	166070168	166070169	COSV100597505	0	-	G	A	COSN30455916

Having these coordinates, we can intersect them with coding genes, e.g. to find the number of COSMIC mutations per gene

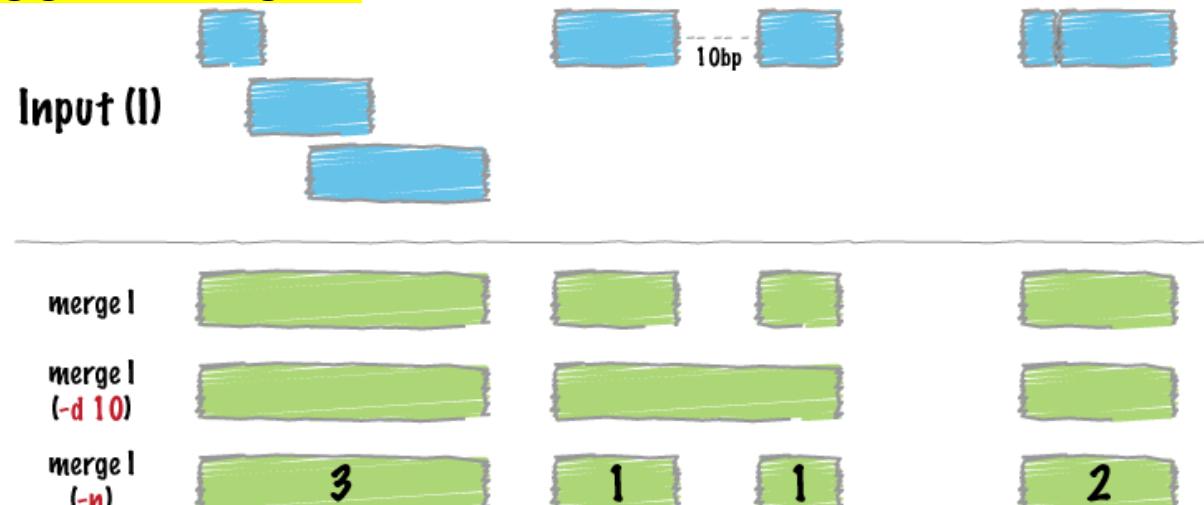
Genomic arithmetic using bedtools

Intersecting genomic regions



<https://bedtools.readthedocs.io/en/latest/content/tools/intersect.html>

Merging genomic regions



<https://bedtools.readthedocs.io/en/latest/content/tools/merge.html>

Genomic arithmetic using bedtools

Intersecting genomic regions

FileA		
chr1	10000	20000
chr2	15000	30000
chrX	1000	2000
chr23	1500	2500

FileB		
chr1	50000	80000
chr2	10000	17000
chrX	1500	3000
chr21	1000	2500



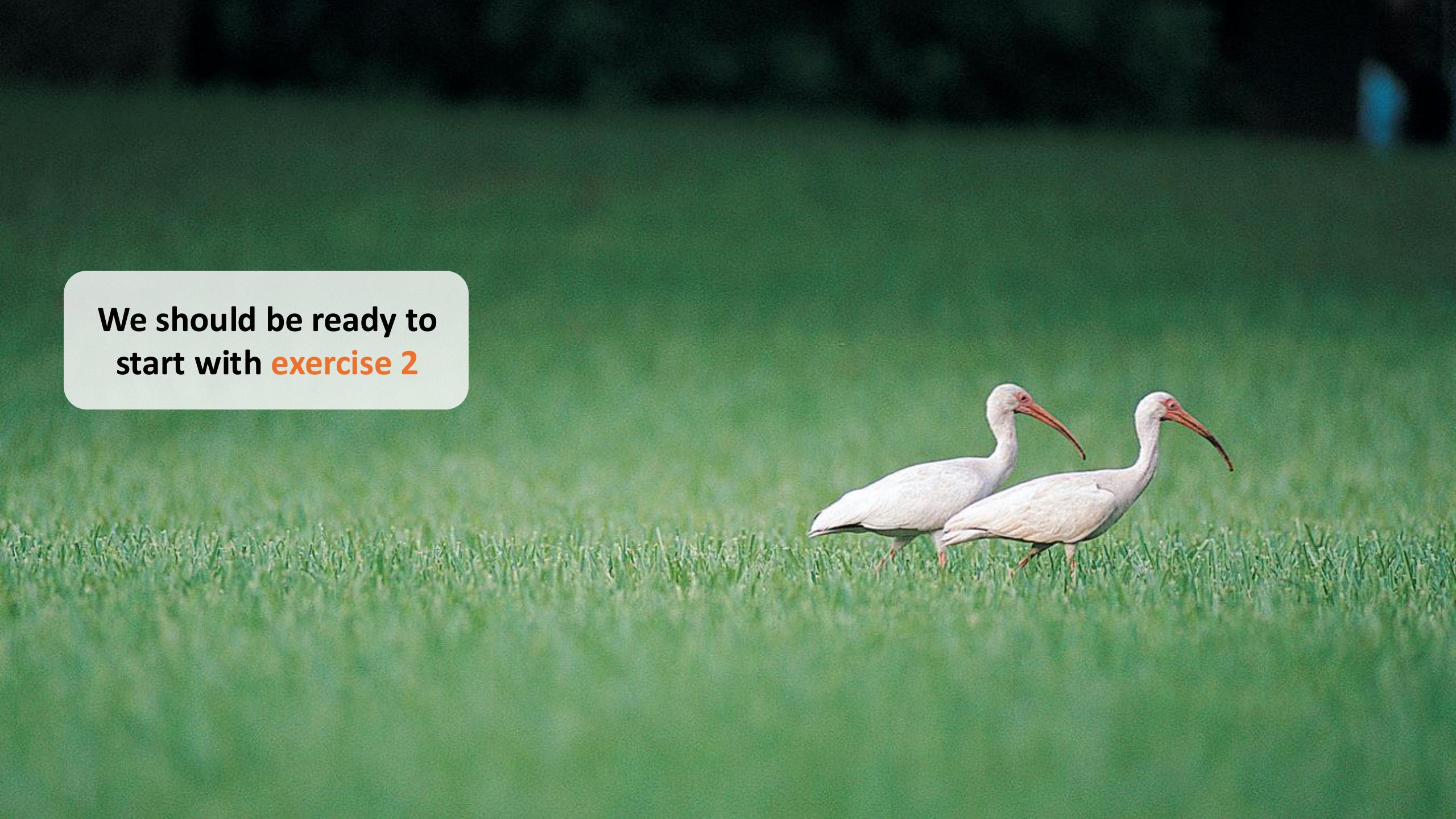
FileC		
chr2	15000	17000
chrX	1500	2000

Merging genomic regions

FileA		
chr1	10000	20000
Chr1	15000	35000
chr2	15000	30000
chrX	1000	2000
chr23	1500	2500



FileB		
Chr1	10000	35000
chr2	15000	30000
chrX	1000	2000
chr23	1500	2500

A photograph of two white ibises standing in a field of tall green grass. They are facing right, showing their long, thin, slightly downward-curving bills. The background is a soft-focus green.

We should be ready to
start with exercise 2