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:

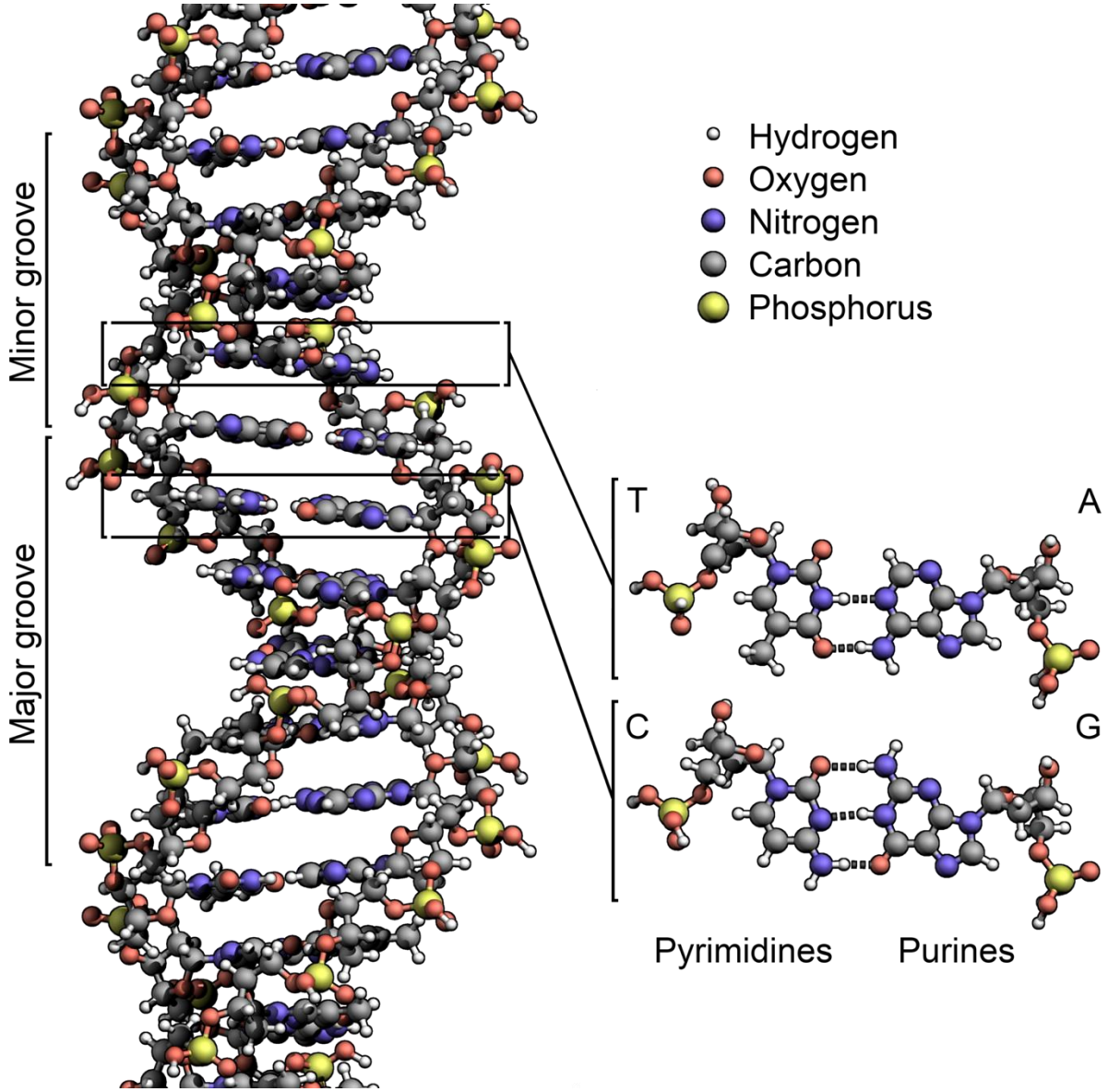
- Which chromosome has the most coding genes?
- Which coding gene has the most coding splice variants?
- Which coding gene has the most non-coding variants?

SYLVESTER COMPREHENSIVE CANCER CENTER
UNIVERSITY OF MIAMI HEALTH SYSTEM

NCI Designated Cancer Center

@SylvesterCancer

Genome organization: from biology to computer bytes



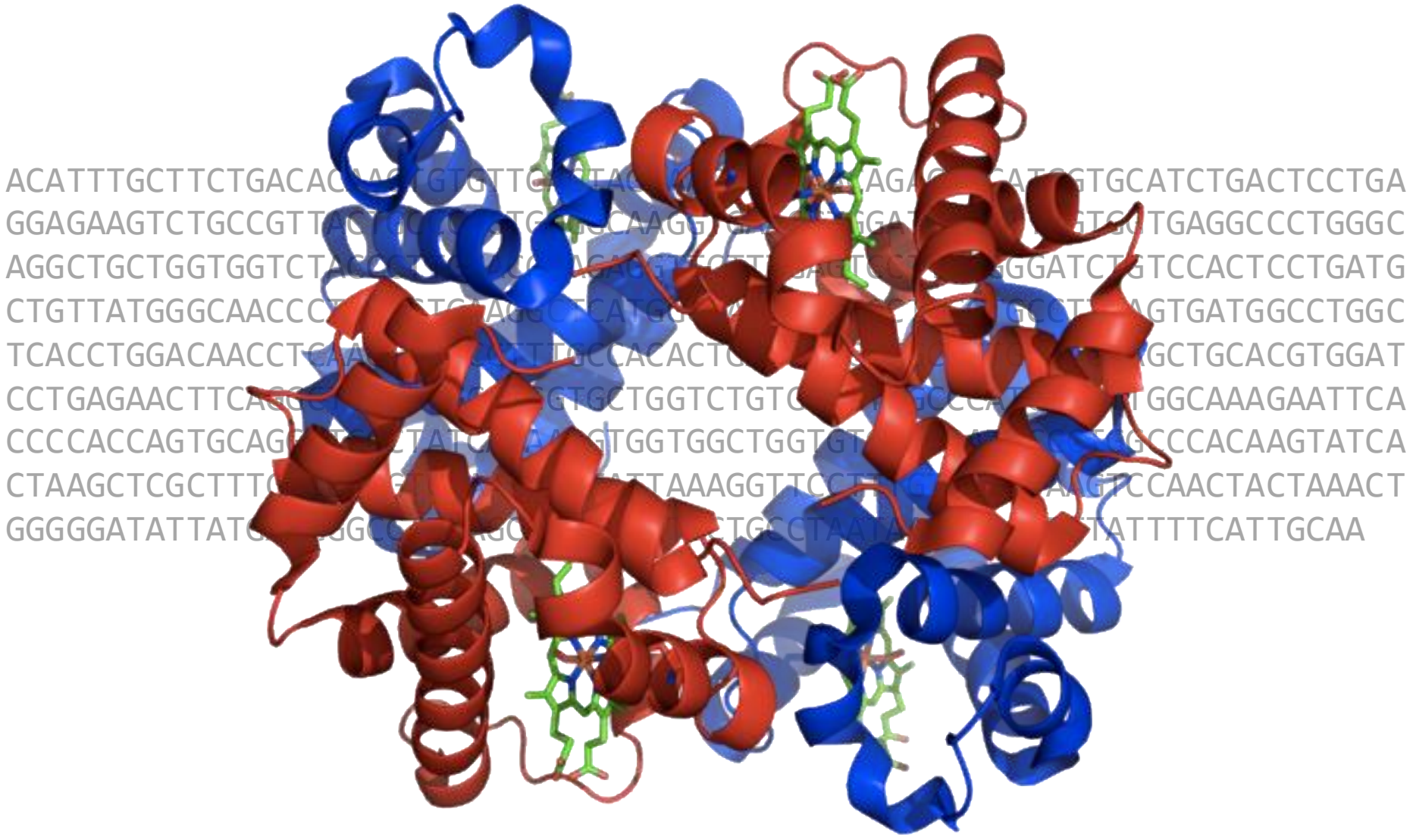
By Zephyris - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=15027555

DNA is a sequence of four letters

ACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCATCTGACTCCTGA
GGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGC
AGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATG
CTGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGC
TCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGAT
CCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCA
CCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATCA
CTAAGCTCGCTTTCTTGCTGTCCAATTTCTATTAAAGGTTCCTTTGTTCCCTAAGTCCAACTACTAAACT
GGGGGATATTATGAAGGGCCTTGAGCATCTGGATTCTGCCTAATAAAAAACATTTATTTTCATTGCAA

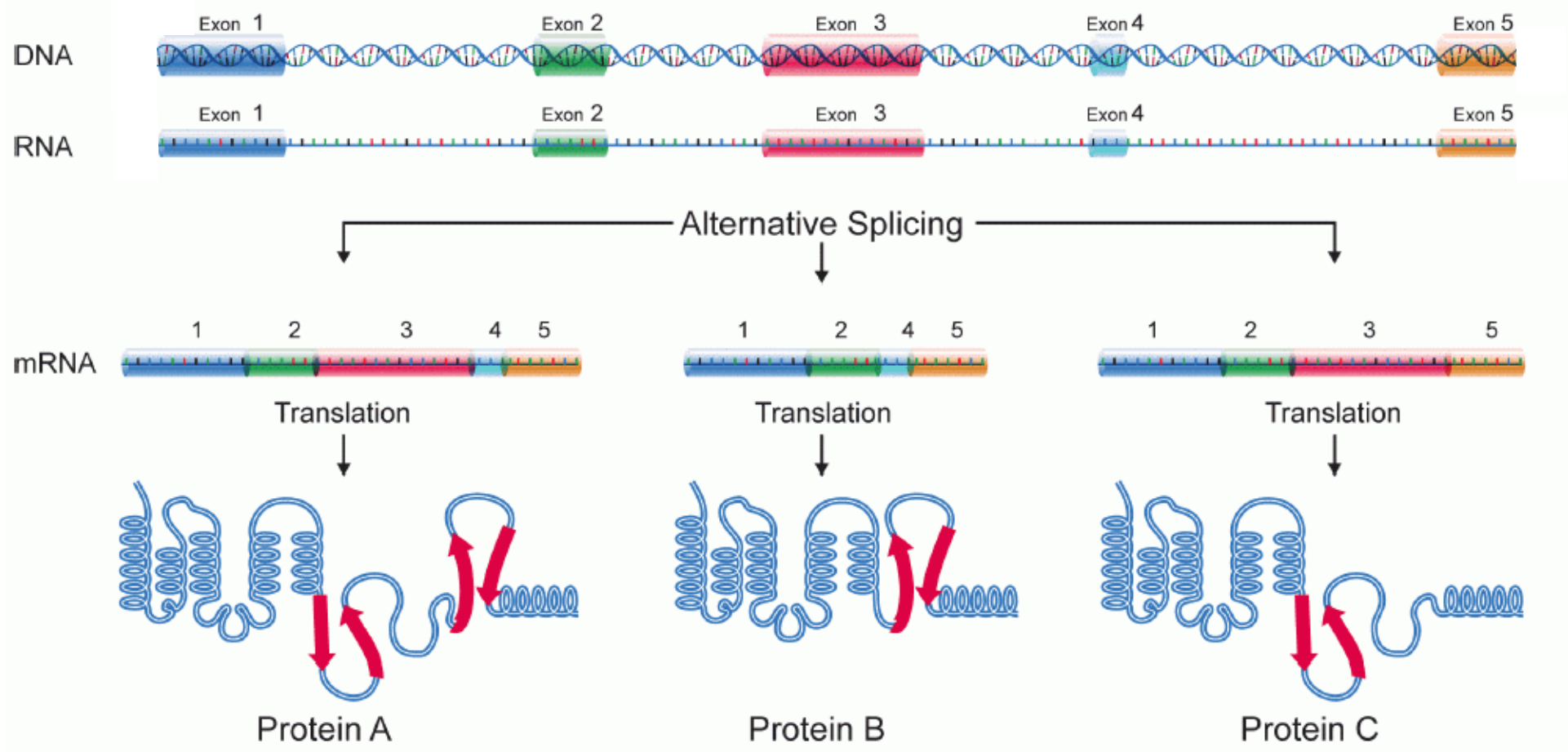
+ 6.4 billion letters

DNA is a sequence of four letters



By Zephyris at the English language Wikipedia, CC BY-SA 3.0,
https://commons.wikimedia.org/w/index.php?curid=2300973

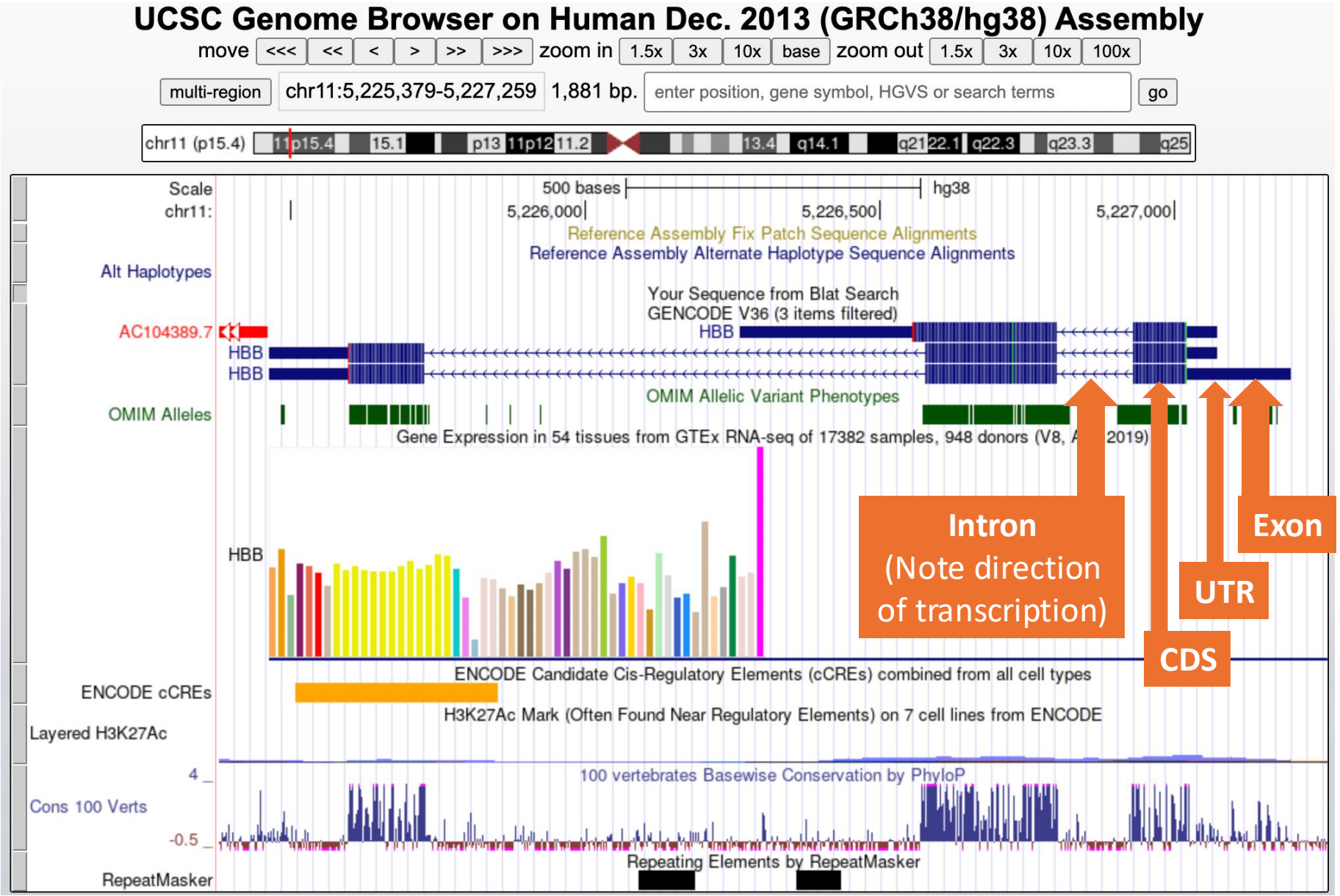
We need to account for the complexity of molecular biology



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

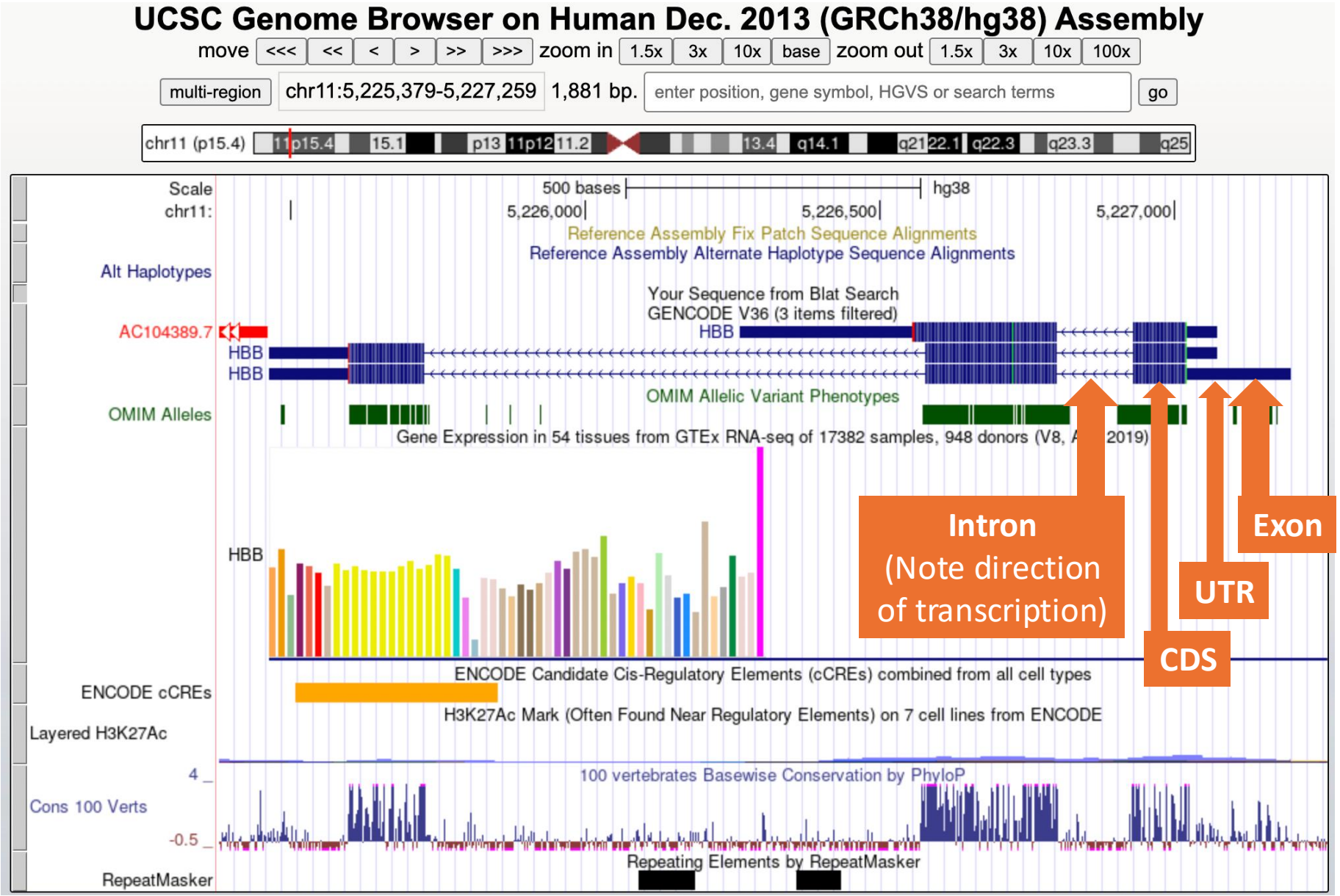
The above mRNAs are protein-coding transcripts. There are a lot of transcripts that are not translated into proteins and are therefore non-coding

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?hgg_gene=ENST00000335295.4&hgg_chrom=chr11&hgg_start=5225463&hgg_end=5227071&hgg_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

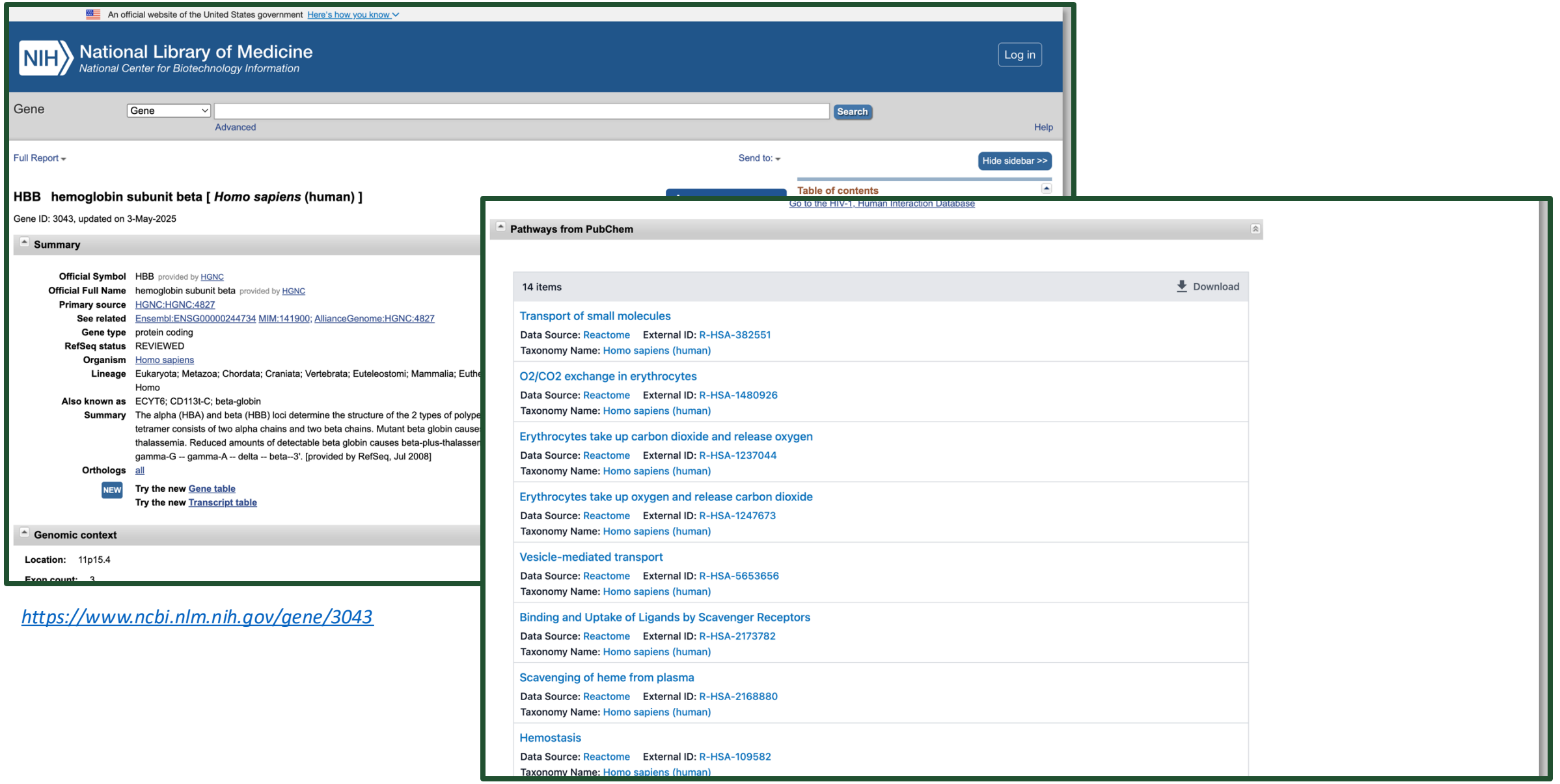
We know a LOT of things, we can access a LOT of information

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

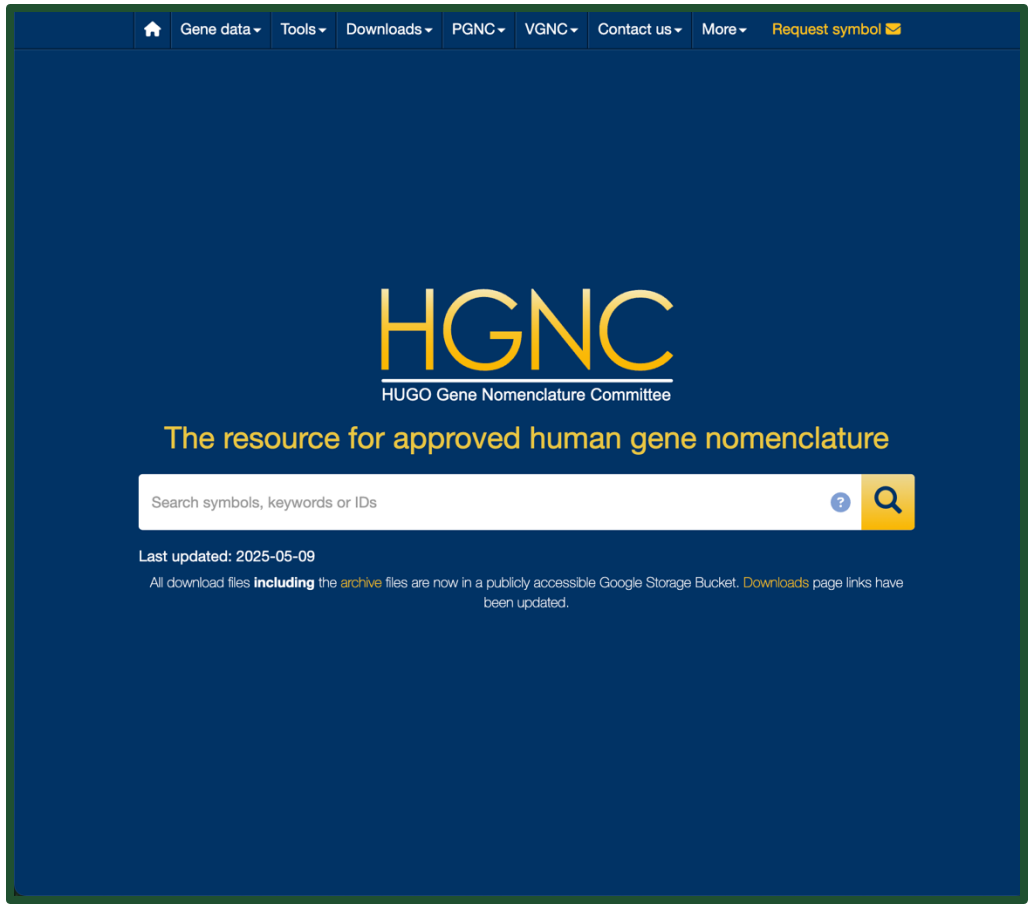
http://www.ensembl.org/Homo_sapiens/Info/Annotation

DNA | Gene NCBI database



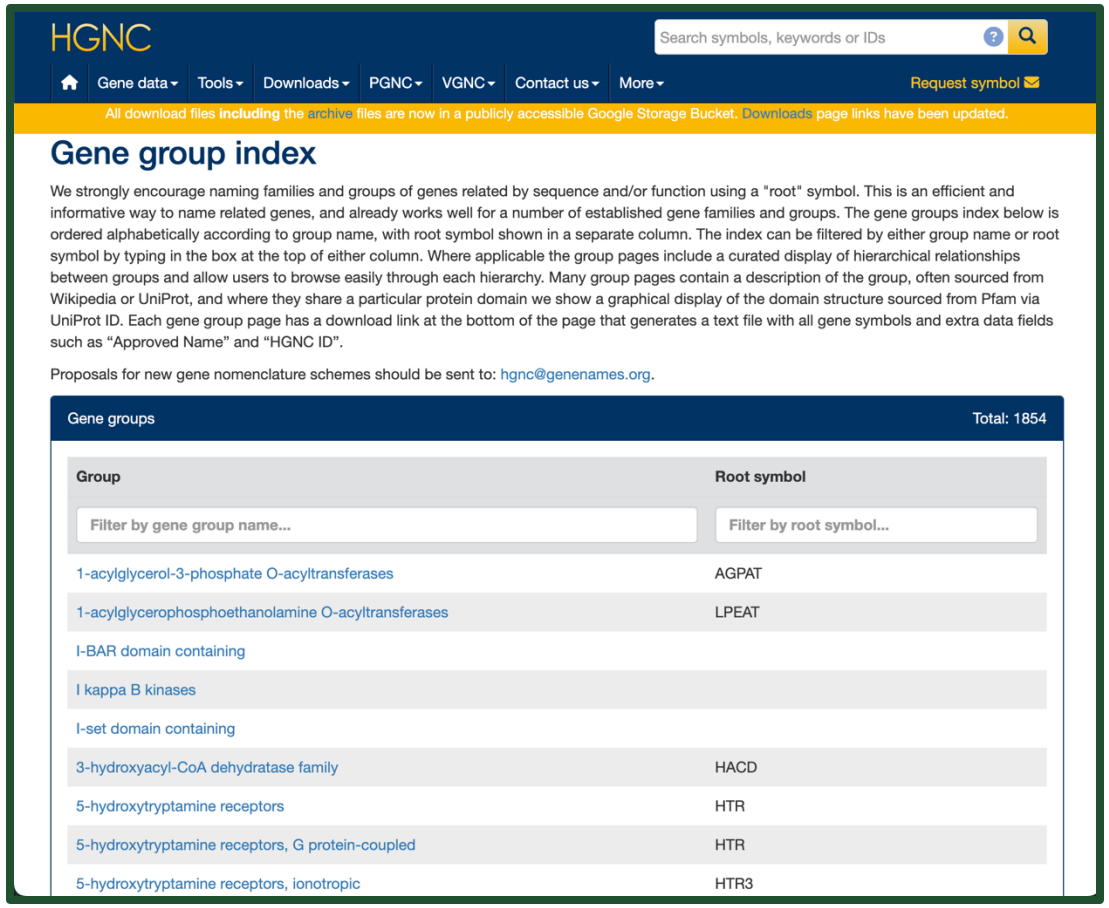
https://www.ncbi.nlm.nih.gov/gene/3043

DNA | Official gene names and gene groups

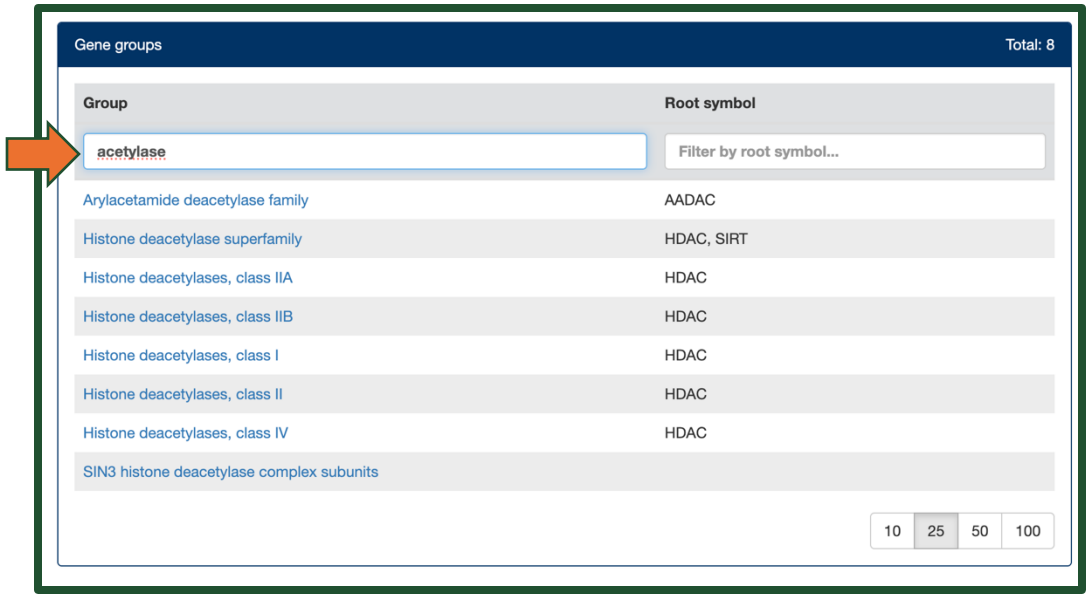


https://www.genenames.org/

Gene data > Gene group reports



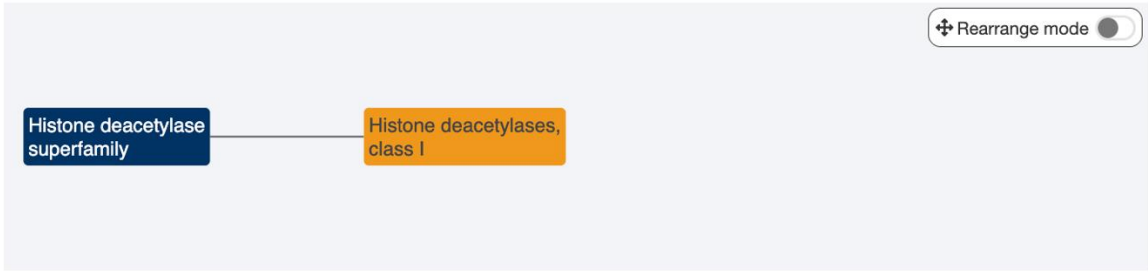
DNA | Official gene names and gene groups



Gene group: Histone deacetylases, class I (HDAC)

A subgroup of: "Histone deacetylase superfamily (HDAC, SIRT)"

Gene group hierarchy map



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

Genes contained within the group: 4

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 ...

DNA | GENCODE: mapping gene locations, splice variants and more

GENCODE

Human Mouse How to access data FAQ Documentation About us

Human

More about GENCODE Human

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		

GENCODE

Human Mouse How to access data FAQ Documentation About us

The GENCODE Project: Encyclopædia of genes and gene variants

Background

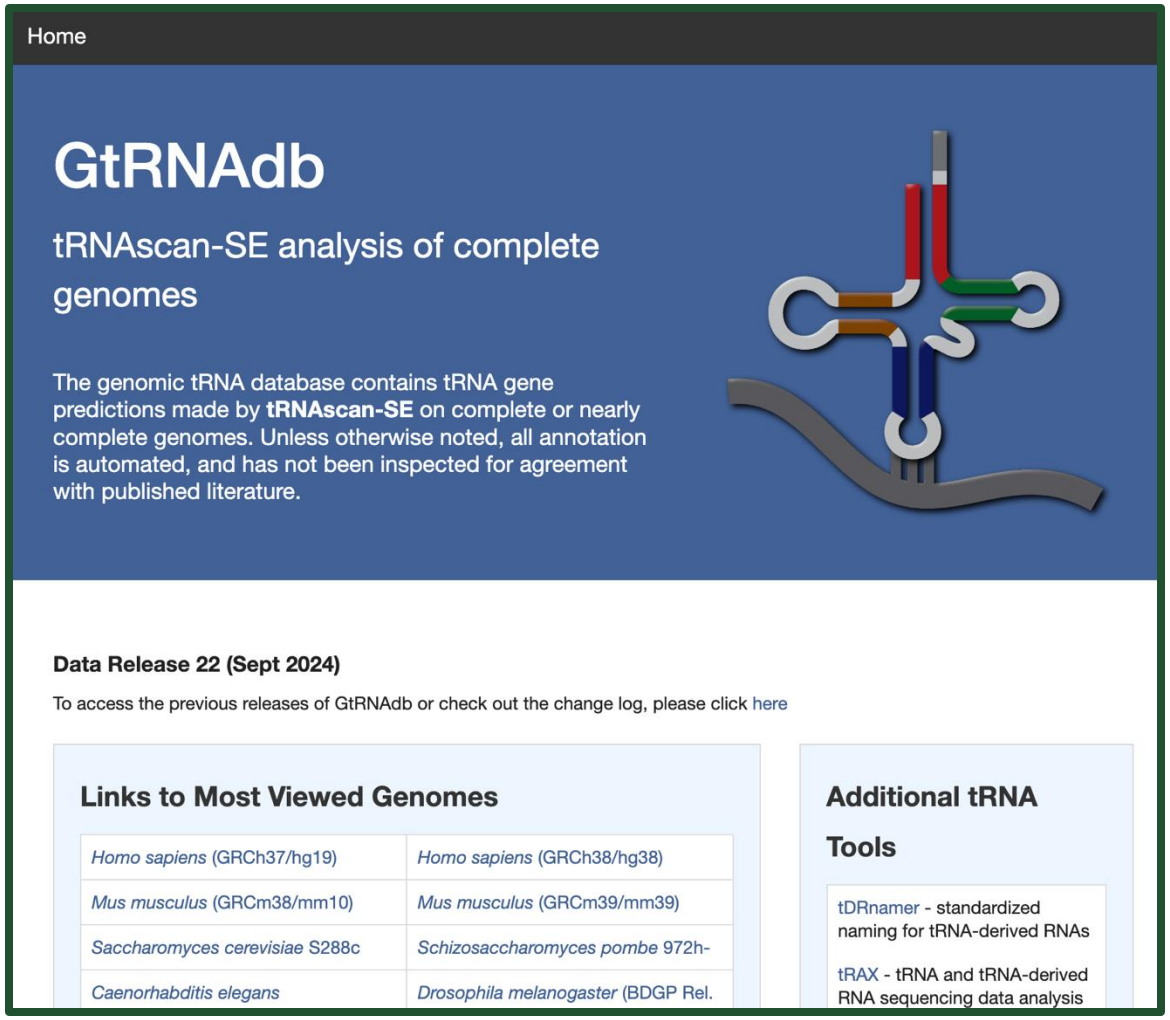
The National Human Genome Research Institute (NHGRI) launched a public research consortium named ENCODE, the Encyclopedia Of DNA Elements, in September 2003, to carry out a project to identify all functional elements in the human genome sequence. After a successful pilot phase on 1% of the genome, the 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.

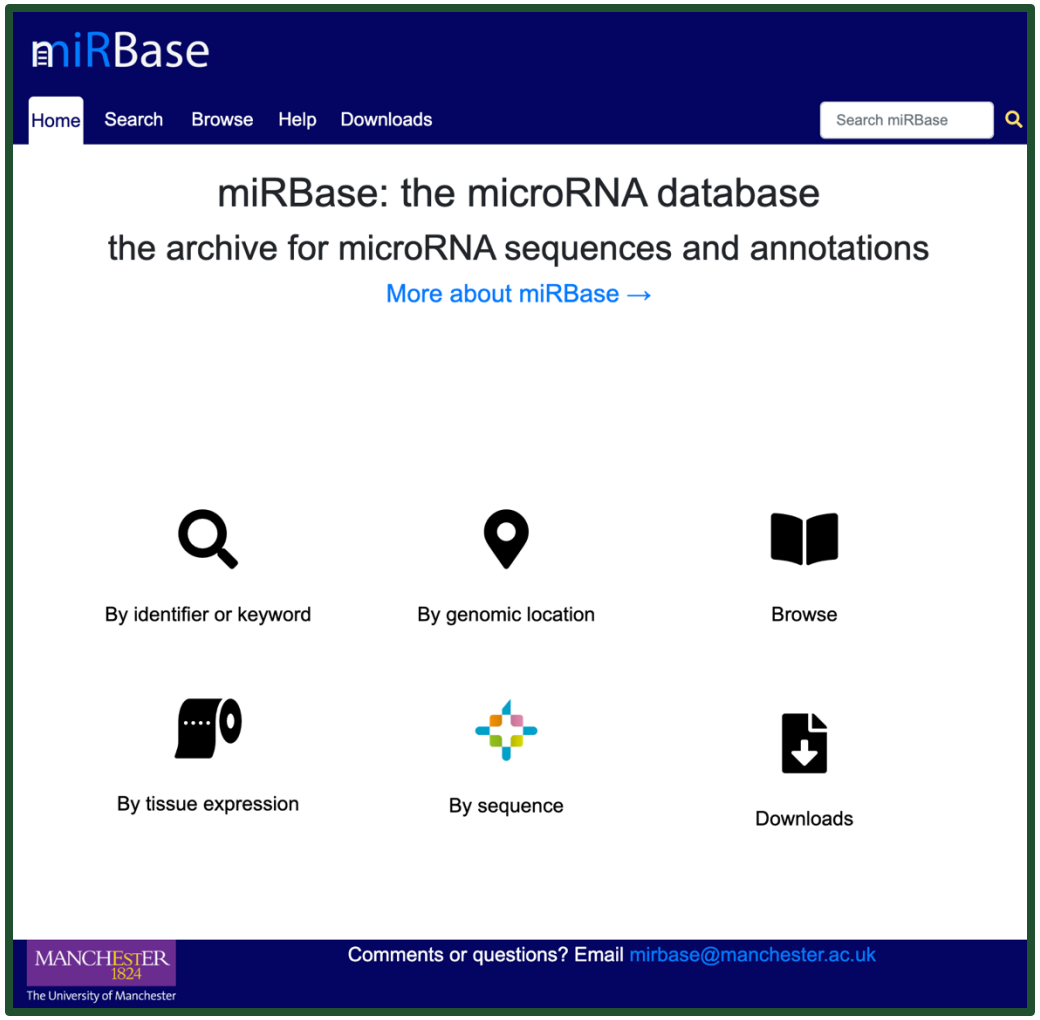
GENCODE gene annotation is used by a large number of international consortia.

https://www.gencodegenes.org/

RNA | Some commonly used non-coding RNA databases

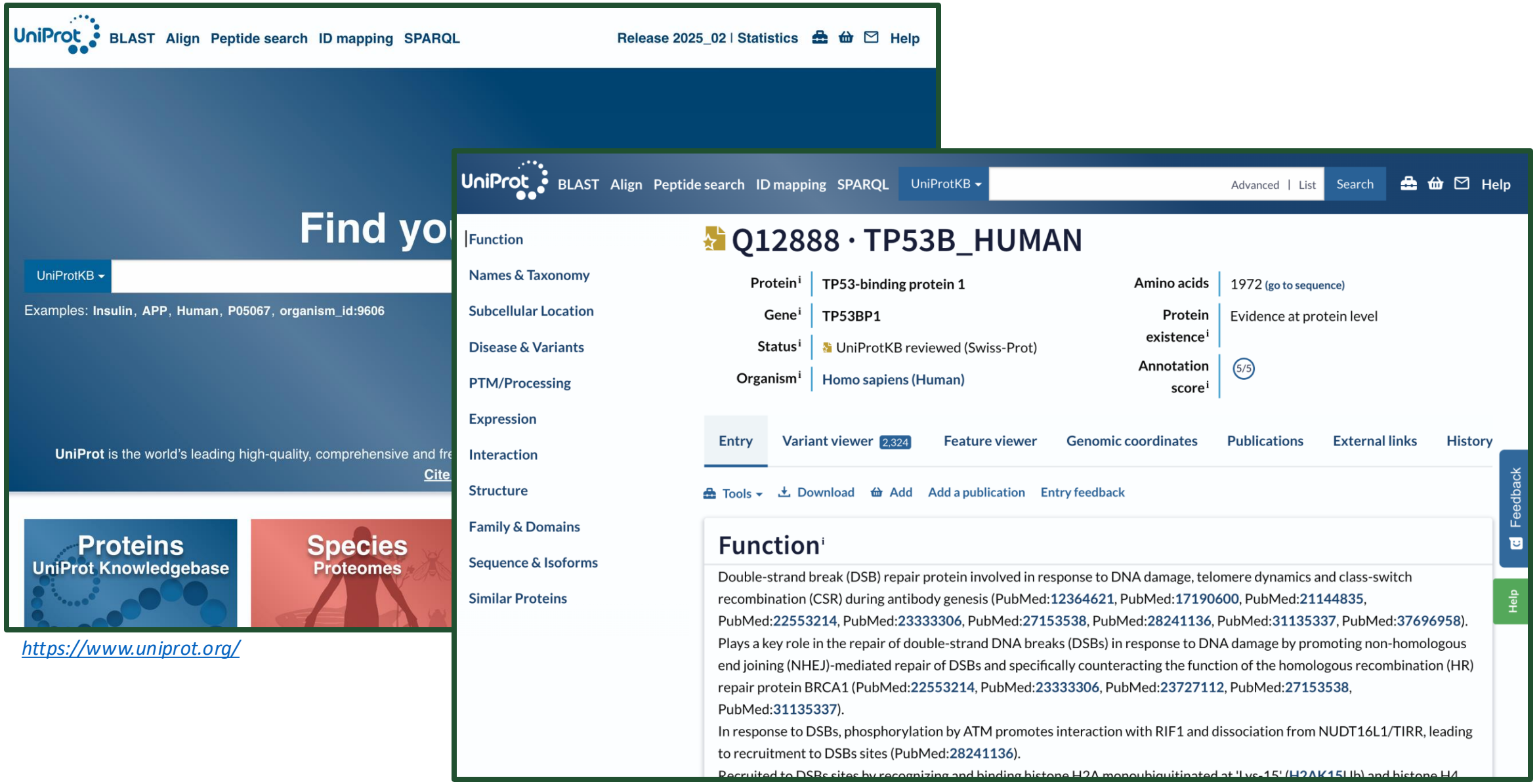


https://gtrnadb.ucsc.edu/



https://www.mirbase.org/

PROTEIN | UniProt: the Universal Protein Knowledgebase



https://www.uniprot.org/

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

KEGG Databases Tools Auto annotation Kanehisa Lab

KEGG PATHWAY Datab

Wiring diagrams of molecular inte

KEGG2 PATHWAY BRITE MODULE KO GENES

Select prefix: map | Organism

Enter keywords

Pathway Maps

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

1. **Metabolism**
   Global/overview Carbohydrate Energy Lipid Nuc
   Cofactor/vitamin Terpenoid/PK Other secondary m
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

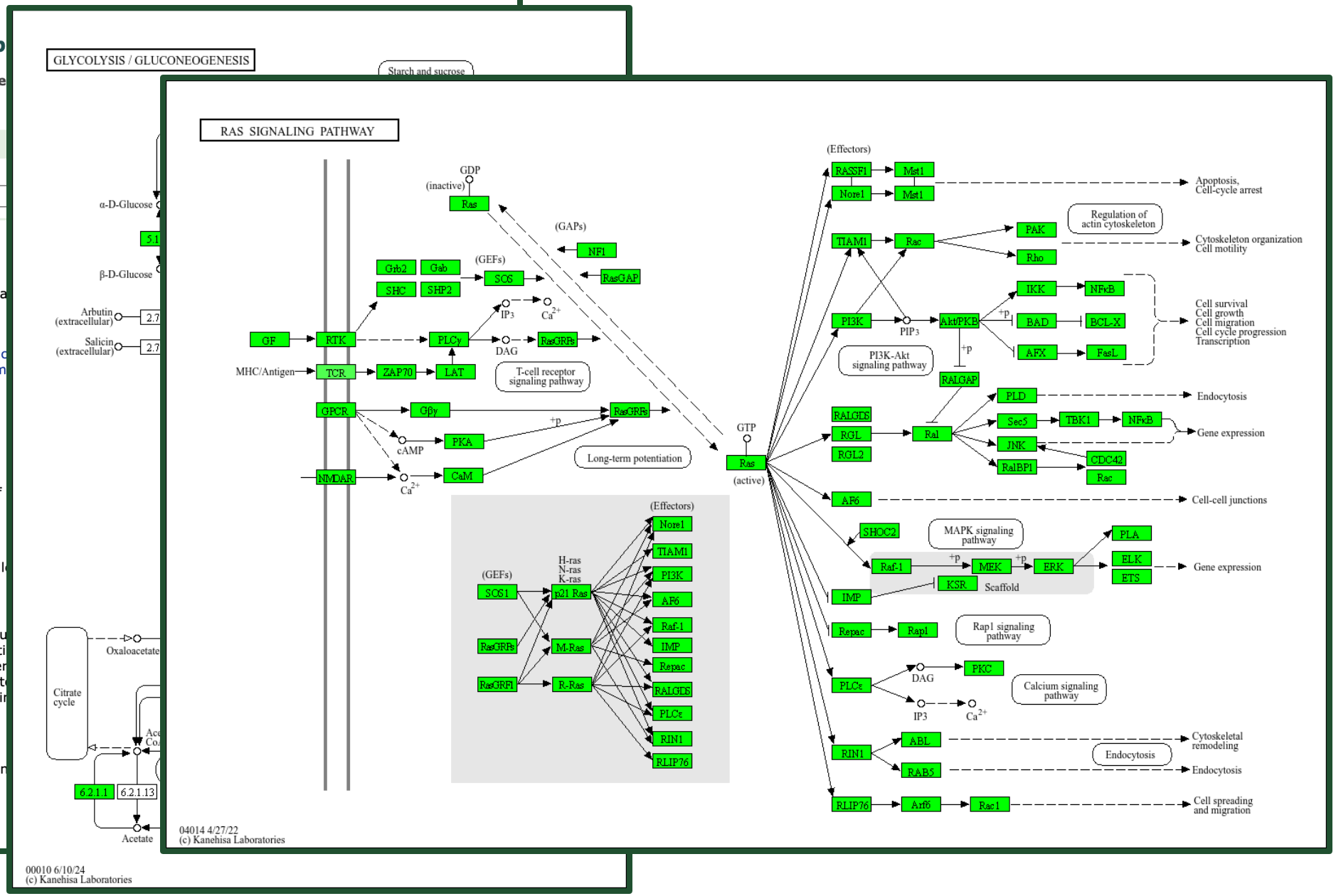
Each pathway map is identified by the combination of 2-4 l has the following meaning:

map	manually drawn reference pathwayAnnotation
ko	reference pathway highlighting KOsidentifiers
ec	reference metabolic pathway highlighting EC nu
rn	reference metabolic pathway highlighting reacti
<org>	organism-specific pathway generated by conver
vg	viruses pathway generated by converting KOs t
vx	viruses extended pathway generated by includir

and the numbers starting with the following:

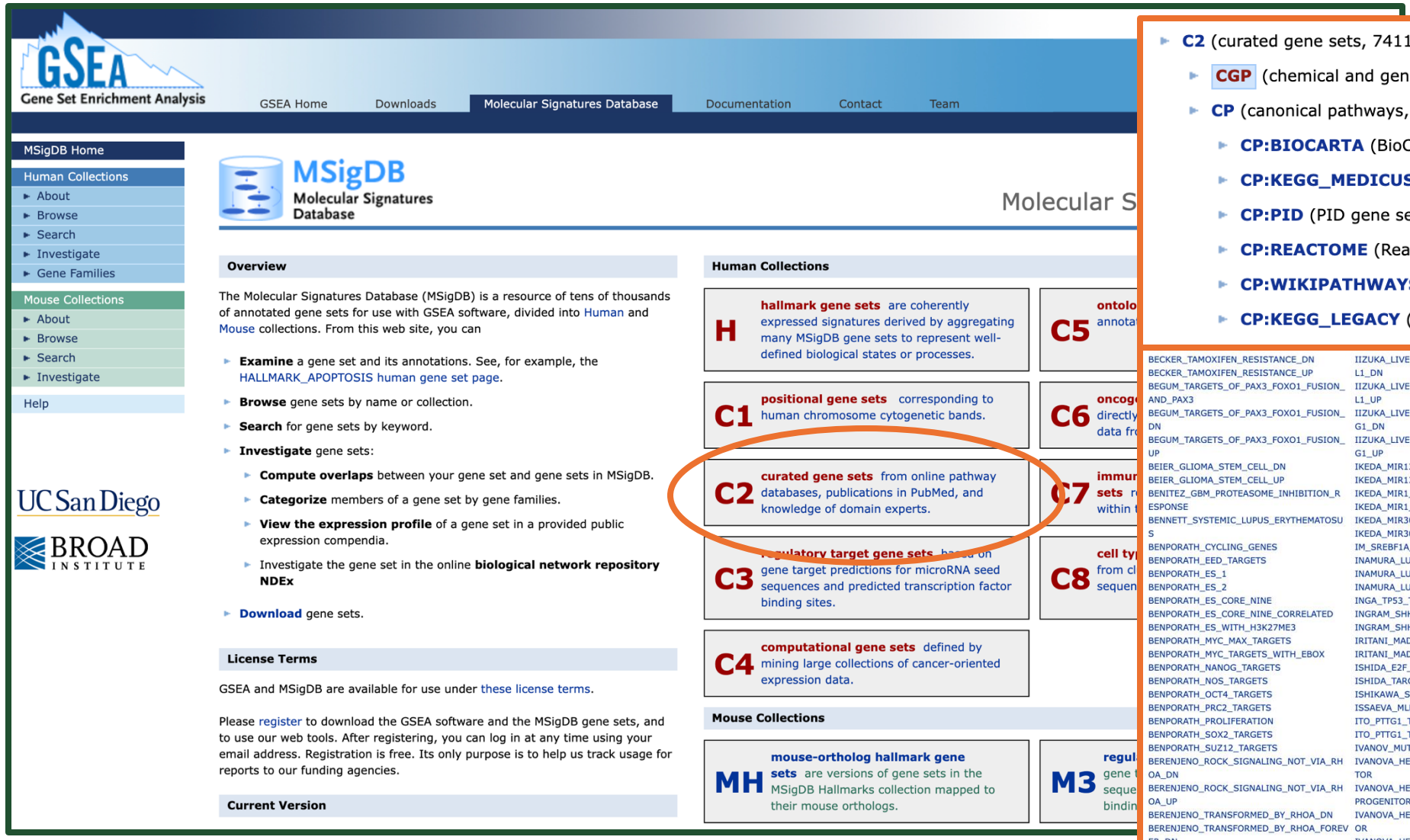
011	global map (lines linked to KOs)
012	overview map (lines linked to KOs)
013	higher (multi-organism) level overview map (lin
010	chemical structure map (no KO expansion)
07	drug structure map (no KO expansion)
other	regular map (boxes linked to KOs)

are used for different types of maps.



https://www.kegg.jp/

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



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

- **C2** (curated gene sets, 7411 gene sets)
  - **CGP** (chemical and genetic pertubations, 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
BECKER_TAMOXIFEN_RESISTANCE_UP
BEGUM_TARGETS_OF_PAX3_FOXO1_FUSION_AND_PAX3
BEGUM_TARGETS_OF_PAX3_FOXO1_FUSION_DN
BEGUM_TARGETS_OF_PAX3_FOXO1_FUSION_UP
BEIER_GLIOMA_STEM_CELL_DN
BEIER_GLIOMA_STEM_CELL_UP
BENITEZ_GBM_PROTEASOME_INHIBITION_RESPONSE
BENNETT_SYSTEMIC_LUPUS_ERYTHEMATOSUS
BENPORATH_CYCLING_GENES
BENPORATH_EED_TARGETS
BENPORATH_ES_1
BENPORATH_ES_2
BENPORATH_ES_CORE_NINE
BENPORATH_ES_CORE_NINE_CORRELATED
BENPORATH_ES_WITH_H3K27ME3
BENPORATH_MYC_MAX_TARGETS
BENPORATH_MYC_TARGETS_WITH_EBOX
BENPORATH_NANOG_TARGETS
BENPORATH_NOS_TARGETS
BENPORATH_OCT4_TARGETS
BENPORATH_PRC2_TARGETS
BENPORATH_PROLIFERATION
BENPORATH_SOX2_TARGETS
BENPORATH_SUZ12_TARGETS
BERENJENO_ROCK_SIGNALING_NOT_VIA_RHOA_DN
BERENJENO_ROCK_SIGNALING_NOT_VIA_RHOA_UP
BERENJENO_TRANSFORMED_BY_RHOA_DN
BERENJENO_TRANSFORMED_BY_RHOA_FOREVER_DN
BERENJENO_TRANSFORMED_BY_RHOA_FOREVER_UP
BERENJENO_TRANSFORMED_BY_RHOA_REVERSIBLY_DN
BERENJENO_TRANSFORMED_BY_RHOA_REVERSIBLY_UP
BERENJENO_TRANSFORMED_BY_RHOA_UP
BERGER_MBD2_TARGETS
BERNARD_PPAPDC1B_TARGETS_DN
IIZUKA_LIVER_CANCER_PROGRESSION_L0_L1_DN
IIZUKA_LIVER_CANCER_PROGRESSION_L0_L1_UP
IIZUKA_LIVER_CANCER_PROGRESSION_L1_G1_DN
IIZUKA_LIVER_CANCER_PROGRESSION_L1_G1_UP
IKEDA_MIR133_TARGETS_DN
IKEDA_MIR133_TARGETS_UP
IKEDA_MIR1_TARGETS_DN
IKEDA_MIR1_TARGETS_UP
IKEDA_MIR30_TARGETS_DN
IKEDA_MIR30_TARGETS_UP
IM_SREBF1A_TARGETS
INAMURA_LUNG_CANCER_SCC_DN
INAMURA_LUNG_CANCER_SCC_SUBTYPES_UP
INAMURA_LUNG_CANCER_SCC_UP
INGA_TP53_TARGETS
INGRAM_SHH_TARGETS_DN
INGRAM_SHH_TARGETS_UP
IRITANI_MAD1_TARGETS_DN
IRITANI_MAD1_TARGETS_UP
ISHIDA_E2F_TARGETS
ISHIDA_TARGETS_OF_SYT_SSX_FUSIONS
ISHIKAWA_STING_SIGNALING
ISSAEVA_MLL2_TARGETS
ITO_PTTG1_TARGETS_DN
ITO_PTTG1_TARGETS_UP
IVANOV_MUTATED_IN_COLON_CANCER
IVANOVA_HEMATOPOIESIS_EARLY_PROGENITOR
IVANOVA_HEMATOPOIESIS_INTERMEDIATE_PROGENITOR
IVANOVA_HEMATOPOIESIS_LATE_PROGENITOR
IVANOVA_HEMATOPOIESIS_MATURE_CELL
IVANOVA_HEMATOPOIESIS_STEM_CELL
IVANOVA_HEMATOPOIESIS_STEM_CELL_AND_PROGENITOR
IVANOVA_HEMATOPOIESIS_STEM_CELL_LONG_TERM
IVANOVA_HEMATOPOIESIS_STEM_CELL_SHORT_TERM
IVANOVSKA_MIR106B_TARGETS
IWANAGA_CARCINOGENESIS_BY_KRAS_DN
PYEON_CANCER_HEAD_AND_NECK_VS_CERVICAL_DN
PYEON_CANCER_HEAD_AND_NECK_VS_CERVICAL_UP
PYEON_HPV_POSITIVE_TUMORS_DN
PYEON_HPV_POSITIVE_TUMORS_UP
QI_HYPOXIA
QI_HYPOXIA_TARGETS_OF_HIF1A_AND_FOXA2
QI_PLASMACYTOMA_DN
QI_PLASMACYTOMA_UP
QUINTENS_EMBRYONIC_BRAIN_RESPONSE_TO_IR
RADAEVA_RESPONSE_TO_IFNA1_DN
RADAEVA_RESPONSE_TO_IFNA1_UP
RADMACHER_AML_PROGNOSIS
RAFFEL_VEGFA_TARGETS_DN
RAFFEL_VEGFA_TARGETS_UP
RAGHAVACHARI_PLATELET_SPECIFIC_GENES
RAHMAN_TP53_TARGETS_PHOSPHORYLATED
RAMALHO_STEMNESS_DN
RAMALHO_STEMNESS_UP
RAMASWAMY_METASTASIS_DN
RAMASWAMY_METASTASIS_UP
RAMJAUN_APOPTOSIS_BY_TGFB1_VIA_MAPK1_DN
RAMJAUN_APOPTOSIS_BY_TGFB1_VIA_MAPK1_UP
RAMJAUN_APOPTOSIS_BY_TGFB1_VIA_SMAD4_DN
RAMJAUN_APOPTOSIS_BY_TGFB1_VIA_SMAD4_UP
RAMPON_ENRICHED_LEARNING_ENVIRONMENT_EARLY_DN
RAMPON_ENRICHED_LEARNING_ENVIRONMENT_EARLY_UP
RAMPON_ENRICHED_LEARNING_ENVIRONMENT_LATE_UP
RANKIN_ANGIOGENIC_TARGETS_OF_VHL_HIF2A_DN
RANKIN_ANGIOGENIC_TARGETS_OF_VHL_HIF2A_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

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** ▾

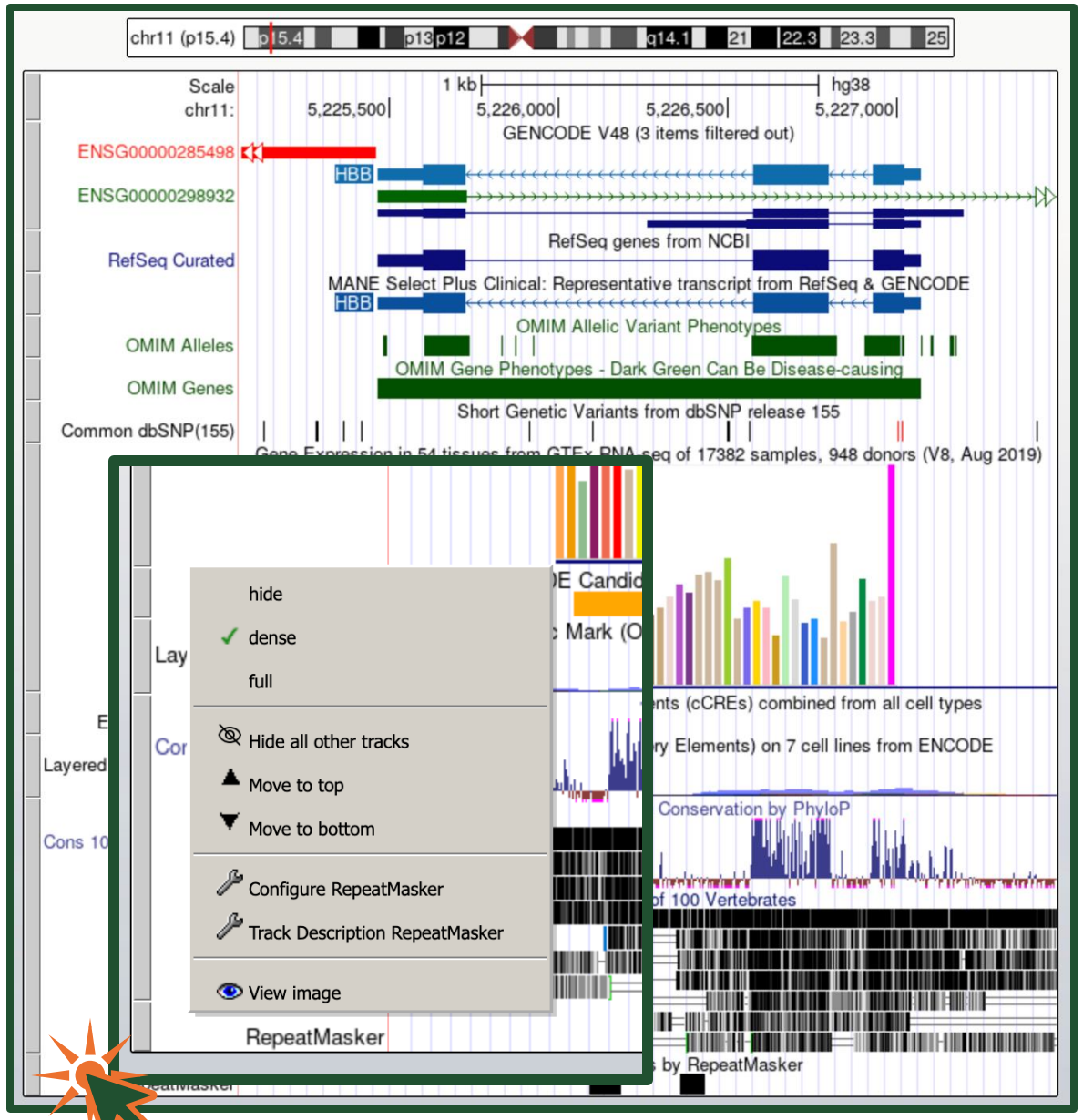
JOURNAL ARTICLE

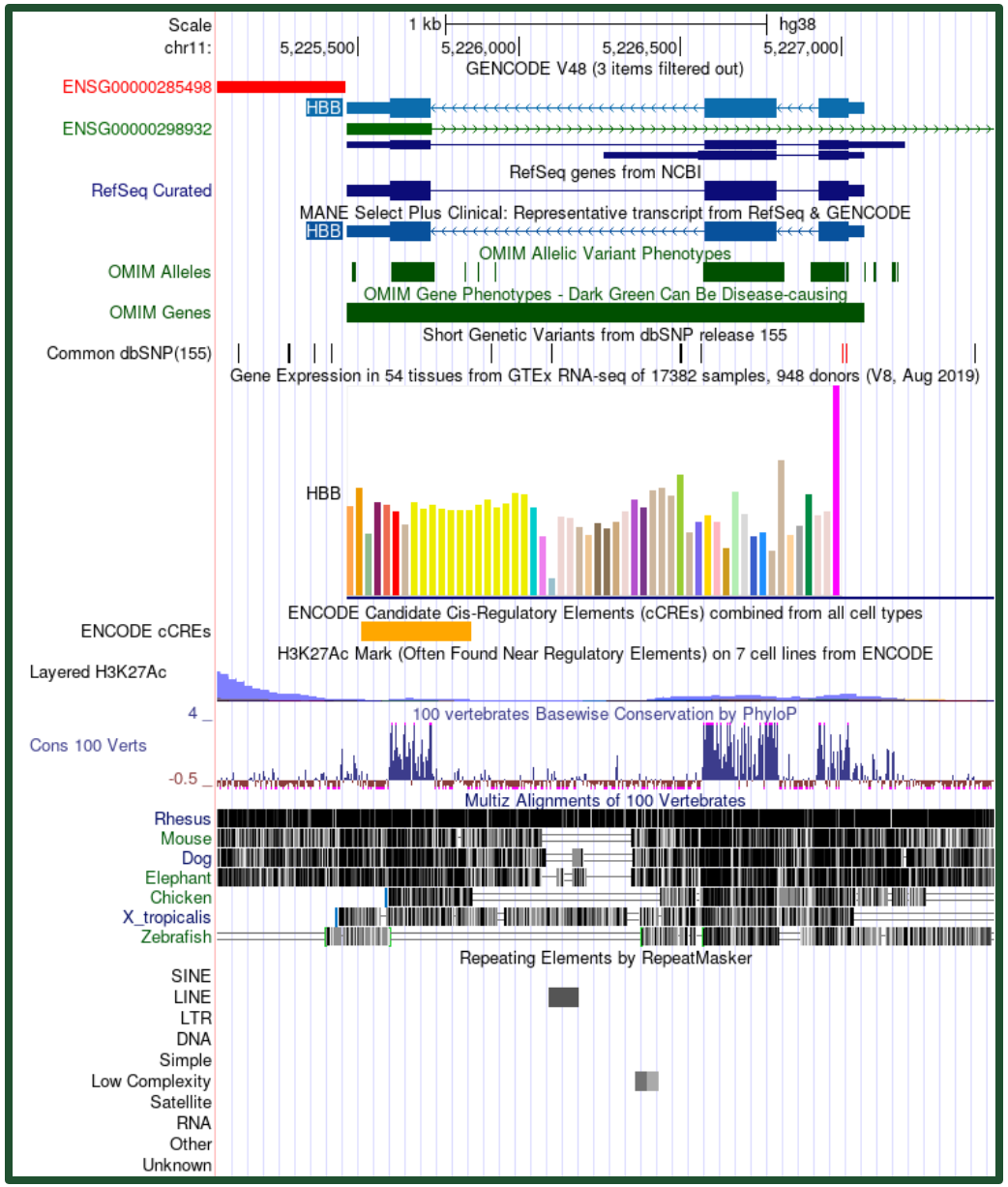
VISTA Enhancer browser: an updated database of tissue-specific developmental enhancers

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 ✉

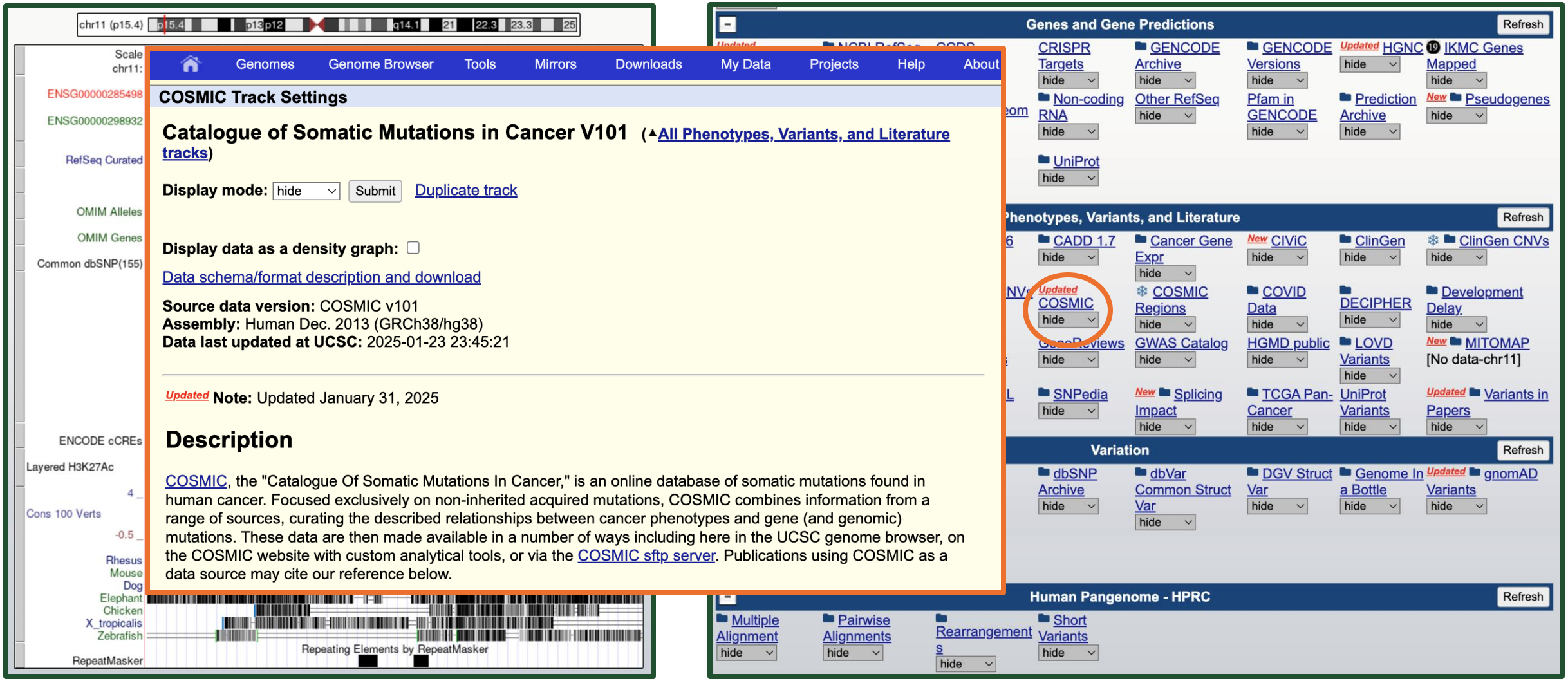
Nucleic Acids Research, Volume 53, Issue D1, 6 January 2025, Pages D324–D330, https://doi.org/10.1093/nar/gkae940
Published: 29 October 2024 **Article history** ▾

The UCSC Genome Browser contains a rich collection of data tracks

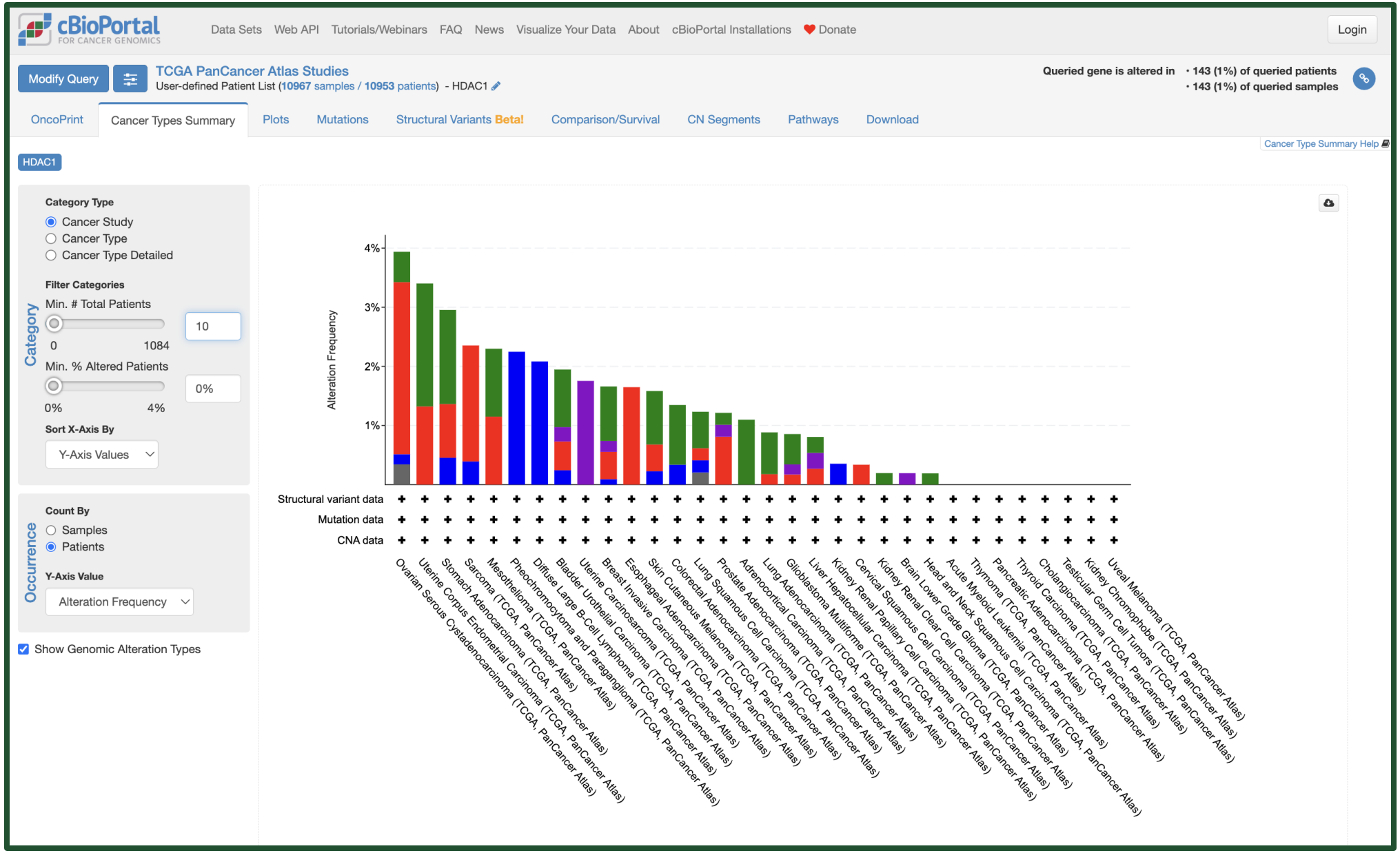




The UCSC Genome Browser contains a rich collection of data tracks



cBioPortal includes cancer-relevant genomic information



https://www.cbioportal.org/results/cancerTypesSummary?case_set_id=all&gene_list=HDAC1&cancer_study_list=5c8a7d55e4b046111fee2296

Mutation

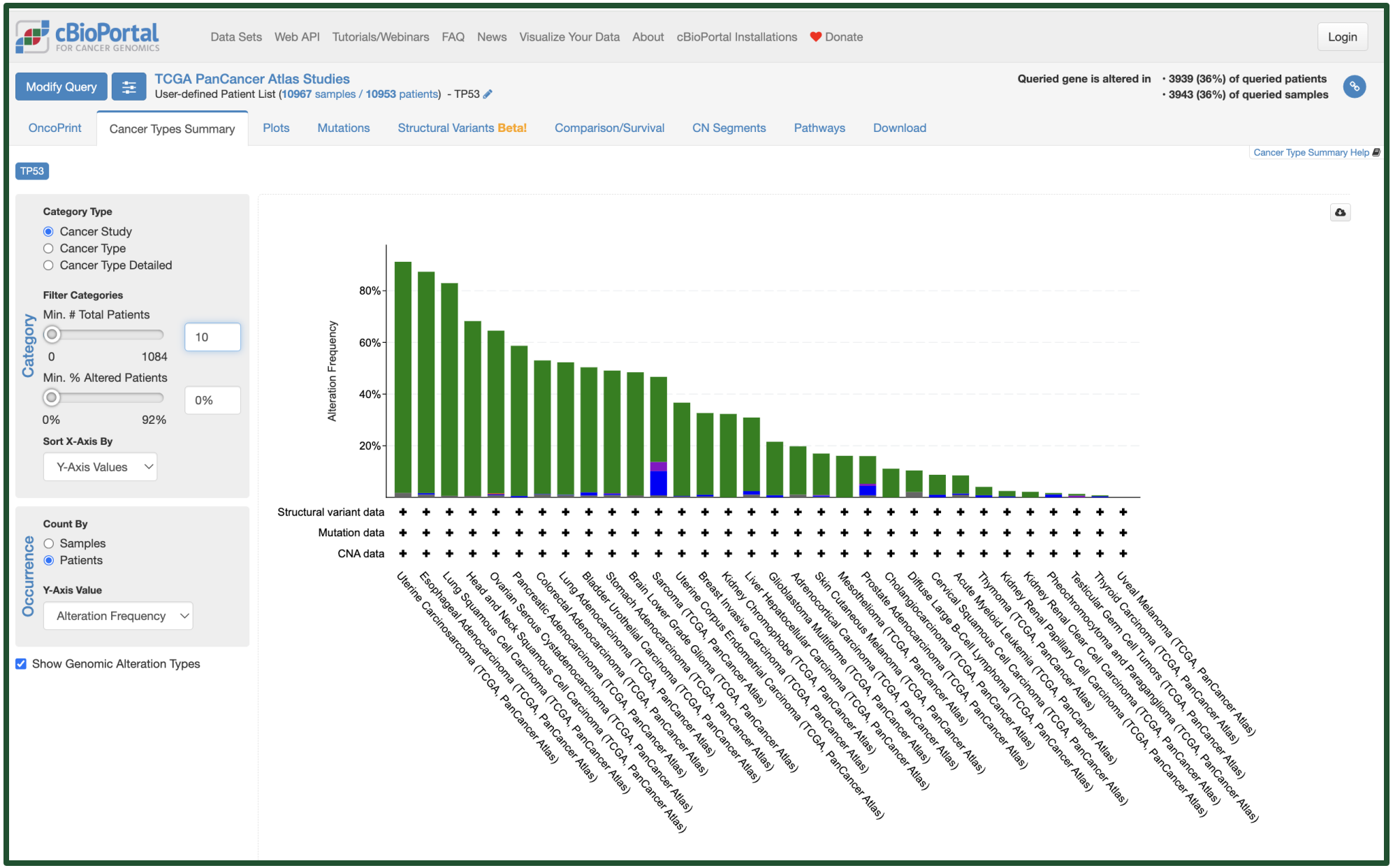
Amplification

Deep Deletion

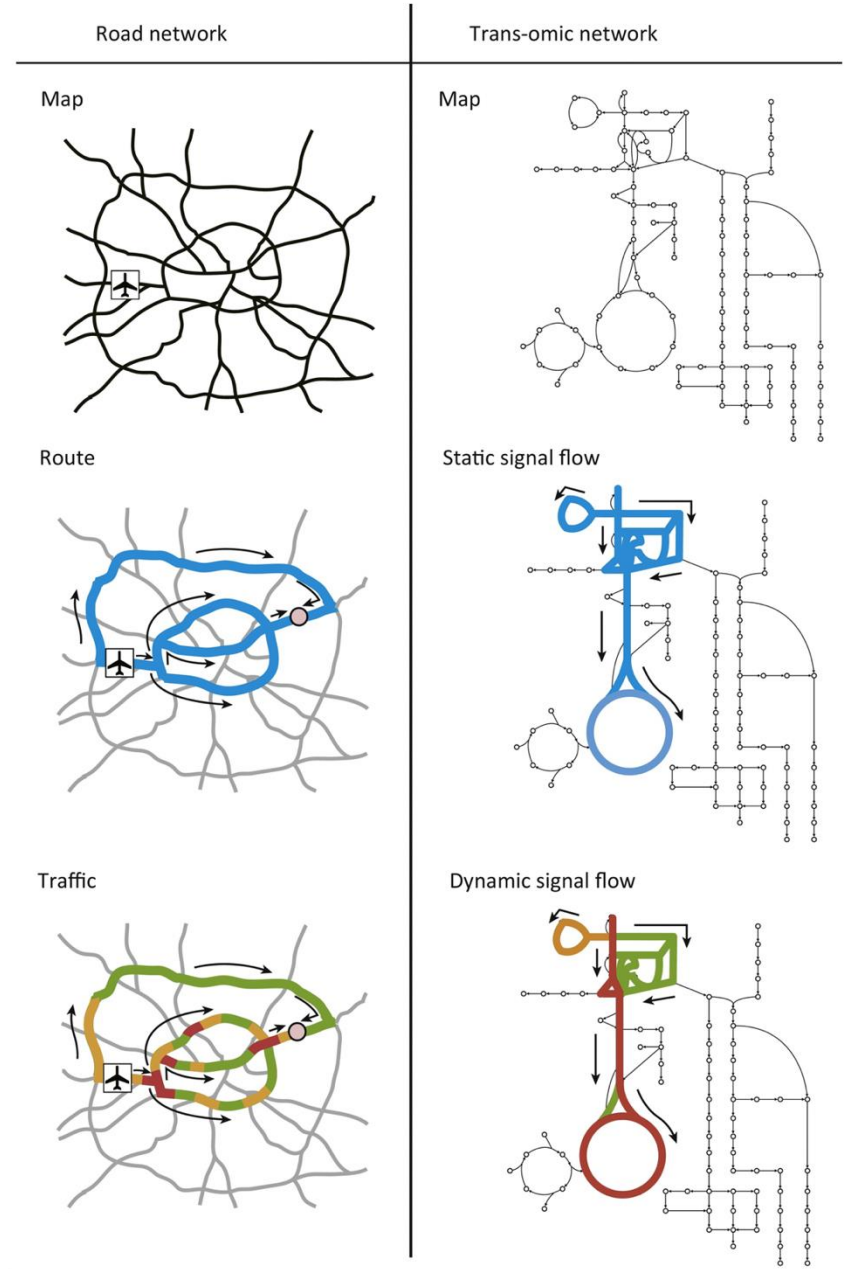
Multiple Alterations

Structural Variant

cBioPortal includes cancer-relevant genomic information



cBioPortal includes cancer-relevant genomic information



Yugi et al. (2016) Trends Biotechnol 34(4):276-290

Genome
What reactions <u>are able to</u> happen under any physiological conditions

Epigenome, Transcriptome
What reactions <u>seem to be</u> active under specific physiological conditions

Proteome, Metabolome
What reactions <u>are</u> 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.

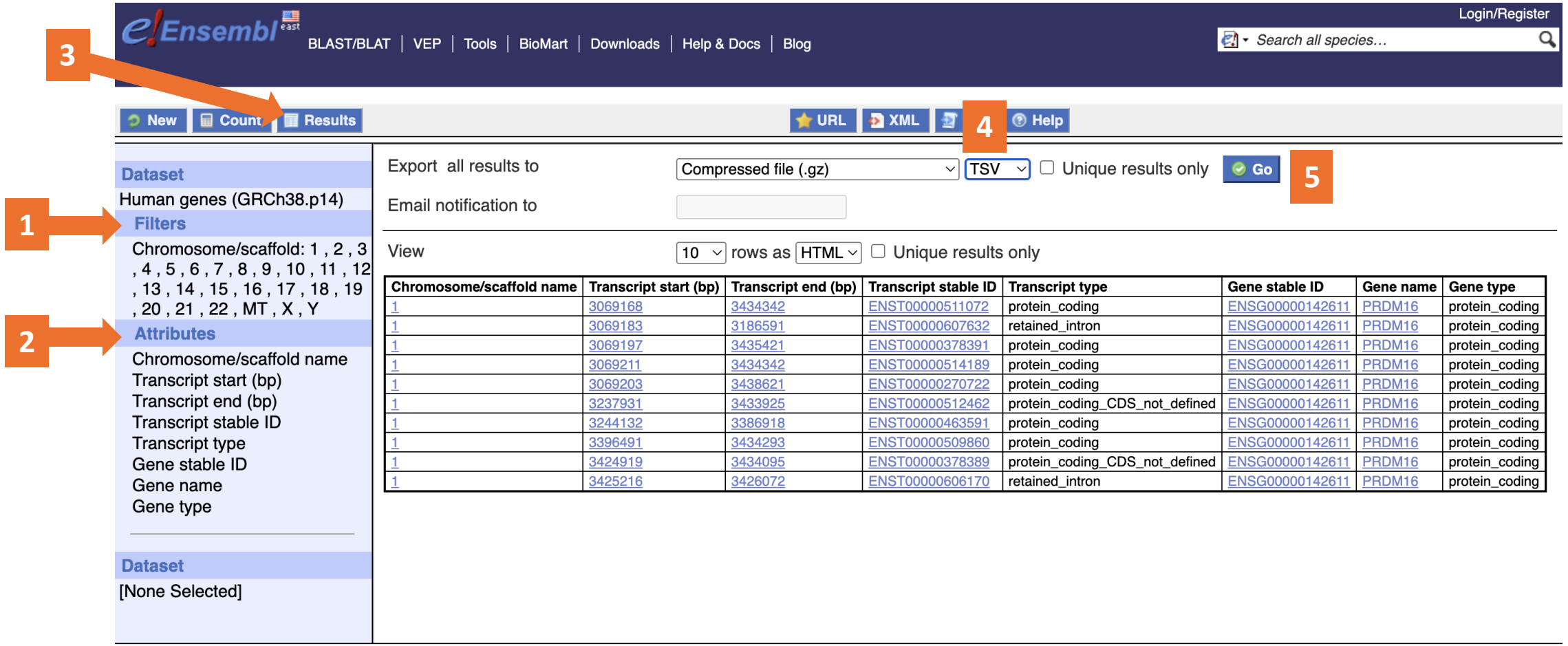
  From this page you can access all human genes.
  Clicking on “Attributes” you can choose what information you want to see.
  Clicking on “Filtering” you can filter the information.

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



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

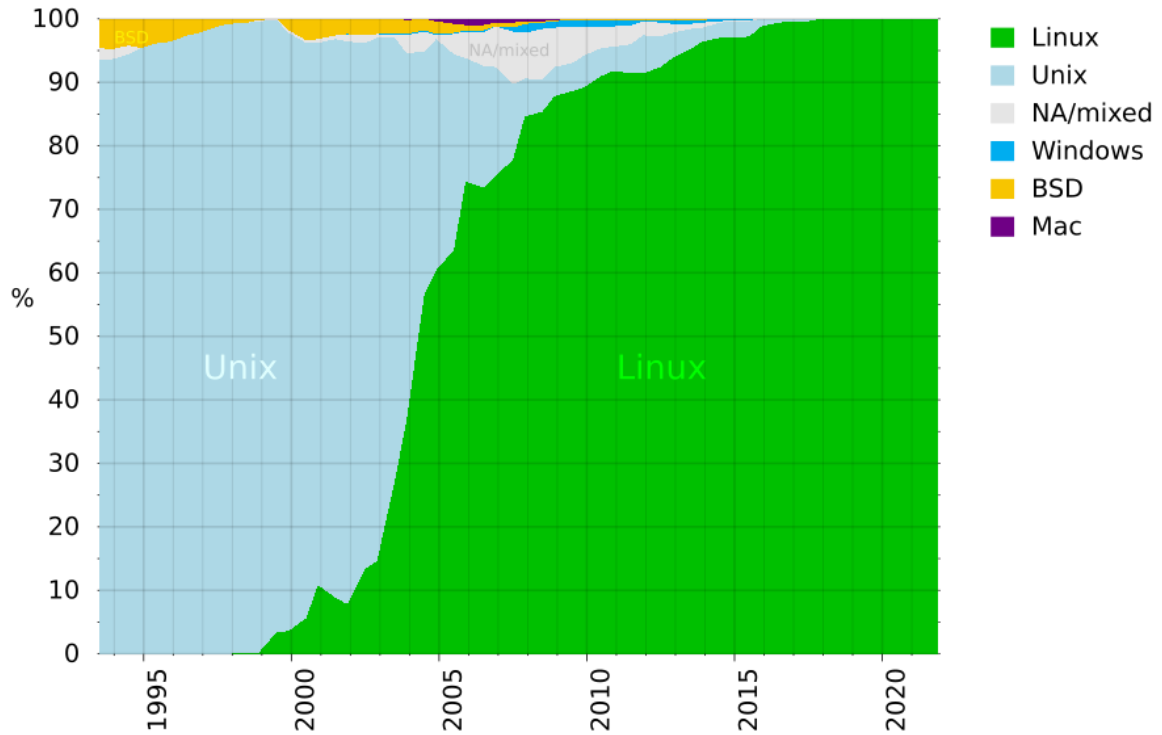
1. Click “Filters” and select these chromosomes
   (you will find them under “REGION”)
2. Click “Attributes” and select these ones in **<u>this specific order</u>**
   (you will find them under “GENE”)
3. Click ”Results”
4. Select “Compressed file (.gz)”
5. Click “Go” and save the file in the directory 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!

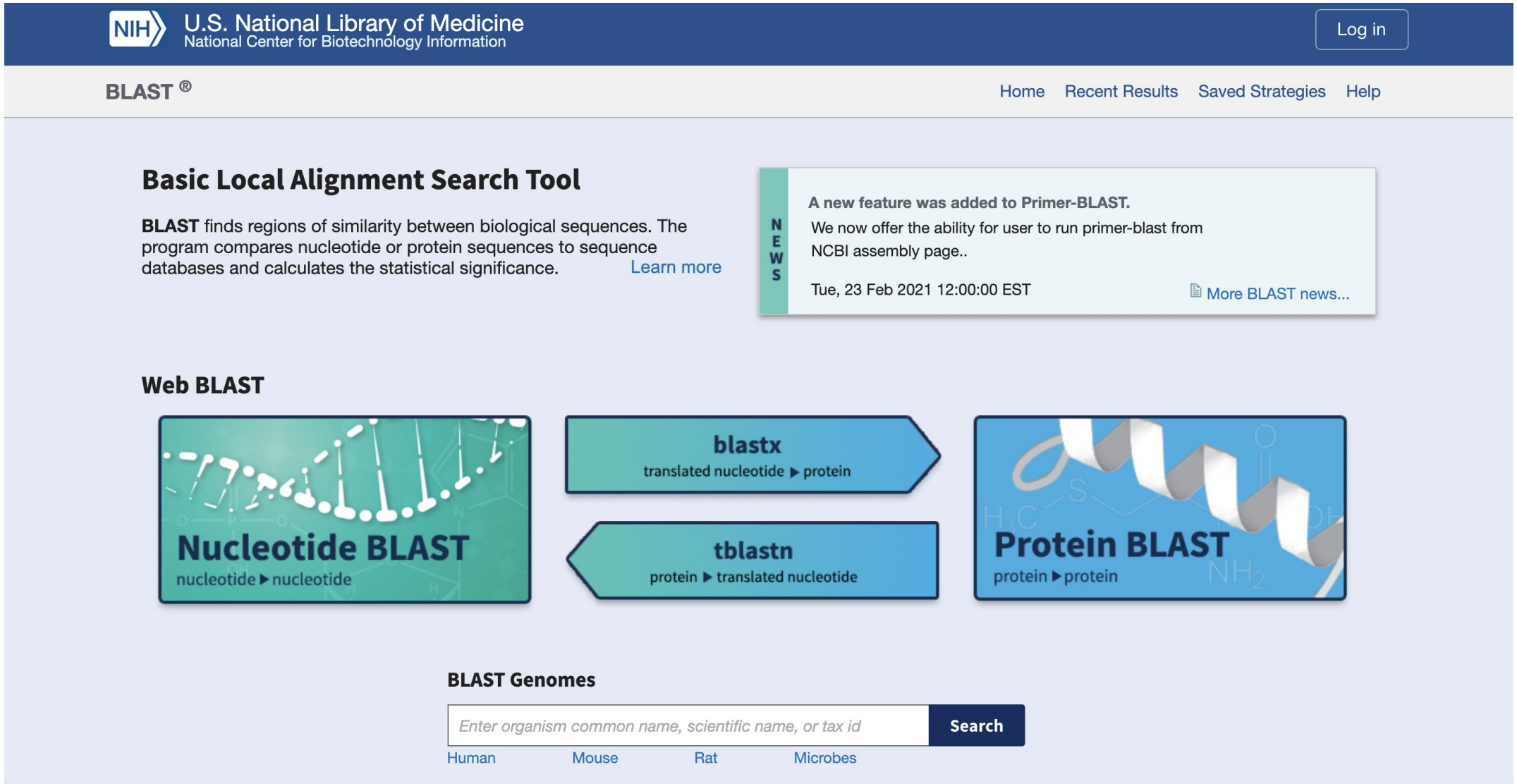
An online tutorial: https://linuxcommand.org/

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



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

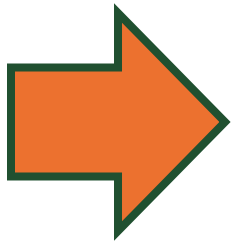
Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Homo sapiens hemoglobin subunit beta (HBB), mRNA	Homo sapiens	1160	1160	100%	0.0	100.00%	628	NM_000518.5
PREDICTED: Pan troglodytes hemoglobin subunit beta (HBB), mRNA	Pan troglodytes	1155	1155	100%	0.0	99.84%	748	XM_508242.4
PREDICTED: Pan paniscus hemoglobin subunit beta (LOC100976465), mRNA	Pan paniscus	1149	1149	100%	0.0	99.68%	646	XM_003819029.3
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
Human messenger RNA for beta-globin	Homo sapiens	1140	1140	99%	0.0	99.52%	626	V00497.1
PREDICTED: Gorilla gorilla gorilla hemoglobin subunit beta (LOC101126932), mRNA	Gorilla gorilla go…	1133	1133	100%	0.0	99.20%	753	XM_019036164.1
Homo sapiens hemoglobin beta mRNA, complete cds	Homo sapiens	1133	1133	99%	0.0	99.36%	647	AY509193.1
PREDICTED: Pongo abelii hemoglobin subunit beta (HBB), mRNA	Pongo abelii	1098	1098	99%	0.0	98.25%	627	XM_002822127.4
PREDICTED: Nomascus leucogenys hemoglobin subunit beta (HBB), mRNA	Nomascus leuc…	1088	1088	100%	0.0	97.93%	753	XM_004090649.3
PREDICTED: Hylobates moloch hemoglobin subunit beta (HBB), mRNA	Hylobates moloch	1077	1077	100%	0.0	97.61%	753	XM_032166808.1
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
Homo sapiens hemoglobin beta subunit variant (HBB) mRNA, complete cds	Homo sapiens	1064	1064	92%	0.0	99.83%	579	AF181989.1
Homo sapiens beta globin chain variant (HBB) mRNA, complete cds	Homo sapiens	1057	1057	92%	0.0	99.48%	581	AF349114.1
PREDICTED: Trachypithecus francoisi hemoglobin subunit beta (LOC117081249), mRNA	Trachypithecus f…	1038	1038	100%	0.0	96.51%	749	XM_033207068.1
PREDICTED: Colobus angolensis palliatus hemoglobin, beta (HBB), mRNA	Colobus angole…	1038	1038	100%	0.0	96.50%	753	XM_011963601.1
PREDICTED: Rhinopithecus roxellana hemoglobin subunit beta (LOC104662339), mRNA	Rhinopithecus r…	1033	1033	100%	0.0	96.35%	749	XM_010363344.2
PREDICTED: Rhinopithecus bieti hemoglobin subunit beta (LOC108521664), mRNA	Rhinopithecus b…	1033	1033	100%	0.0	96.35%	749	XM_017861621.1
Homo sapiens hemoglobin beta chain (HBB) mRNA, complete cds	Homo sapiens	1027	1027	89%	0.0	99.64%	562	AF117710.1
Human sickle beta-hemoglobin mRNA	Homo sapiens	1024	1024	91%	0.0	98.44%	576	M25113.1
PREDICTED: Mandrillus leucophaeus hemoglobin subunit beta (LOC105535916), transcript variant X2, mRNA	Mandrillus leuco…	1022	1022	100%	0.0	96.02%	753	XM_0119751
PREDICTED: Piliocolobus tephrosceles hemoglobin subunit beta (HBB), mRNA	Piliocolobus tep…	1016	1016	100%	0.0	95.87%	748	XM_0232096

Fe

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

Query file

```
>MyQuery
ACATTTGCTTCTGACACAACTGTGTTCACTAG
CAACCTCAAACAGACACCATGGTGCATCTGAC
TCCTGAGGAGAAGTCTGCCGTTACTGCCCTGT
GGGGCAAGGTGAACGTGGATGAAGTTGGTGGT
GAGGCCCTGGGCAGGCTGCTGGTGGTCTACCC
TTGGACCCAGAGGTTCTTTGAGTCCTTTGGGG
ATCTGTCCACTCCTGATGCTGTTATGGGCAAC
CCTAAGGTGAAGGCTCATGGCAAGAAAGTGCT
CGGTGCCTTTAGTGATGGCCTGGCTCACCTGG
ACAACCTCAAGGGCACCTTTGCCACACTGAGT
GAGCTGCACTGTGACAAGCTGCACGTGGATCC
TGAGAACTTCAGGCTCCTGGGCAACGTGCTGG
TCTGTGTGCTGGCCCATCACTTTGGCAAAGAA
TTCACCCCACCAGTGCAGGCTGCCTATCAGAA
AGTGGTGGCTGGTGTGGCTAATGCCCTGGCCC
ACAAGTATCACTAAGCTCGCTTTCTTGCTGTC
CAATTTCTATTAAAGGTTCCTTTGTTCCCTAA
GTCCAACTACTAAACTGGGGGATATTATGAAG
GGCCTTGAGCATCTGGATTCTGCCTAATAAAA
AACATTTATTTTCATTGCAA
```

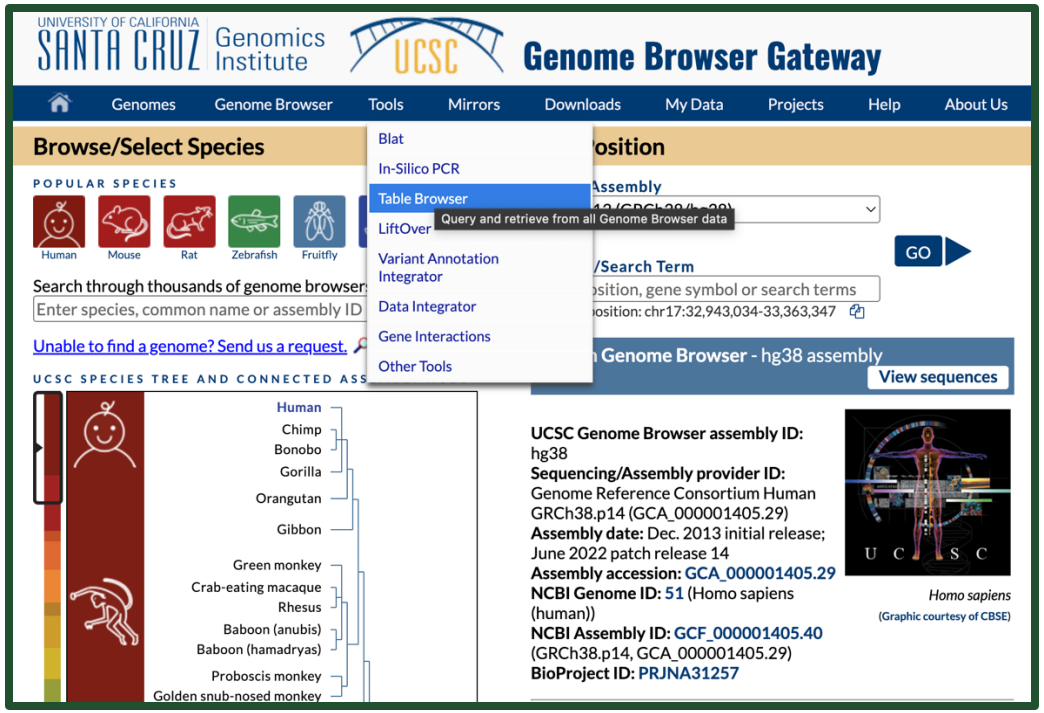


Database

```
>chr1
ACAACTGTGTTCACTAGCAAACATTTGCTTCTGAC……
>chr2
ACATTTGCTTCTGACACAACTGTGTTCACTAGCAA……
……
>chrX
TCTGACACAACTGTGTTCACTAGCAAACATTTGCT……
```

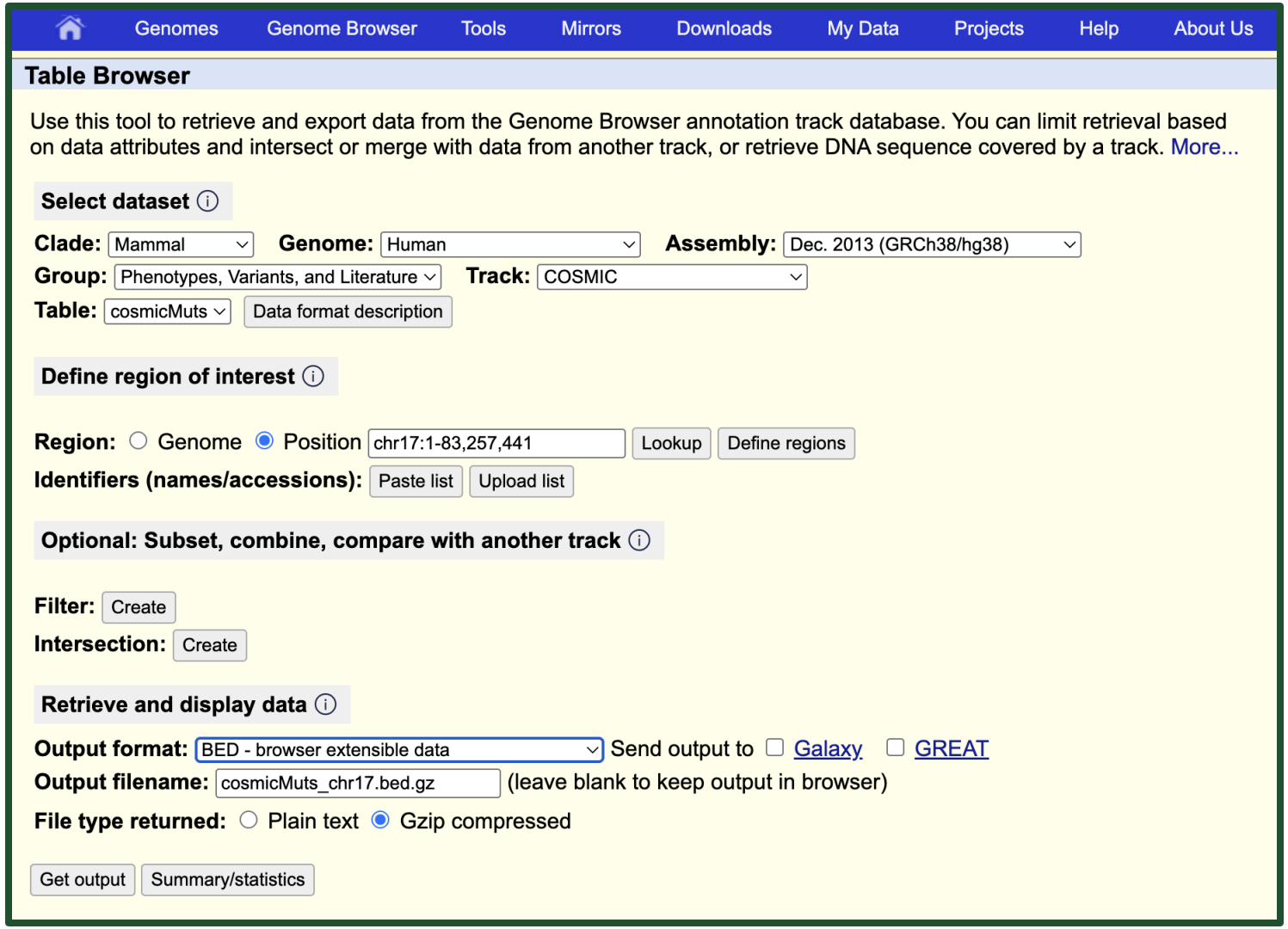
Fasta files (.fa) are used to store sequences in a simple format: Lines starting with “>” contain the name, description or any other information about the sequence and are followed by the sequence (in one or multiple lines) until the next “>” (if any).

Table Browser: querying annotation tables



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

How to access the COSMIC mutations from the Table Browser



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

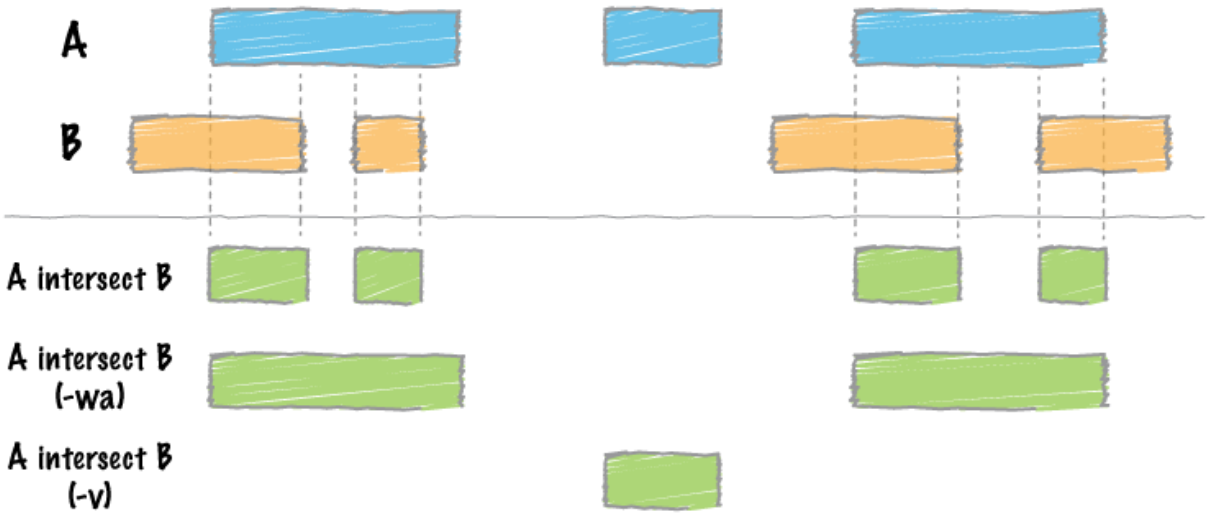
Sample Rows

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

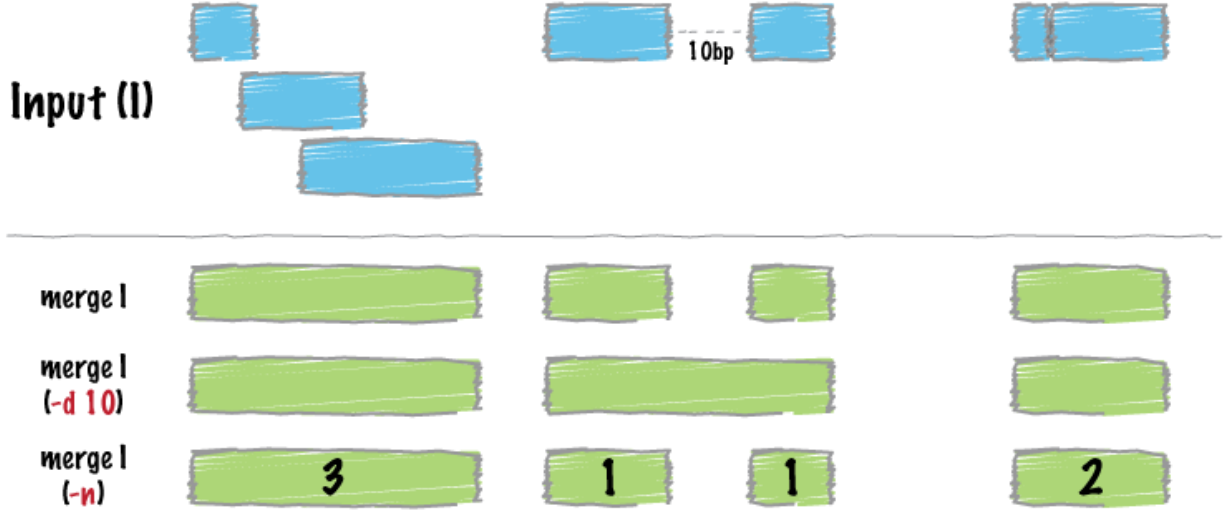
Genomic arithmetic using bedtools

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

We should be ready to start with exercise 2