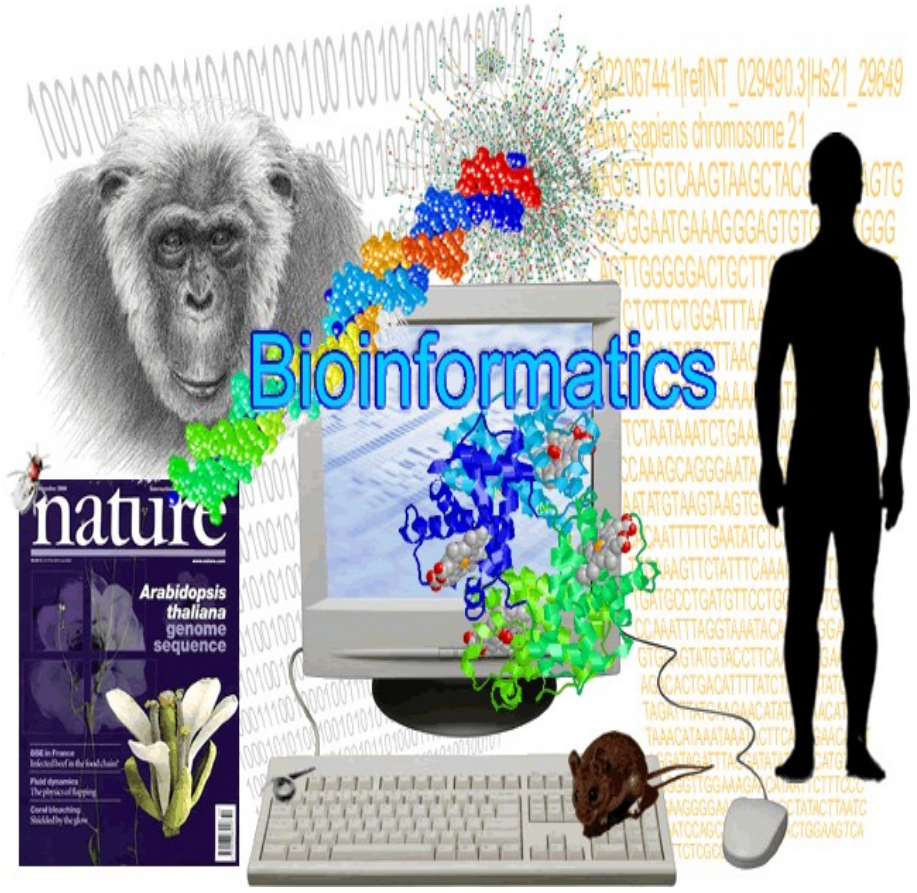
# Databases in molecular biology

Bioinformatics Course UEB-VHIR
November 2023

**Mireia Ferrer[1], Álex Sánchez[1,2], Esther Camacho[1], Berta Miró[1]**
[1] Unitat d'Estadística i Bioinformàtica (UEB) VHIR
[2] Departament de Genètica Microbiologia i Estadística, UB

# Information in the omics era




- Massive quantities of information (not necessarily “big data”)
- Open-access
- For this information to be accessible it must be properly stored.
- Access to information
  - Must be fast
  - Must be flexible
- This has been made possible
  - Creating databases
  - Distributing them through the web

# Biological Databases

- **Definition:** *libraries* of life sciences information, collected from scientific experiments, published literature, high-throughput experiment technology and computational analysis
- **What are they for?**
  - Storage of information
  - Data organization
  - Access information
  - Knowledge discovery
- There are many different general and specialized databases.
  - Large list published yearly in *NAR* : 1645 in 2022!
    - https://www.oxfordjournals.org/nar/database/c/
    - https://academic.oup.com/nar/article/50/D1/D1/6495890

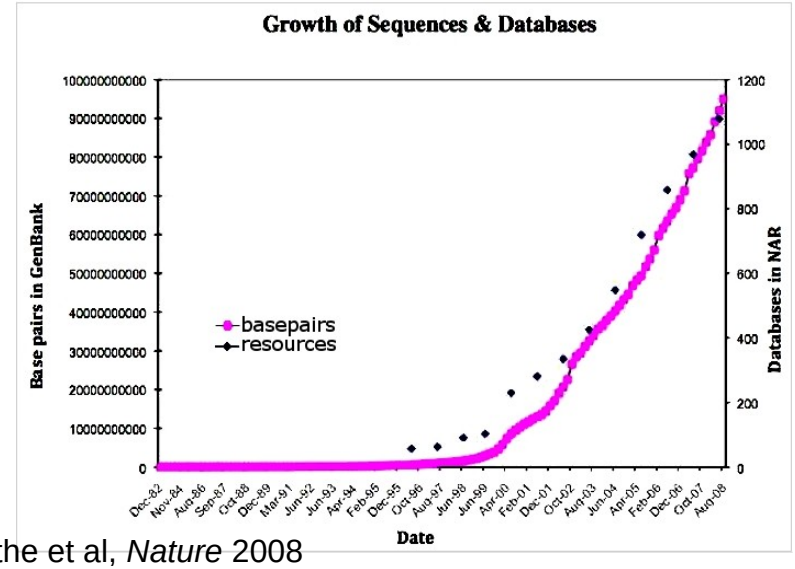
# Biological Databases

## NAR Database Summary Paper Category List

Nucleotide Sequence Databases
RNA sequence databases
Protein sequence databases
Structure Databases
Genomics Databases (non-vertebrate)
Metabolic and Signaling Pathways
Human and other Vertebrate Genomes
Human Genes and Diseases
Microarray Data and other Gene Expression Databases
Proteomics Resources
Other Molecular Biology Databases
Organelle databases
Plant databases
Immunological databases
Cell biology

- Nucleotide Databases
  - ASD
  - ATD
  - EMBL-Bank
  - EMBL CDS
  - Ensembl
  - Genome Reviews
  - IMGT/HLA
- Protein Databases
  - CSA
  - GOA
  - IntAct
  - IntEnz
  - InterPro
- Microarray Databases
  - ArrayExpress
  - MIAME
- Literature Databases
  - MEDLINE
  - OMIM
  - Patent Abstracts
  - more...

# Challenges




Lathe et al, *Nature* 2008

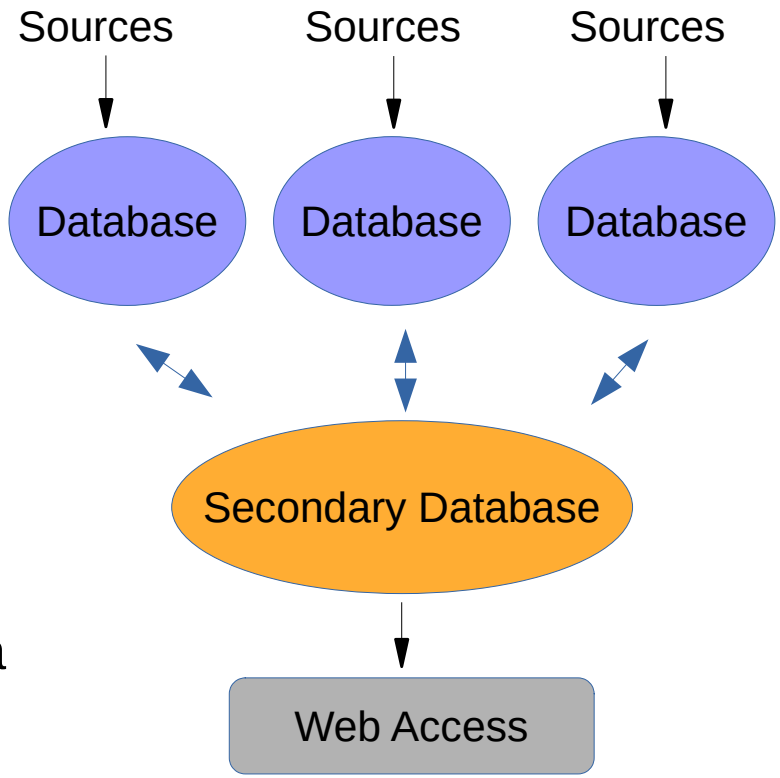
This large number of databases, though extremely useful, can lead to its own issues of redundancy and lack of integration.

- **Structure/Integrate information**
- **Annotation and Curation**
- **Centralize data management**

# I. Structuring and Integrating the information

# Integrating the information

- **Primary databases:**
  - often hold only one type of specific data which is stored in their own archive.
  - upload new data from experiments and update entries
- **Secondary databases:**
  - use other databases as their source of information.
  - often already process or analyze the data to get new results.




**Different formats and models for structuring the data**

# Integrating the information

- The quality of the information in a database is closely related to its structure
- This aspect is also crucial for its efficiency and accessibility.
- There are many different types of file formats depending on the type of information they contain / source of data
- Most are based on the flat-file format (text file)

FASTA format (sequences)

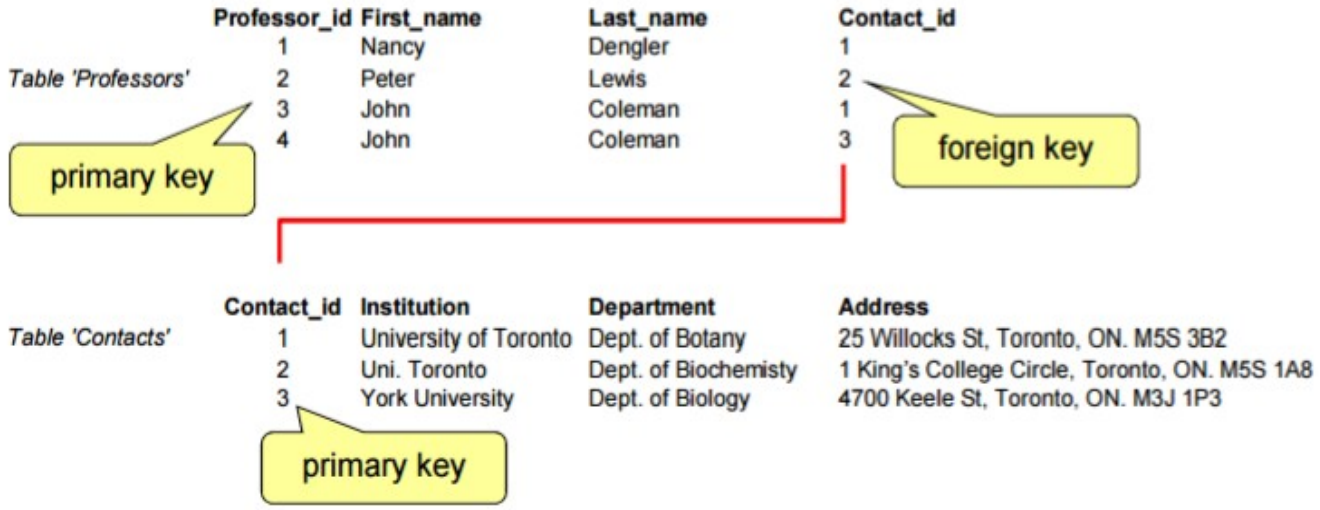
Identifier Description

sequence

```
>l36435 Mus Musculus basic domain/leucine zipper transcription factor mRNA, complete cds.
gcgccgccgcgtccccagacaaaggcttggccggcggccccggcccgctgcgccctcggctccccgcctccc
cggcttgccgctcttcgccccgcgtttggctcggcgcgtcccggccggccgcaaagttttccccgcggcag
cggcggctgagcctcgcttttagcgatggccgcggagctgagcatggggcaagagctgcccaccagcccgct
ggccatggagtacgtcaacgacttcgaccttctcaagttcgacgtgaagaaggagcccctggggcgcgcgga
gcgtccgggccggccatgcacacgcctgcagcctgctggctcggtgtcgtccaccccgctcagcactccgtg
```

# Integrating the information

- Different models exist to relate/integrate the information (Relational, Hierarchical, Networks...).
- Most common model: Relational databases
  - flat-file format (text file)
  - Many tables linked to each other: cross-referencing through a key (common) field (unique identifier)



# Integrating the information

- In many databases an entry can be identified in 2 (ore more!) different ways:
  - **Identifier** ("locus" in GenBank, "entry name" in UniProt): is a string of letters and digits. May change if the database curators decide that is no longer appropriate.
  - **Accession code (number)**: is a number (possibly with a few characters in front) that uniquely identifies an entry in its database. It is supposed to be stable.
  - **Versions and Gene Indices**: The same accession number may be associated with a different GI if a newer or corrected sequence is submitted. (+info)

---

Example: human gene ADH6

GenBank

```
LOCUS       AH001409                2625 bp    DNA     linear   PRI 10-JUN-2016
DEFINITION  Homo sapiens alcohol dehydrogenase 6 (ADH6) gene, complete cds.
ACCESSION   AH001409 M68895 M84402 M84403 M84404 M84405 M84406 M84407 M84408
            M84409
VERSION     AH001409.2
KEYWORDS    .
SOURCE      Homo sapiens (human)
```

UniProt

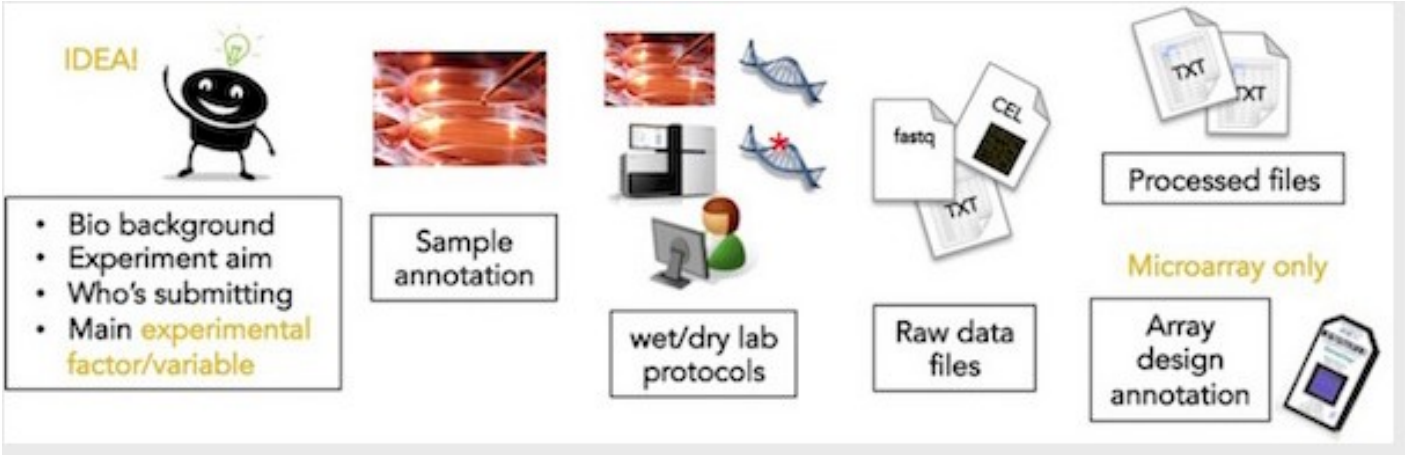
| Entry | Entry name | | Protein names | Gene names |
|---|---|---|---|---|
| P28332 | ADH6_HUMAN | | **Alcohol dehydrogenase 6** | **ADH6** |

# II. Data Annotation and Curation

# Data Annotation

- Different levels of annotation
  - Data: annotation of sequences/genomes (chromosome position, gene function, …)
  - Metadata: information for an experiment, identification of samples, …
- Collaborative efforts to provide as much information about the data
- The **Minimum Information Standard** is a set of guidelines for reporting data

MIAME (Minimum Information About a Microarray Experiment)

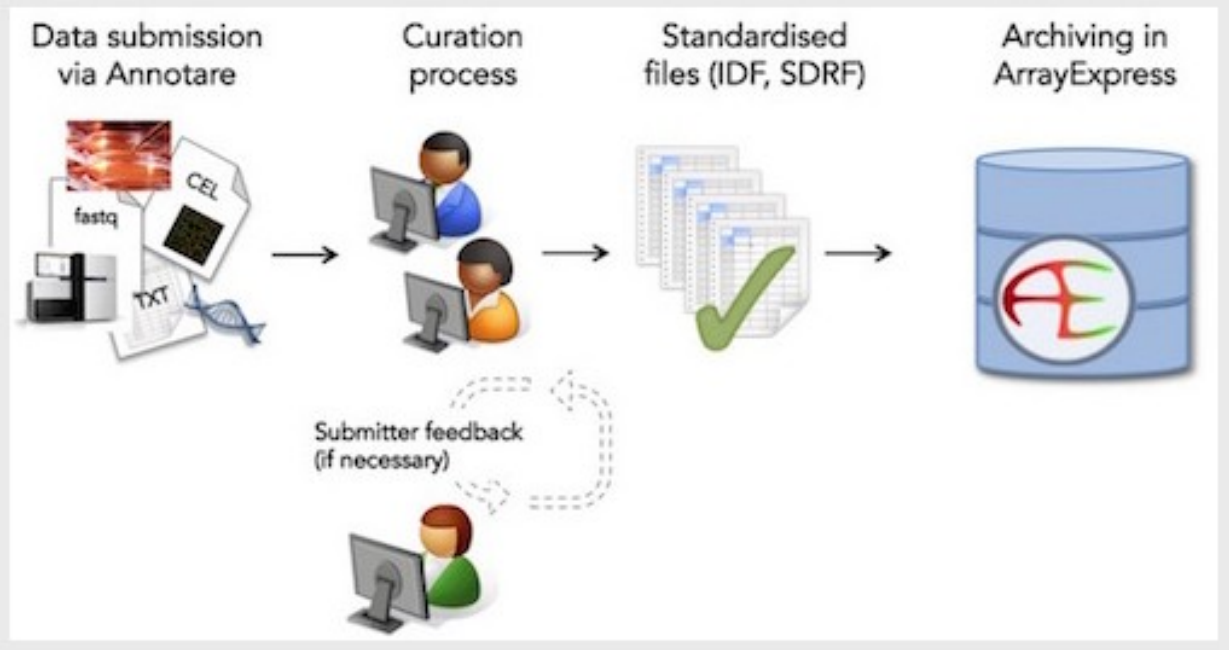



Source: https://www.ebi.ac.uk/arrayexpress/submit/overview.html

- Benefits:
  - Ensures the verification, interpretation and reproducibility of data
  - Facilitates the creation of structured databases and development of analysis

# Data Curation

- It involves annotation, publication and presentation of the data such that the value of the data is maintained over time, and the data remains available for reuse and preservation.
- May be done by database experts or experts of the scientific community.

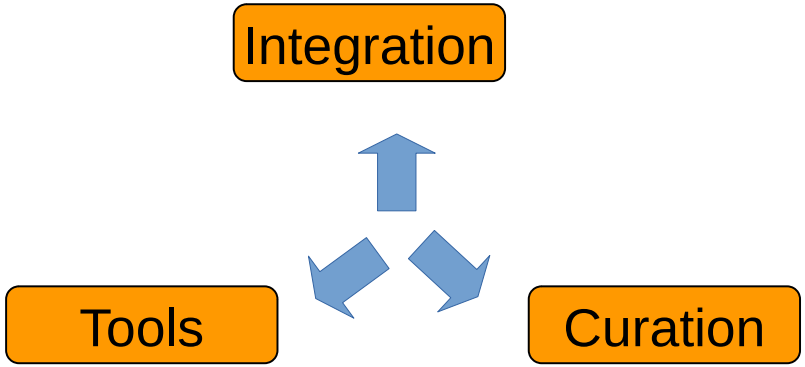



Source: https://www.ebi.ac.uk/arrayexpress/submit/overview.html

# III. Centralizing data management

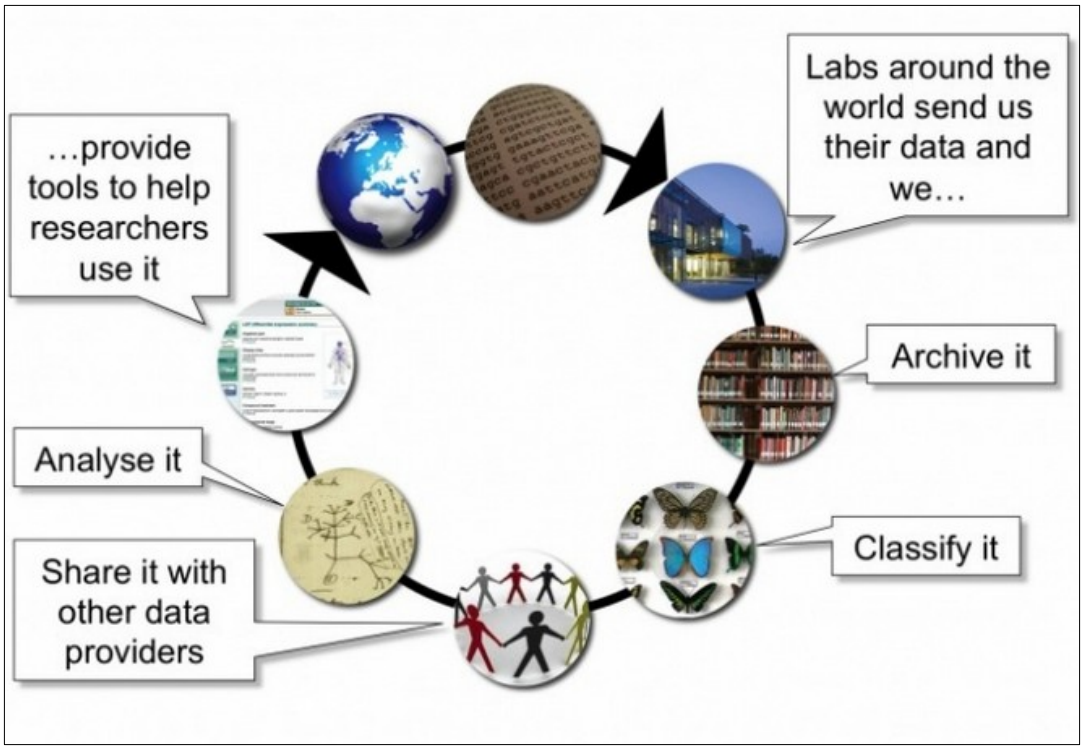
# Centralizing data management

- General
  - **Resource providers**
- Subject-specific
  - **Collaborative projects**
  - **Multi-omics repositories**
  - **Genomic Browsers**



# Resource providers

- Big organizations that act as *hubs* that provide transparent access to data sources.




Source: https://www.ebi.ac.uk

# Resource providers

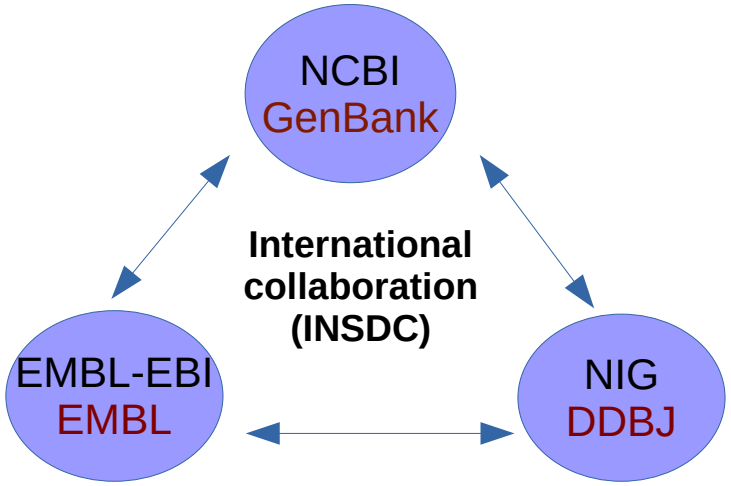
NCBI

- NCBI Home
- Resource List (A-Z)
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

EMBL-EBI

- DNA & RNA
- Gene Expression
- Structures
- Systems
- Ontologies
- Literature
- Proteins
- Chemical biology
- Cross domain

- Provide integrated access to databases
- Classification according to multiple criteria
- Primary databases may be common or specific
- Example: nucleotide DB are daily synchronized



# Resource providers


- Amino Acid Explorer
- Assembly Archive
- Basic Local Alignment Search Tool (BLAST)
- Batch Entrez
- BioAssay Services
- BLAST Link (BLink)
- BLAST Microbial Genomes
- BLAST RefSeqGene
- CDTree
- Cn3D
- COBALT
- Concise Microbial Protein BLAST
- Conserved Domain Architecture Retrieval Tool (CDART)
- Conserved Domain Search Service (CD Search)
- Digital Differential Display (DDD)
- Electronic PCR (e-PCR)
- Frequency-weighted Link (FLink)


**Clustal Omega**

Multiple sequence alignment of DNA or protein sequences. Clustal Omega replaces the older ClustalW alignment tools.

Multiple sequence alignment

**InterProScan**

InterProScan searches sequences against InterPro's predictive protein signatures.

Protein feature detection | Sequence motif recognition

**BLAST [protein]**

Fast local similarity search tool for protein sequence databases.

Sequence similarity search

**BLAST [nucleotide]**

Fast local similarity search tool for nucleotide sequence databases.

Sequence similarity search

**HMMER**

Fast sensitive protein homology searches using profile hidden Markov models (HMMs) for querying against both sequence and HMM target databases.

Sequence similarity search | Protein function analysis

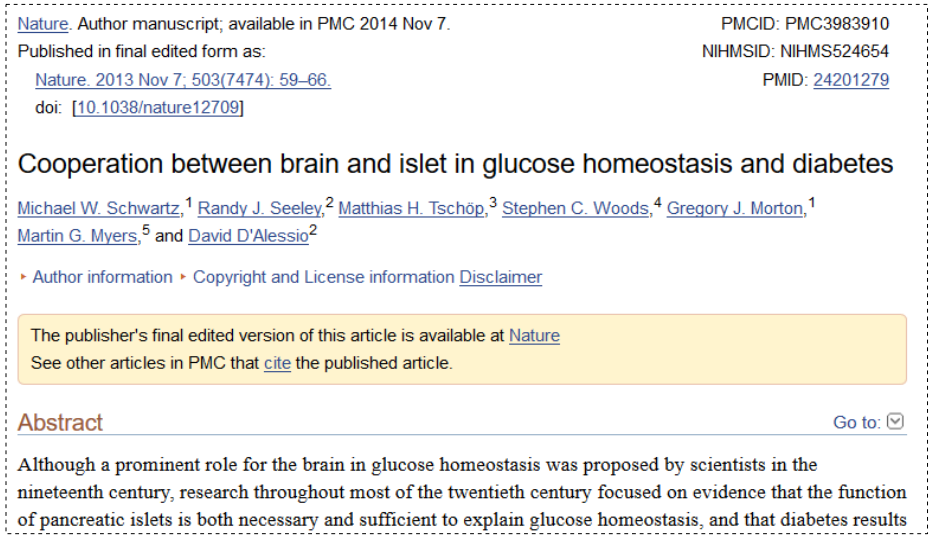
- Provide a wide variety of data analysis tools that allow users to explore, manipulate, align, visualize and evaluate biological data.

# Examples of Databases

# Examples of Databases

## Literature DB

- Contain different types of bibliographic information (articles, reviews, books, patents...). Not only peer-reviewed!
- PubMed (NCBI): references and abstracts on life and biomedical sciences
- Europe PMC (EBI-EMBL): a full-text literature database for life sciences
- ArXiv: repository of electronic pre-prints after moderation
- Patent databases (eg. EPO) can be accessed from EBI-search
- Biocatalogue: provides a curated catalog of life-sciences web services



Nature. Author manuscript; available in PMC 2014 Nov 7.
PMCID: PMC3983910
Published in final edited form as:
NIHMSID: NIHMS524654
Nature. 2013 Nov 7; 503(7474): 59–66.
PMID: 24201279
doi: [10.1038/nature12709]

Cooperation between brain and islet in glucose homeostasis and diabetes

Michael W. Schwartz,[1] Randy J. Seeley,[2] Matthias H. Tschöp,[3] Stephen C. Woods,[4] Gregory J. Morton,[1] Martin G. Myers,[5] and David D'Alessio[2]

Author information · Copyright and License information Disclaimer

The publisher's final edited version of this article is available at Nature
See other articles in PMC that cite the published article.

Abstract Go to:

Although a prominent role for the brain in glucose homeostasis was proposed by scientists in the nineteenth century, research throughout most of the twentieth century focused on evidence that the function of pancreatic islets is both necessary and sufficient to explain glucose homeostasis, and that diabetes results

# Examples of Databases

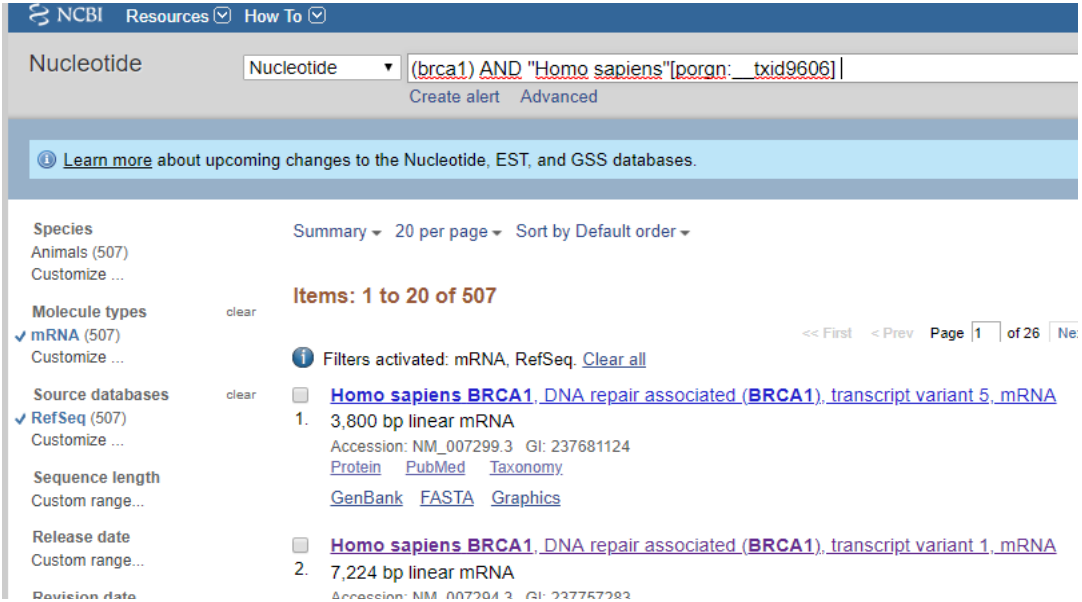
## Taxonomic DB

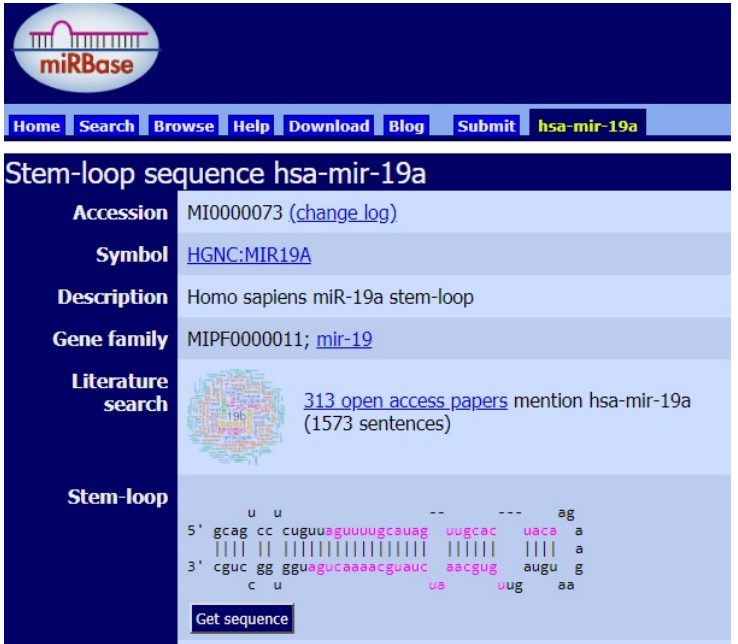
- Contain information about the classification of organisms, mainly from molecular data
- Taxonomy DB: curated classification and nomenclature for all of the organisms in the public sequence databases.
- This represents about 10% of described species

# Examples of Databases

## Nucleotide DB

- Contain DNA / RNA (coding or non-coding) sequences from all organisms
- Primary DB: GenBank (NCBI) / ENA (EMBL-EBI) / DDBJ (NIG)
- **RefSeq** (NCBI) Project: maintains and curates a publicly available database of annotated genomic, transcript, and protein sequence records.
- Nucleotide: collection from several DB (GenBank, RefSeq, TPA, PDB...)
- miRBase: database of published miRNA sequences and annotation.



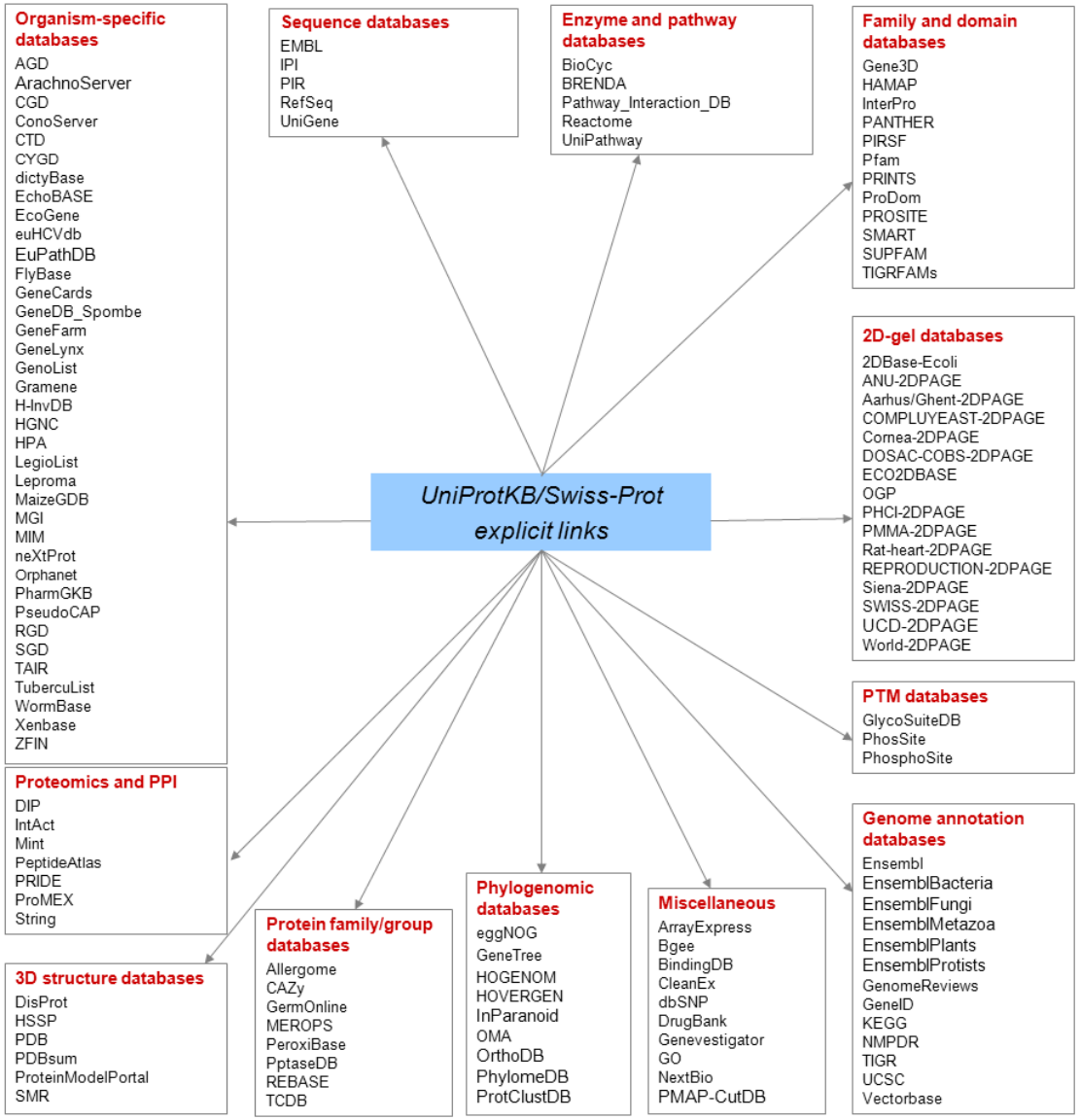



# Examples of Databases

## Protein DB

- Contain data from protein sequences, structures. Predicted / experimental.
- Protein (NCBI) / UniProtKB: collection of protein **sequences** from several sources:
  - translations from annotated coding regions (GenBank, RefSeq.../TrEMBL)
  - Records from SwissProt, PIR, PRF, and PDB.
- InterPro: integrates information from protein **family and domain** DB like Pfam, PROSITE, CDD, ...
- PDB: contains **3D structural data** of large biological molecules (proteins, nucleic acids). Typically obtained by X-ray crystallography, NMR spectroscopy or cryo-electron microscopy.
- IntAct: a curated DB of **molecular interactions**

# Examples of Databases




https://web.expasy.org/docs/userman.html

# Practicum

To warm up...

## Querying databases to answer biological questions

**1-** Using PubMed Advanced Search, look for a *review* paper published in *Nature* on *colorectal cancer* and authored by *David SS*

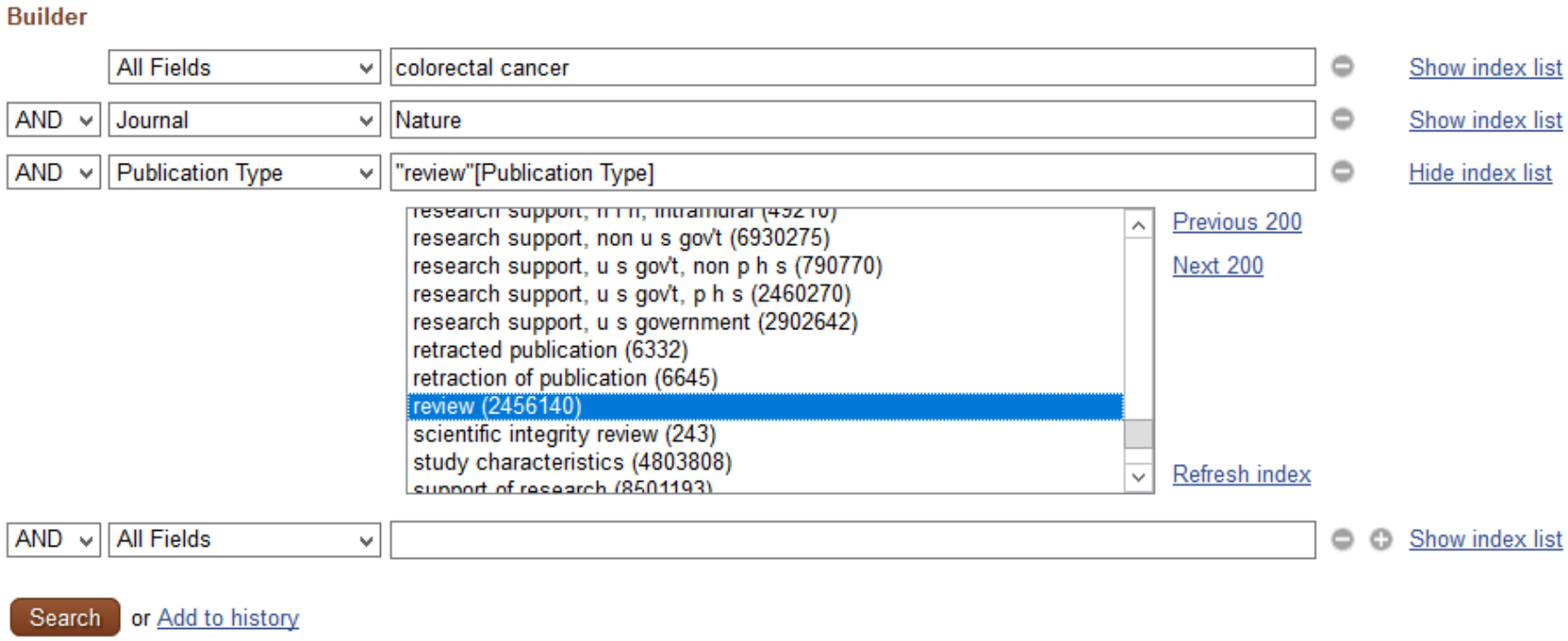
**2-** In the abstract, the authors mention a gene associated to the disease. Find a *well annotated mRNA* sequence for this gene using the Nucleotide DB

**3-** Look for MUTYH human protein in UniProtKB

- Identify protein sequence, motifs and 3D structure
- With which proteins interacts according to *IntAct DB*?

# Practicum

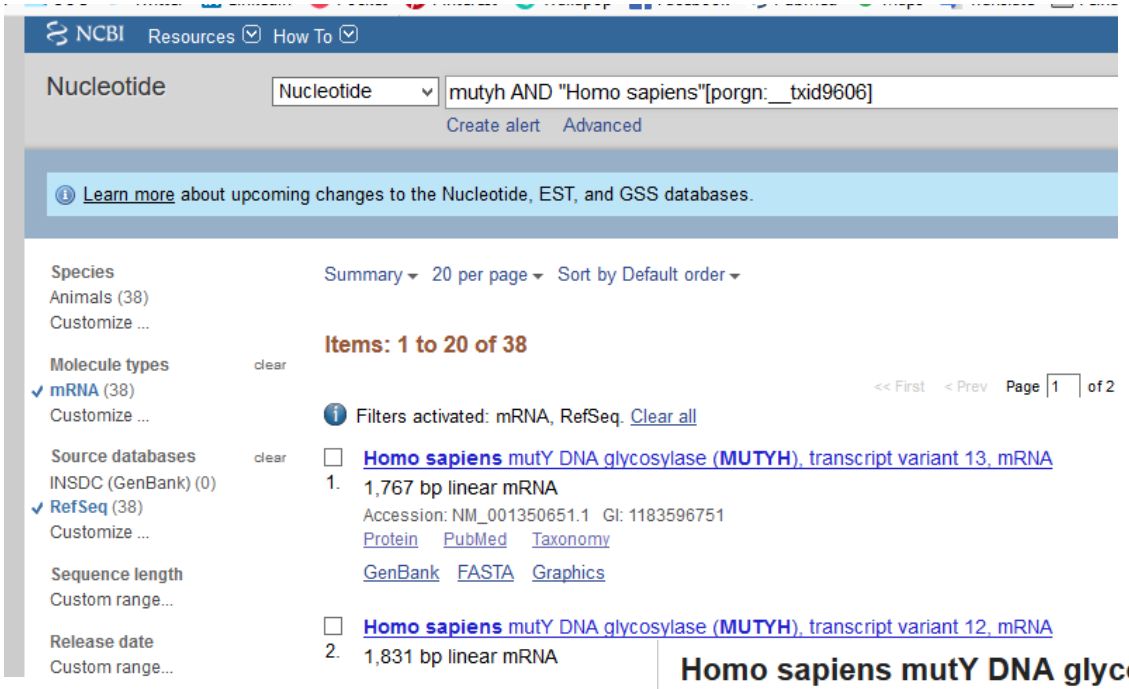
**1-** Using PubMed Advanced Search, look for a *review* paper published in *Nature* on *colorectal cancer* and authored by *David SS*

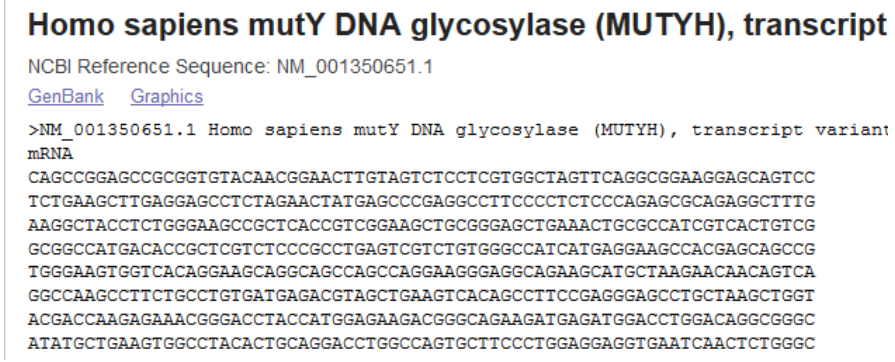


# Practicum

**2-** In the abstract, the authors mention a gene associated to the disease. Find a *well annotated mRNA* sequence for this gene using the Nucleotide DB

Using filters

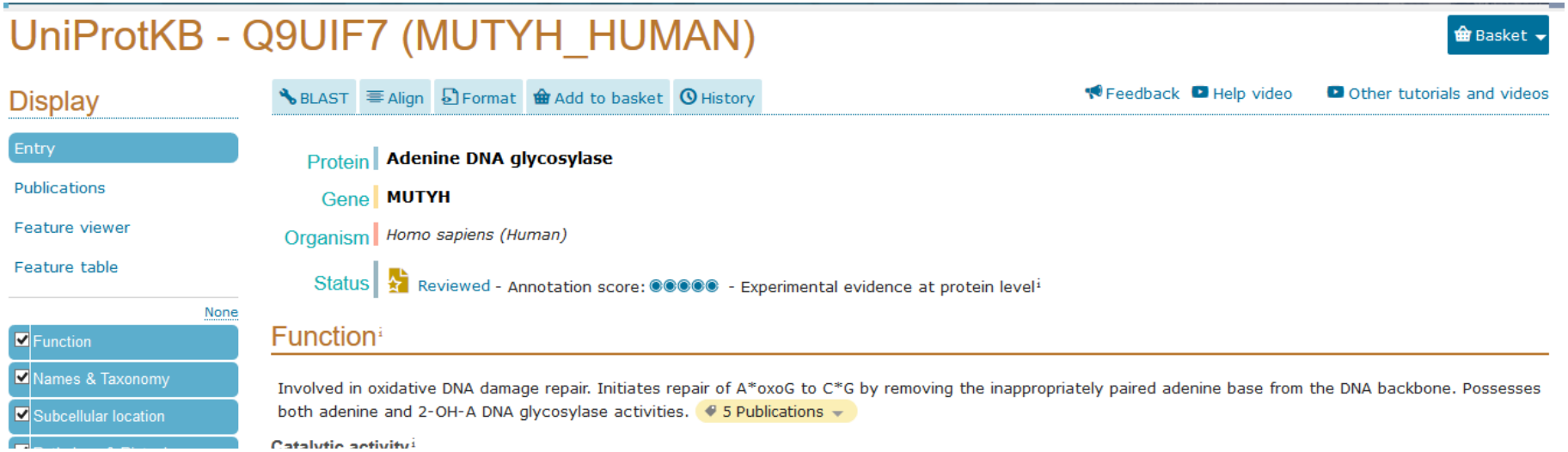


# Practicum

**3-** Look for MUTYH human protein in UniProtKB

- Identify protein sequence, motifs and 3D structure
- With which proteins interacts according to IntAct DB?



# Practicum

## Interaction

### Binary interactions

| With | Entry | #Exp. | IntAct | Notes |
|---|---|---|---|---|
| AGTRAP | Q6RW13 | 3 | EBI-10321956,EBI-741181 | |

### Protein-protein interaction databases

| | |
|---|---|
| BioGrid | 110681, 11 interactors |
| DIP | DIP-41972N |
| IntAct | Q9UIF7, 15 interactors |
| MINT | Q9UIF7 |
| STRING | 9606.ENSP00000361170 |

## Structure

| PDB Entry | Method | Resolution | Chain | Positions | Links |
|---|---|---|---|---|---|
| **1X51** | NMR | | A | 356-497 | PDBe RCSB PDB PDBj PDBsum |
| **3N5N** | X-ray | 2.30 Å | X/Y | 76-362 | PDBe RCSB PDB PDBj PDBsum |

1 notificació

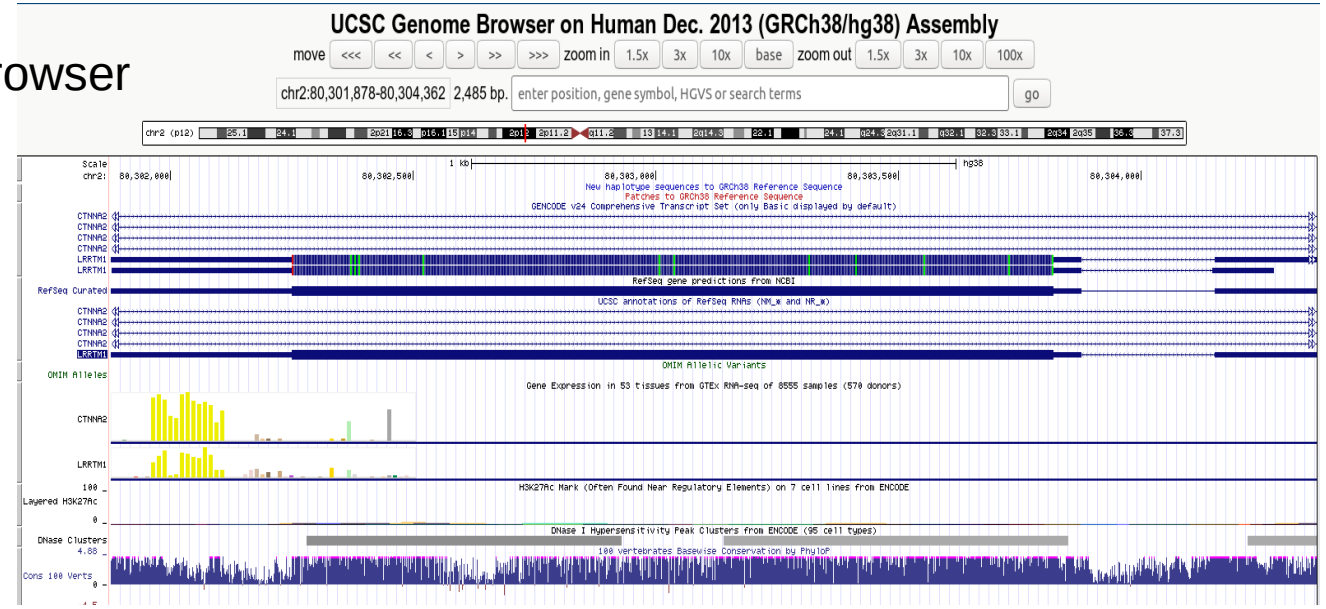
### Family and domain databases

| | |
|---|---|
| CDD | cd03431 DNA_Glycosylase_C, 1 hit<br>cd00056 ENDO3c, 1 hit |
| Gene3D | 1.10.1670.10, 1 hit |
| InterPro | View protein in InterPro<br>IPR011257 DNA_glycosylase<br>IPR004036 Endonuclease-III-like_CS2<br>IPR003651 Endonuclease3_FeS-loop_motif<br>IPR004035 Endouclease-III_FeS-bd_BS<br>IPR003265 HhH-GPD_domain<br>IPR000445 HhH_motif<br>IPR023170 HTH_base_excis_C<br>IPR029119 MutY_C<br>IPR015797 NUDIX_hydrolase-like_dom_sf<br>IPR000086 NUDIX_hydrolase_dom |
| Pfam | View protein in Pfam<br>PF00633 HHH, 1 hit<br>PF00730 HhH-GPD, 1 hit<br>PF14815 NUDIX_4, 1 hit |
| SMART | View protein in SMART<br>SM00478 ENDO3c, 1 hit<br>SM00525 FES, 1 hit |
| SUPFAM | SSF48150 SSF48150, 1 hit<br>SSF55811 SSF55811, 1 hit |
| PROSITE | View protein in PROSITE<br>PS00764 ENDONUCLEASE_III_1, 1 hit<br>PS01155 ENDONUCLEASE_III_2, 1 hit<br>PS51462 NUDIX, 1 hit |

# Examples of Databases

## Genomic databases

- Organize information on genomes including sequences, maps, chromosomes, assemblies, and annotations
- Species-specific genome databases (eg. Mouse Genome Informatics)
- Genome Browsers: provide tools for visualization and integrative genomic analysis
  - NCBI Genome Data Viewer
  - UCSC Genome Browser
  - EBI's Ensembl



# Examples of Databases

## Gene Expression Databases

- Contain gene expression data derived from microarray, next-generation sequencing, and other forms of high-throughput functional genomics data submitted by the research community
- Gene Expression Omnibus (GEO) / ArrayExpress (EBI)
- Sequence Read Archive (SRA): stores raw sequencing data and alignment information from high-throughput sequencing platforms
- GTEx
- Expression Atlas: provides gene expression results for different organisms, including metazoans and plants. Expression profiles of tissues from Human Protein Atlas, GTEx and FANTOM5, and of cancer cell lines from ENCODE, CCLE and Genentech projects can be explored.

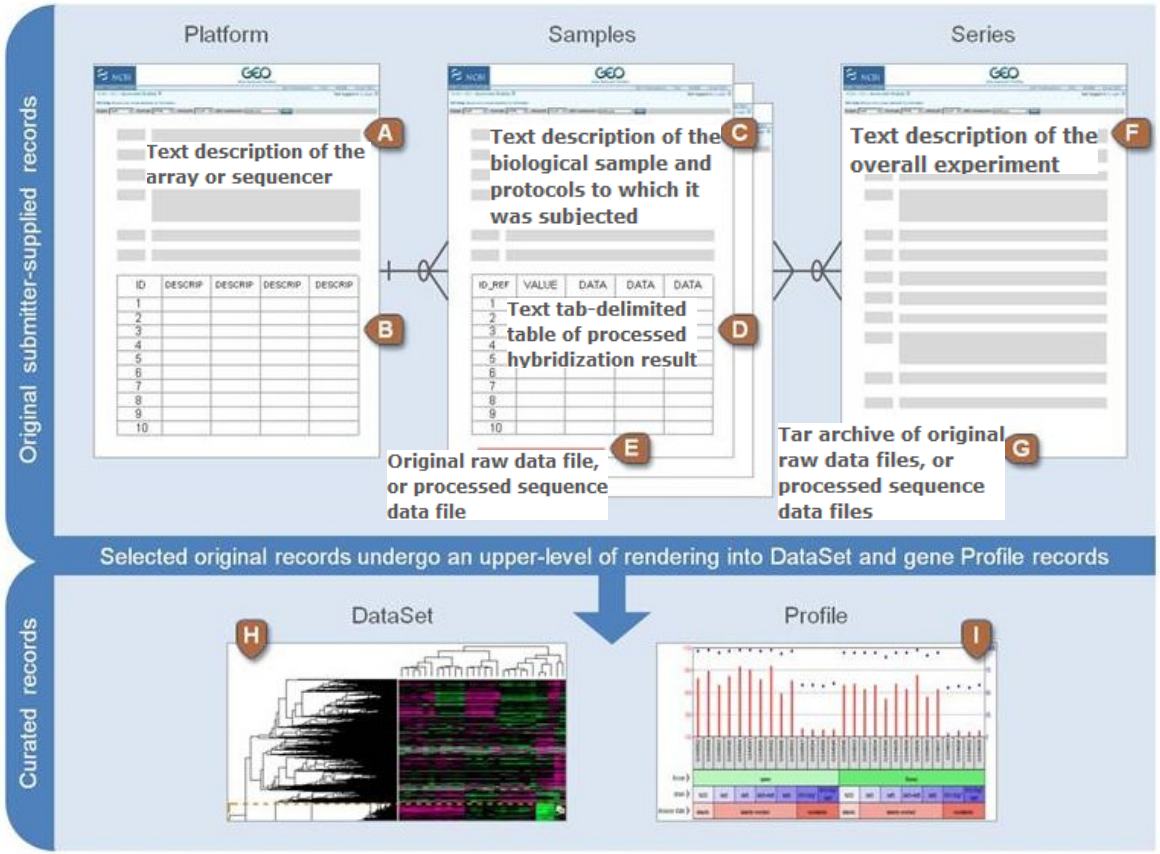
# Practicum

## Retrieving data from GEO

- The Gene Expression Omnibus (GEO) is an international public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomics data submitted by the research community.

- Data organization:
  - Platform (GPLxxx)
  - Samples (GSMxxx)
  - Series (GSExxx)
  - Datasets (curated) (GSDxxx)
  - Profiles (curated)

See some examples



# Practicum

## Gene Expression Omnibus (GEO)

- Queries can be performed for datasets or gene expression profiles
  - GEO Datasets: stores original submitter-supplied study descriptions as well as curated gene expression DataSets.
    - GEO Series (GSEXXX): original submitter-supplied record that summarizes a study
    - GEO Datasets (GDSXXX): represents a collection of biologically- and statistically-comparable samples processed using the same platform.

Example with GDS browser:

| DataSet | Title | Organism(s) | Platform | Series | Samples |
|---|---|---|---|---|---|
| GDS6248 | Diet-induced obesity model: liver | *Mus musculus* | GPL6887 | GSE39549 | 51 |
| GDS6247 | Diet-induced obesity model: white adipose tissue | *Mus musculus* | GPL6887 | GSE39549 | 40 |
| GDS6177 | Acute alcohol consumption effect on whole blood (control... | *Homo sapiens* | GPL570 | GSE20489 | 25 |

# Retrieving data from GEO

**Series GSE39549** Query DataSets for GSE39549

| | |
|---|---|
| Status | Public on Mar 01, 2014 |
| Title | Time-course microarrays reveal early activation of the immune transcriptome and adipokine dysregulation leads to fibrosis in visceral adipose depots during diet-induced obesity |
| Organism | Mus musculus |
| Experiment type | Expression profiling by array |
| Summary | Time-course analysis of adipocyte gene expression profiles response to high fat diet. The hypothesis tested in the present study was that in diet-induced obesity, early activation of TLR-mediated inflammatory signaling |
| Overall design | Total RNA obtained from isolated epididymal and mesenteric adipose tissue of C57BL/6J mice fed normal diet or high fat diet for 2, 4, 8, 20 and 24weeks |
| Contributor(s) | Kwon E, Choi M |
| Platforms (1) | GPL6887 Illumina MouseWG-6 v2.0 expression beadchip |
| Samples (91) | GSM971546 Mice fed Normal diet for 2weeks rep1 |
| ⊞ More... | GSM971547 Mice fed Normal diet for 2weeks rep2 |
| | GSM971548 Mice fed Normal diet for 2weeks rep3 |
| **Relations** | |
| BioProject | PRJNA171109 |

Analyze with GEO2R

| Download family | Format |
|---|---|
| SOFT formatted family file(s) | SOFT |
| MINiML formatted family file(s) | MINiML |
| Series Matrix File(s) | TXT |

| Supplementary file | Size | Download | File type/resource |
|---|---|---|---|
| GSE39549_Matrix_non-normalized_EPI.txt.gz | 8.4 Mb | (ftp)(http) | TXT |
| GSE39549_Matrix_non-normalized_MES.txt.gz | 2.9 Mb | (ftp)(http) | TXT |
| GSE39549_RAW.tar | 15.8 Mb | (http)(custom) | TAR |

*Raw data is available on Series record*
*Processed data included within Sample table*

Study information

Platform used (data table with annotation of probes)

Samples

SOFT file can hold both data tables and descriptive information for multiple Platforms, Samples, and/or Series.

Series matrix with info for all samples and raw/processed data

Info on data files

# Retrieving data from GEO

| | |
|---|---|
| Source name | Adipose tissue of mice |
| Organism | Mus musculus |
| Characteristics | strain: C57BL/6J<br>treatment protocol: Normal diet<br>time: 2weeks<br>age: 7 weeks<br>tissue: epididymal adipose tissue |
| Treatment protocol | C57BL/6J mice were fed a high-fat diet (HFD) or normal diet (ND) and sacrificed at 5 time-points (2, 4, 8, 20 and 24 weeks) over 24 weeks. |
| Extracted molecule | total RNA |
| Extraction protocol | RNA was extracted with Trizol reagent, followed by clean-up and DNase I treatment with QIAGEN RNeasy mini kit in accordance with the prescribed protocol provided with the kit. Quality control was performed with Agilent Bioanalyser. |
| Label | biotin |
| Label protocol | Biotinylated cRNA were prepared with the Ambion MessageAmp kit for Illumina arrays |
| Hybridization protocol | Standard Illumina hybridization protocol |
| Scan protocol | Standard Illumina scanning protocol |
| Description | Sample name: E2N1<br>replicate 1 |
| Data processing | Raw data were extracted using the software provided by the manufacturer (Illumina BeadStudio v3.1.3 (Gene Expression Module v3.3.8).<br>The data were normalised by quantile method using ArrayAssist® |

Sample specifications (identification, protocol, source...)

| **Data table header descriptions** | |
|---|---|
| **ID_REF** | |
| **VALUE** | normalized signal |

| **Data table** | |
|---|---|
| **ID_REF** | **VALUE** |
| ILMN_2417611 | 7.1251793 |
| ILMN_2762289 | 6.838682 |
| ILMN_2896528 | 12.505199 |
| ILMN_2721178 | 11.040463 |
| ILMN_2458837 | 6.5777817 |

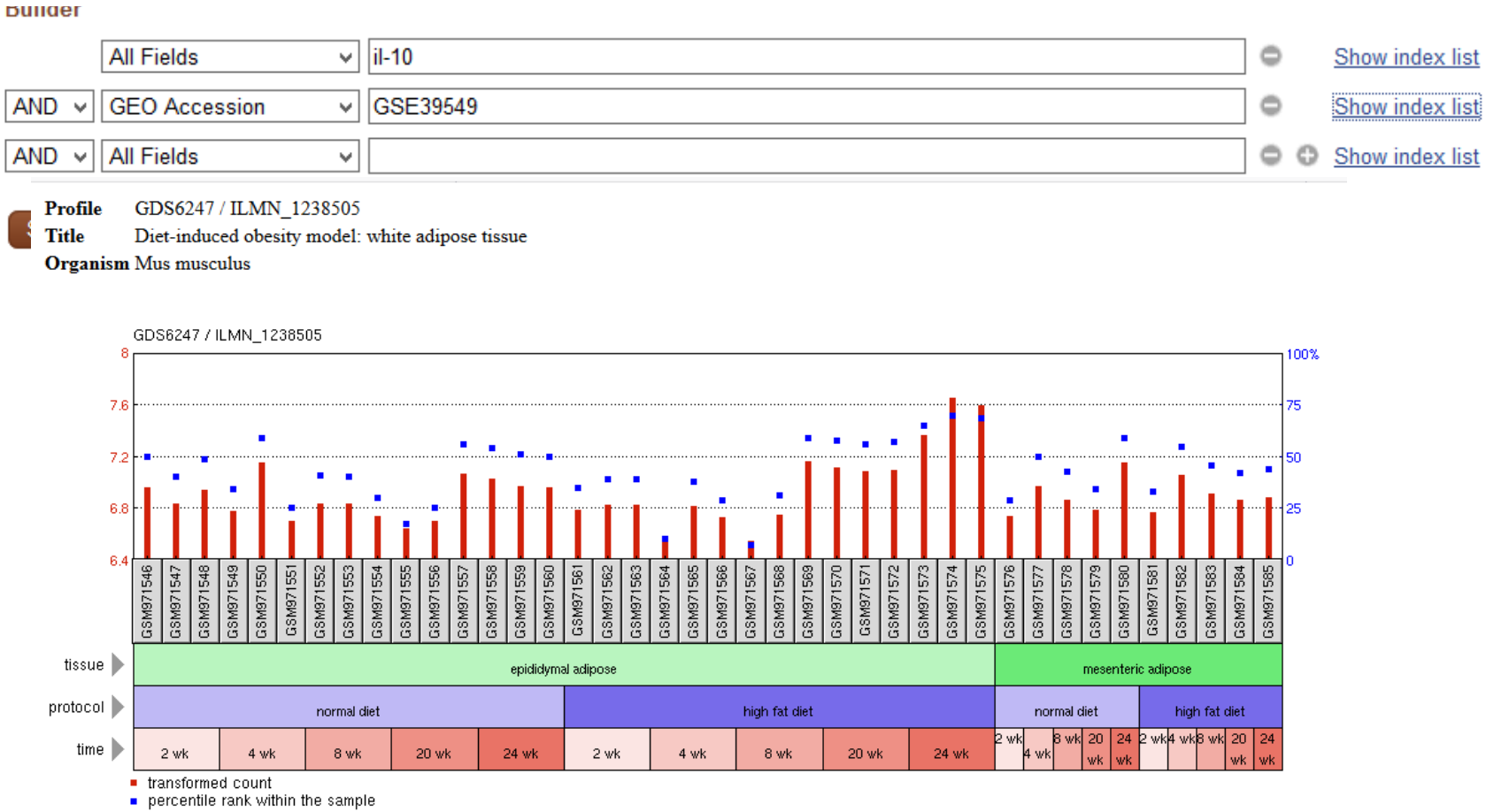
Data table with normalized expression values

# Practicum

## Retrieving data from GEO

- GEO Profiles: stores individual gene expression profiles from curated DataSets.

Example: search for expression profile of IL-10



# Examples of Databases

- **Functional annotations**
  - Gene Ontology (GO): unify the representation of gene and gene product attributes across all species
  - KEGG / Reactome: integrates genomic, chemical and systemic functional information
  - Gene Cards
- **Terapeuthic targets**
  - Therapeutic targets database
  - PharmGKB : pharmacogenomics (impact of genetics on drug response)
- **Disease-related**
  - DisGeNet: genes and variants associated to human diseases
  - TCGA, COSMIC (Cancer)

# Examples of databases

- **And many more that can be found in / accessed from:**
  - Large list published yearly in NAR
  - Journal-recommended repositories for publication *(Nature)*
  - Resource providers portals (NCBI / EBI-EMBL)
  - Integrative projects (disease-specific / organism-specific)
  - Genome Browsers (genome-oriented)

# Subject-specific repositories and collaborative projects

# Subject-specific repositories

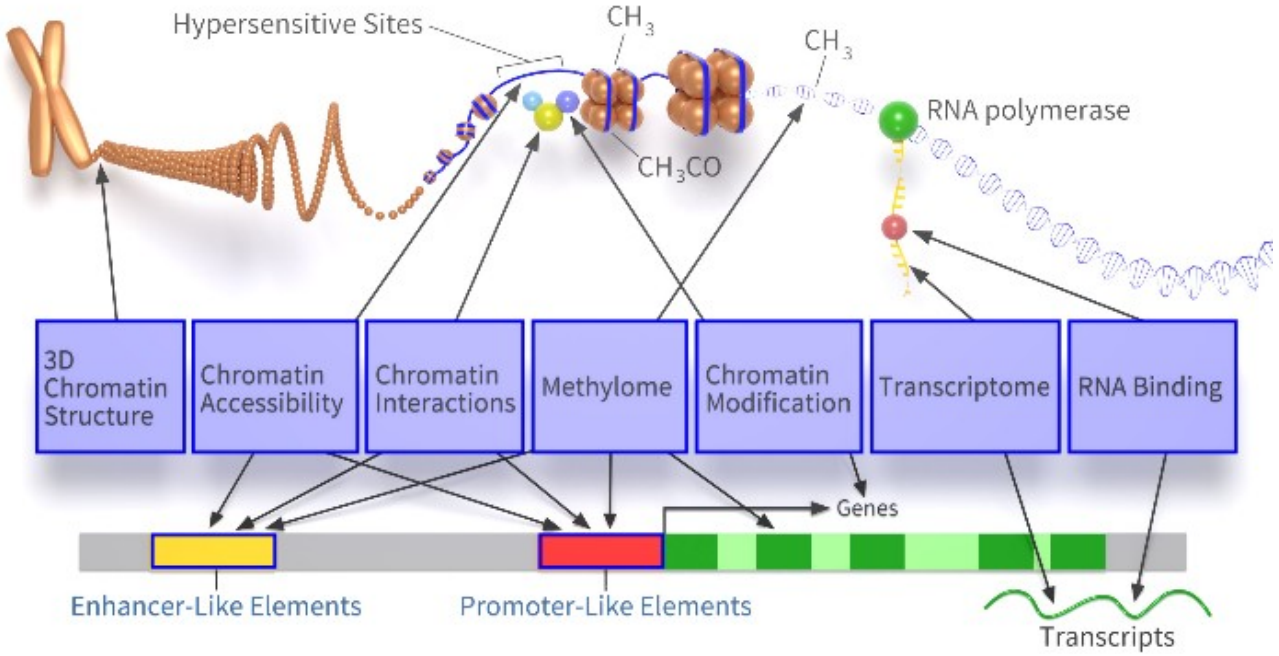
- Collective initiatives: summing efforts from different researchers from different sites of the world.
- **Integrate** different types of data (clinical, multi-omics,...) related to a **specific subject** (or sub-subject)
  - A disease
  - An organism
  - A biological question, tissue, …
- Provide curated, standardized, **high-quality** datasets for public research
  - Raw and/or Processed data
- Provide specialized visualization and analysis tools through their web sites

# Subject-specific repositories

## ENCODE

https://www.encodeproject.org/

- The Encyclopedia of DNA Elements (ENCODE) Consortium is an ongoing international collaboration of research groups funded by the NHGRI.
- Intended as a follow-up to the Human Genome Project, it aims to identify all functional elements in the human genome (genes, transcripts, miRNA, regulatory elements, etc) employing a variety of assays and methods.

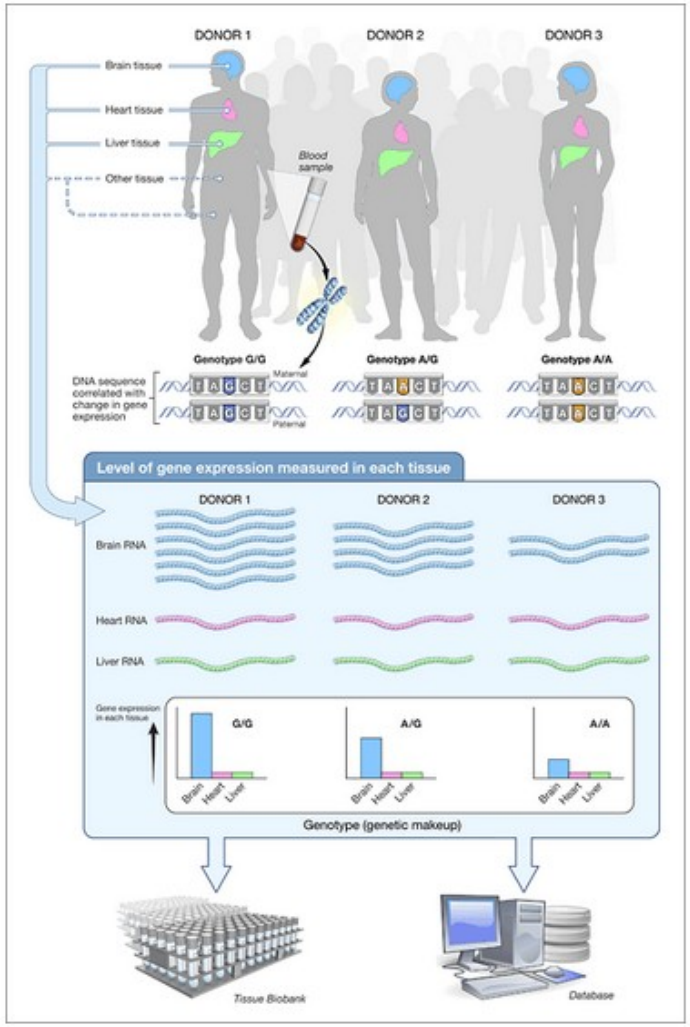


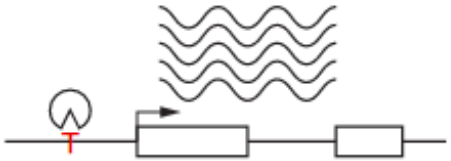
# Subject-specific repositories

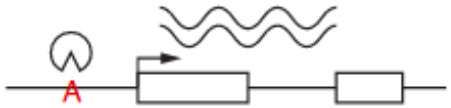
## Genotype-Tissue Expression (GTEx) Project

https://gtexportal.org/home/

- Aims to provide the scientific community with a public resource to study tissue-specific gene expression and regulation and its relationship to genetic variation across individuals.
- On-going project
- Samples from 54 non-diseased tissues across nearly 1000 individuals who were also densely genotyped.
- Variations in gene expression that are highly correlated with genetic variation can be identified as expression quantitative trait loci, or eQTLs.



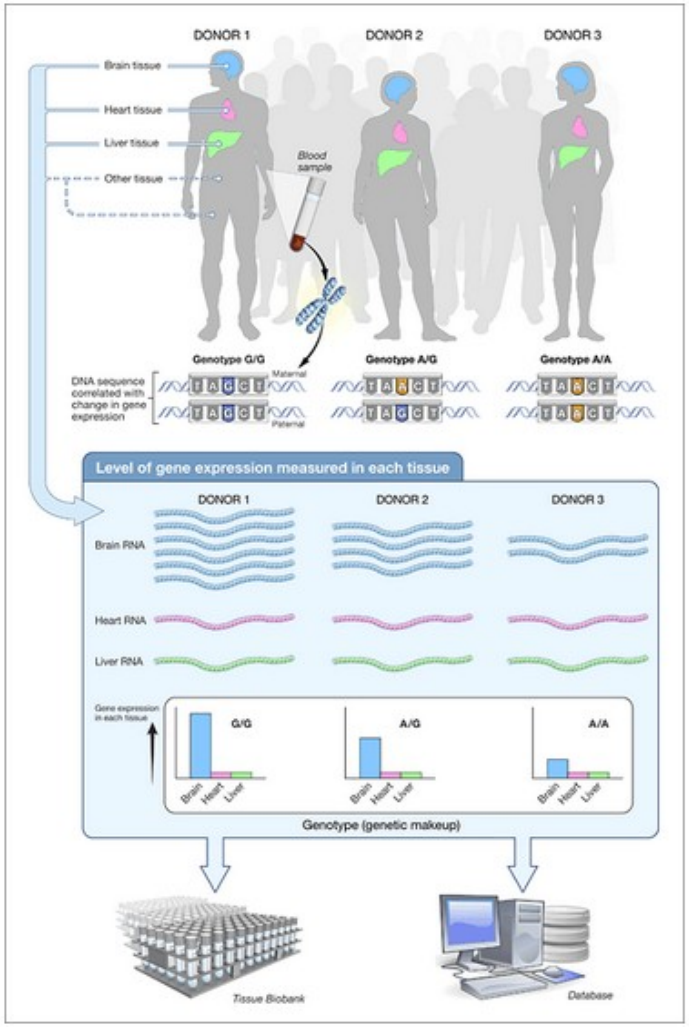



# Subject-specific repositories

## Genotype-Tissue Expression (GTEx) Project

https://gtexportal.org/home/



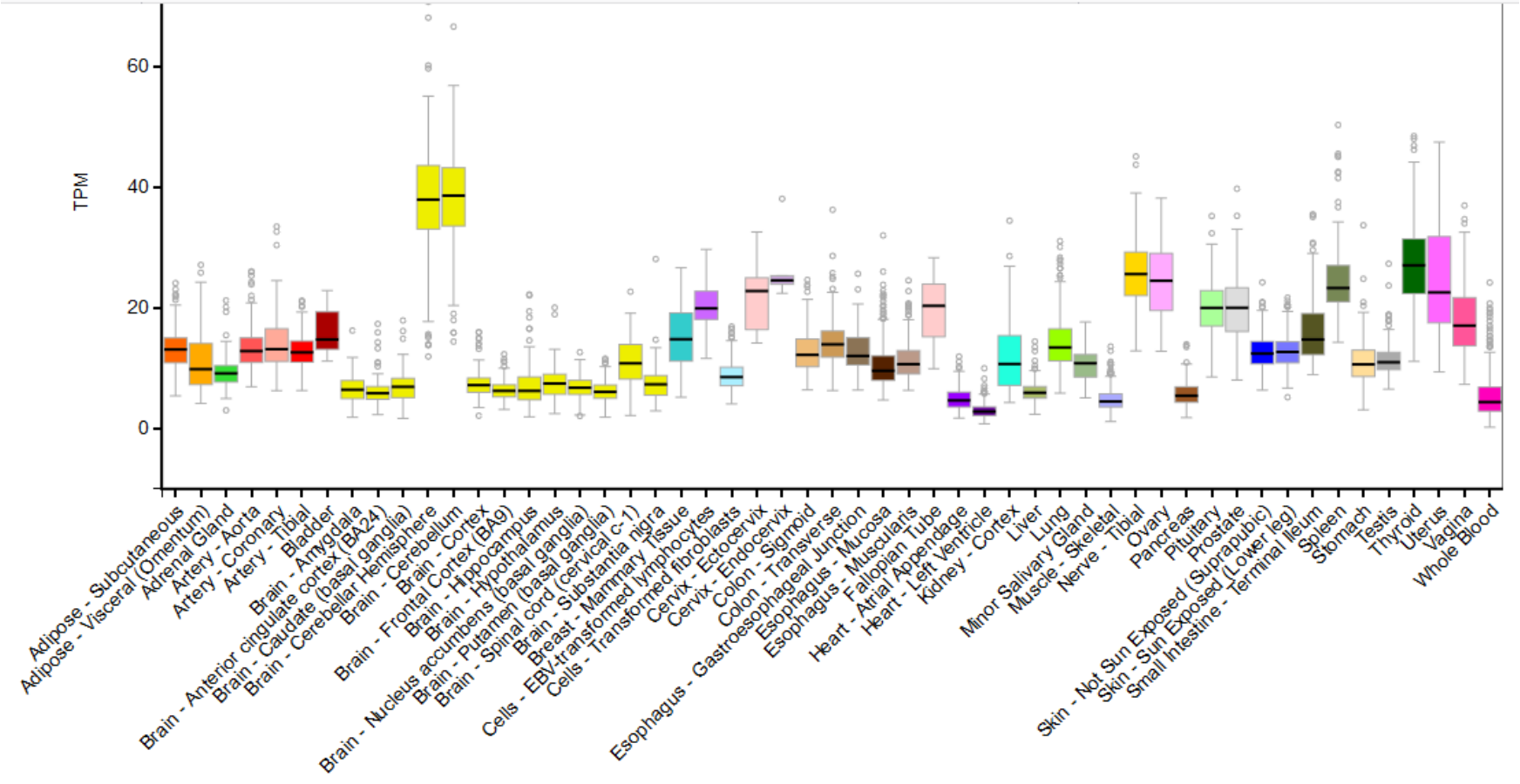

- Types of data provided:
  - Gene/transcript expression in tissues
  - variants associated to gene expression (eQTLs and sQTLs)
  - histology images
  - Patient/sample metadata

- Available datasets:
  https://www.gtexportal.org/home/datasets
- Summary statistics:
  https://gtexportal.org/home/tissueSummaryPage

# Subject-specific repositories

## Genotype-Tissue Expression (GTEx) Project

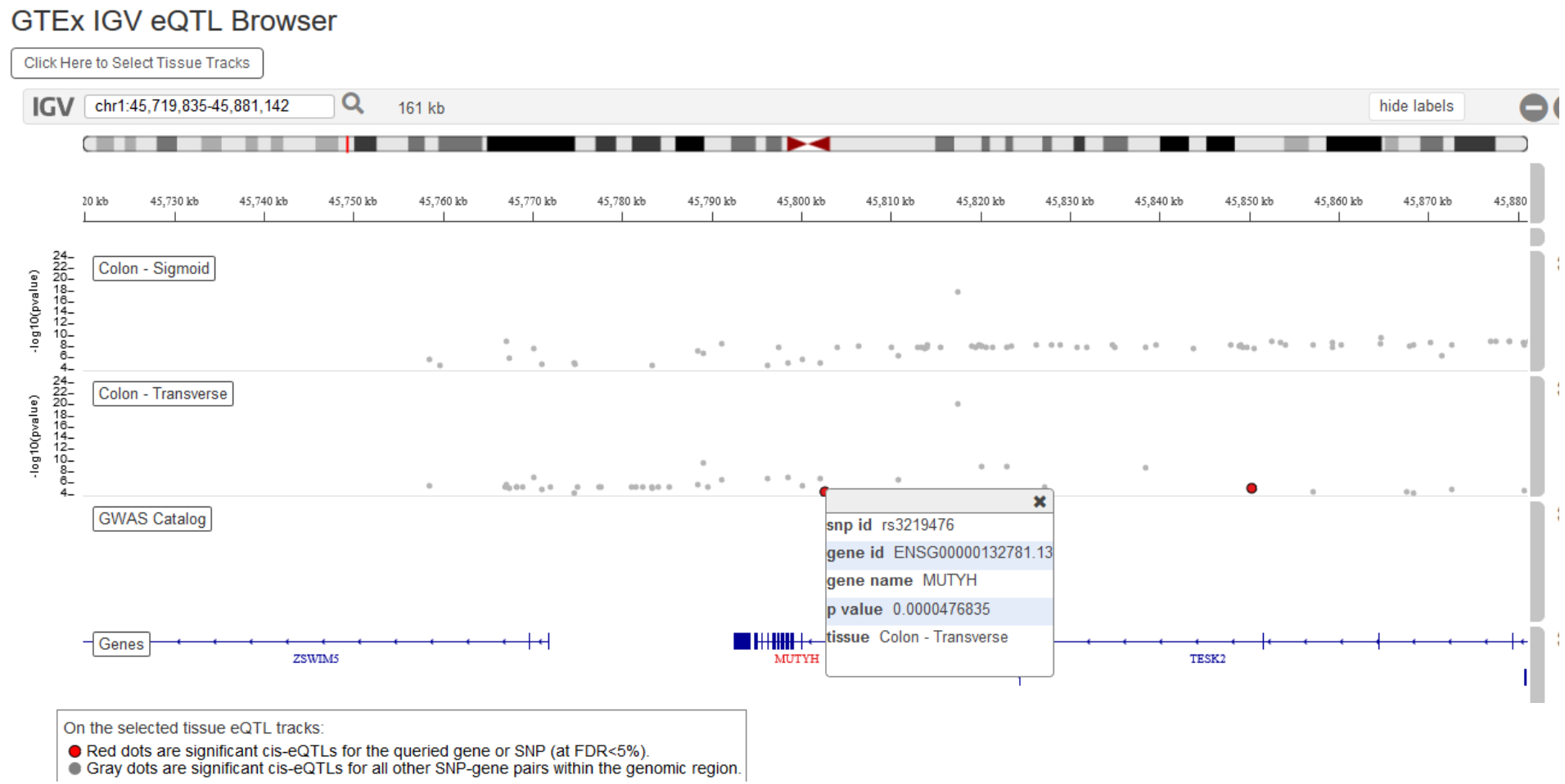
Example: Normal tissue expression profile of *mutyh* gene



# Subject-specific repositories

## Genotype-Tissue Expression (GTEx) Project

Example: Looking for SNPs associated to changes in *mutyh* expression



# Subject-specific repositories

## Examples

- Disease-related multi-omics repositories (eg. cancer)

**Table 1.** List of multi-omics data repositories.

| DATA REPOSITORY | WEB LINK | DISEASE | TYPES OF MULTI-OMICS DATA AVAILABLE |
|---|---|---|---|
| The Cancer Genome Atlas (TCGA) | https://cancergenome.nih.gov/ | Cancer | RNA-Seq, DNA-Seq, miRNA-Seq, SNV, CNV, DNA methylation, and RPPA |
| Clinical Proteomic Tumor Analysis Consortium (CPTAC) | https://cptac-data-portal.georgetown.edu/cptacPublic/ | Cancer | Proteomics data corresponding to TCGA cohorts |
| International Cancer Genomics Consortium (ICGC) | https://icgc.org/ | Cancer | Whole genome sequencing, genomic variations data (somatic and germline mutation) |
| Cancer Cell Line Encyclopedia (CCLE) | https://portals.broadinstitute.org/ccle | Cancer cell line | Gene expression, copy number, and sequencing data; pharmacological profiles of 24 anticancer drugs |
| Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) | http://molonc.bccrc.ca/aparicio-lab/research/metabric/ | Breast cancer | Clinical traits, gene expression, SNP, and CNV |
| TARGET | https://ocg.cancer.gov/programs/target | Pediatric cancers | Gene expression, miRNA expression, copy number, and sequencing data |
| Omics Discovery Index | https://www.omicsdi.org | Consolidated data sets from 11 repositories in a uniform framework | Genomics, transcriptomics, proteomics, and metabolomics |

Source: Subramanian et al 2020, Bioinformatics and Biology Insights

# Subject-specific repositories

## The Cancer Genome Atlas (TCGA)

https://cancergenome.nih.gov/

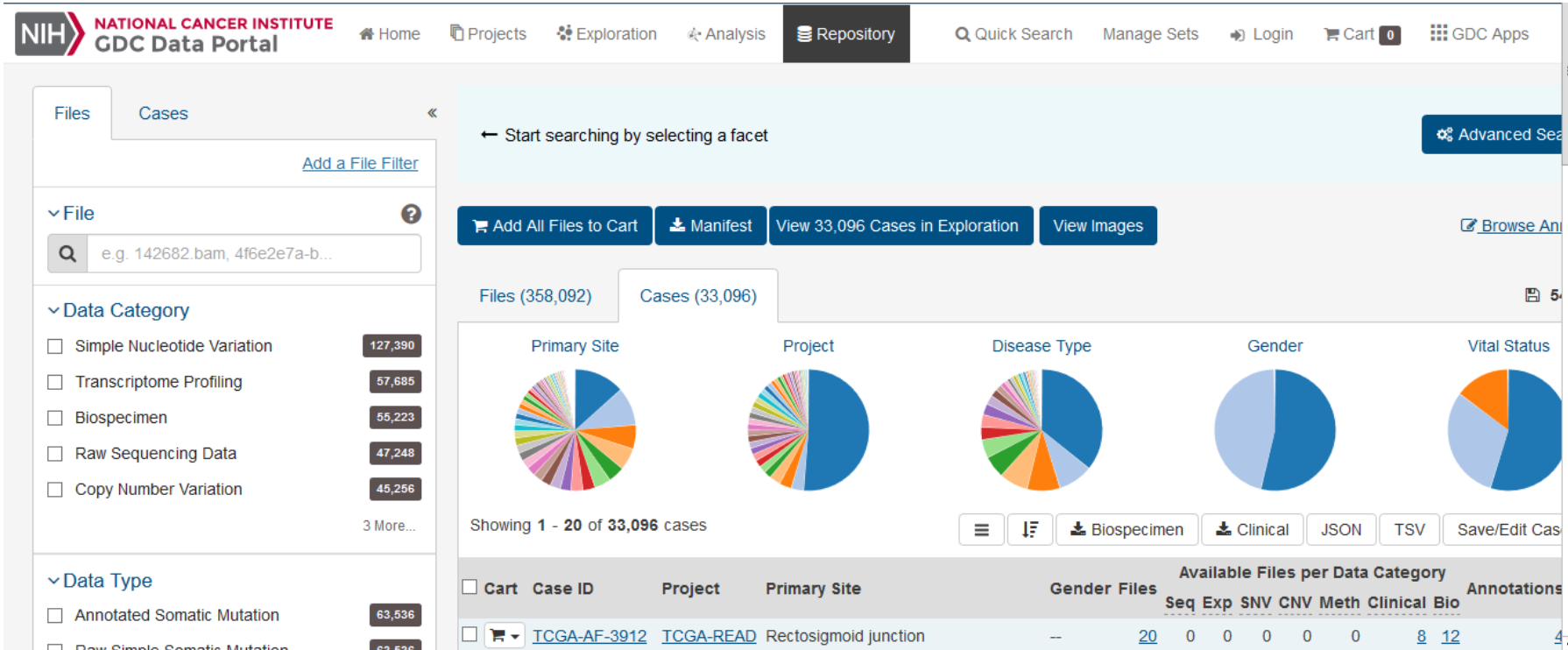
- Collaboration between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) that has generated comprehensive, multi-dimensional maps of the key genomic changes in 33 types of cancer.
- High-quality tumor and matched normal samples from over 11,000 patients collected over 12 years (ended in 2015). The data collected includes:
  - Clinical information about participants
  - Metadata about the samples
  - Histopathology slide images from sample portions
  - Molecular information derived from the samples (e.g. mRNA/miRNA expression, protein expression, copy number, etc.)

# Subject-specific repositories

## The Cancer Genome Atlas (TCGA)

https://cancergenome.nih.gov/

- Data is available through the Genomic Data Commons (GDC) portal,
  - receives, processes, and distributes genomic, clinical, and biospecimen data from cancer research programs
  - Provides web-based analysis and visualization tools



# Subject-specific repositories

## The Cancer Genome Atlas (TCGA)

https://cancergenome.nih.gov/
Data in https://portal.gdc.cancer.gov/

- To take into account when downloading data from the GDC/TCGA:
  - Different procedures to process the data (workflows)

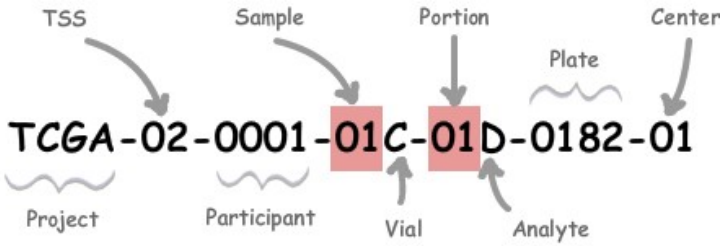
    For documentation on procedures: https://docs.gdc.cancer.gov/
  - Two available sources to download GDC data:
    - GDC Legacy Archive: provides access to an unmodified copy of data that was previously stored in CGHub and in the TCGA Data Portal hosted by the TCGA Data Coordinating Center (DCC), in which uses as references GRCh37 (hg19) and GRCh36 (hg18).
    - GDC harmonized database: data available was harmonized against GRCh38 (hg38) using GDC Bioinformatics Pipelines which provides methods to the standardization of biospecimen and clinical data.

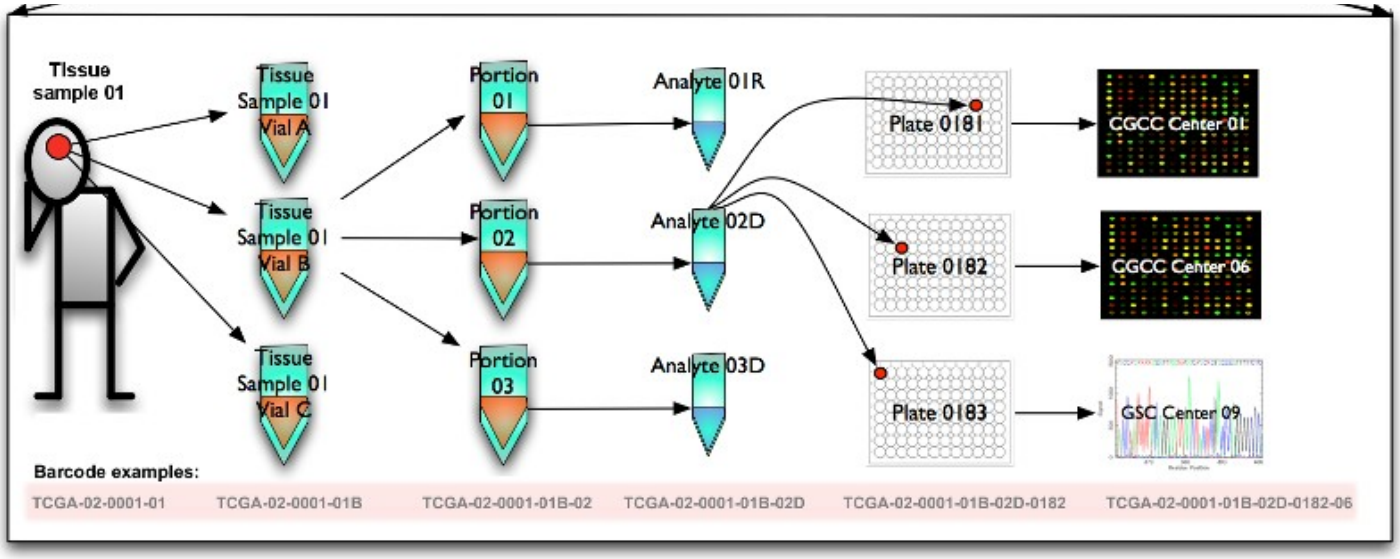
# Subject-specific repositories

## The Cancer Genome Atlas (TCGA)

https://cancergenome.nih.gov/
Data in https://portal.gdc.cancer.gov/

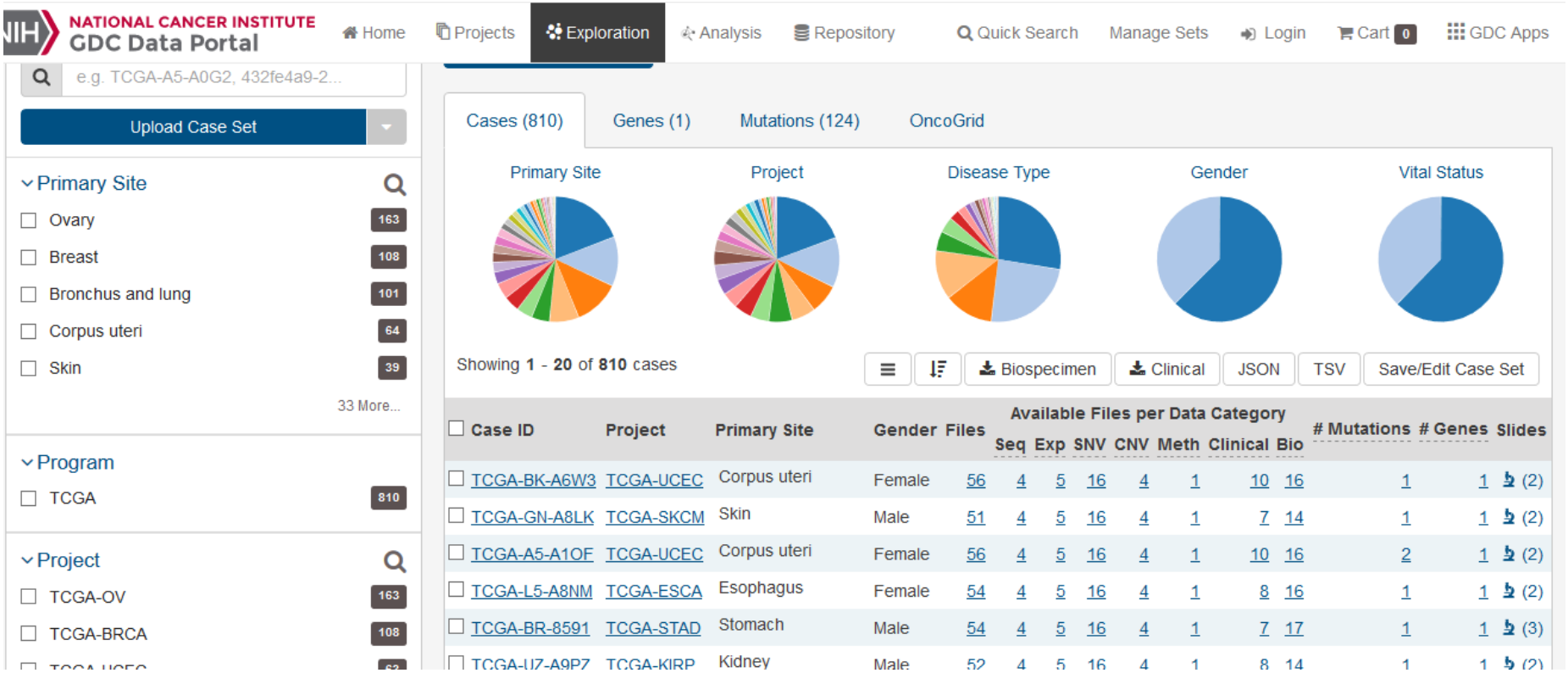
- To take into account when downloading data from the GDC/TCGA:
  - TCGA sample identification (barcode)



# Subject-specific repositories

## The Cancer Genome Atlas (TCGA)



# Subject-specific repositories


https://www.omicsdi.org/

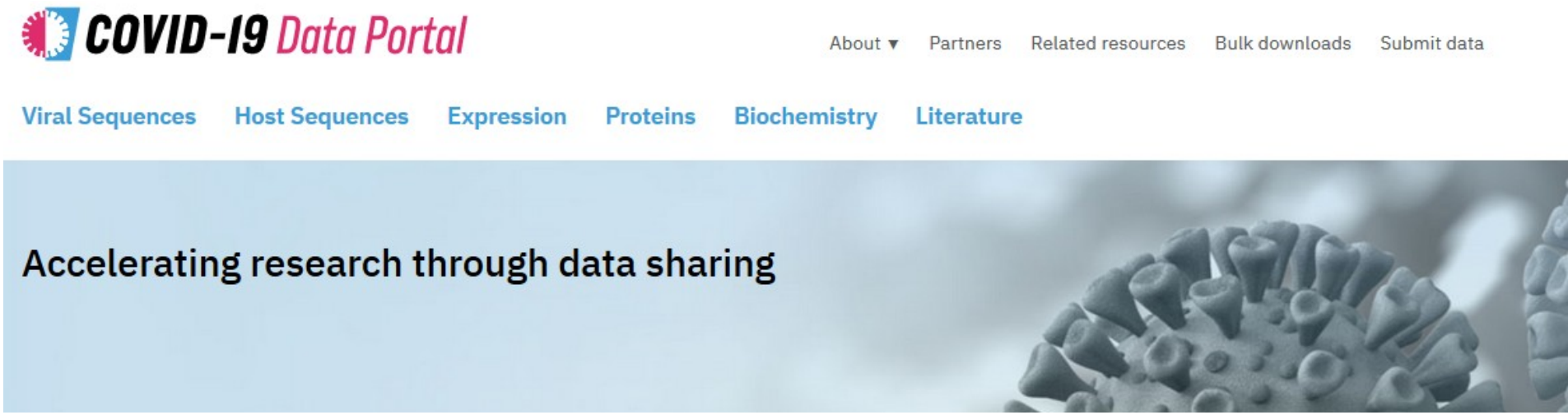
# Subject-specific repositories

## Examples

- Many other projects including
  - 1000 Genomes Project: the largest public catalogue of human variation and genotype data.
  - Human Cell Atlas: aims to create comprehensive reference maps of all human cells
  - NIH Roadmap Epigenomics Mapping Consortium: public resource of human epigenomic data

# Subject-specific repositories

## Examples




- COVID-19 has put a spotlight on open science and open repositories to improve the discovery and access to research outputs.
- Many repositories initiatives arised to act as central hub for data management
- Challenges related to:
  - copyright, embargoes and licenses attached to resources
  - metadata and data curation
  - Infrastructure and connectivity

# Subject-specific repositories

- Not all is human! Model organisms-based resources:

| Organism | Scientific name | Database (link) |
|---|---|---|
| Baker's yeast | *Saccharomyces cerevisiae* | Saccharomyces Genome Database |
| Fission yeast | *Schizosaccharomyces pombe* | PomBase |
| Clawed frog | *Xenopus* | Xenbase |
| Fruitfly | *Drosophila melanogaster* | FlyBase |
| Mouse | *Mus musculus* | Mouse Genome Informatics |
| Nematode | *Caenorhabditis elegans* | WormBase |
| Rat | *Rattus norvegicus* | Rat Genome Database |
| Social amoeba | *Dictyostelium discoideum* | DictyBase |
| Thale cress | *Arabidopsis thaliana* | The Arabidopsis Information Resource |
| Maize | *Zea mays ssp. mays* | MaizeGDB |
| Zebrafish | *Danio rerio* | Zebrafish Information Network |
| Yeast | *Candida albicans* | CGD |
| Bacteria | *Escherichia coli* | EcoCyc |

Source: Wikipedia

# Tools for exploiting database information

# Tools

- What are they for?
  - Search of information (eg. *Entrez*)
  - Finding/comparing sequences (eg. *BLAST*)
  - Data exploration and visualization (eg. *Genome Browsers*)
  - Manipulating and analyzing data
  - Make predictions
  - Knowledge discovery (data mining)
  - Downloading/Exporting data
- Can be accessed through
  - Web interface from resource providers, databases, projects or subject-specific repositories
  - Software (eg. *R, Cytoscape*)
  - Platforms (eg. *Galaxy*)

Vall d'Hebron Institut de Recerca VHIR

UNITAT D'ESTADÍSTICA I BIOINFORMÀTICA

# Tools

NCBI

- Amino Acid Explorer
- Assembly Archive
- Basic Local Alignment Search Tool (BLAST)
- Batch Entrez
- BioAssay Services
- BLAST Link (BLink)
- BLAST Microbial Genomes
- BLAST RefSeqGene
- CDTree
- Cn3D
- COBALT
- Concise Microbial Protein BLAST
- Conserved Domain Architecture Retrieval Tool (CDART)
- Conserved Domain Search Service (CD Search)
- Digital Differential Display (DDD)
- Electronic PCR (e-PCR)
- Frequency-weighted Link (FLink)

EMBL-EBI

### Clustal Omega

Multiple sequence alignment of DNA or protein sequences. Clustal Omega replaces the older ClustalW alignment tools.

Multiple sequence alignment

### InterProScan

InterProScan searches sequences against InterPro's predictive protein signatures.

Protein feature detection | Sequence motif recognition

### BLAST [protein]

Fast local similarity search tool for protein sequence databases.

Sequence similarity search

### BLAST [nucleotide]

Fast local similarity search tool for nucleotide sequence databases.

Sequence similarity search

### HMMER

Fast sensitive protein homology searches using profile hidden Markov models (HMMs) for querying against both sequence and HMM target databases.

Sequence similarity search | Protein function analysis

# Tools

## The Entrez Search and Retrieval System

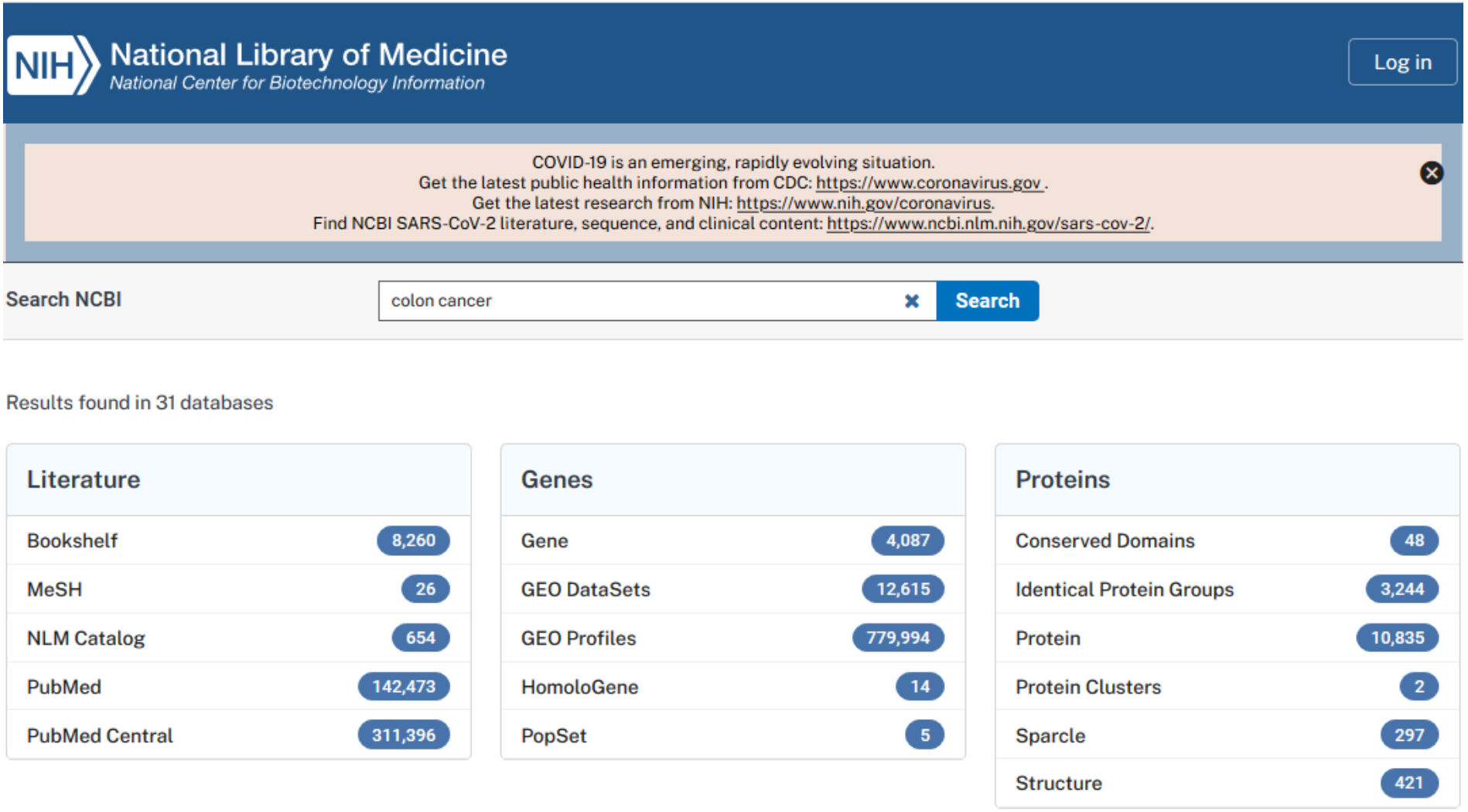
- Text-based search and retrieval system used at NCBI for all of its major databases
- All databases indexed by Entrez can be searched via a single query string. This returns a unified results page, that shows the number of hits for the search in each of the databases, which are also links to actual search results for that particular database.
- Supports boolean operators (AND, OR, NOT, “”, *)
- Use tags to limit parts of the search statement to particular fields.

```
term [field] OPERATOR term [field]
```

- Start with a general query and refine it progressively using Filters/Limits
- For individual databases, the Advanced Search and Limits pages assist greatly in the construction of complex queries.

# Practicum

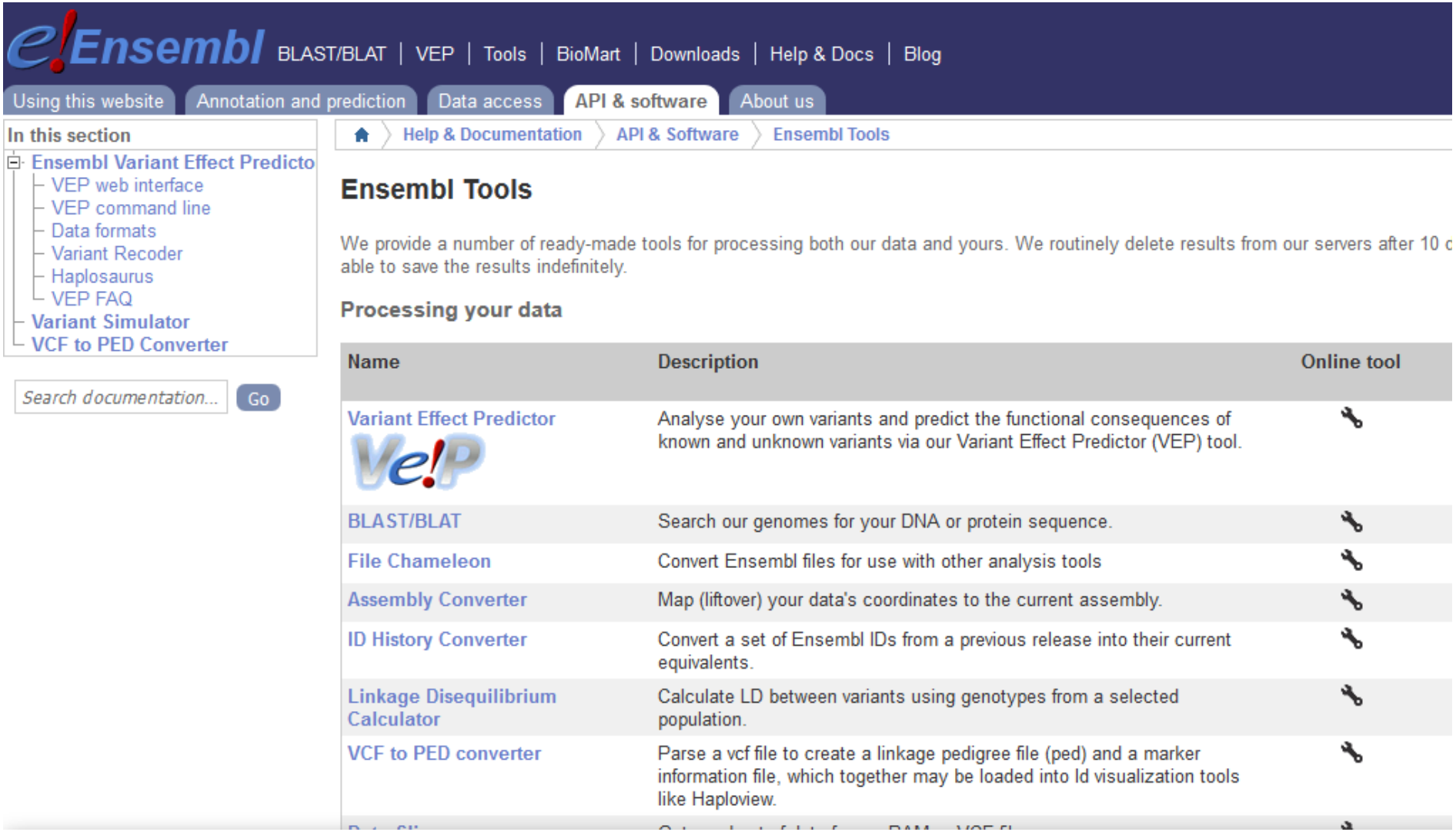
## Example of database cross-search with Entrez



Vall d'Hebron Institut de Recerca VHIR

UNITAT D'ESTADÍSTICA I BIOINFORMÀTICA

# Tools




https://www.ensembl.org/info/docs/tools/index.html

# Tools

Vall d'Hebron Institut de Recerca VHIR

UNITAT D'ESTADÍSTICA I BIOINFORMÀTICA

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Search:

Home | Install | Help | Developers | About

Home » BiocViews

## All Packages

**Bioconductor version 3.12 (Release)**

Autocomplete biocViews search:

- Software (1974)
  - AssayDomain (791)
  - BiologicalQuestion (822)
  - Infrastructure (456)
  - ResearchField (902)
    - BiomedicalInformatics (62)
    - CellBiology (54)
    - Cheminformatics (13)
    - ComparativeGenomics (8)
    - Epigenetics (63)
    - Epitranscriptomics (1)
    - FunctionalGenomics (53)
    - Genetics (200)
    - ImmunoOncology (447)
    - Lipidomics (11)
    - MathematicalBiology (8)
    - Metabolomics (74)

**Packages found under FunctionalGenomics:**

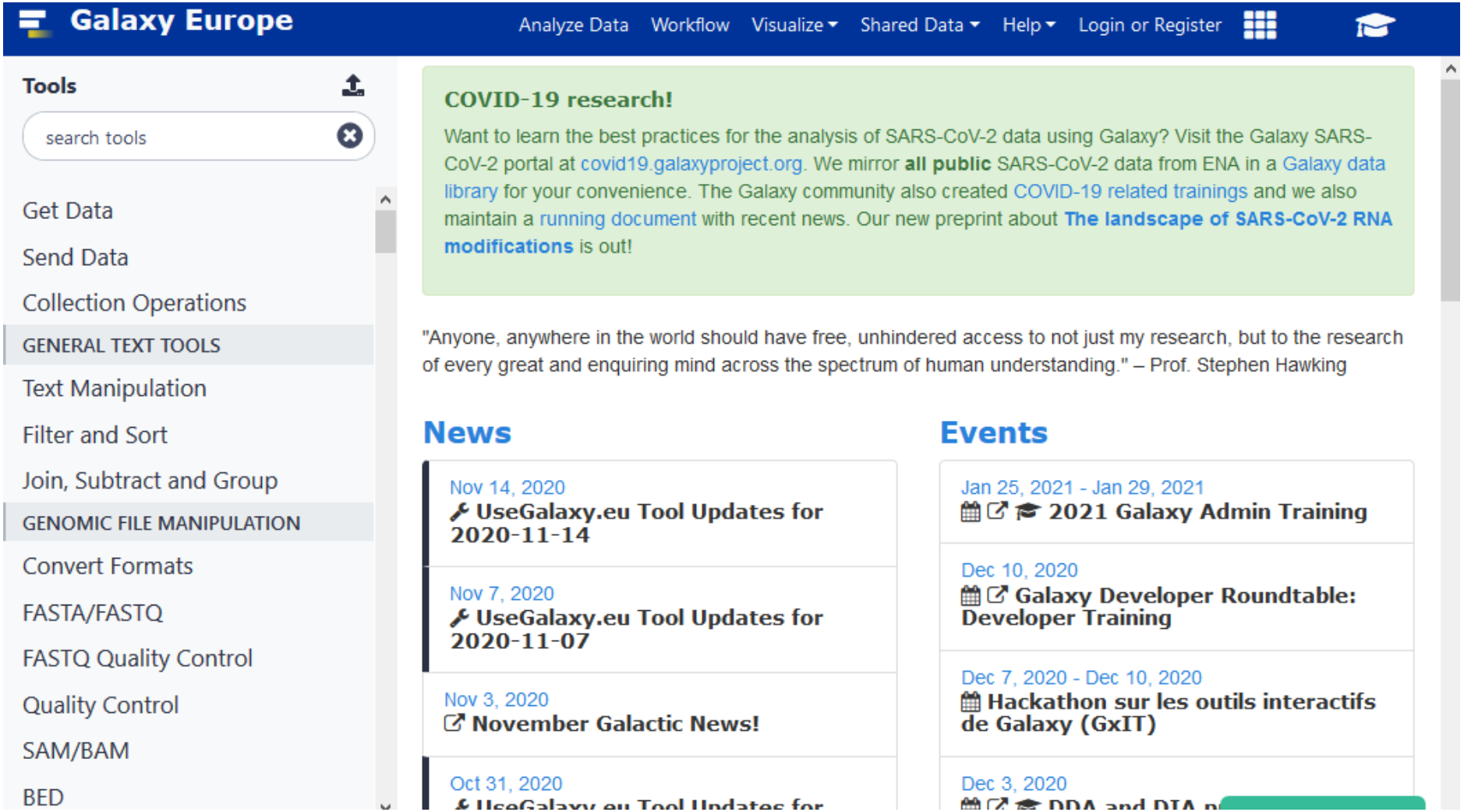
Rank based on number of downloads: lower numbers are more frequently downloaded.

Show All entries  Search table:

| Package | Maintainer | Title | Rank |
|---|---|---|---|
| limma | Gordon Smyth | Linear Models for Microarray Data | 14 |
| edgeR | Yunshun Chen, Gordon Smyth, Aaron Lun, Mark Robinson | Empirical Analysis of Digital Gene Expression Data in R | 23 |
| maftools | Anand Mayakonda | Summarize, Analyze and Visualize MAF Files | 112 |
| tximeta | Michael Love | Transcript Quantification Import with Automatic Metadata | 162 |
| DiffBind | Rory Stark | Differential Binding Analysis of ChIP-Seq Peak Data | 165 |
| annotatr | Raymond G. Cavalcante | Annotation of Genomic Regions to Genomic Annotations | 274 |
| variancePartition | Gabriel E. Hoffman | Quantify and interpret divers of variation in multilevel gene expression experiments | 278 |

https://bioconductor.org/packages/3.12/BiocViews.html#___FunctionalGenomics

# Tools




https://usegalaxy.eu/

# Tools

Cytoscape App Store | Submit an App | Search the App Store | Sign In

All Apps

## Categories

- collections
- data visualization
- network generation
- network analysis
- graph analysis
- online data import
- automation
- integrated analysis
- clustering
- systems biology
- utility
- enrichment analysis
- visualization
- data integration

## Newest Releases

Get Started with the App Store »

**DKernel** 3.0+
DKernel uses Diffusion Kernel algorithm to propagate sub-

**PathLinker** 3.0+
Reconstructs signaling pathways from protein interaction networks

**XlinkCyNET** 3.0+
XlinkCyNET generates residue-to-residue connections provided by

**IntAct App** 3.0+
BETA: Build molecular interaction networks from IntAct database.

**MCODE** 3.0+
Clusters a given network based on topology to find densely

**OmniPath** 3.0+
OmniPath: literature curated human signaling pathways

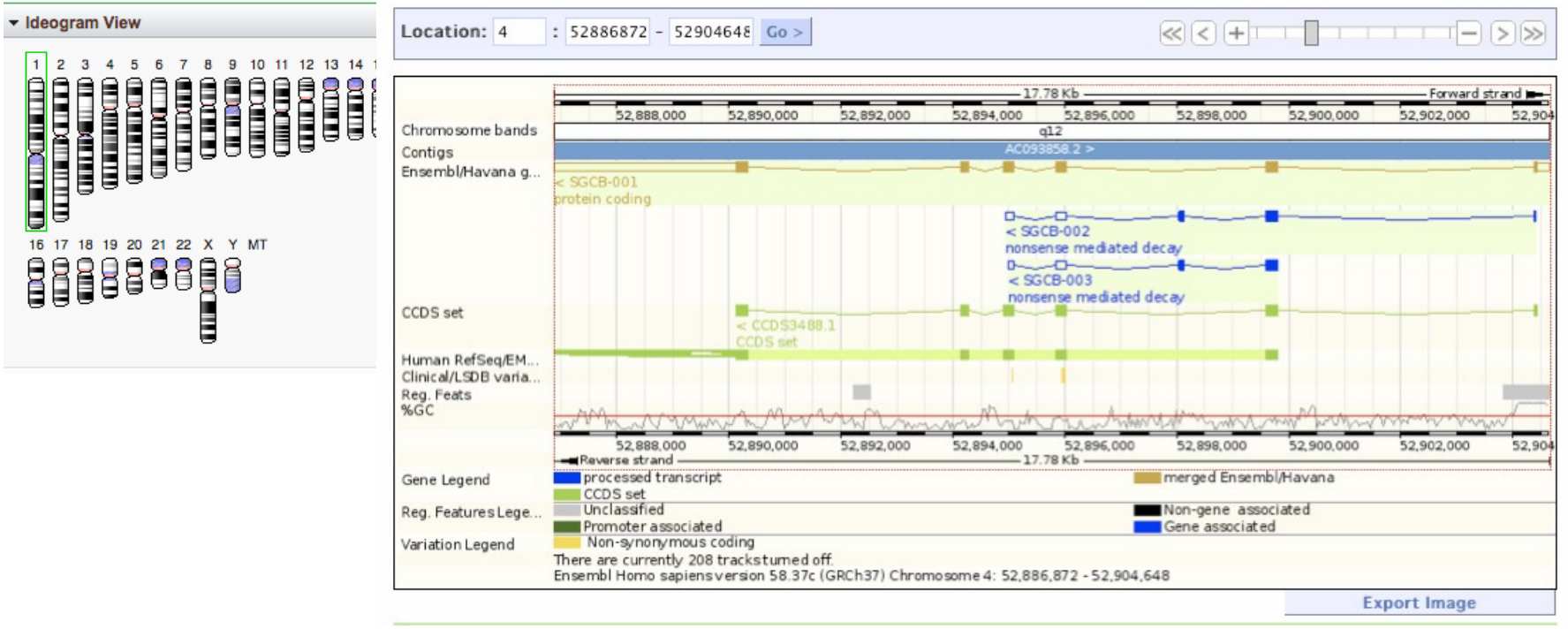
more newest releases »

https://apps.cytoscape.org/apps/

# Genome Browsers

# Genome Browsers

- The advent of the Human Genome Project and subsequent projects to sequence genomes of other species and multiple individuals has driven the need for tools that can visualize vast amounts of genomics data.
- Genome Browsers bring together information from multiple resources, using the genome as a base for this annotation.
- Provide a graphic interface for visualization and integrative genomics analyses.



# Genome Browsers

- A wealth of biological data can be viewed, downloaded and compared:
  - Genes
    - Genomic location
    - Gene model structures: exons, introns, UTRs
    - Transcripts: mRNA, splice variants, pseudogenes, non-coding RNA,...
    - Protein(s)
    - Links to other sources of information
  - Cytogenetic bands
  - Polymorphic markers
  - Genetic variation: SNPs, deletions/insertions, short tandem repeats,...
  - Repetitive sequences
  - Expressed Sequence Tags (ESTs)
  - cDNAs or mRNAs from related species
  - Regions of sequence homology

# Genome Browsers

- Popular Genome Browsers:
  - NCBI Genome Data Viewer
  - EBI's Ensembl
  - UCSC Genome Browser

# Genome Browsers

- Important concepts to be aware of:
  - Genome of reference
    - A reference genome (also known as a reference assembly) is a *digital* nucleic acid sequence database, assembled by scientists as a representative example of the set of genes in one idealized individual organism of a species.
    - Reference genomes are typically used as a guide on which new genomes are built, enabling them to be assembled much more quickly and cheaply than the initial Human Genome Project.
    - The reference provides a good approximation of the DNA of any single individual. But in regions with high allelic diversity the reference genome may differ significantly from other individuals
      - sets of *alternate loci* are assembled alongside the reference locus
    - There are reference genomes for multiple species of viruses, bacteria, fungus, plants, and animals.

# Genome Browsers

- Important concepts to be aware of:
  - Genome of reference
    - The Genome Reference Consortium (GRC) is a collaborative effort between different institutes charged to maintain and improve reference genomes for human and some model organisms (mouse, zebrafish and chicken)
    - New assemblies are released every *X* years integrating improvements in sequence
      - closing gaps
      - fixing misrepresentations in the sequence
      - correcting sequences
    - The coordinates of your favorite gene in one assembly may not be the same as those in the next release of the assembly!
      - Always be aware of which version you are using
      - NCBI provides a Genome Remapping Service

# Genome Browsers

- Important concepts to be aware of:
  - Genome of reference
    - To date, the major assembly releases for human, mouse, zebrafish, and chicken are GRCh38, GRCm38, GRCz11, and GRCg6a, respectively.

For human:

| Release name | Date of release | Equivalent UCSC version |
|---|---|---|
| GRCh38 | Dec 2013 | hg38 |
| GRCh37 | Feb 2009 | hg19 |
| NCBI Build 36.1 | Mar 2006 | hg18 |
| NCBI Build 35 | May 2004 | hg17 |
| NCBI Build 34 | Jul 2003 | hg16 |

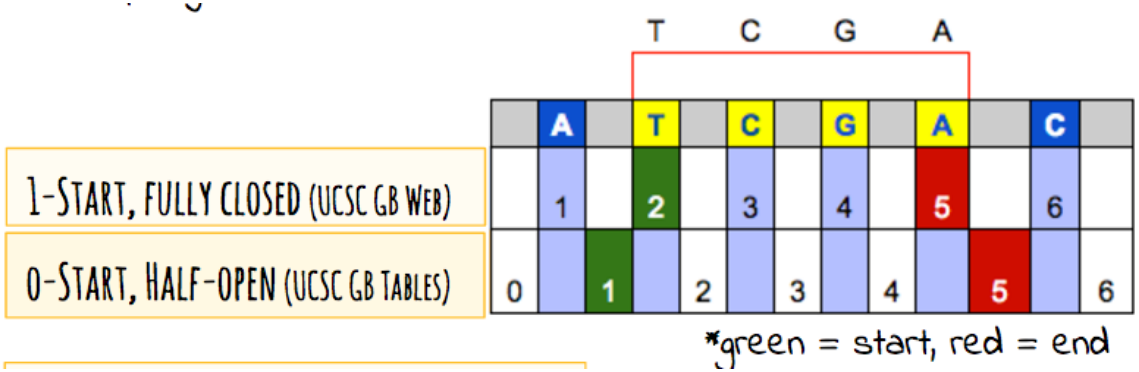
For mouse:

| Release name | Date of release | Equivalent UCSC version |
|---|---|---|
| GRCm38 | Dec 2011 | mm10 |
| NCBI Build 37 | Jul 2007 | mm9 |
| NCBI Build 36 | Feb 2006 | mm8 |
| NCBI Build 35 | Aug 2005 | mm7 |
| NCBI Build 34 | Mar 2005 | mm6 |

- There are "minor" assembly updates in the form of genome *patches* which either correct errors in the assembly or add additional alternate loci *without changing chromosome coordinates.* (e.g. GRCh37.p1).
- The assembly **accession.version** is an unambiguous identifier for the assembly and should always be included in publications.

# Genome Browsers

- Important concepts to be aware of:
  - Genome of reference (assembly version)
  - Genomic coordinate systems:
    - **base** coordinate system: the first nucleotide counts as position **1**
    - **interbase** coordinate system: the first nucleotide counts as position **0**
      - allows to represent features that occur between nucleotides (like a splice site)
      - simpler arithmetic for computing the length of features (length=end-start)
      - more rational conversion of coordinates from the + to the - strand




→ TCGA is 2-5 (both start and end included)

→ TCGA is 1-5 (start included, end excluded)

# Genome Browsers

**The "Position" format** (referring to the "1-start, fully-closed" system as coordinates are "positioned" in the browser)

- Written as: chr1:127140001-127140001
- No spaces.
- Includes punctuation: a colon after the chromosome, and a dash between the start and end coordinates.
- When in this format, the assumption is that the coordinate is 1-start, fully-closed.

**The "BED" format** (referring to the "0-start, half-open" system)

- Written as: chr1 127140000 127140001
- Spaces between chromosome, start coordinate, and end coordinate.
- No punctuation.
- When in this format, the assumption is that the coordinates are 0-start, half-open.

# Genome Browsers

- Important concepts to be aware of:
  - Genome of reference (assembly version)
  - Genomic coordinate systems:
    - Most genome annotation portals (e.g. NCBI or Ensembl), bioinformatics software (e.g. BLAST) and annotation file formats (e.g. GFF, BED) use the base coordinate system
    - The UCSC genome browser uses both systems:
      - the base coordinate system (1-based, fully-closed) is used in the UCSC genome browser display
      - the interbase coordinate (0-based, half-open) is used in their tools and file formats

**Table 2. SNP coordinates in web browser (1-start) vs table (0-start)**

| rs782519173 (hg38) | Start | End |
|---|---|---|
| Positioned in web browser: 1-start, fully-closed | 133255708 | 133255708 |
| Stored in table: 0-start, half-open | 133255707 | 133255708 |

# Genome Browsers

- Popular Genome Browsers:
  - NCBI Genome Data Viewer
  - EBI's Ensembl
  - UCSC Genome Browser
  - IGV from Broad Institute
- Similar though may differ in:
  - Presentation
  - Species represented
  - Source of annotations / links to other resources
  - Tools

# Genome Browsers

## Ensembl

https://www.ensembl.org/index.html

- The project began in 1999, with the completion of the HGP, as a joint project between the EMBL-EBI and the Sanger Centre (now all moved to the EMBL-EBI).
- As of Ensembl release 106 (April 2022), more than 250 species are supported (mainly vertebrate species)
- Sister project: EnsemblGenomes for non-chordates
  - Since its establishment in 2009, the resource has grown rapidly and now contains over 1,400 eukaryotic and 44,000 prokaryotic genomes.
- Genome assemblies are retrieved from other institutes/consortia (eg. NCBI for human, mouse genomes)

# Genome Browsers

## Ensembl

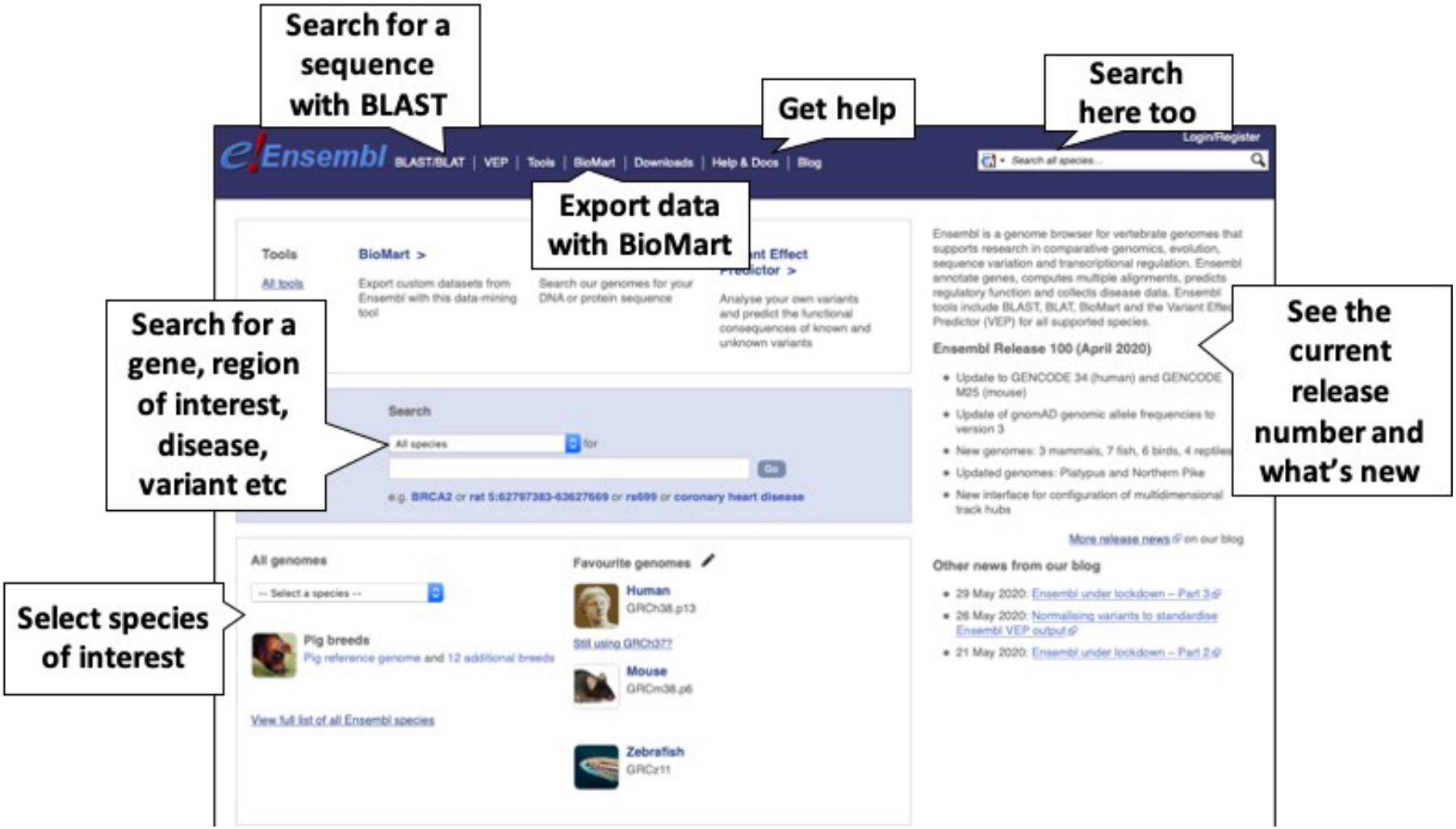
https://www.ensembl.org/index.html

- There are a number of ways to access Ensembl data:
  - From the website
  - Using BioMart tool to quickly obtain tables of gene information
  - Programmatically

# Genome Browsers

## Ensembl

https://www.ensembl.org/index.html



# Genome Browsers

## Ensembl

https://www.ensembl.org/index.html

- The navigation of the Ensembl website is organised into tabs, or main pages. The location, gene, transcript, variant and regulation tabs allow data browsing at that level.

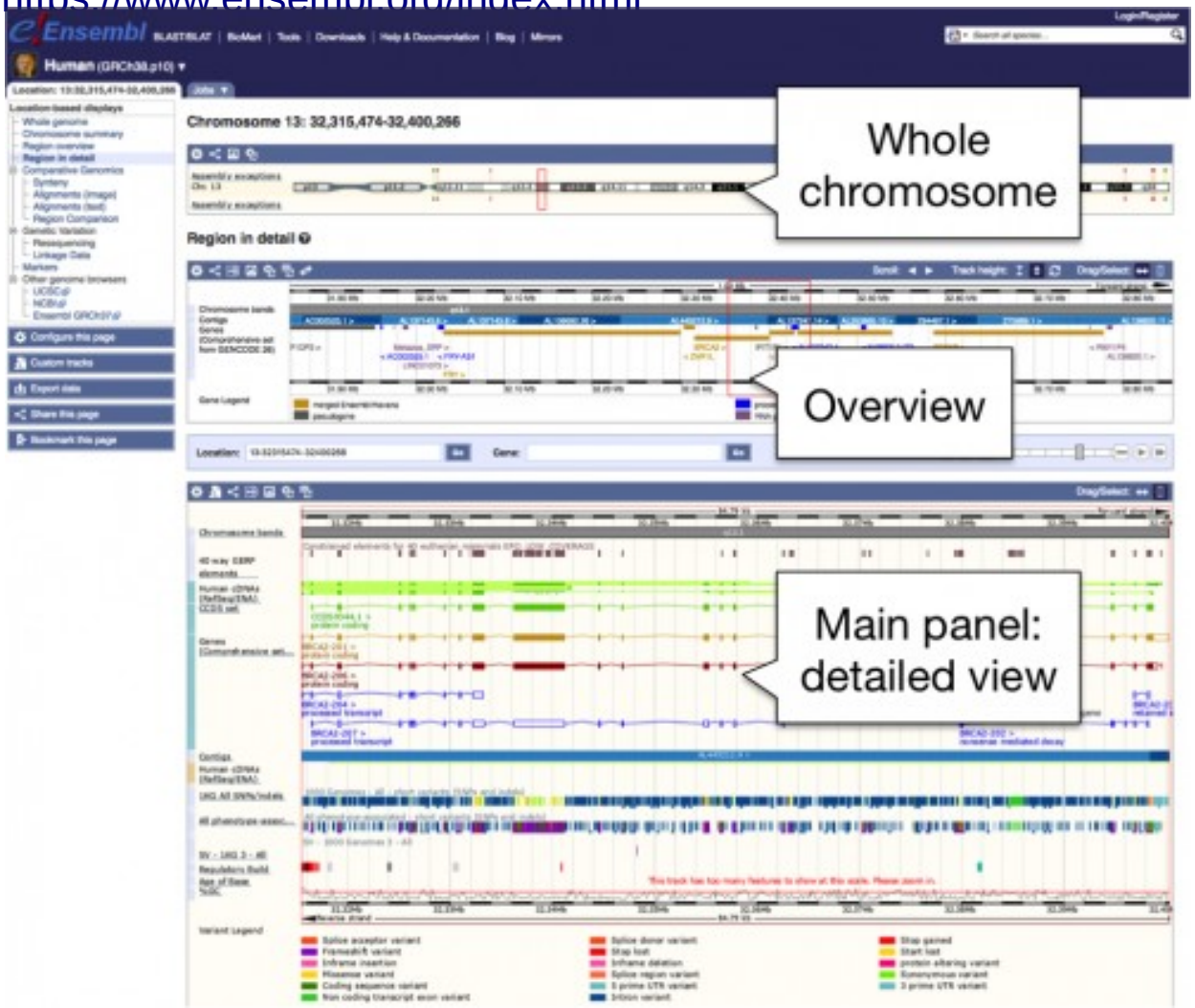

- The available annotations (tracks) to be displayed are configured through the 'Configure this page' button

# Genome Browsers

## Ensembl

https://www.ensembl.org/index.html

- The location tab's main display is the 'region in detail', which shows a region of the genome up to 1 Mb long in a highly customisable view

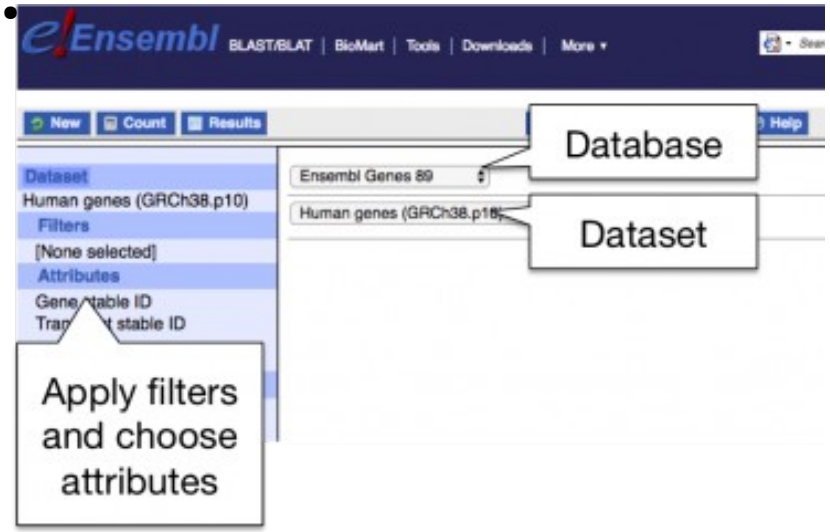


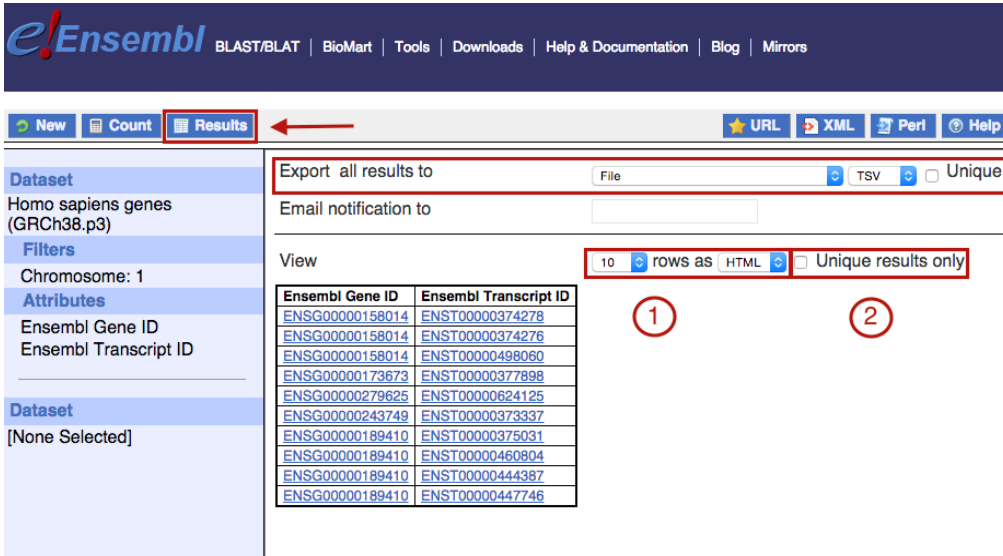
# Genome Browsers

## Ensembl

https://www.ensembl.org/index.html

- BioMart tool allows to:
  - Search and quickly generate tables of information.
  - Export data in different formats (FASTA, html, csv, tsv, xls ...)
  - 'Translate' one ID type into another (eg. an Ensembl gene ID to an NCBI RefSeqID)
  - Also in R! (see biomaRt Bioconductor's package )



# Genome Browsers

## Ensembl

https://www.ensembl.org/index.html

- Other tools:


The Variant Effect Predictor is our most popular tool. Enter in transcript or genomic coordinates to determine the effect of sequence variation on transcripts and proteins. A dbSNP identifier will be given in the output, if there is a matching one.

The Assembly Converter allows coordinates from an older genome sequence to be updated to new coordinates (and vice-versa). As genomes are sequenced, the improved technology allows current genome sequence to be more accurate, containing fewer gaps and fewer mistakes. Using the most recent genome version or assembly is advised. Ensembl, the UCSC genome browser, and NCBI Genome Data Viewer strive to show all annotation on the newest assembly possible, once the genome sequence is released to the public.

ID History converter displays IDs that are in the current version of Ensembl. Start with a list of old IDs, and see which ones are still used, and which ones have been 'retired', or changed into a different ID. Though Ensembl IDs are stable (a gene or transcript should always have the same ID), the ID can change if one gene is split into two, or two genes that were erroneously split in a previous release are fused together into one.

# Genome Browsers

## Ensembl

https://www.ensembl.org/index.html

- Resources for training:
  - Free Online EMBL course on using Ensembl genome browser
  - Free Online EMBL course on using EnsemblGenomes

# Genome Browsers

## UCSC Genome Browser

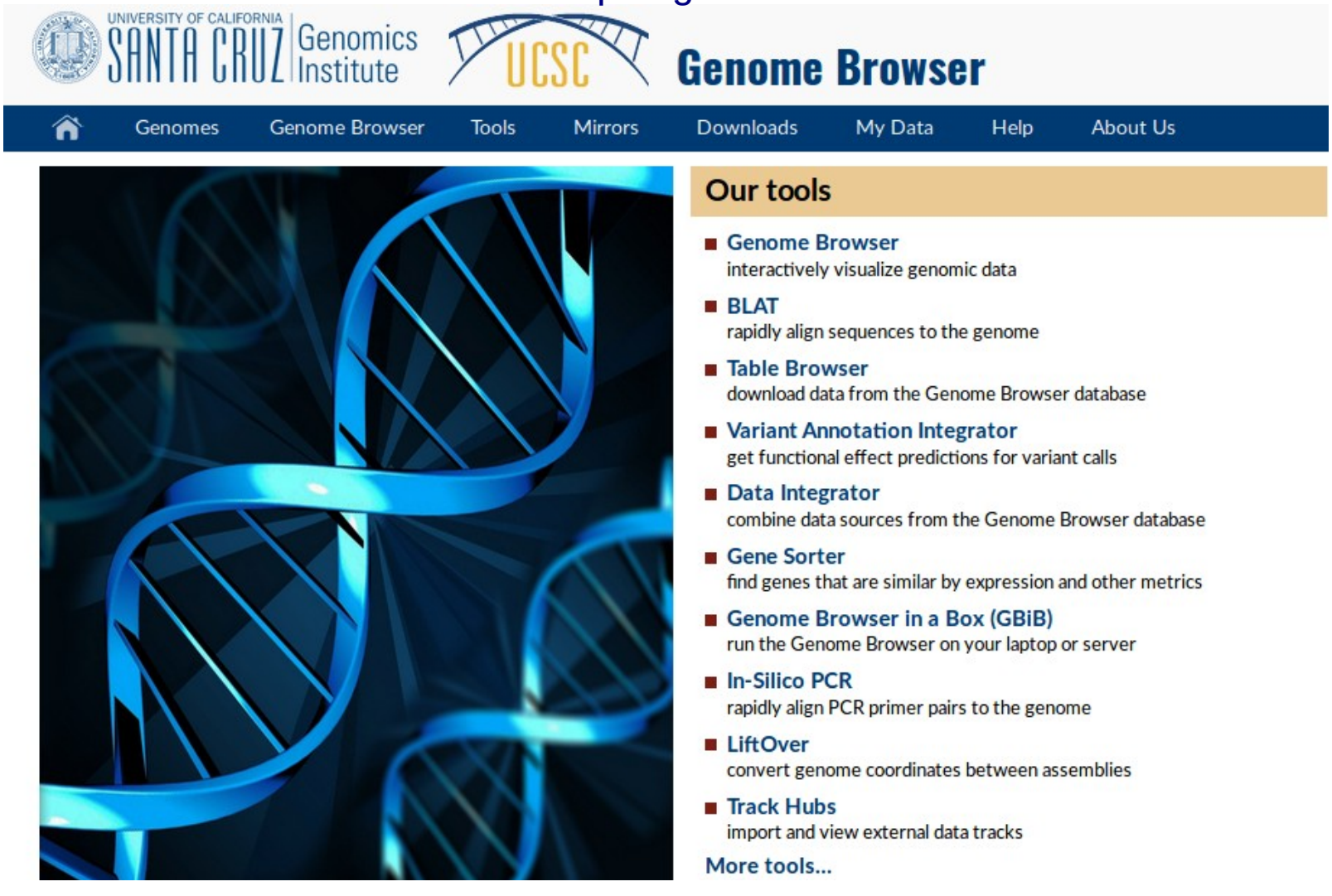
https://genome.ucsc.edu/

- Developed and maintained by the Genome Bioinformatics Group, within the UCSC Genomics Institute.
- It began as a resource for the distribution of the initial fruits of the Human Genome Project. Funded by the Howard Hughes Medical Institute and the NHGRI, the browser offered a graphical display of the first full-chromosome draft assembly of human genome sequence.
- In the ensuing years, the website has grown to include a broad collection of vertebrate and model organism assemblies and annotations, along with a large suite of tools for viewing, analyzing and downloading data.

# Genome Browsers

## UCSC Genome Browser

https://genome.ucsc.edu/

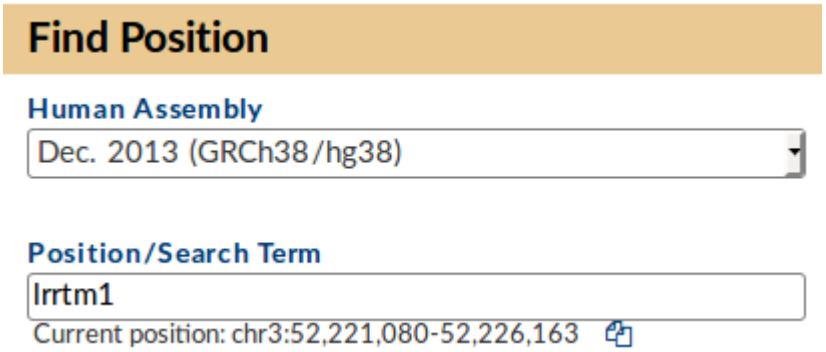


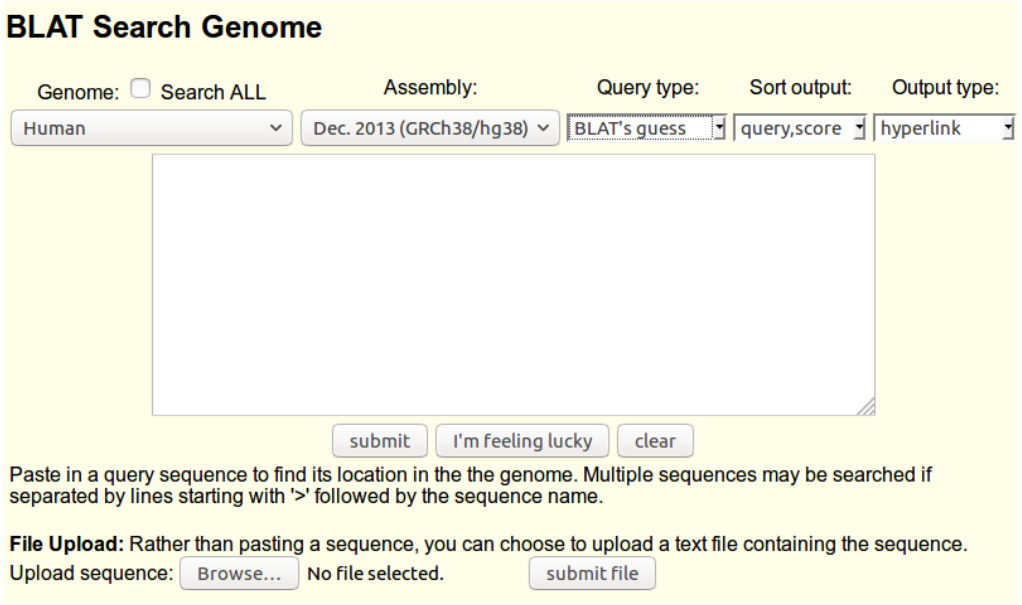
# Genome Browsers

## UCSC Genome Browser

https://genome.ucsc.edu/

- Different search options:
  a) By gene/transcript/protein name, symbol or ID: LRRTM1
  b) By Chromosome number or region: chr11:1038475-1075482
  c) By Keywords: kinase, receptor
  d) By sequence (BLAT tool)
  e) By track type (Track search)



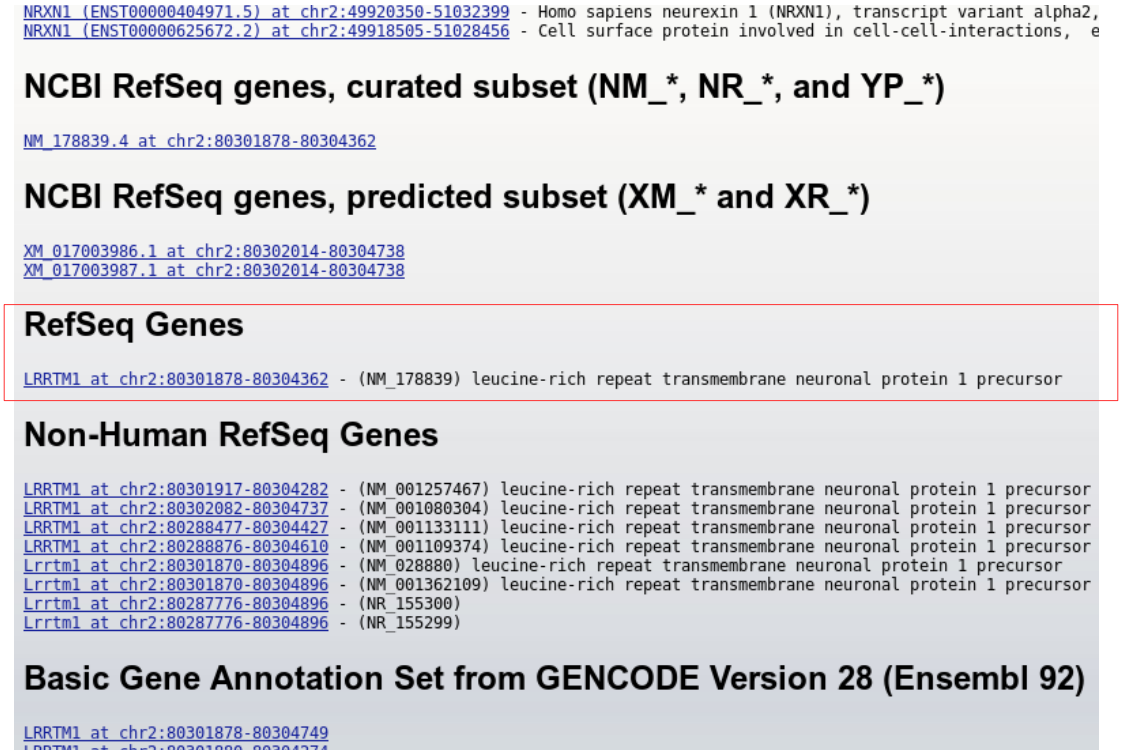



# Genome Browsers

## UCSC Genome Browser

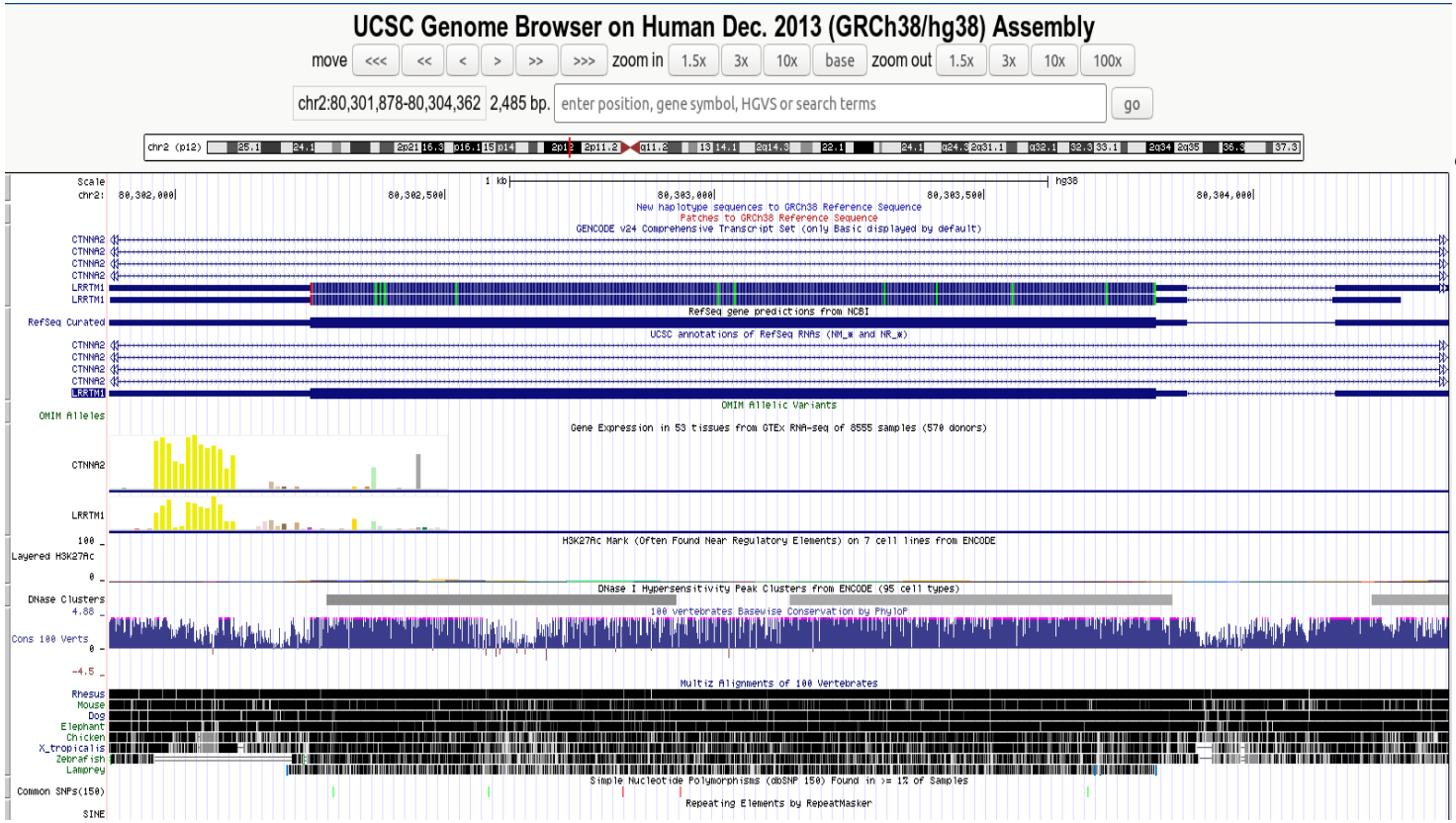
https://genome.ucsc.edu/

- Multiple results depending on available annotations:
  a) UCSC Genes
  b) RefSeq Genes
  c) Non-human RefSeq Genes: orthologs of the gene in other species
  d) ENCODE Gencode
  e) Human mRNA: annotated transcripts of the gene



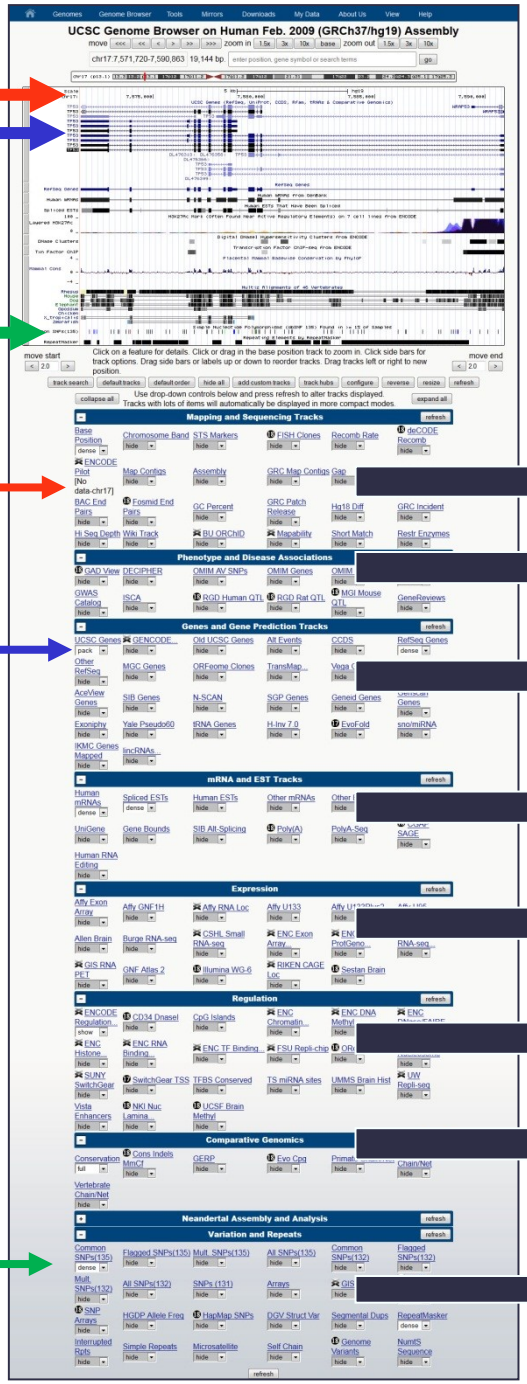
# Genome Browsers

- Visualizing results: Genome Viewer and Tracks settings




- Genomic location is shown along with data annotations that link out to additional data and databases.

Tracks info and options




# Genome Viewer

# Tracks (data type)

**Mapping and Sequencing Tracks**

**Phenotype and Disease Tracks**

**Genes and Gene Prediction Tracks**

***(including sno/miRNA data)***

**mRNA and EST Tracks**

**Expression** ***(such as microarray)***

**Regulation** ***(including TFBS)***

**Comparative Genomics**

***As a group***

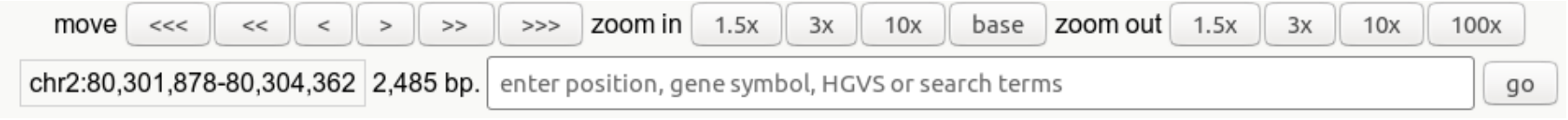
***Individual species***

**Variation and Repeats**

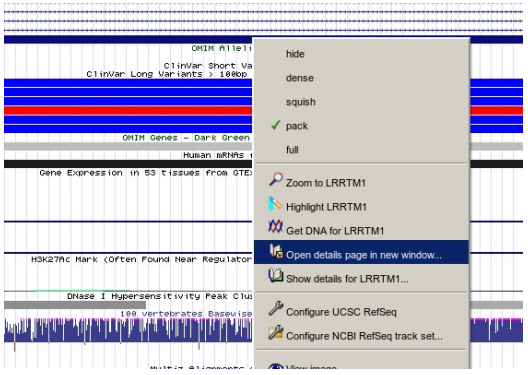
***(including SNPs, copy number variation)***

# Genome Browsers

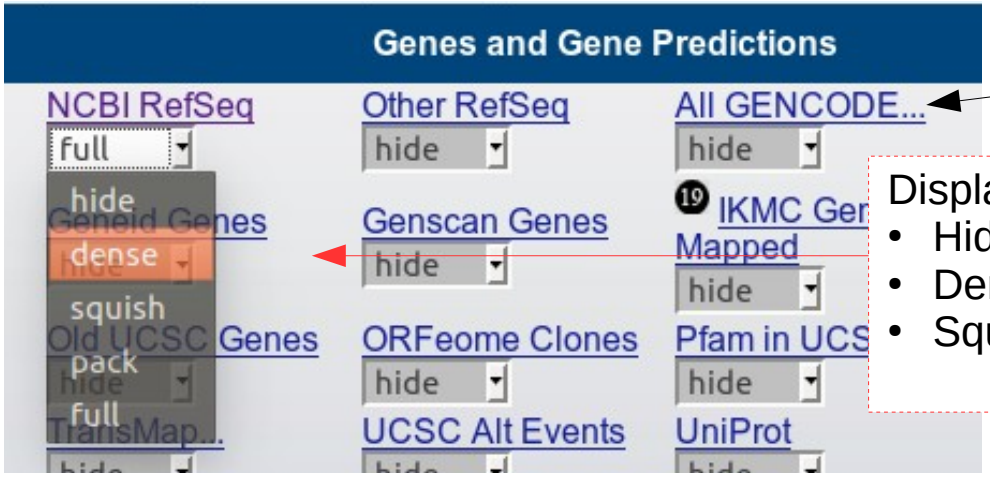
- Change your view or location with controls at the top




- Click on items to view details in new window or right click items to get details




- Change track display modes:
  - Tip: Hide all and then select specific tracks to visualize so you don't get lost



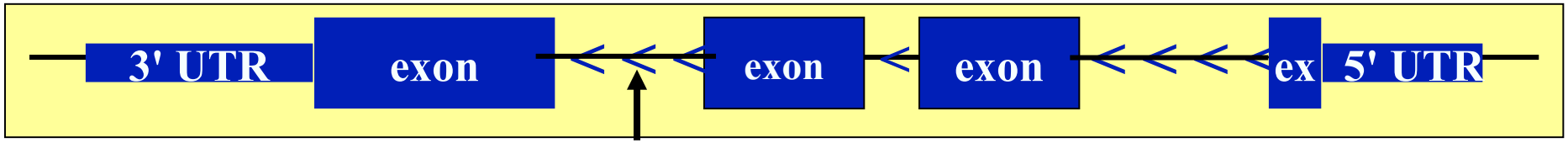

Source of information

Refresh!

Display modes:
- Hide=don't show
- Dense=features collapsed into a single line
- Squish, Pack, Full=features in different lines

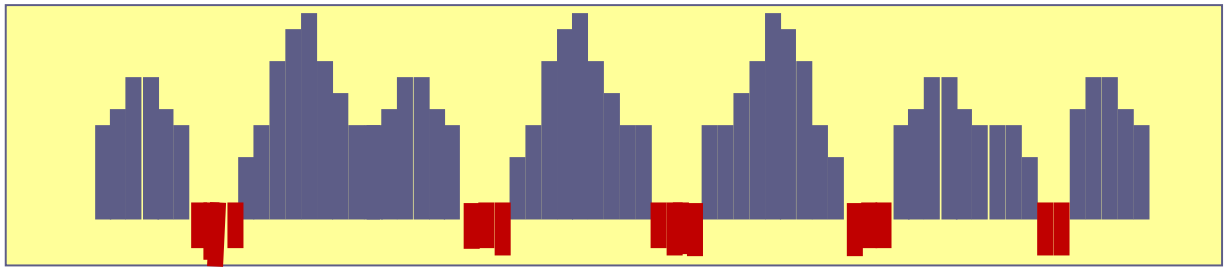
# Genome Browsers

- Some visual clues:



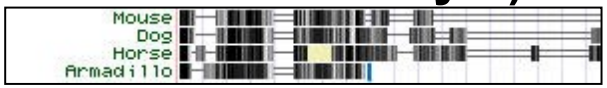

**Intron and direction of transcription <<< or >>>**

Sequence conservation



**height of a blue bar is increased likelihood of conservation,**
**red indicates a likelihood of faster-evolving regions**

**Alignment indications (Conservation pairs: "chain" or "net" style)**
**Alignments = boxes, Gaps = lines**




# Practicum

## Retrieving information with the UCSC Genome Browser

https://genome.ucsc.edu/

**1.** What is the genomic localization of human *lrrtm1* gene?

-chromosome:

-position:

-strand:

**2.** Which genes are in the neighbourhood of this gene?

**3.** How many exons has the gene?

**4.** How many different transcripts do we know of this genomic region?

**5.** Can you find SNPs in this gene?

**6.** In which tissue is this gene mainly expressed?

**7.** Does the protein encoded by this gene have a transmembrane domain?

**8.** Has this gene an ortholog in mouse?

**9.** Use the CDS of human *lrrtm1* gene to localize this gene in mouse genome. (Use BLAT)

# Practicum

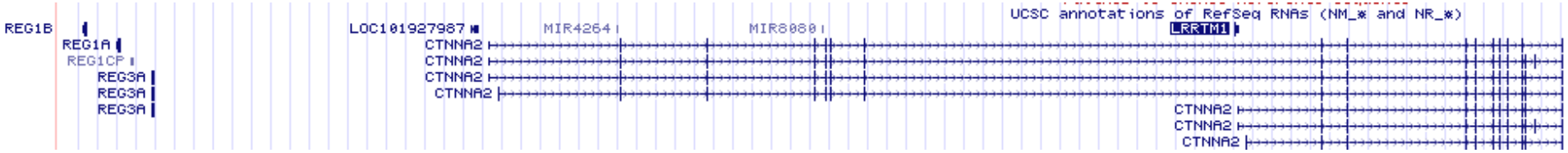
1. What is the genomic localization of human *lrrtm1* gene?

   - Click on the gene to see the information

**Position:** chr2:80301878-80304752
**Band:** 2p12
**Genomic Size:** 2875
**Strand:** -
**Gene Symbol:** LRRTM1
**CDS Start:** complete
**CDS End:** complete

2. Which genes are in the neighborhood of this gene?

   - Zoom out in genome viewer



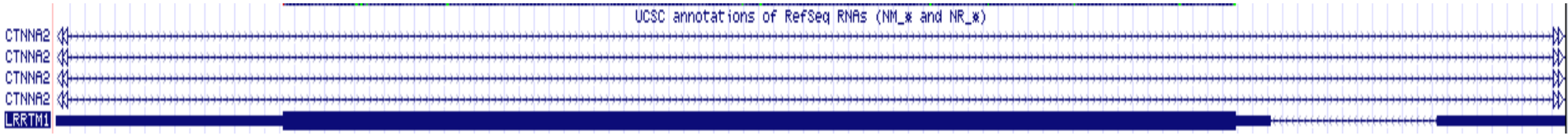

3. How many exons has the gene? 2

   - Place the mouse on the gene
   - Or: click on RefSeq link to open in NCBI-GenBank
   - Or: TableBrowser

# Practicum

4. How many different transcripts do we know of this genomic region? 5

   - Use Genome Viewer or Table Browser for more detailed information

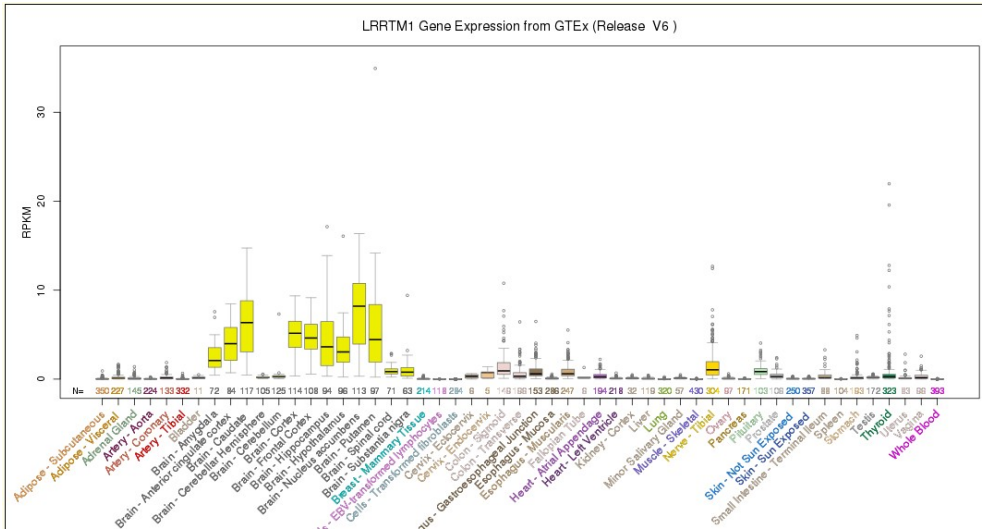



5. Can you find SNPs in this gene? yes

   - Set display mode of track “Variation” > “SNPs” to ON


6. In which tissue is this gene mainly expressed? brain

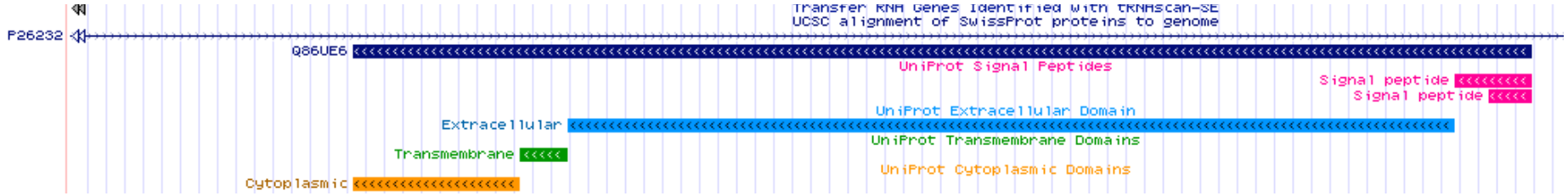
   - Info in track “Expression” > “GTEx”



# Practicum

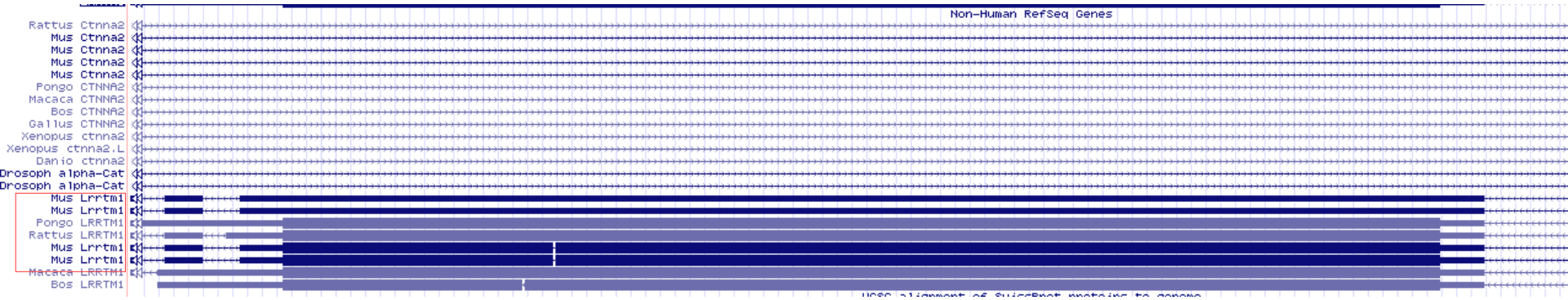
7. Does the protein encoded by this gene have a transmembrane domain?

- Set display mode of track “Gene and Gene Predictions” > “UniProt” to “pack”




8. Has this gene an ortholog in mouse? yes

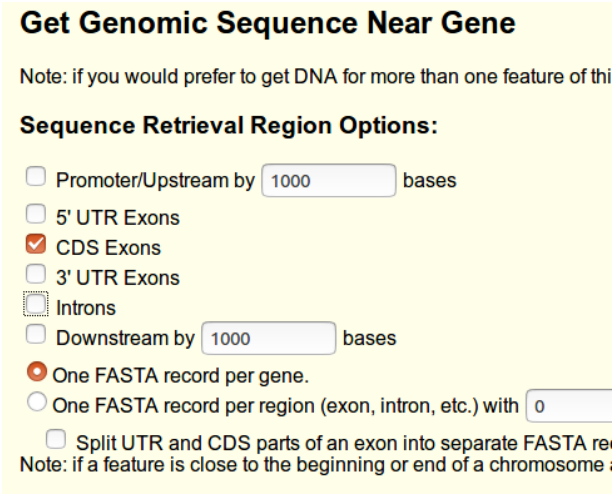
- Set display mode of track “Gene and Gene Predictions” > “Other RefSeq” to “pack”



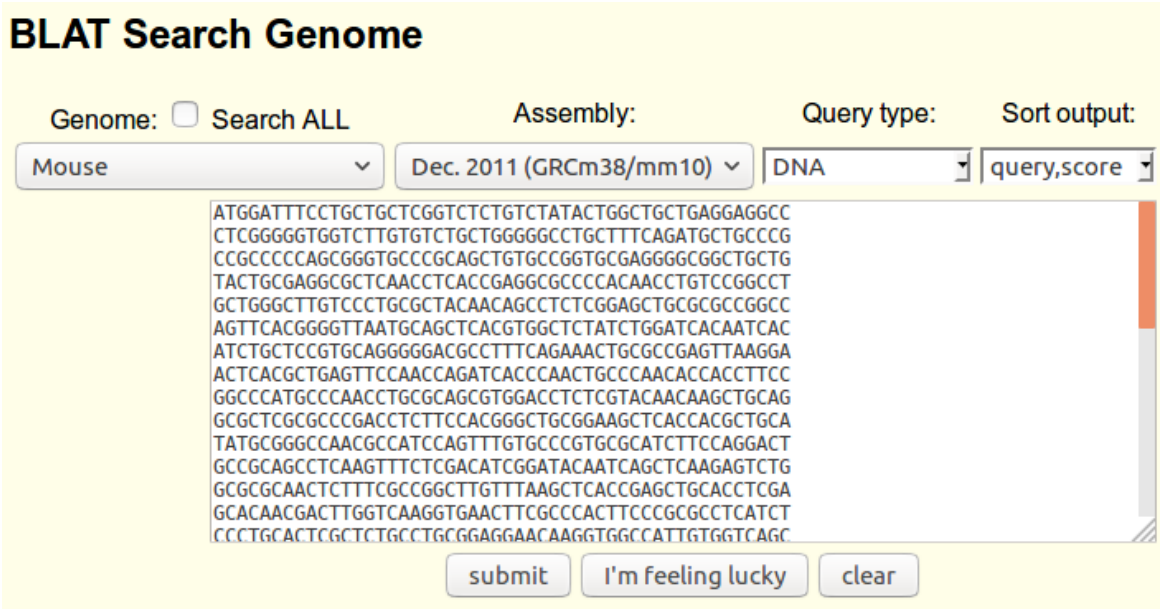
# Practicum

9. Use the CDS of human *lrrtm1* gene to localize this gene in mouse genome. (Use BLAT)

- Get the CDS sequence of human gene:



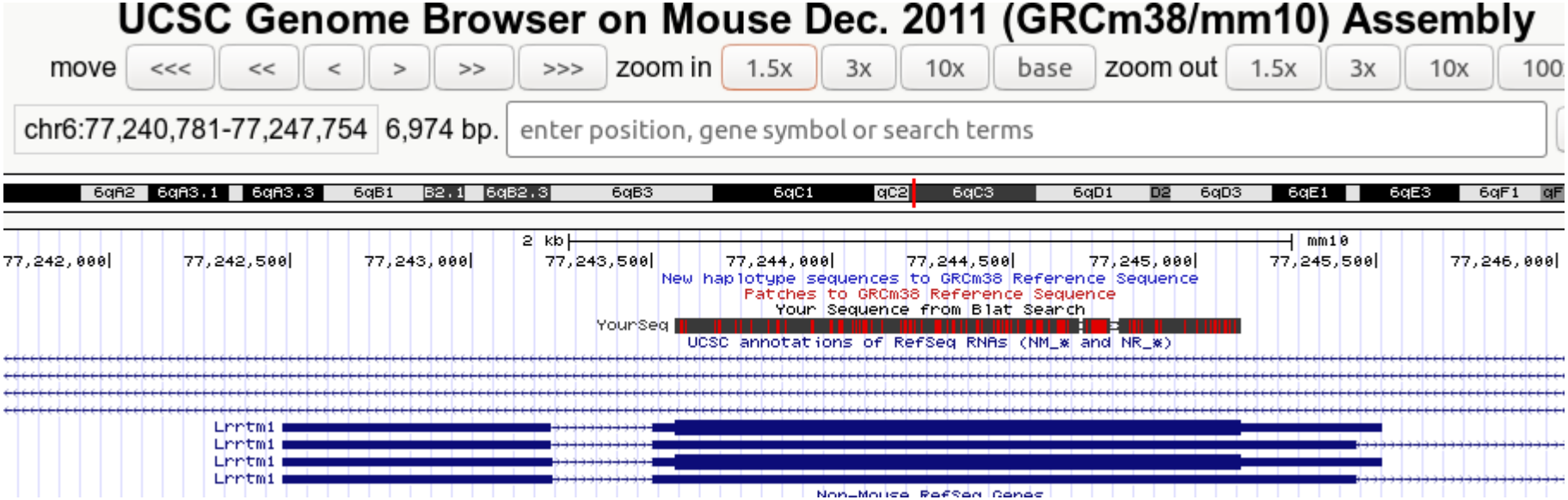

- Use BLAT to align this sequence on Mouse genome:



# Practicum

- Visualize entry with highest score

| ACTIONS | QUERY | SCORE | START | END | QSIZE | IDENTITY | CHROM | STRAND | START | END | SPAN |
|---|---|---|---|---|---|---|---|---|---|---|---|
| browser details | YourSeq | 1305 | 1 | 1569 | 1569 | 92.6% | chr6 | + | 77243562 | 77245130 | 1569 |
| browser details | YourSeq | 30 | 43 | 88 | 1569 | 94.2% | chr12 | - | 20883055 | 20883105 | 51 |
| browser details | YourSeq | 24 | 1304 | 1329 | 1569 | 96.2% | chr11 | - | 106320438 | 106320463 | 26 |
| browser details | YourSeq | 22 | 1399 | 1420 | 1569 | 100.0% | chr10 | - | 66129920 | 66129941 | 22 |
| browser details | YourSeq | 22 | 367 | 389 | 1569 | 100.0% | chr13 | + | 97550711 | 97550734 | 24 |
| browser details | YourSeq | 22 | 23 | 49 | 1569 | 92.4% | chr10 | + | 129969924 | 129969951 | 28 |
| browser details | YourSeq | 22 | 23 | 49 | 1569 | 92.4% | chr10 | + | 129978799 | 129978826 | 28 |
| browser details | YourSeq | 21 | 396 | 426 | 1569 | 83.9% | chr14 | + | 57525552 | 57525582 | 31 |
| browser details | YourSeq | 21 | 848 | 869 | 1569 | 100.0% | chr1 | + | 22046494 | 22046516 | 23 |



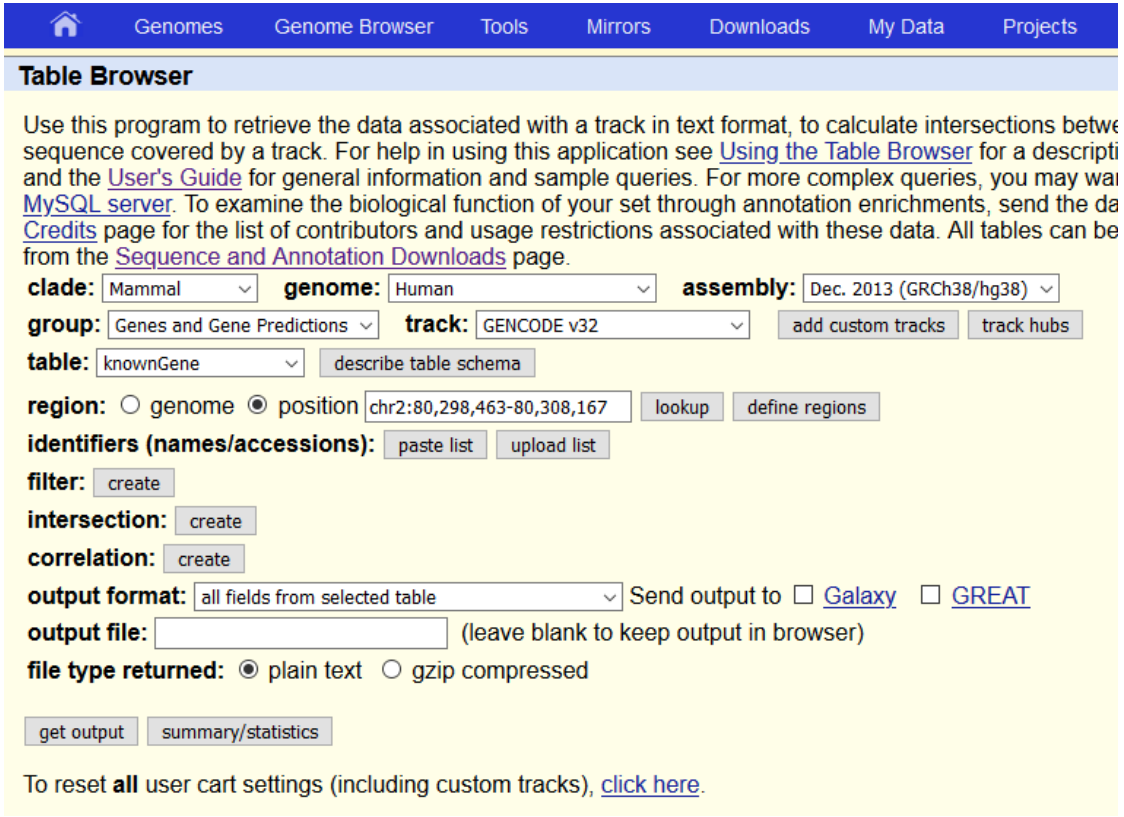
# Genome Browsers

## UCSC Genome Browser

https://genome.ucsc.edu/

- Table Browser tool allows to:
  - Search for genes and annotation
  - Combine queries on multiple tables or tracks
  - Display basic statistics over a dataset
  - Output to results table in convenient format
  - Retrieve sequences
  - Export to external resources

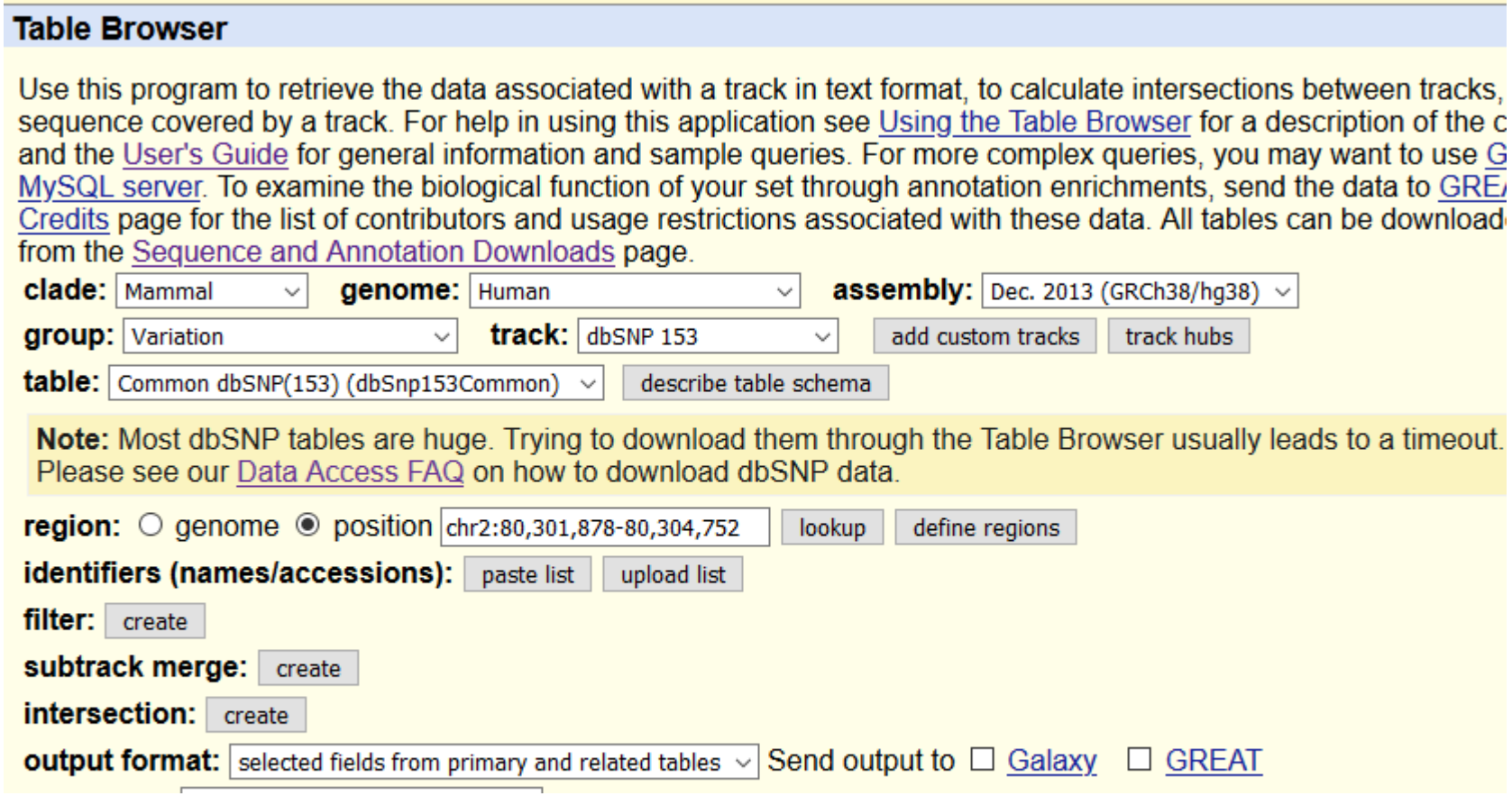
Examples



# Retrieving information with the UCSC Genome Browser

https://genome.ucsc.edu/

**10.** Use the Table Browser to get the list of SNPs in your region

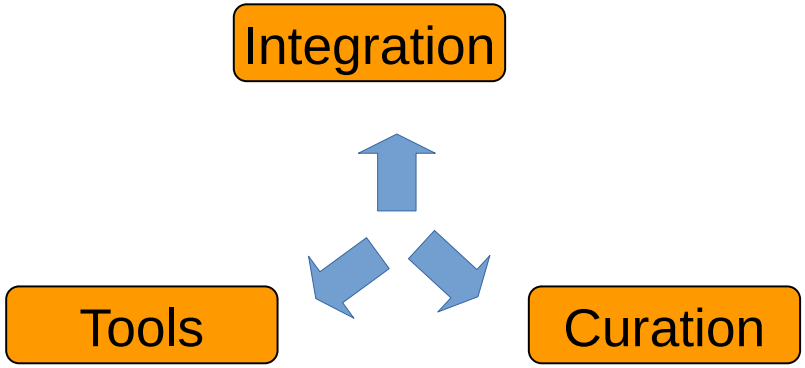


# Summary

- **Databases**
  - data collections
  - many types and diverse information
- **For information to be accessible / useful it must be**
  - Structured
  - Annotated
- **Resource Providers**
  - centers or organizations specialized in storing and maintaining databases
  - centralize data management

# Summary

- General
  - **Resource providers**
- Subject-specific
  - **Collaborative projects**
  - **Multi-omics repositories**
- Bioinformatics tools for exploiting database information
  - **Queries and access to data**
  - **Genomic Browsers**



# Final considerations

- Keeping informed about what you are seeing ensures correct interpretation of results
  - type of information (mRNA, gene, protein, SNP...)
  - source of information (curated, experimental, predicted, annotation, database)
  - specific tutorials: a good beginning
- Don't get overwhelmed: make specific queries, filter output
- When using data for drawing conclusions, appropriate controls may be used that make confidence of your search
  - scores of confidence
  - contrast information