Databases in molecular biology (part II)

Bioinformatics Course UEB-VHIR
November 2020

Ricardo Gonzalo[1], Mireia Ferrer[1], Álex Sánchez[1,2]
Berta Miró[1], Angel Blanco[1,2]

1 Unitat d'Estadística i Bioinformàtica (UEB) VHIR
2 Departament de Genètica Microbiologia i Estadística, UB

Summary

Previously in session 2...

- **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

Previously in session 2...

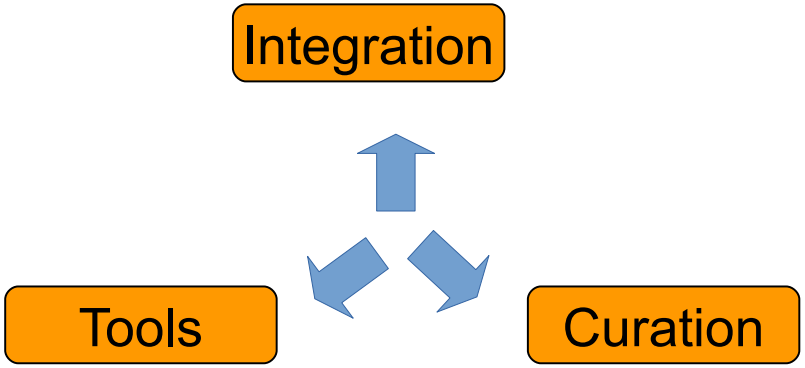
- **Resource Providers**

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

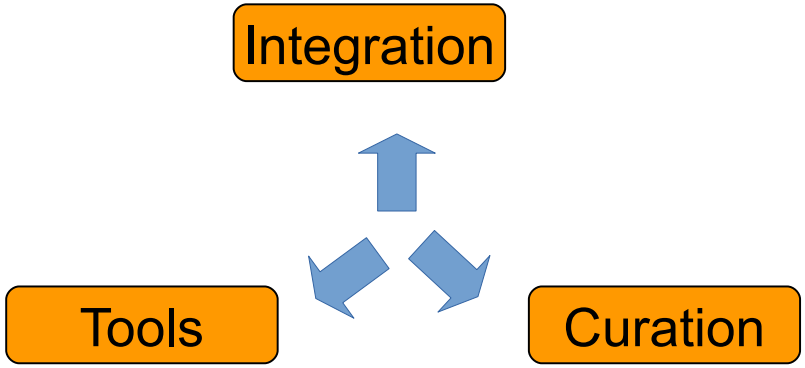
- DNA & RNA
- Gene Expression
- Structures
- Systems
- Ontologies
- Literature
- Proteins
- Chemical biology
- Cross domain



Summary

In this session

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



Subject-specific repositories and collaborative projects

Subject-specific repositories

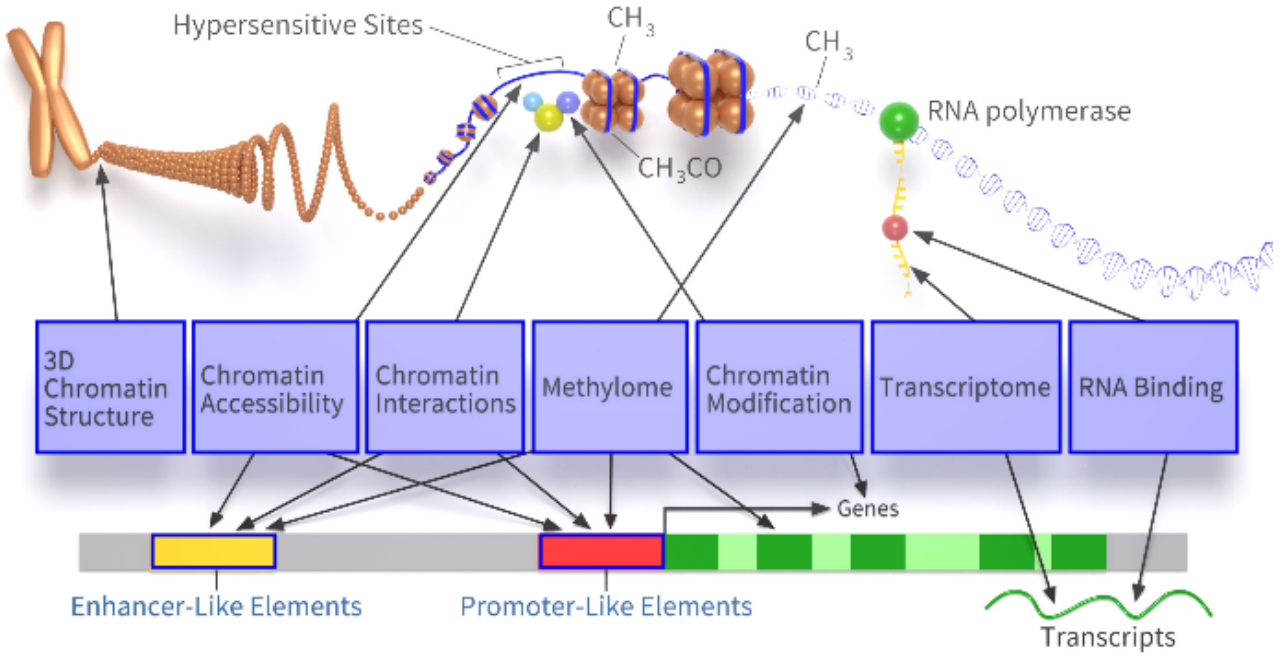
- Collective initiatives: summing efforts from different researchers from different sites of the world.
- **Integrate** different types of data (clinical, multi-omics,...) related to a **specific subject** (or sub-subject)
  - A disease
  - An organism
  - A biological question, tissue, …
- Provide curated, 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.

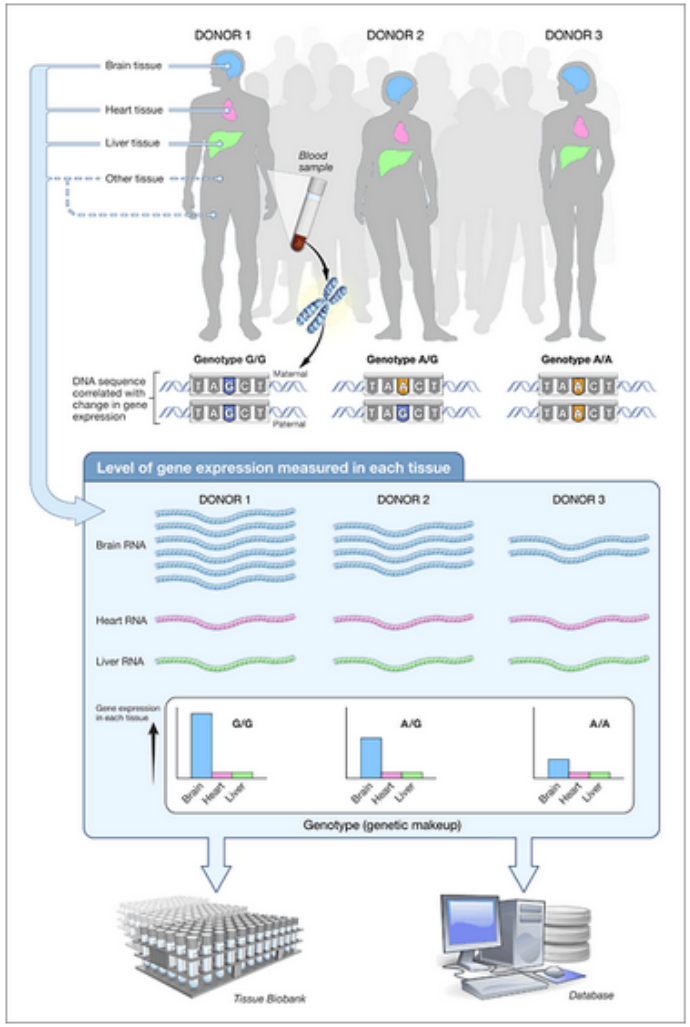


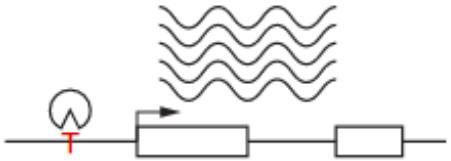
Subject-specific repositories

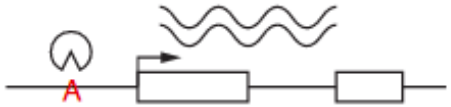
Genotype-Tissue Expression (GTEx) Project

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 53 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.







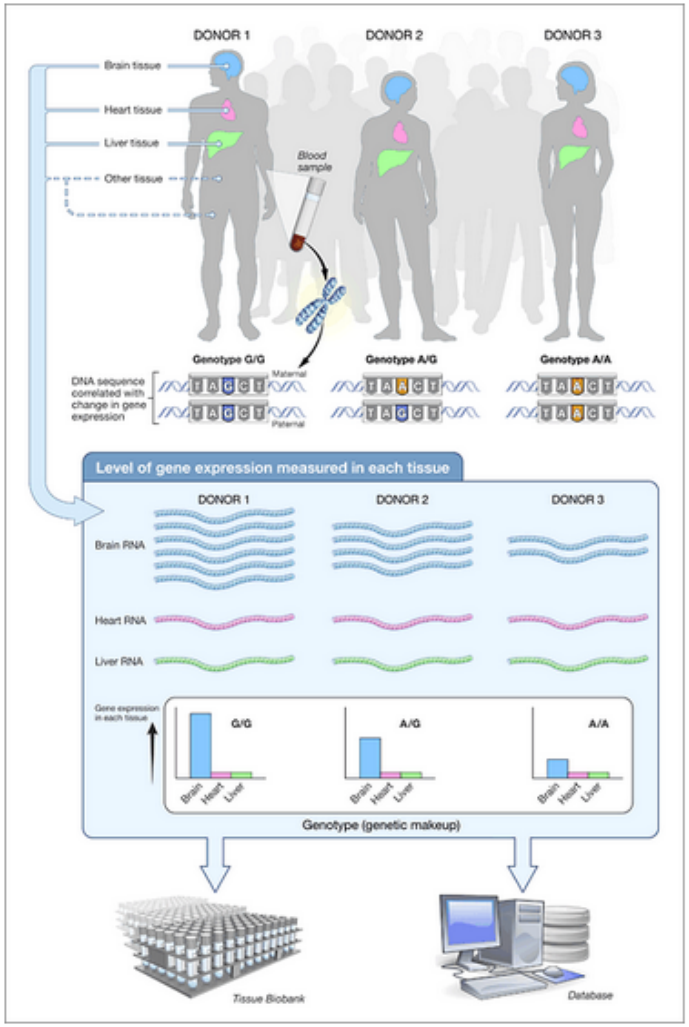
Subject-specific repositories

Genotype-Tissue Expression (GTEx) Project

https://gtexportal.org/home/

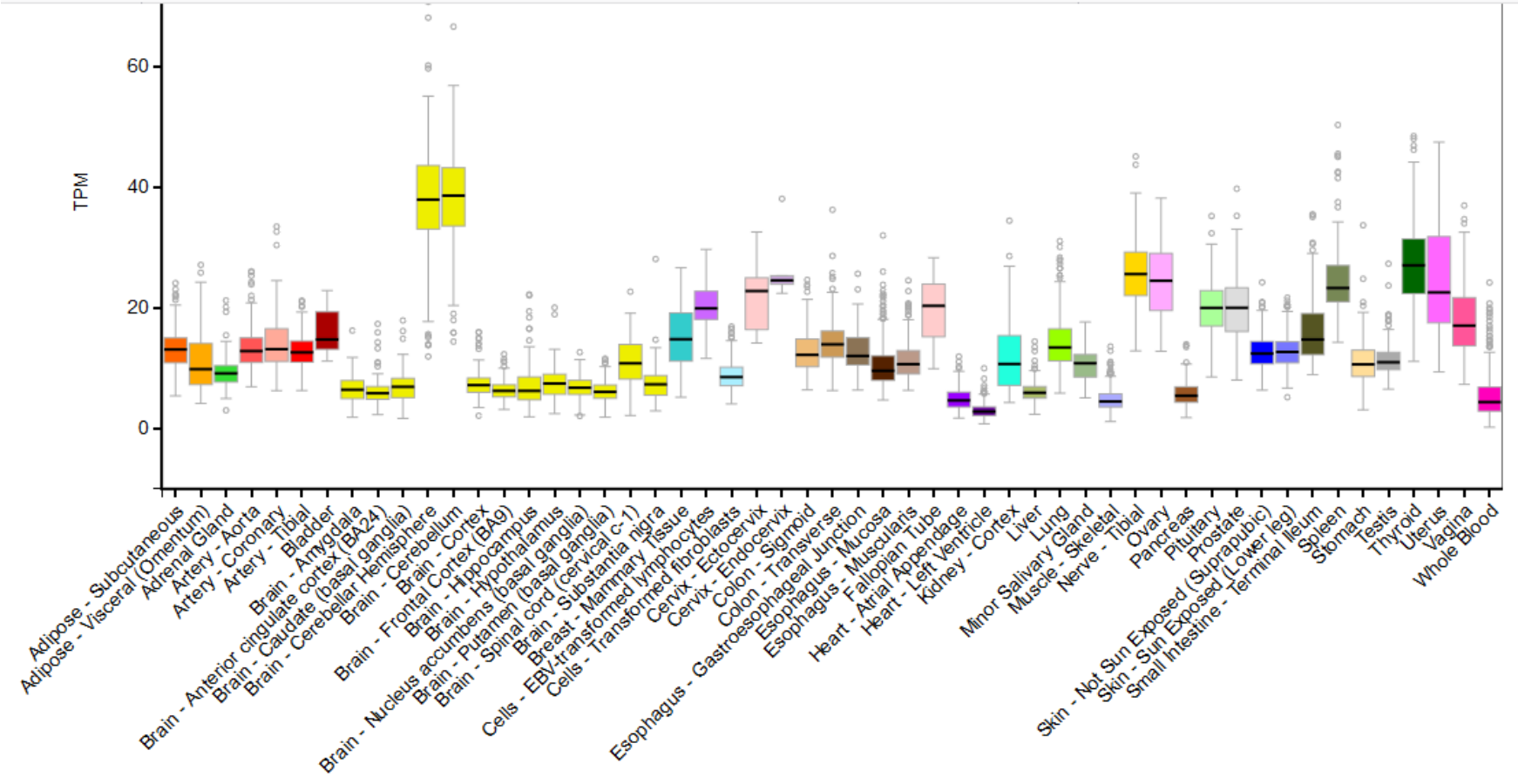
- Types of data provided:
  - Gene/transcript expression in tissues
  - variants associated to gene expression (eQTLs and sQTLs)
  - histology images
  - Patient/sample metadata

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



Genotype-Tissue Expression (GTEx) Project

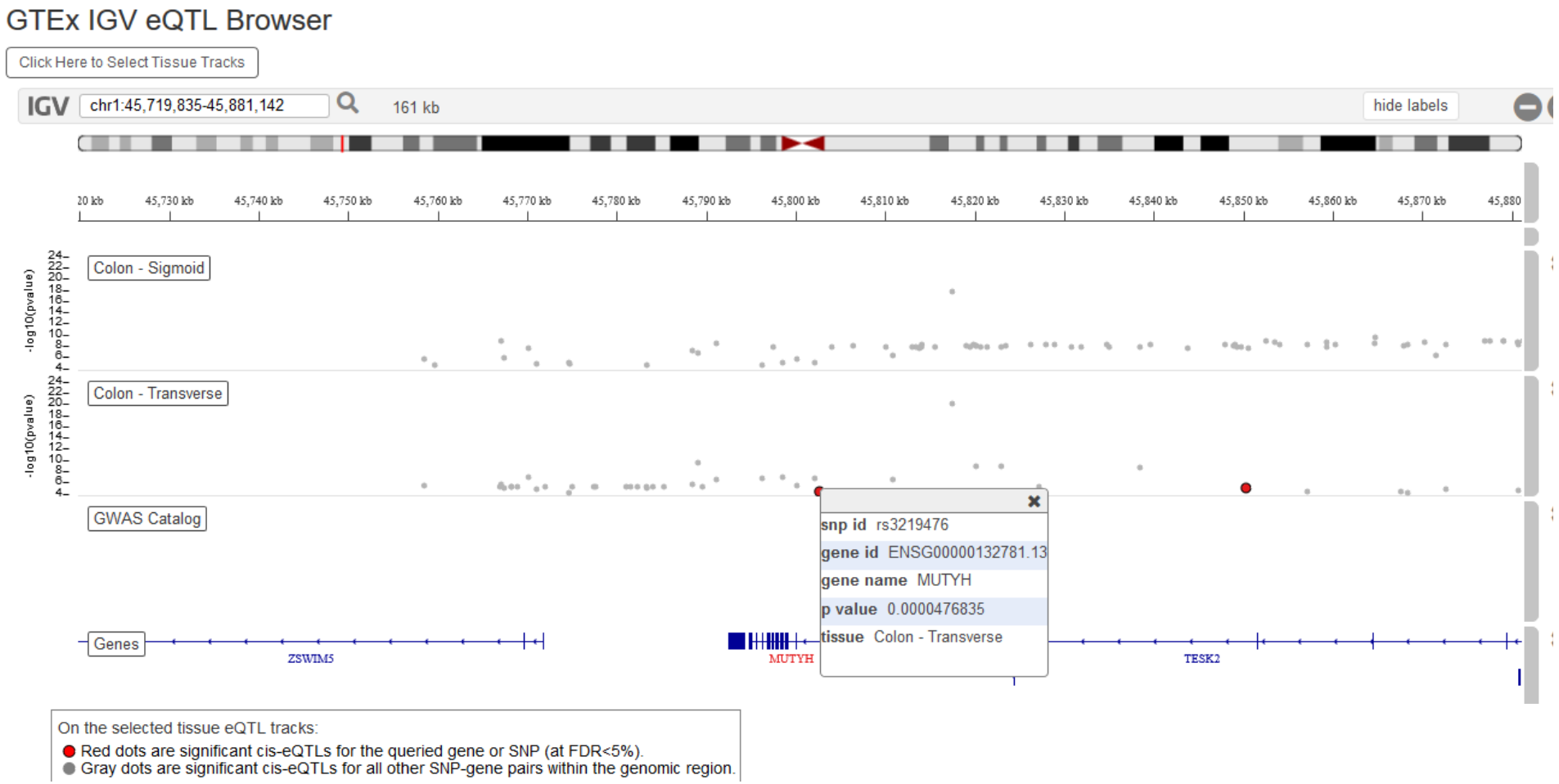
Example: Normal tissue expression profile of *mutyh* gene



Practicum

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/

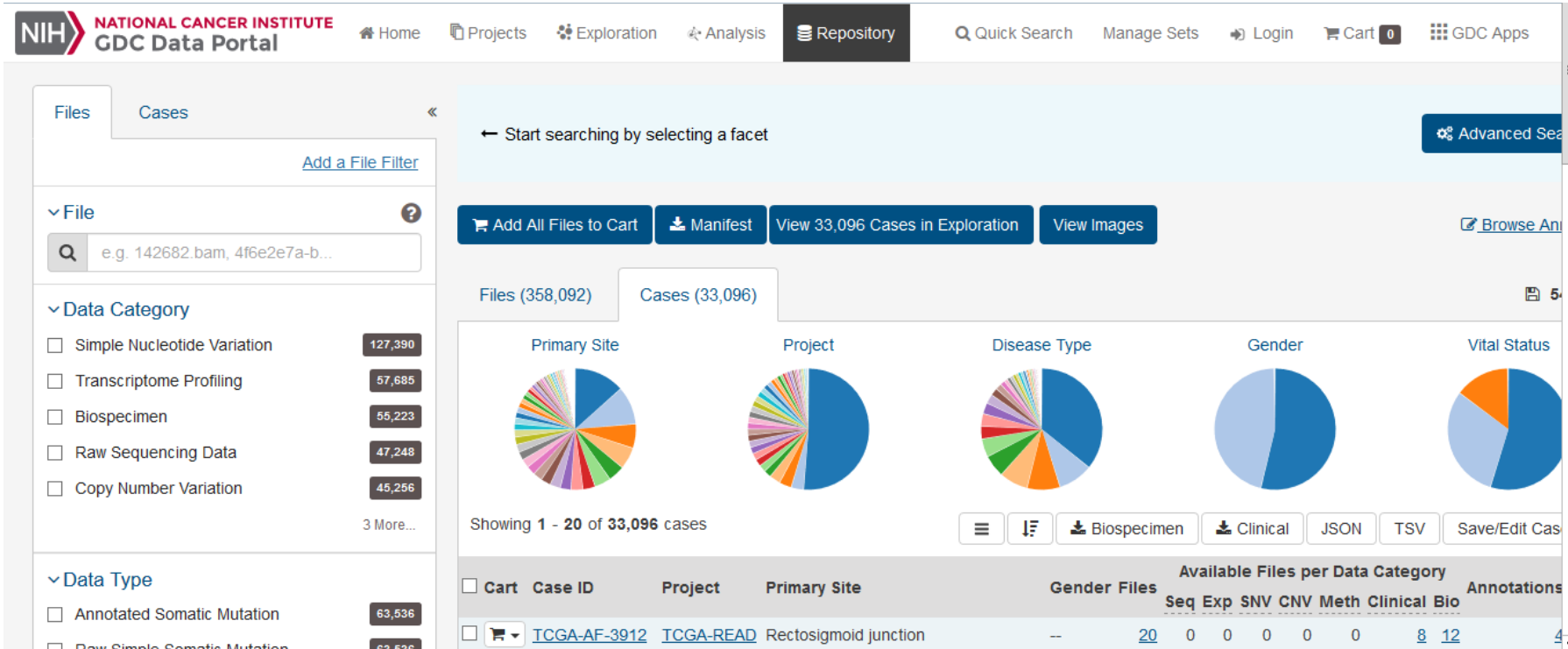
- Collaboration between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) that has generated comprehensive, multi-dimensional maps of the key genomic changes in 33 types of cancer.
- High-quality tumor and matched normal samples from over 11,000 patients collected 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)

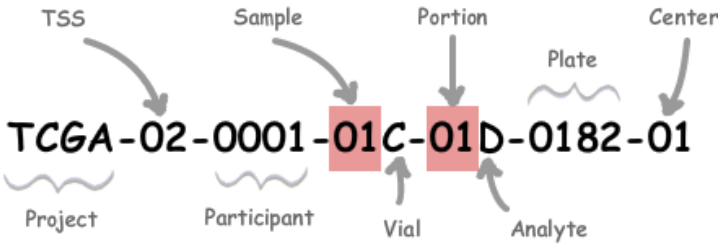
    For documentation on procedures: https://docs.gdc.cancer.gov/
  - Two available sources to download GDC data:
    - GDC Legacy Archive: provides access to an unmodified copy of data that was previously stored in CGHub and in the TCGA Data Portal hosted by the TCGA Data Coordinating Center (DCC), in which 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.

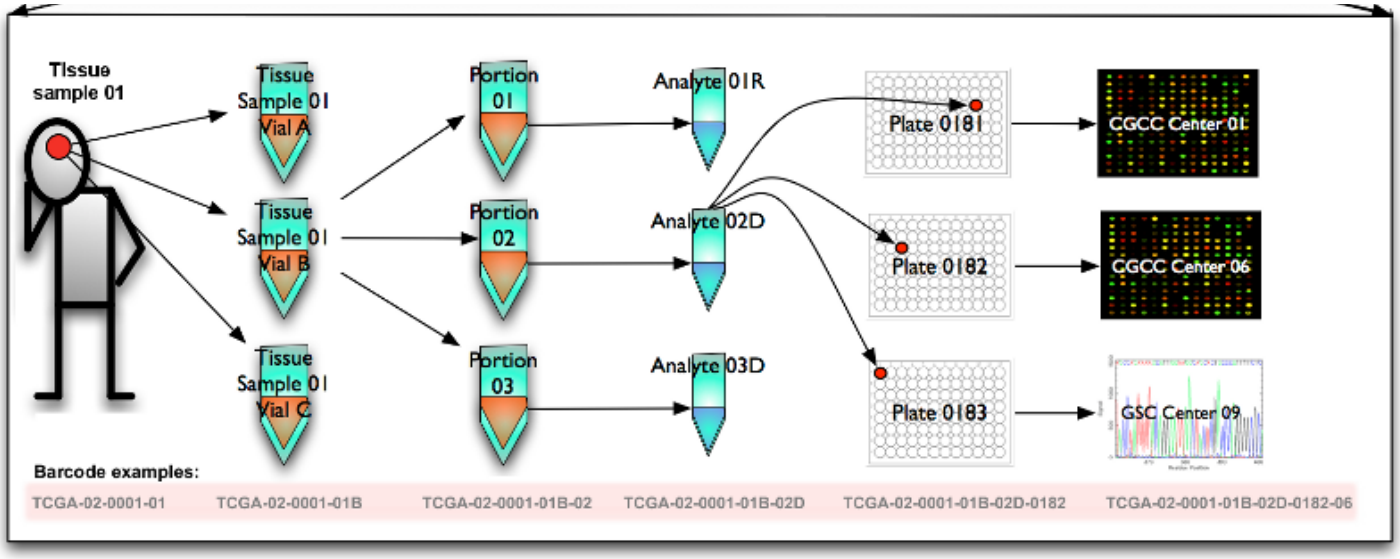
Subject-specific repositories

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)

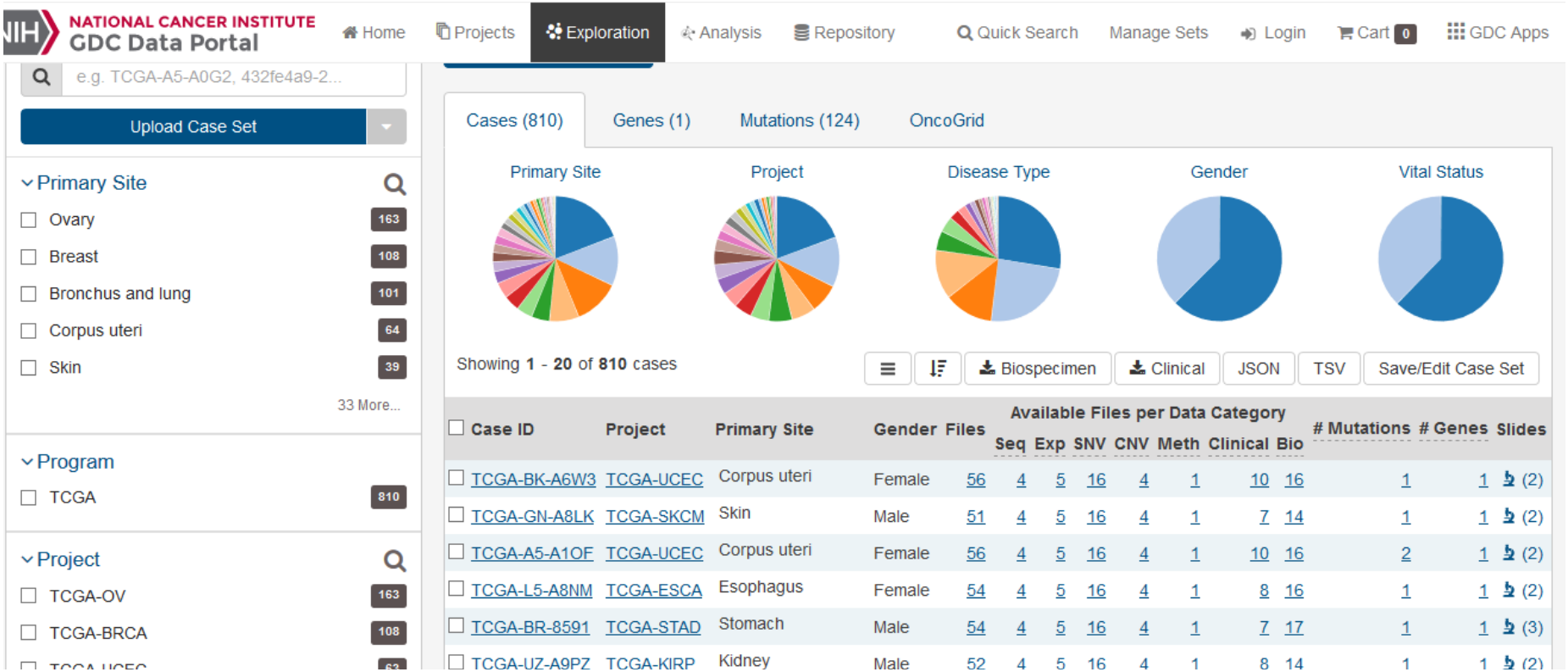




Practicum

The Cancer Genome Atlas (TCGA)

Example: Retrieving known mutations in *mutyh* gene associated to colon cancer



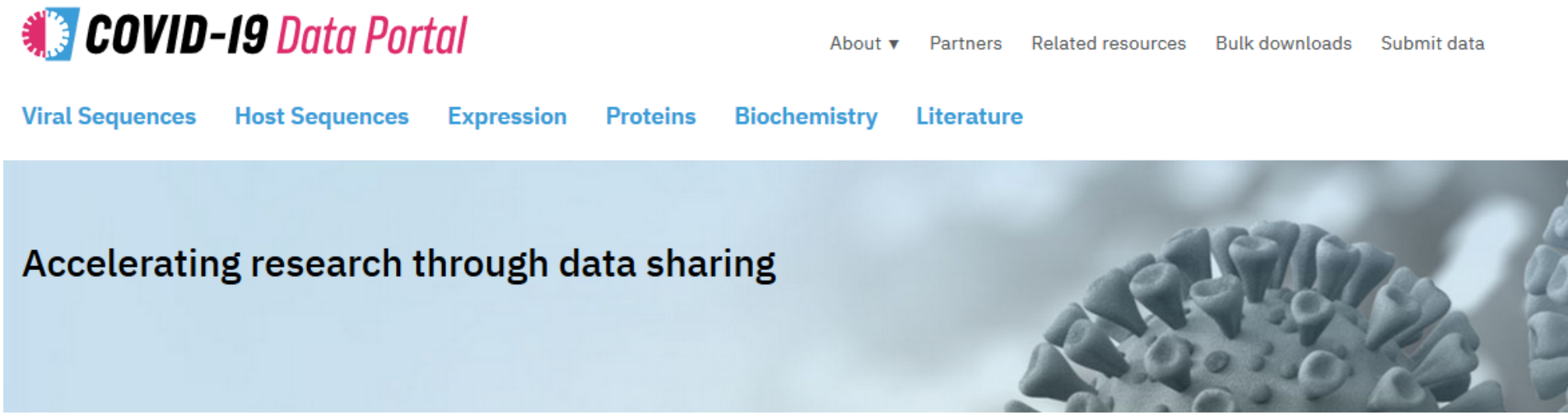
Subject-specific repositories

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

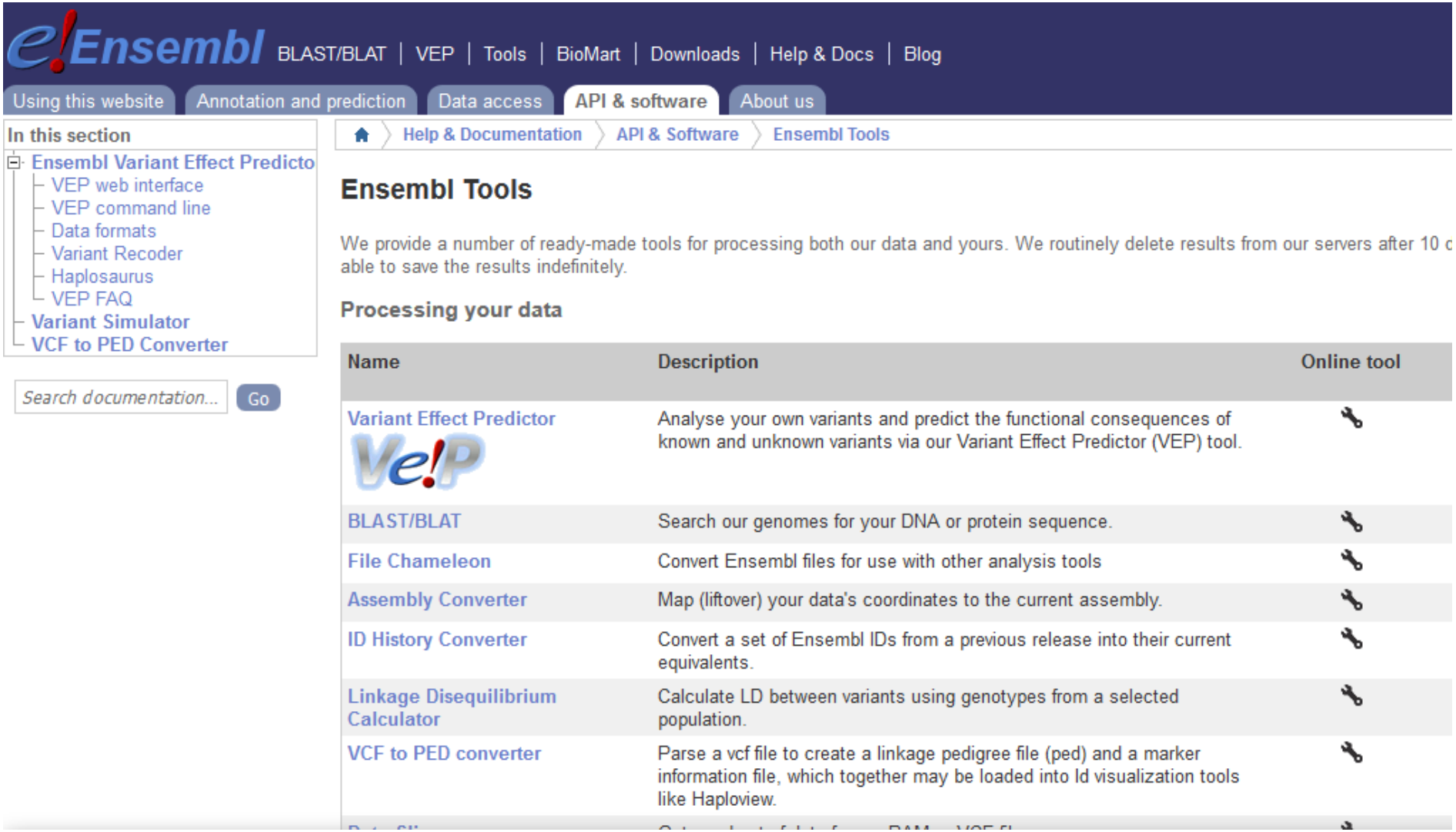
Fast sensitive protein homology searches using profile hidden Markov models (HMMs) for querying against both sequence and HMM target databases.

Sequence similarity search | Protein function analysis

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

Home | Install | Help | Developers | About

Search:

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:

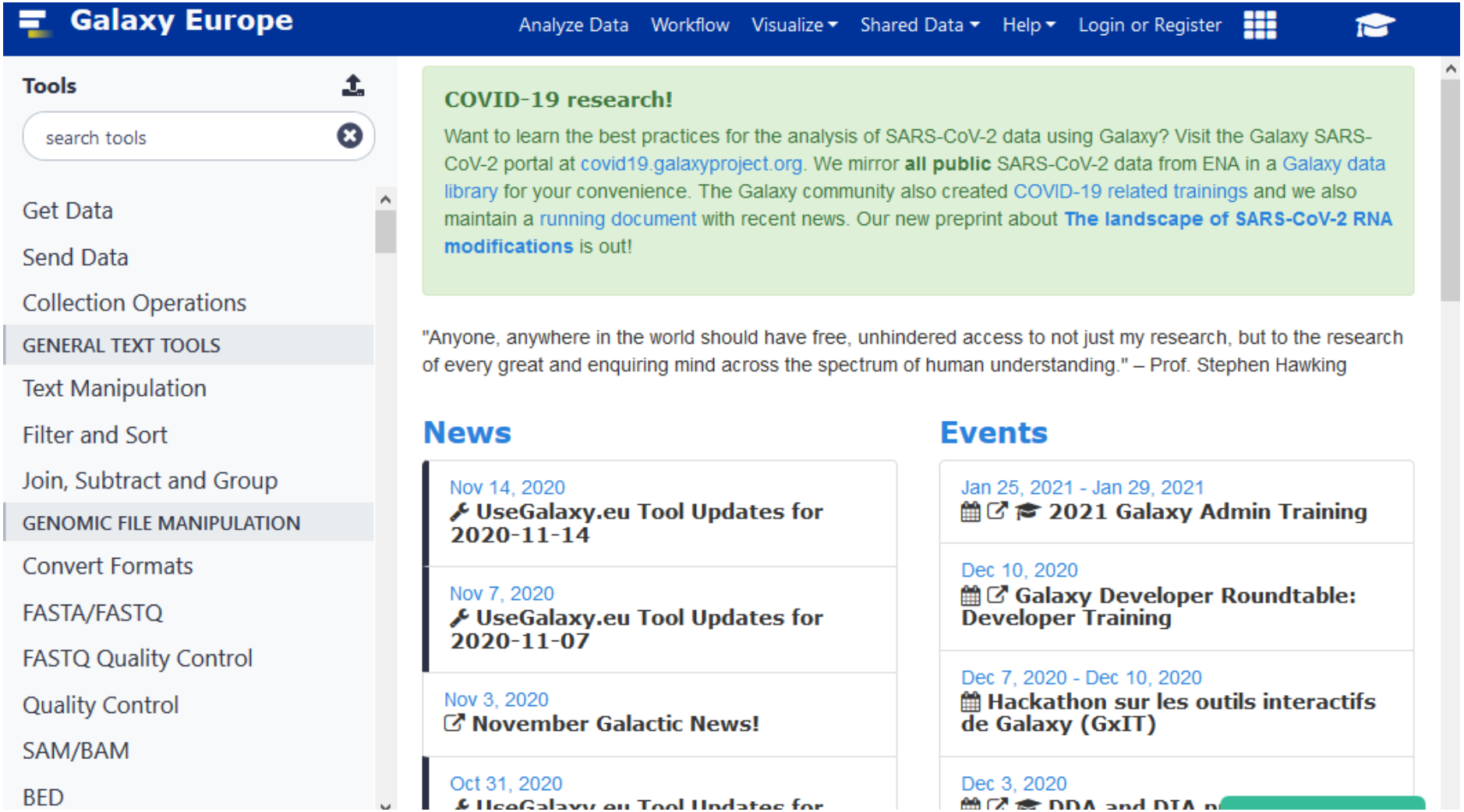
Rank based on number of downloads: lower numbers are more frequently downloaded.

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/

Vall d'Hebron Institut de Recerca VHIR

UNITAT D'ESTADÍSTICA I BIOINFORMÀTICA

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/

Tools

https://www.omicsdi.org/

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.

Your turn!

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*

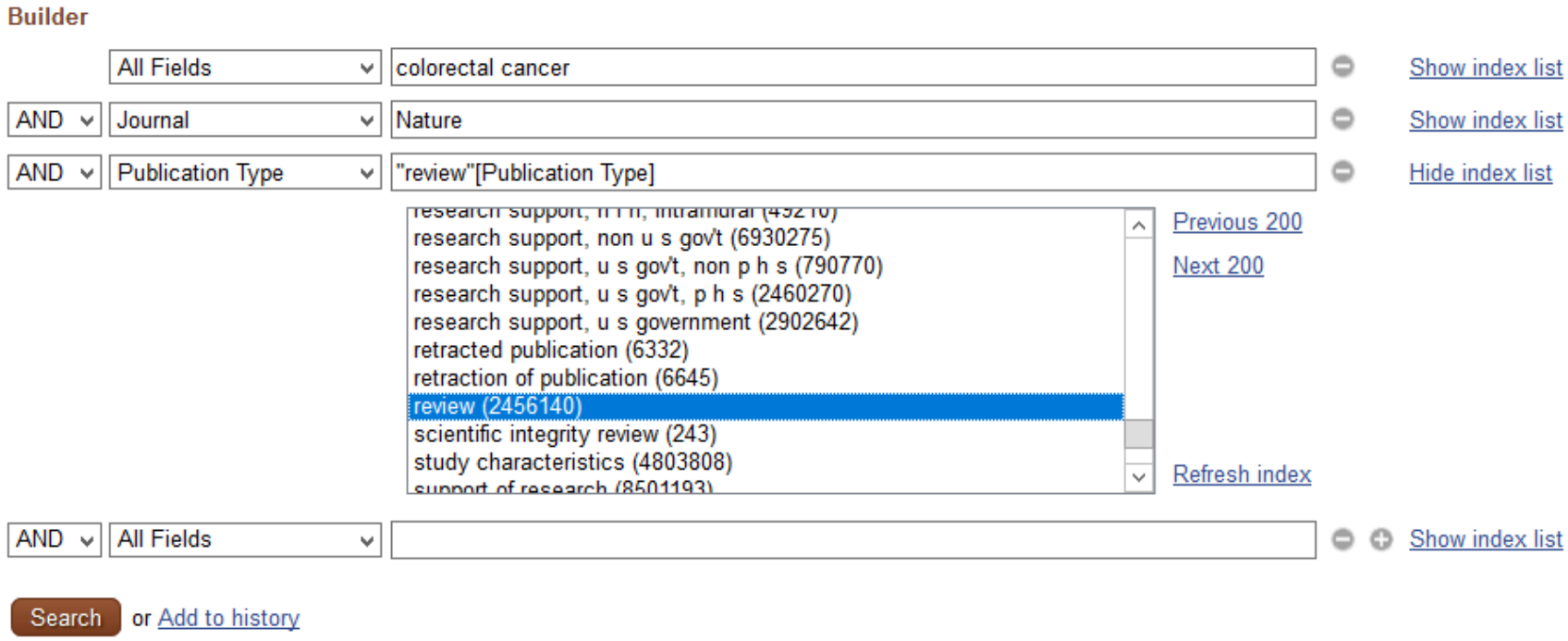
2- In the abstract, the authors mention a gene associated to the disease. Find a *well annotated mRNA* sequence for this gene using the Nucleotide DB

3- Look for MUTYH human protein in UniProtKB

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

Practicum

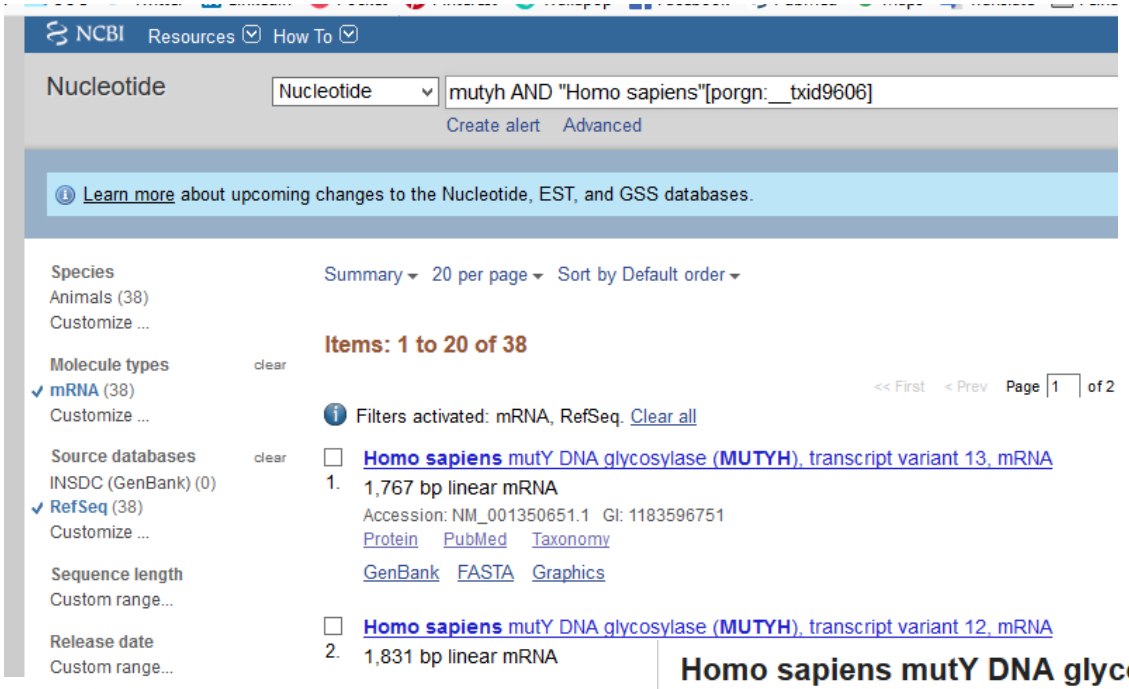
1- Using PubMed Advanced Search, look for a *review* paper published in *Nature* on *colorectal cancer* and authored by *David SS*

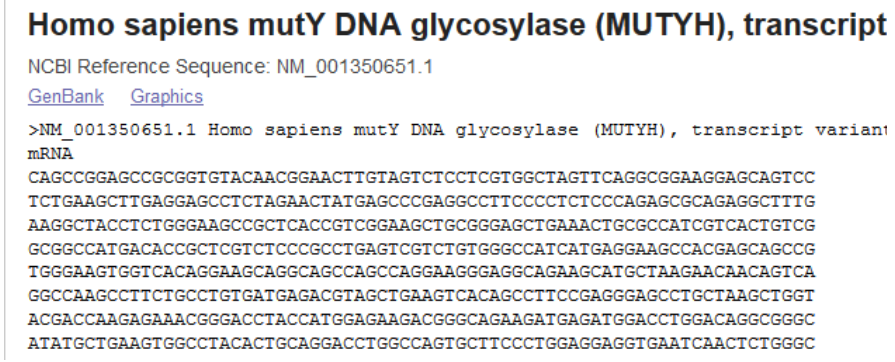


Practicum

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

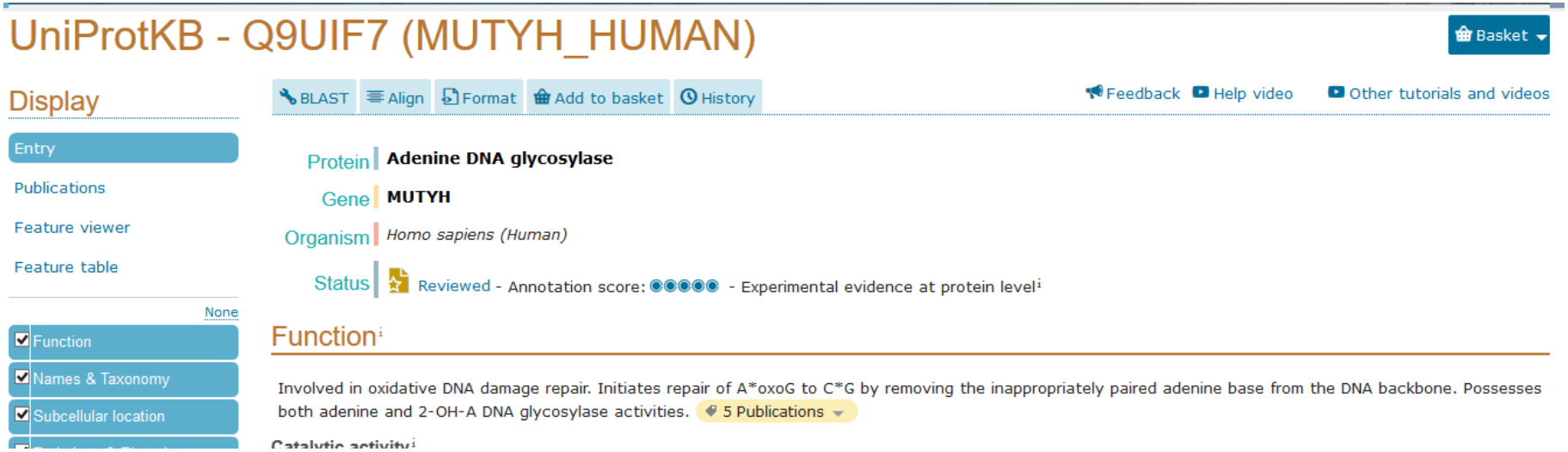




Practicum

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

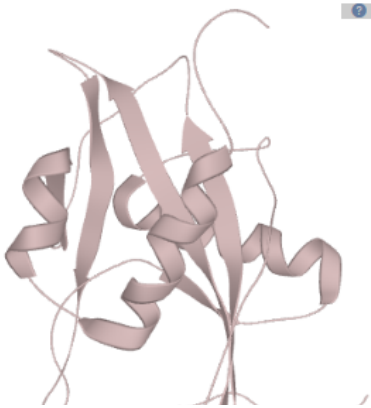
Binary interactions

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



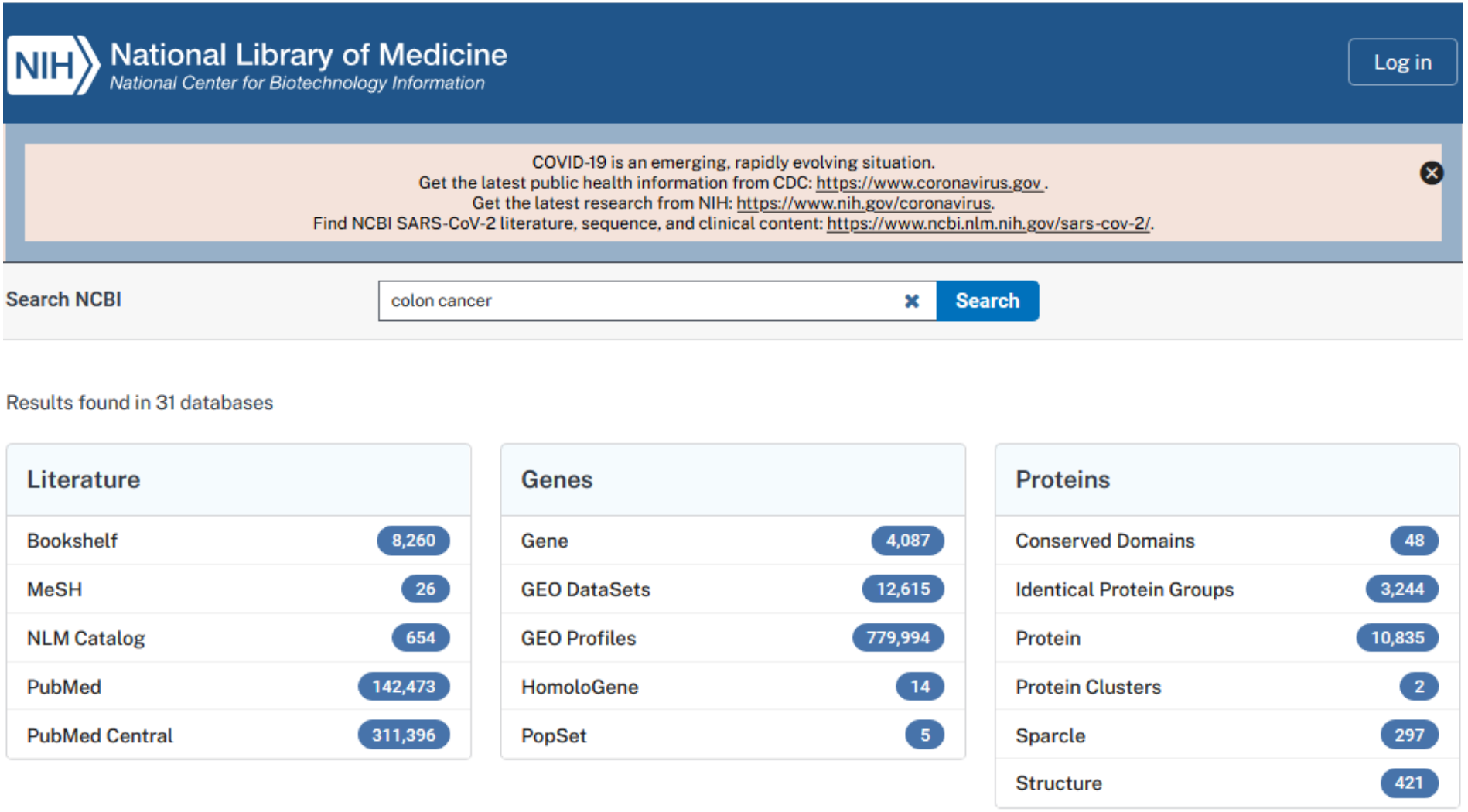
PDB Entry	Method	Resolution	Chain	Positions	Links
1X51	NMR		A	356-497	PDBe RCSB PDB PDBj PDBsum
3N5N	X-ray	2.30 Å	X/Y	76-362	PDBe RCSB PDB PDBj PDBsum

Family and domain databases

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

Practicum

Example of database cross-search with Entrez

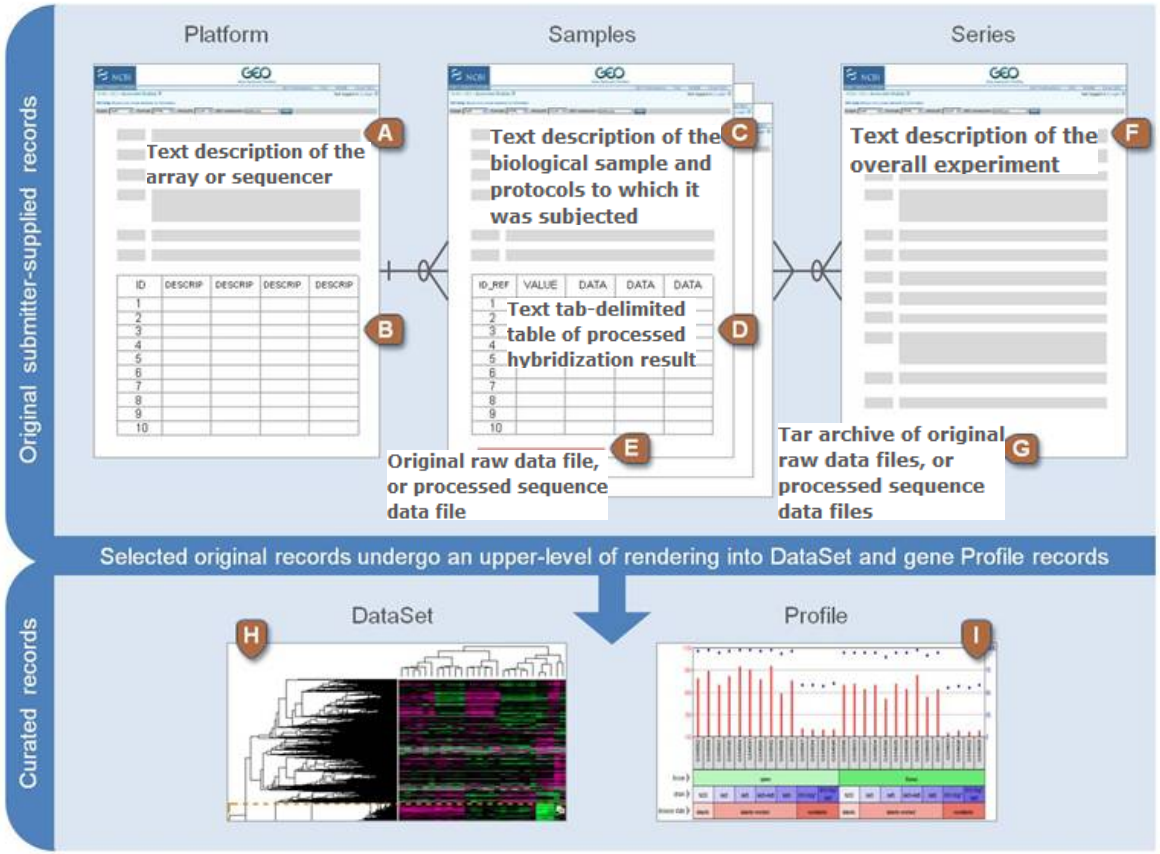


Retrieving data from GEO

- 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



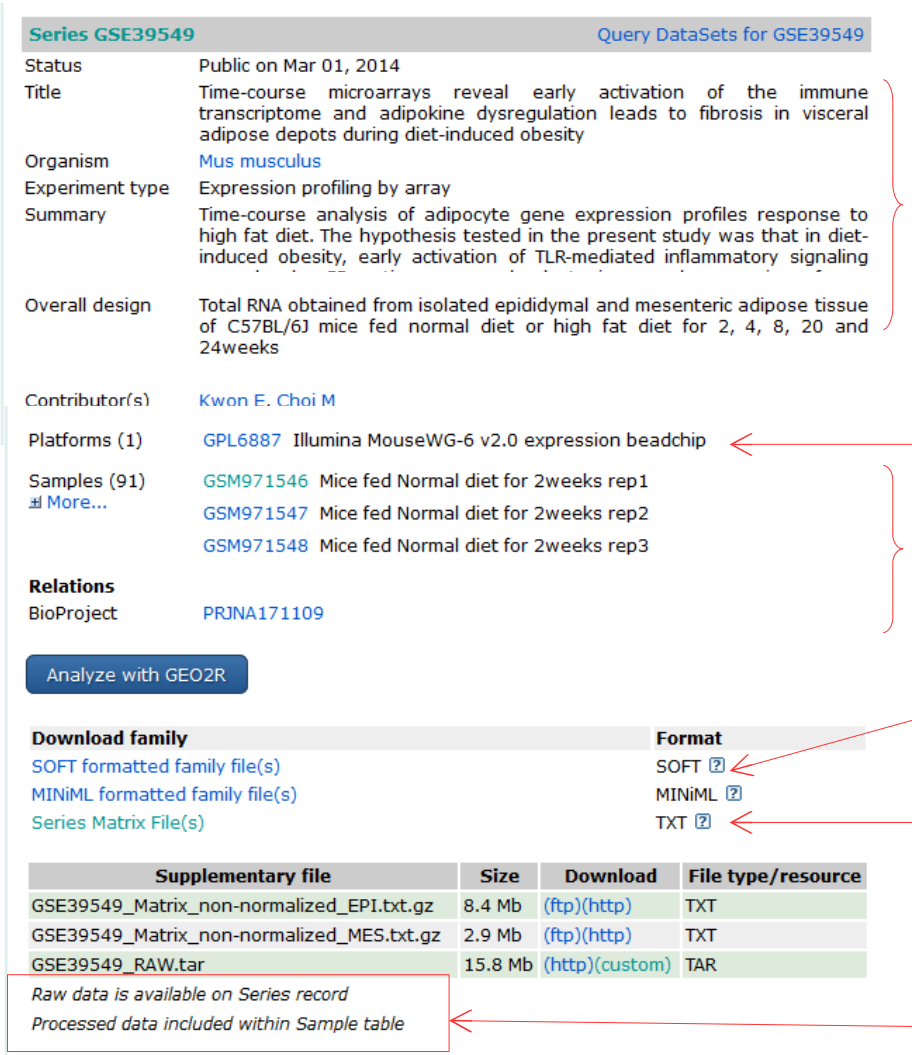
Gene Expression Omnibus (GEO)

- 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
	GSM971547 Mice fed Normal diet for 2weeks rep2
	GSM971548 Mice fed Normal diet for 2weeks rep3
⊞ More...	

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

Practicum

Retrieving data from GEO

Source name	Adipose tissue of mice
Organism	Mus musculus
Characteristics	strain: C57BL/6J treatment protocol: Normal diet time: 2weeks age: 7 weeks 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 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). 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

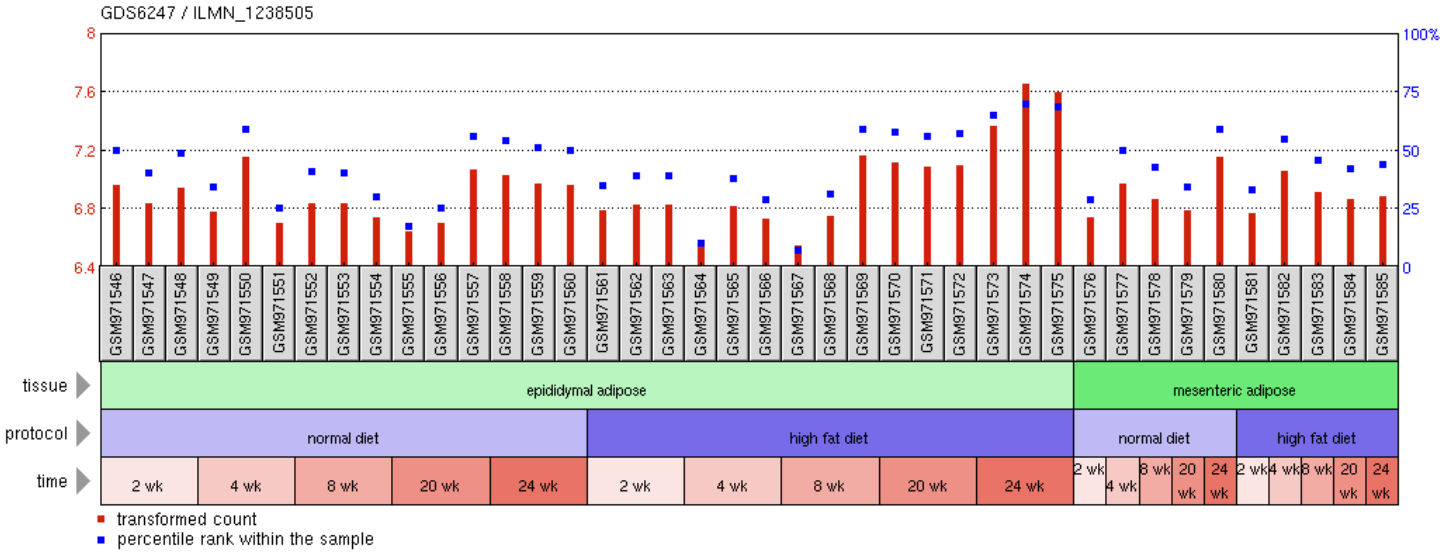
- GEO Profiles: stores individual gene expression profiles from curated DataSets.

Example: search for expression profile of IL-10

Builder

	All Fields	il-10	Show index list
AND	GEO Accession	GSE39549	Show index list
AND	All Fields		Show index list

Profile GDS6247 / ILMN_1238505
Title Diet-induced obesity model: white adipose tissue
Organism Mus musculus



Practicum

Retrieving data from GEO in *R*!

- Getting data from GEO is quite easy using the GEOquery package from Bioconductor
- There is only one command that is needed:

```
getGEO("GEO Accession")
```

- It directs the download and parsing of a GEO SOFT file into an R data structure specifically designed to make access to each of the important parts of the GEO SOFT format easily accessible.

Let's try: *Practicum_geoqueries.Rmd*

Practicum

Downloading data from TCGA using *R*

In this example we will use the *TCGAbiolinks* package from Bioconductor to download data from Colon Adenocarcinoma tumor (TCGA-COAD project). We are interested in the RNA-seq gene expression data from the GDC harmonized database (aligned to hg38).

1. Get a summary of the data available from the project using `TCGAbiolinks:::getProjectSummary(project)` function, where *project* will be the id of your project of interest

   a) How many cases are within this project? How many files are available within this project?

   b) Which data categories are available within this project?

   c) See the data types available for each category using `getSampleFilesSummary(project)` (column names)

Downloading data from TCGA using *R*

2. Make a query using `GDCquery(`*`arguments`*`)` function. The *arguments* will be the filters as used in the TCGA portal:
   - Type `?GDCquery` to see the list of arguments accepted
   - Try it! Remember: We are interested in the *gene expression* data from the GDC *harmonized database* (legacy=FALSE).
3. Get the results table from the query using `getResults(`*`query`*`)` function, where *`query`* will be the query performed above.
   a) How many files (rows) did you get?
   b) Can you identify any columns of interest that could be used to refine the search?

Practicum

Downloading data from TCGA using *R*

4. Make a more specific query by specifying the barcodes of the first 5 samples and the *workflow.type* as *"HTSeq – Counts"*.

5. Download the files from the query using `GDCdownload(query)` function, where `query` will be the query performed above.

Downloading data from TCGA using *R*

6. Prepare a SummarizedExperiment (SE) object from the query using `GDCprepare(query)` function, where *query* will be the query performed above.

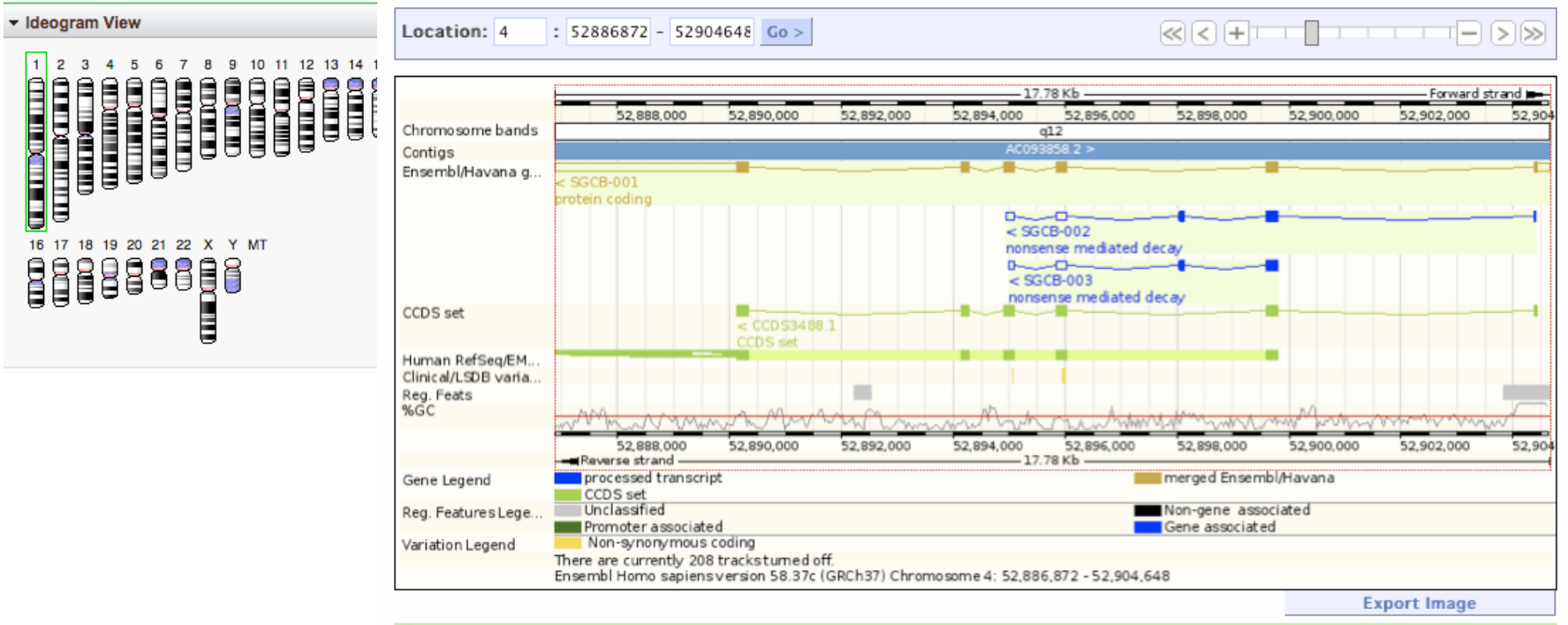
7. Examine the data. A SummarizedExperiment object has three main matrices that can be accessed using the *SummarizedExperiment* package:

   - Sample matrix: stores sample information (clinical data, ...). Can be accessed using `colData(SEdata)` function, where *SEdata* is the SE object created above.
   - Assay matrix: stores molecular data (eg. expression data). Can be accessed using `assay(SEdata)` function, where *SEdata* is the SE object created above.
   - Feature matrix: stores information about the features (gene information). Can be accessed using `rowRanges(SEdata)` function, where *SEdata* is the SE object created above.

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

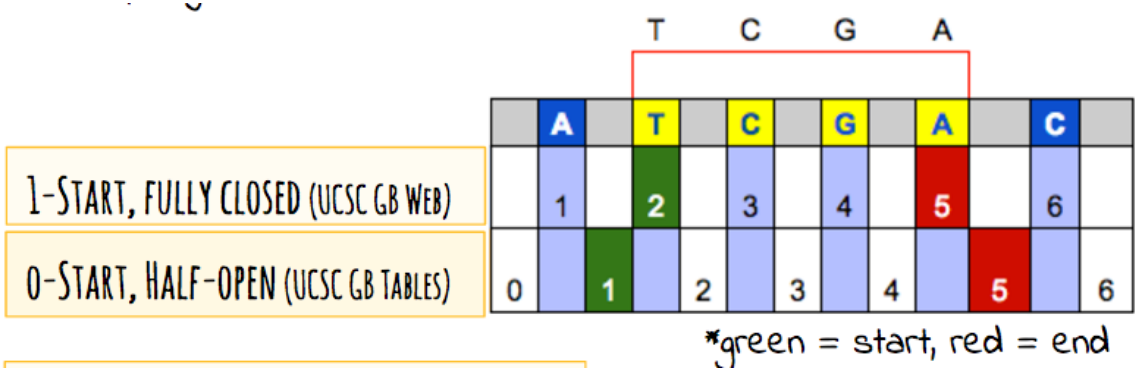
For mouse:

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
- 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 100 (April 2020), over 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

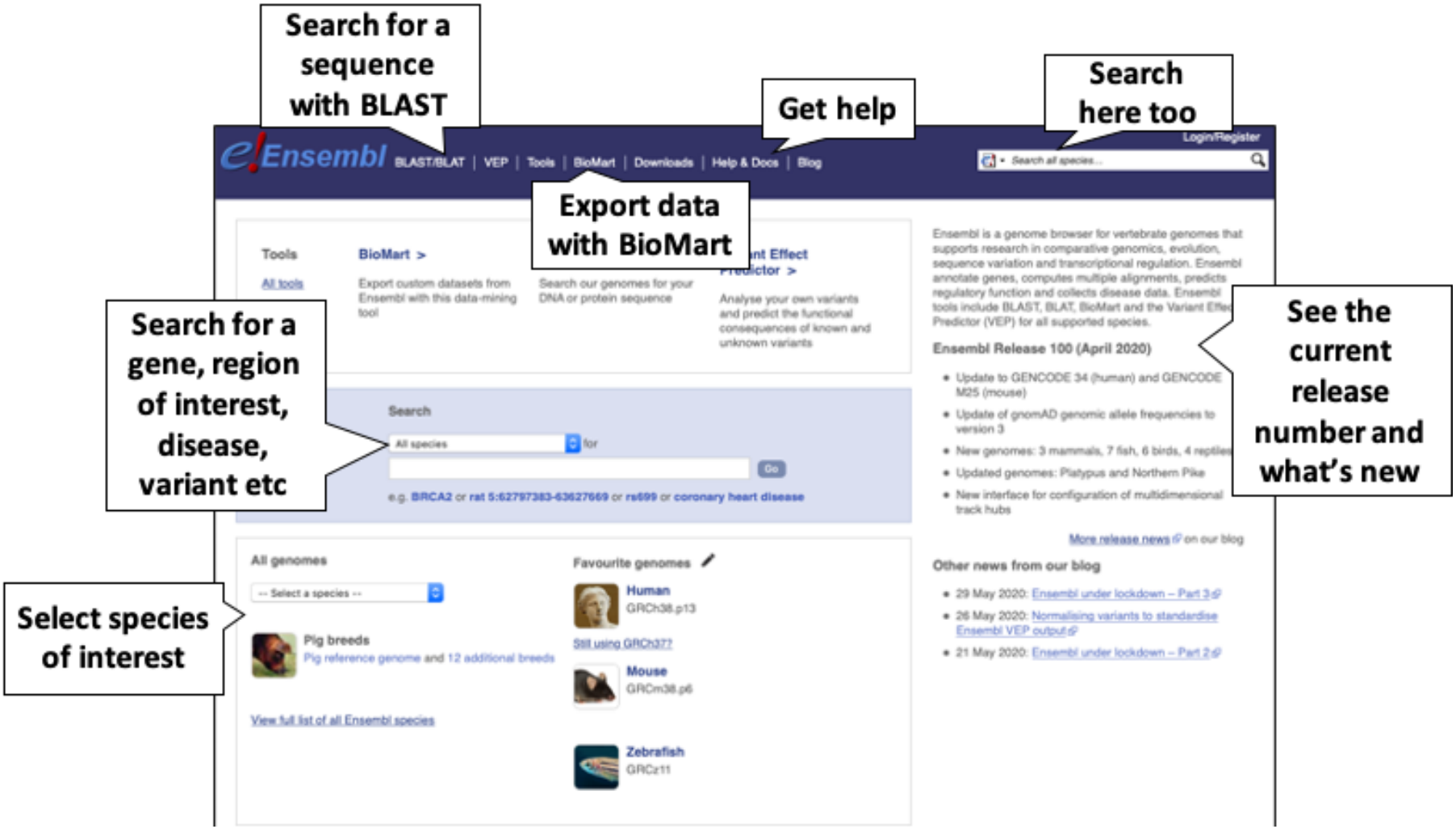
https://www.ensembl.org/index.html

- 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.

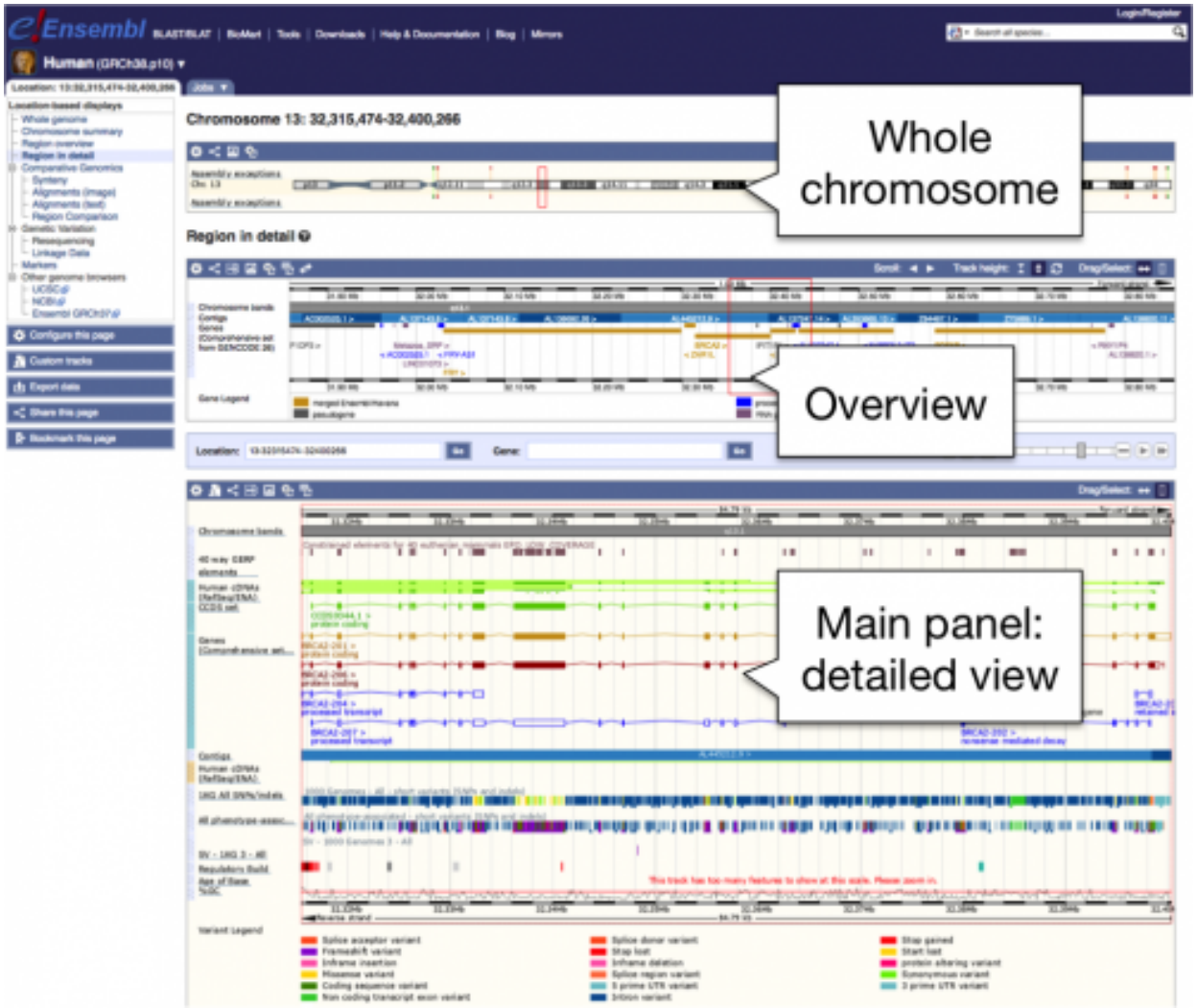
- The available annotations (tracks) to be displayed are configured through the 'Configure this page' button

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

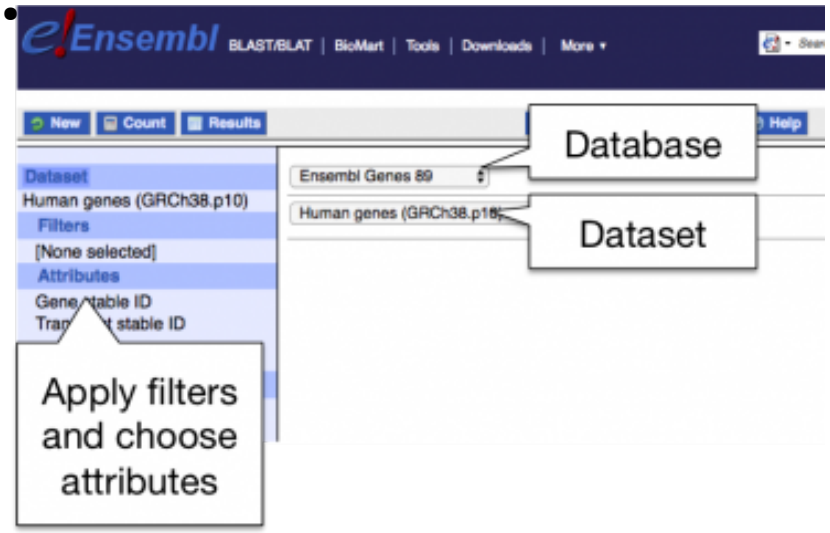


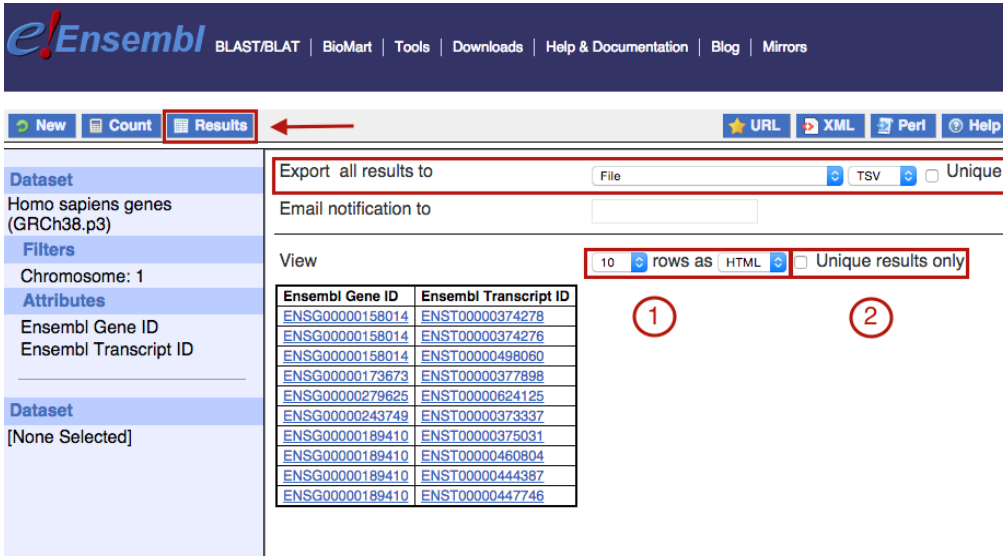
Genome Browsers

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

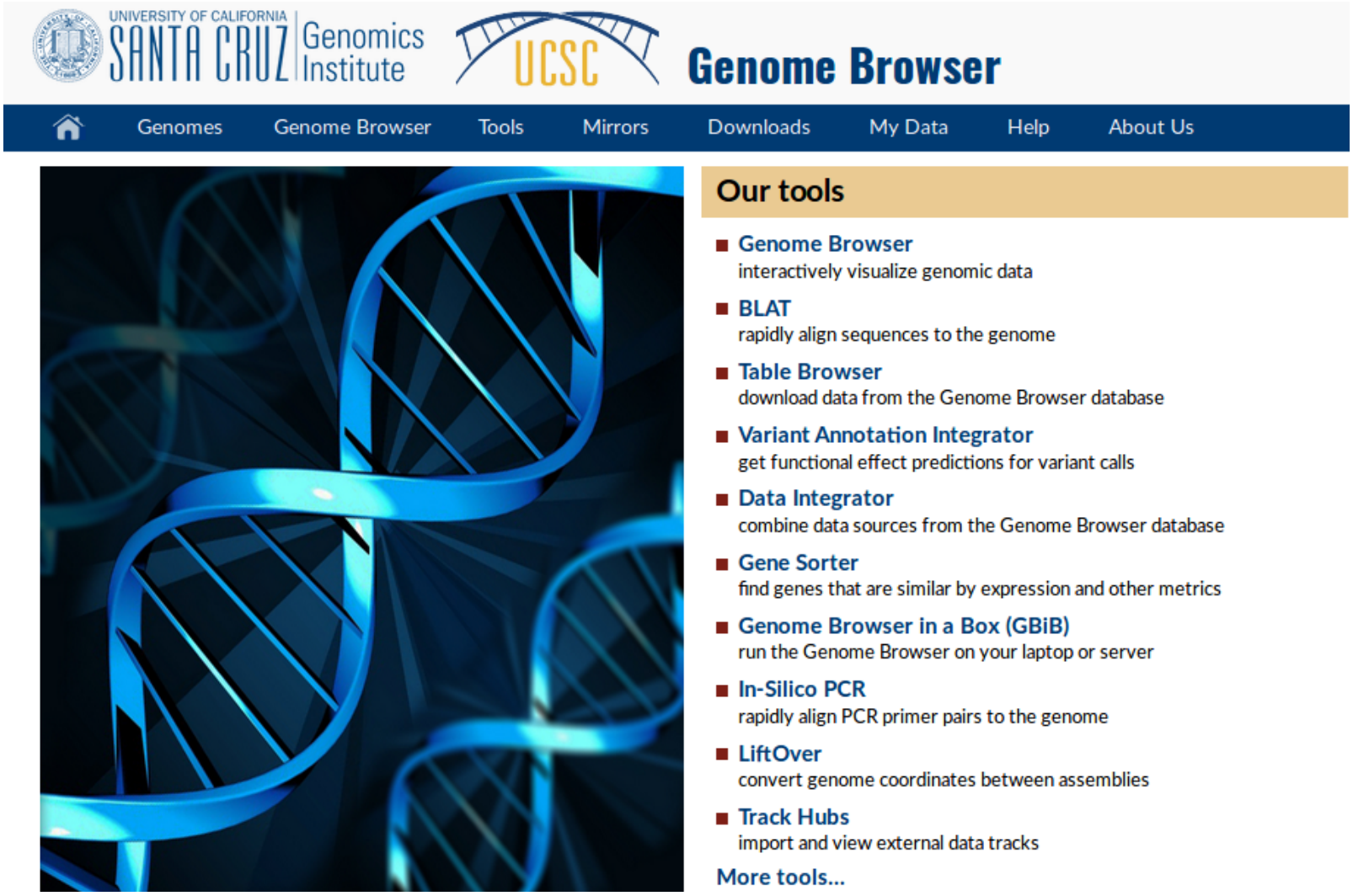
https://genome.ucsc.edu/

- 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/

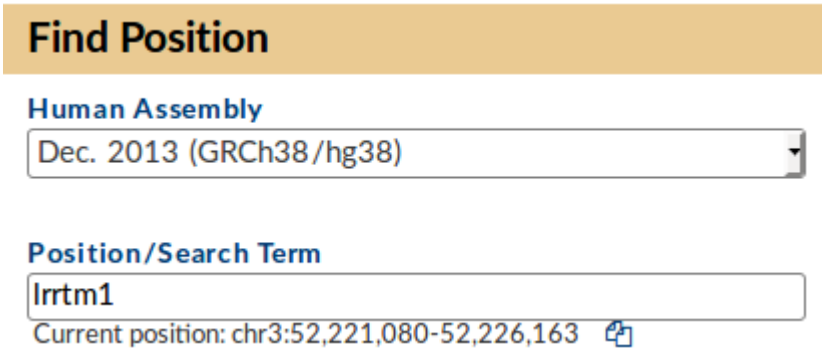


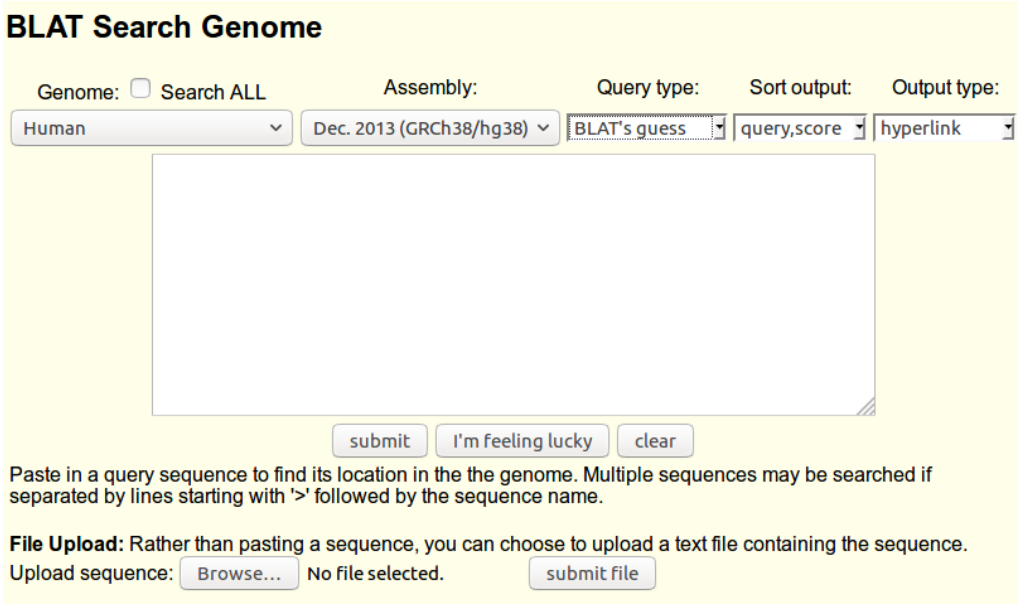
Genome Browsers

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)



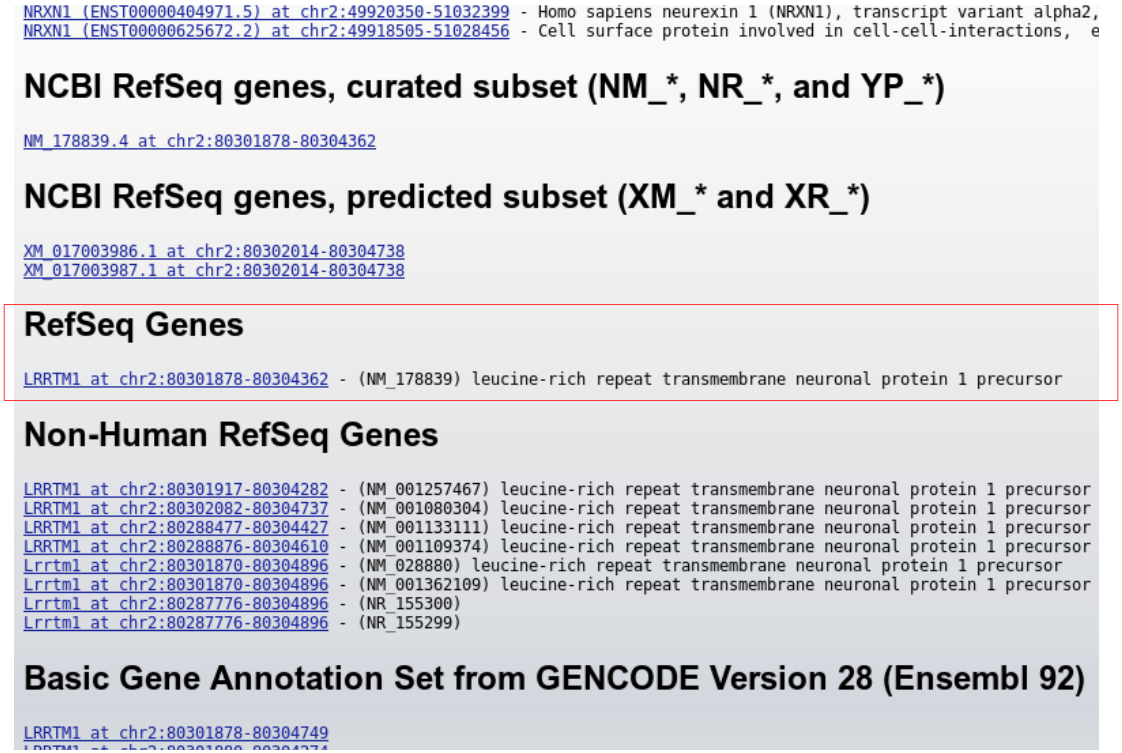


Genome Browsers

UCSC Genome Browser

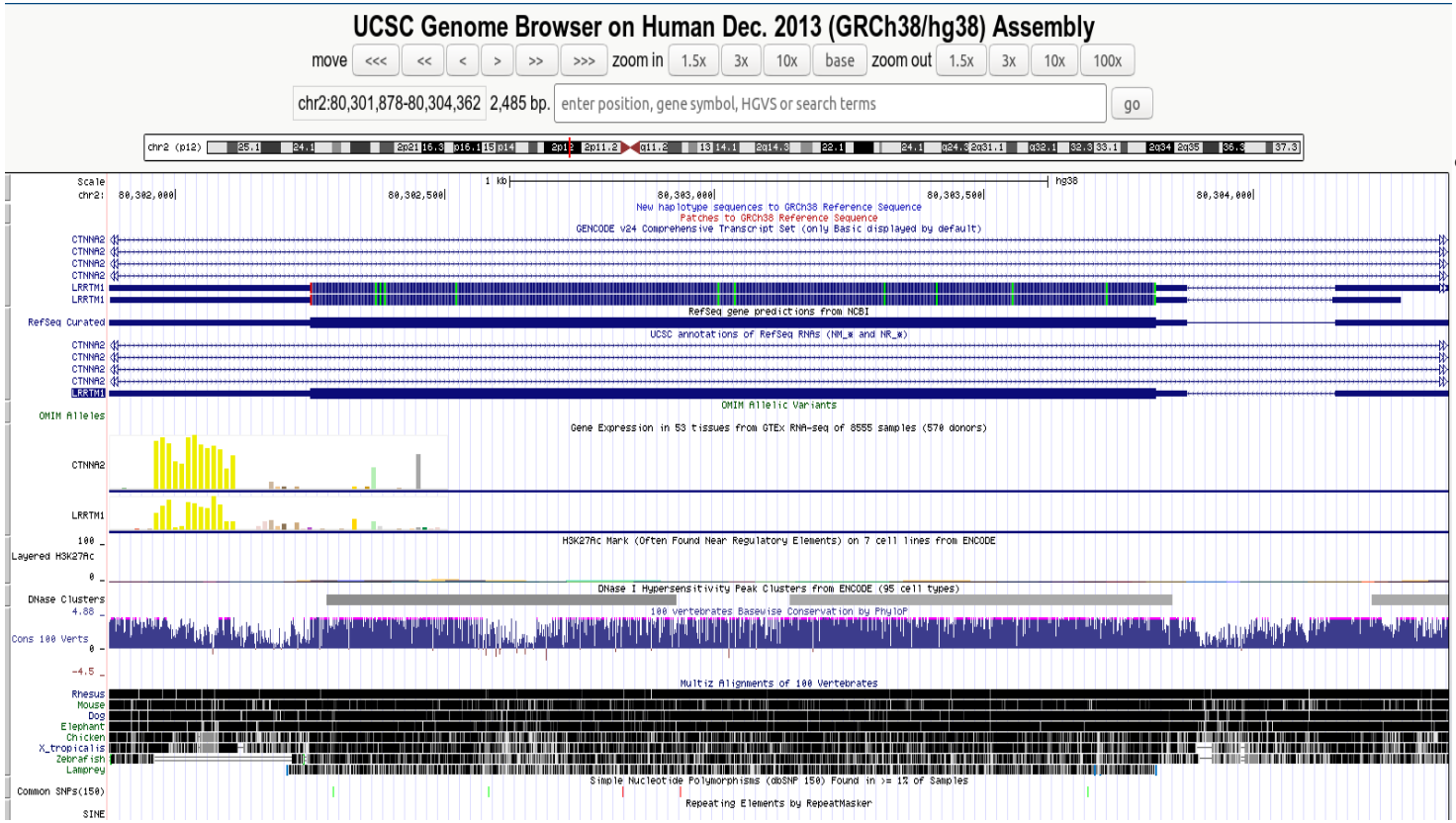
https://genome.ucsc.edu/

- 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



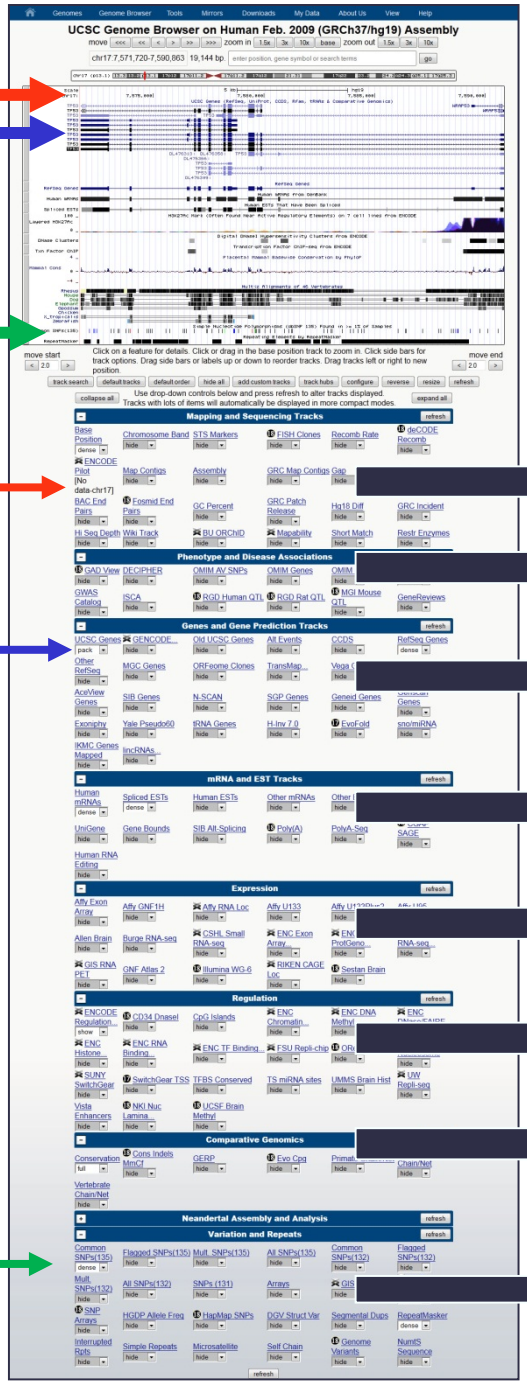
Genome Browsers

- Visualizing results: Genome Viewer and Tracks settings



Tracks info and options

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



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

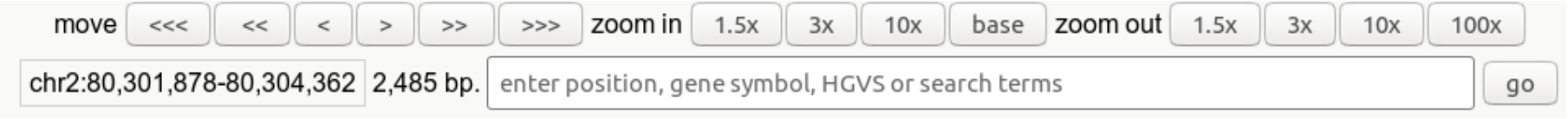
Individual species

Variation and Repeats

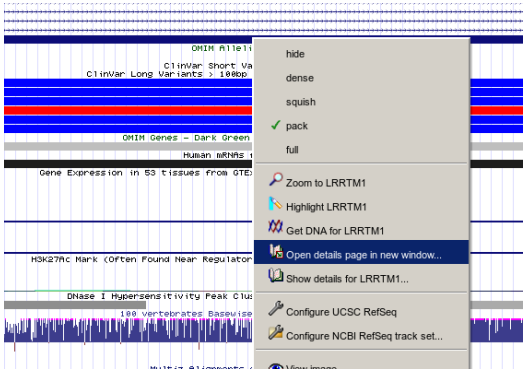
(including SNPs, copy number variation)

Genome Browsers

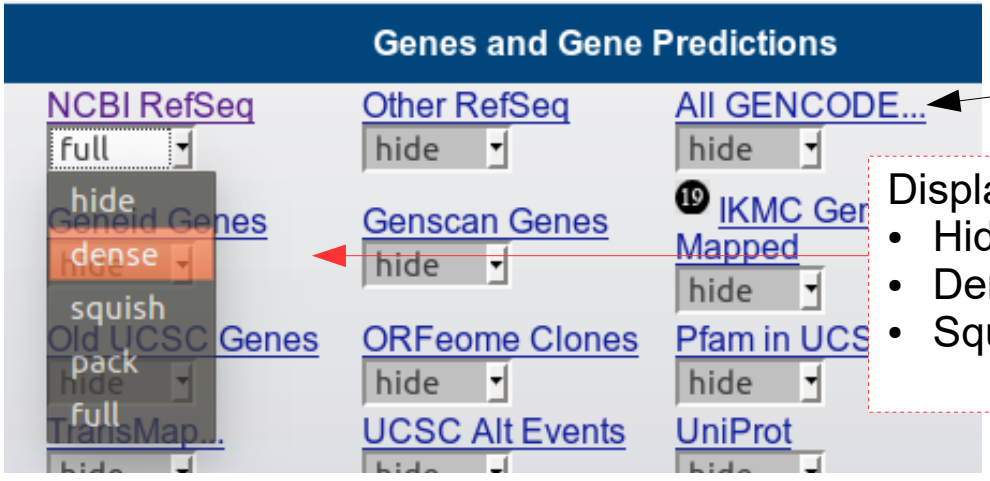
- Change your view or location with controls at the top



- 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



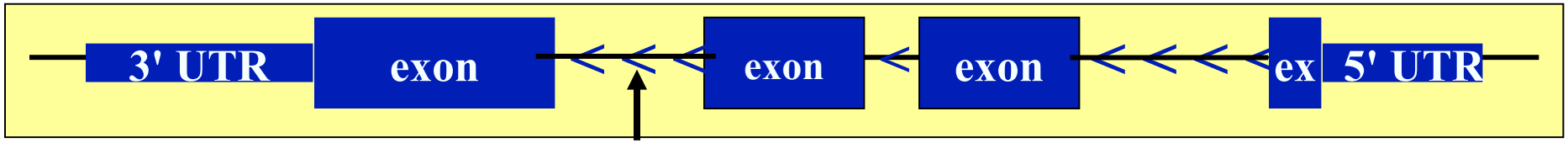
Source of information

Refresh!

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

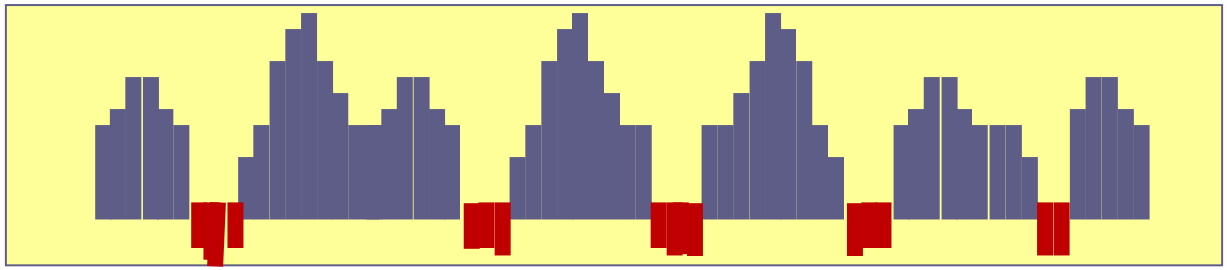
Genome Browsers

- Some visual clues:



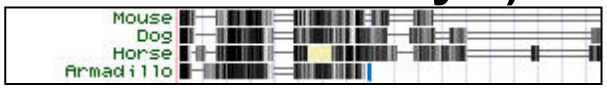
Intron and direction of transcription <<< or >>>

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

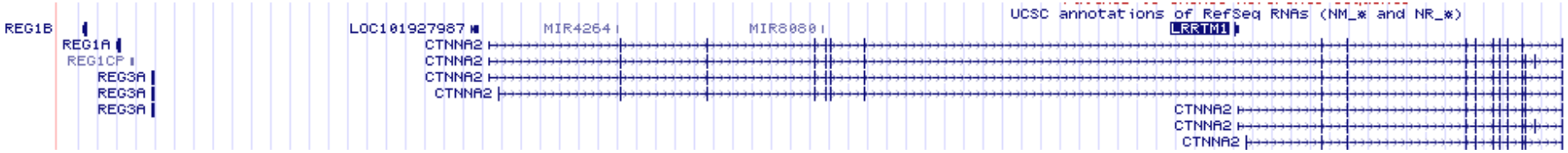
1. What is the genomic localization of human *lrrtm1* gene?

   - 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



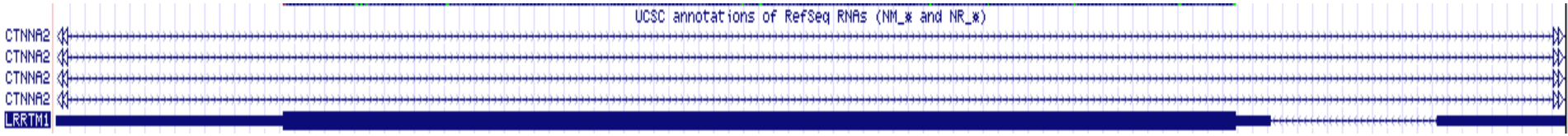
3. How many exons has the gene? 2

   - 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

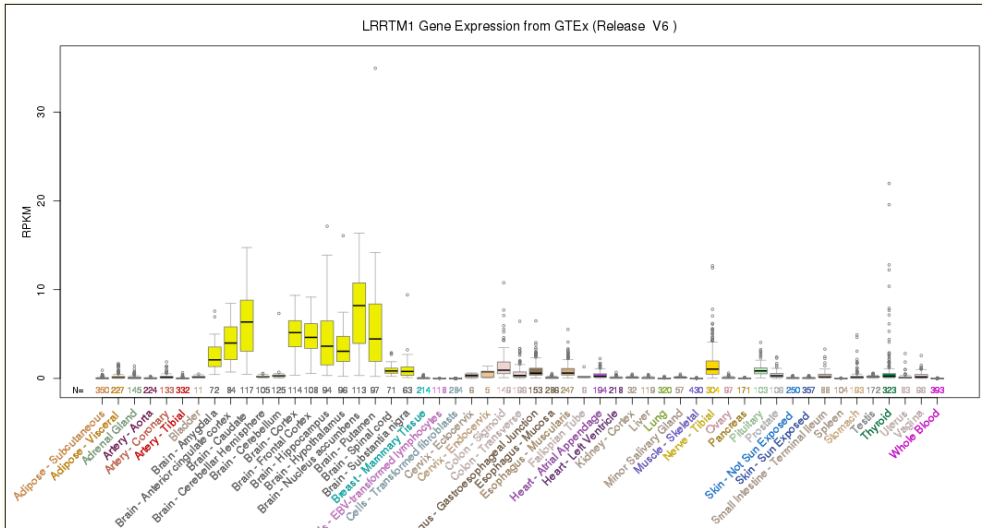


5. Can you find SNPs in this gene? yes

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

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

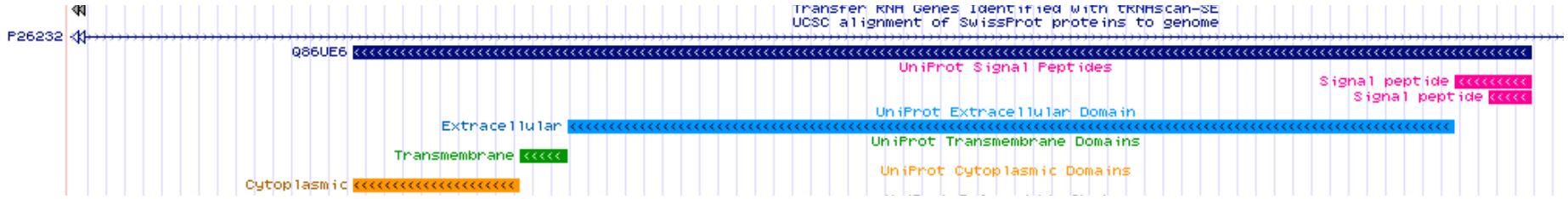
   - Info in track “Expression” > “GTEx”



Practicum

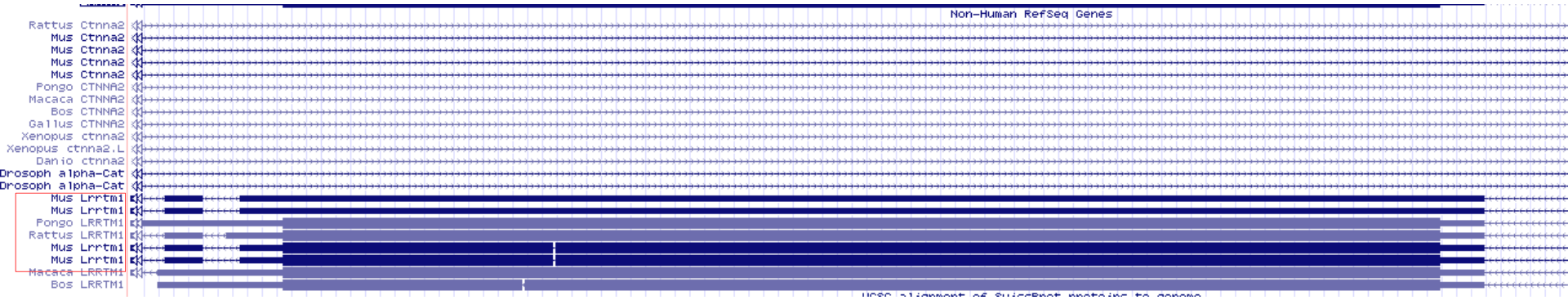
7. Does the protein encoded by this gene have a transmembrane domain?

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



8. Has this gene an ortholog in mouse? yes

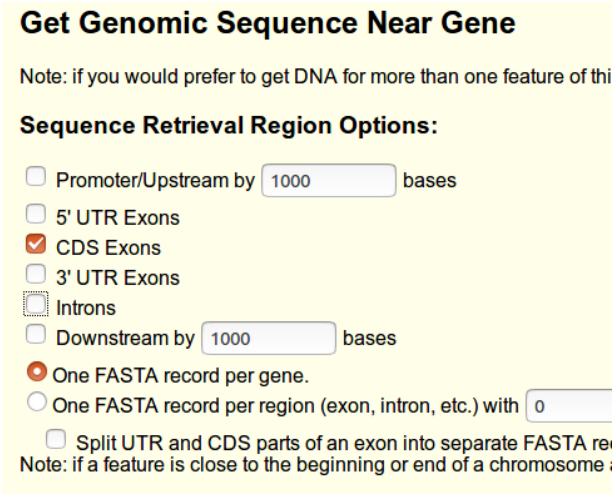
- Set display mode of track “Gene and Gene Predictions” > “Other RefSeq” to “pack”



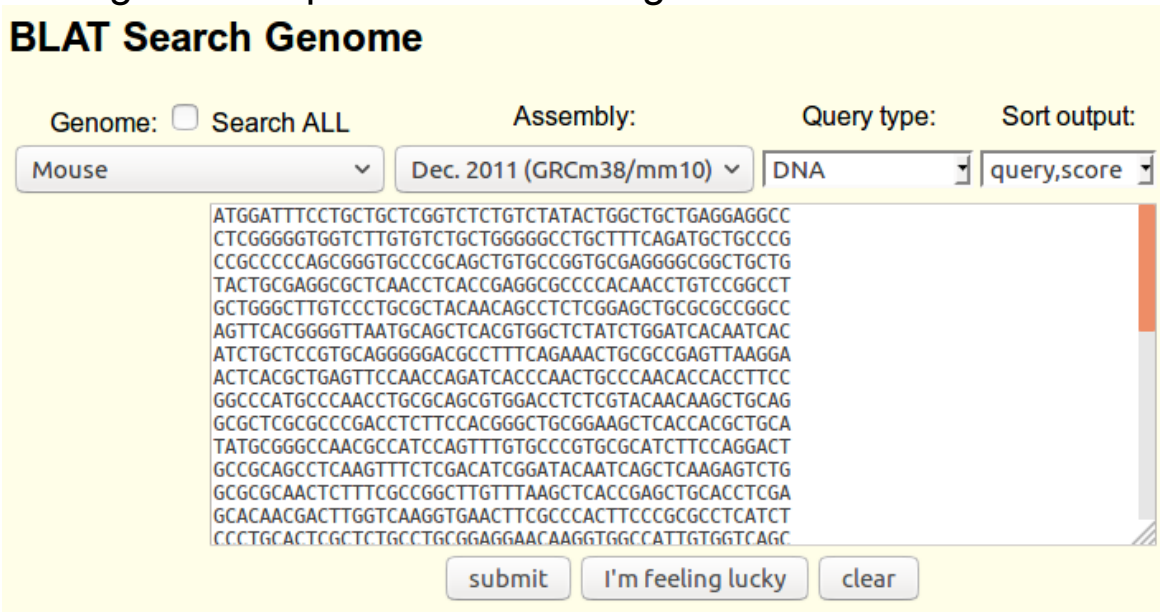
Practicum

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

- Get the CDS sequence of human gene:



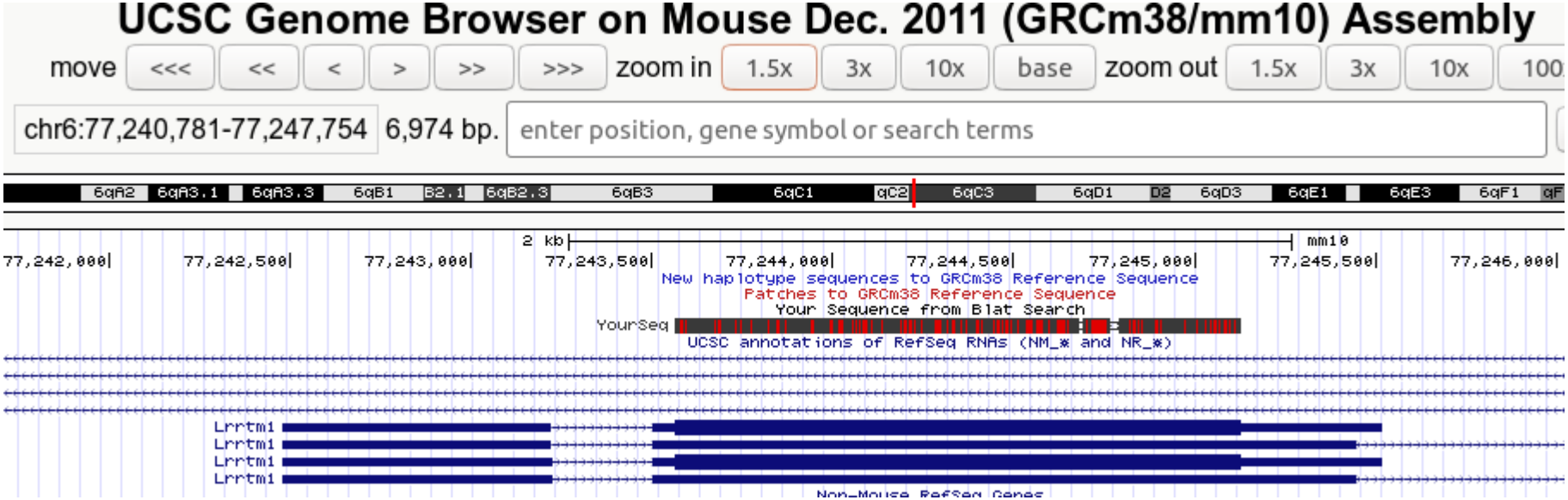
- Use BLAT to align this sequence on Mouse genome:



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

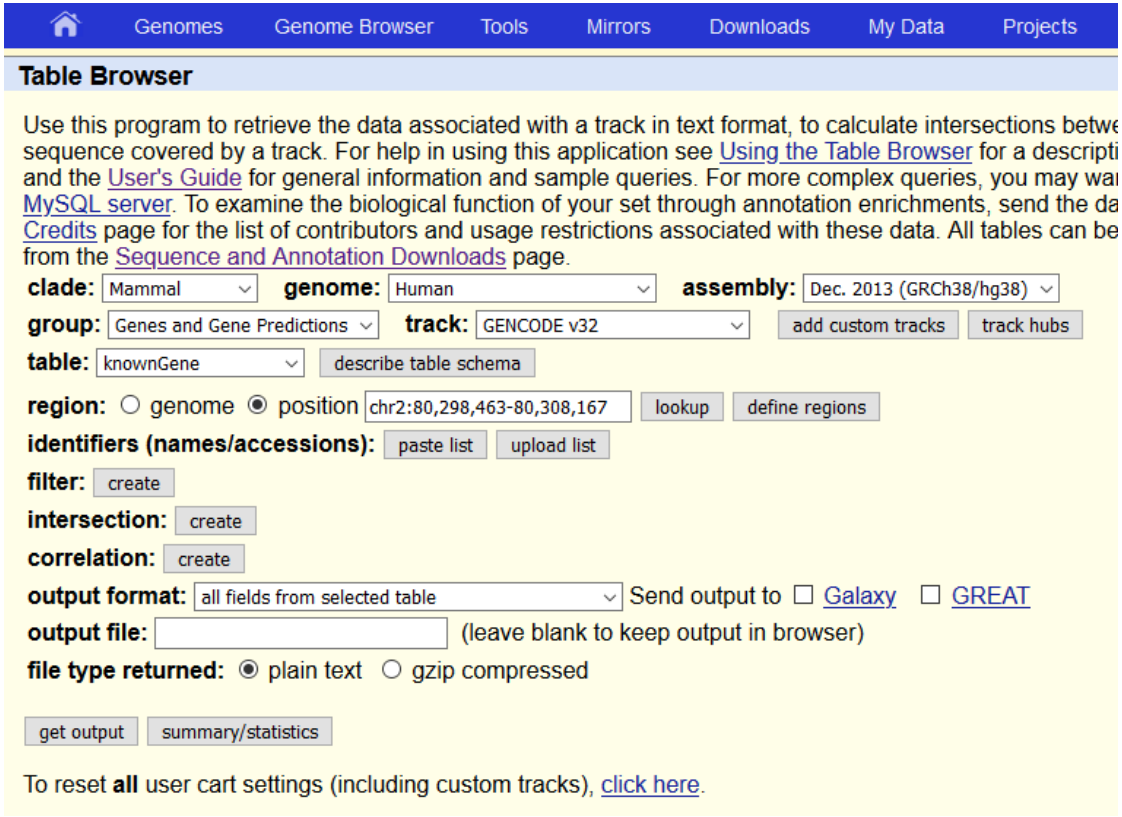


Genome Browsers

UCSC Genome Browser

https://genome.ucsc.edu/

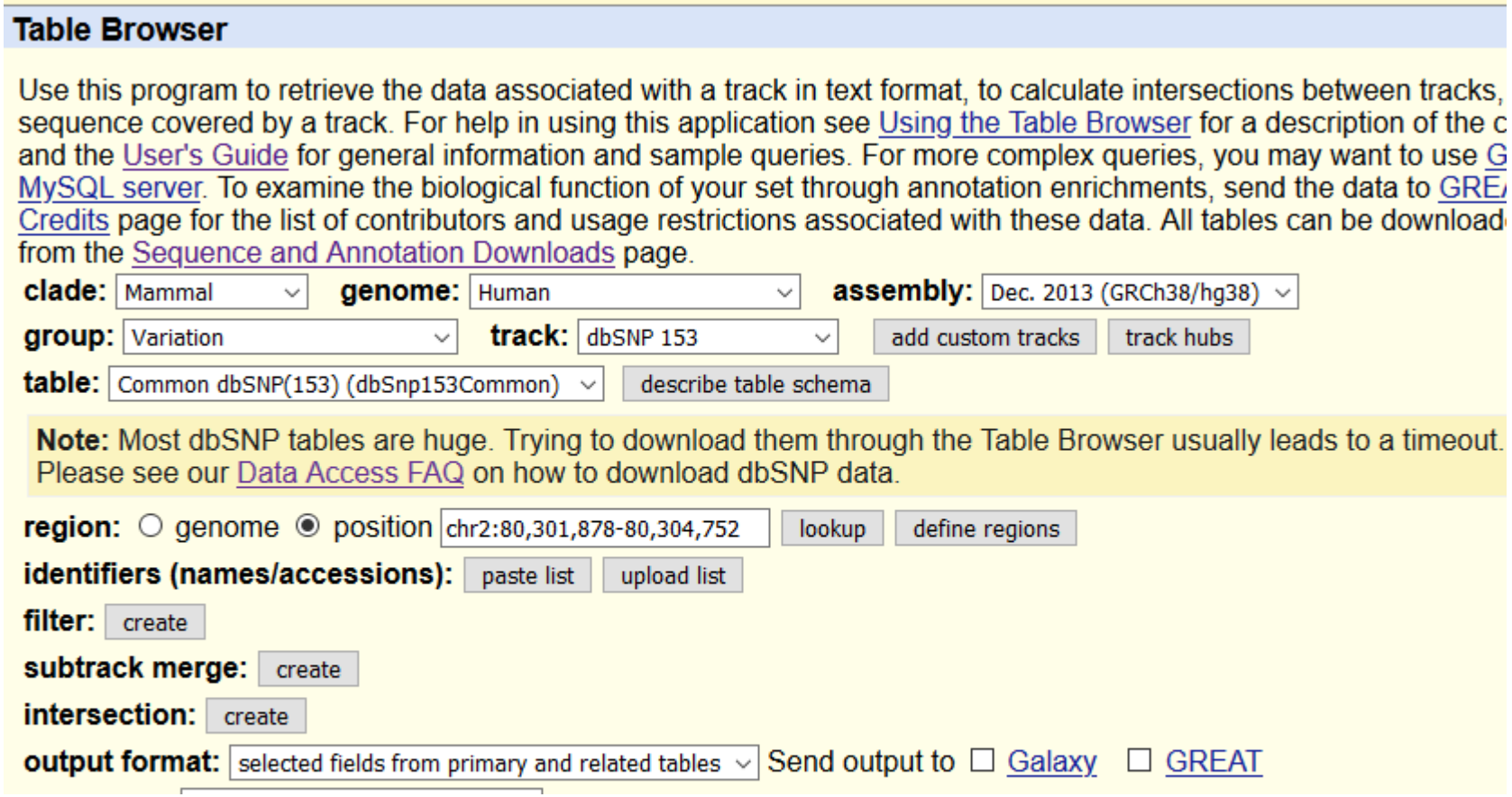
- Table Browser tool allows to:
  - Search for genes and annotation
  - Combine queries on multiple tables or tracks
  - Display basic statistics over a dataset
  - Output to results table in convenient format
  - Retrieve sequences
  - Export to external resources



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



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