

# Databases in molecular biology

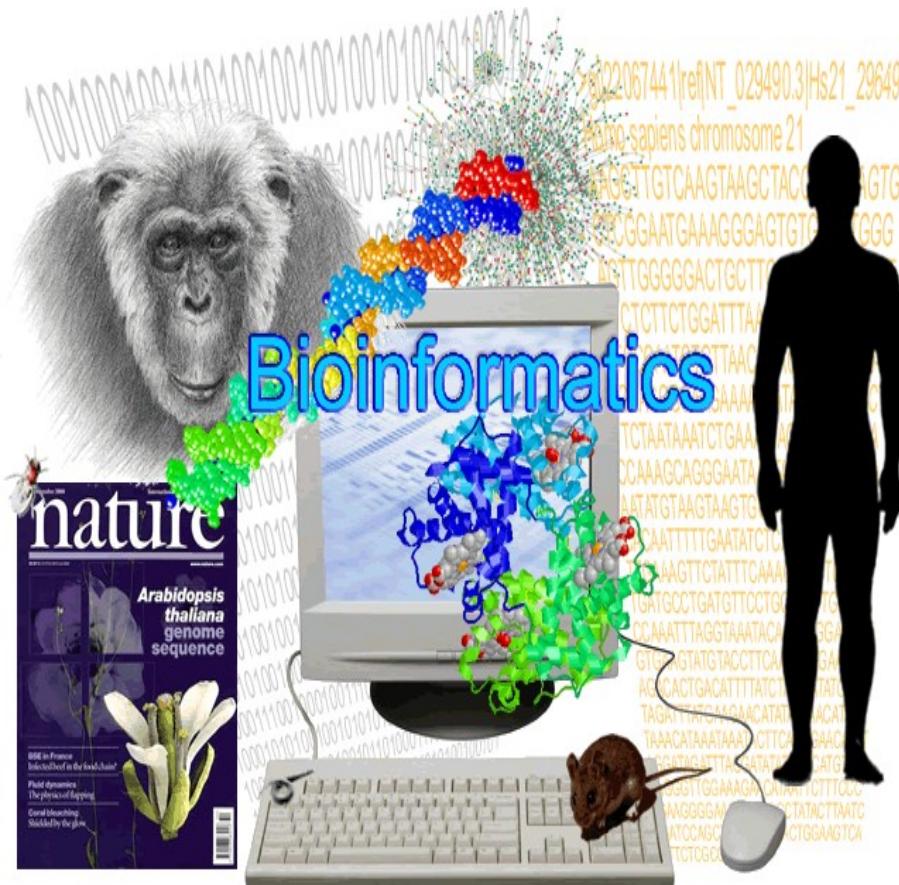
Bioinformatics Course UEB-VHIR  
November 2023

Mireia Ferrer<sup>1</sup>, Àlex Sánchez<sup>1,2</sup>, Esther Camacho<sup>1</sup>, Berta Miró<sup>1</sup>

<sup>1</sup> Unitat d'Estadística i Bioinformàtica (UEB) VHIR

<sup>2</sup> 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](#)

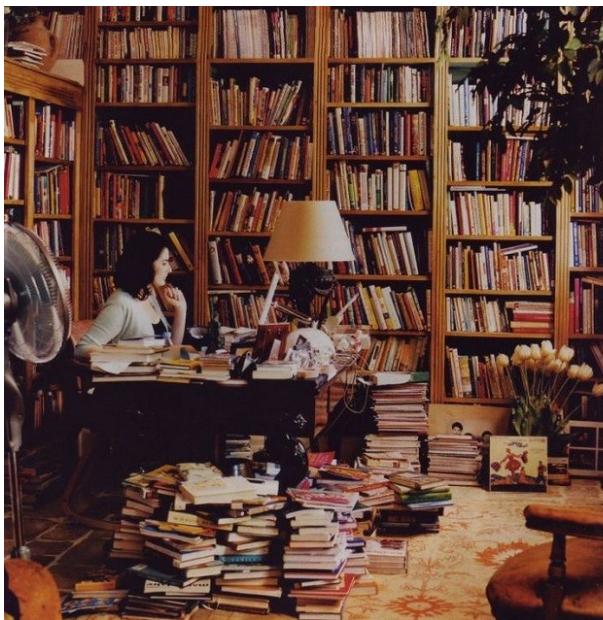
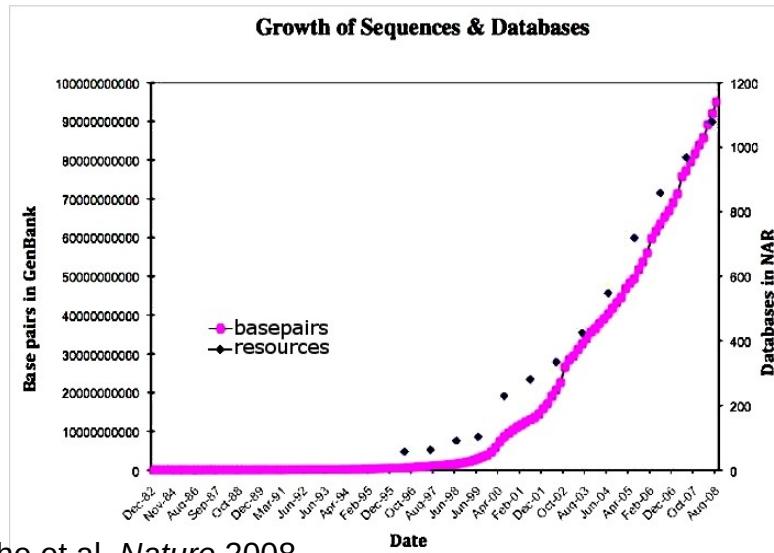
### Microarray Databases

- [ArrayExpress](#)
- [MIAME](#)

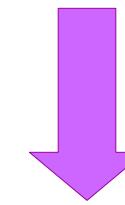
### Literature Databases

- [MEDLINE](#)
- [OMIM](#)
- [Patent Abstracts](#)
- [more...](#)

# Challenges



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

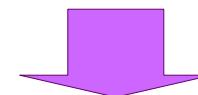
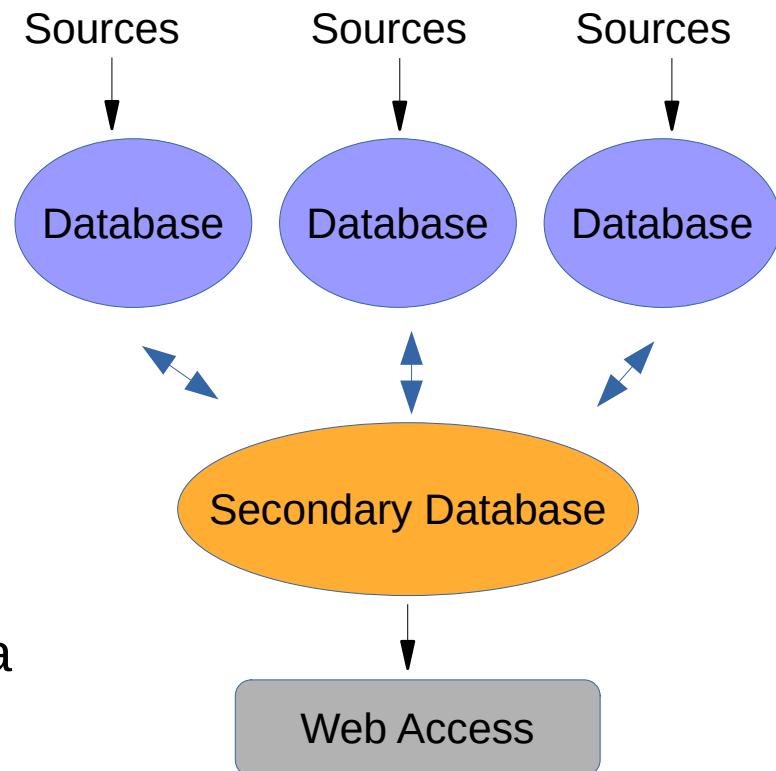
# Integrating the information

- **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

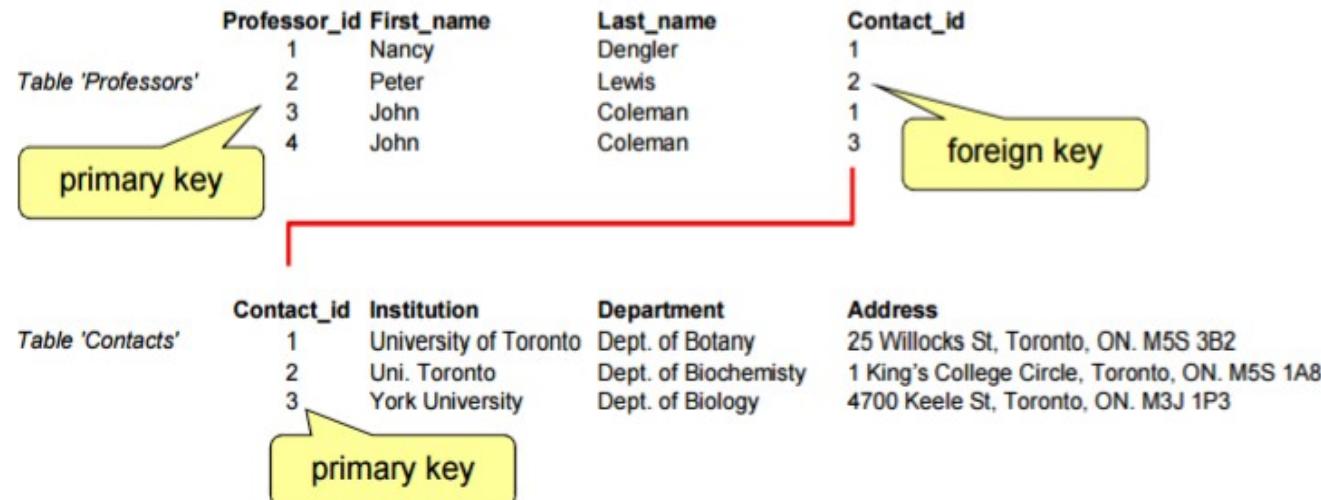
- The quality of the information in a database is closely related to its structure
- This aspect is also crucial for its efficiency and accessibility.
- There are many different types of file formats depending on the type of information they contain / source of data
- Most are based on the flat-file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

## FASTA format (sequences)

Identifier	Description
sequence	<pre>&gt;136435 Mus Musculus basic domain/leucine zipper transcription factor mRIIA, complete cds. gccccggccgcgtccccagacaaaaggcttggccggccggccggccggccgtgcgcctcgctccccgcctcccc cgcttgcgcgtcttcgcccccgctttggctggcgcgtcccggccggccgaaagtttccccgcggcag cgccggctgagcctcgcttttagcgatggccggagctgagcatggggcaagagactgcccaccagccgct ggccatggagtacgtcaacgacttcgaccttctaagttcgacgtgaagaaggagccctggggcgccgga gcgtccggccggccatgcacacgcctgcagcctgctggctgggtcgccaccccgctcagcactccgt</pre>

# 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

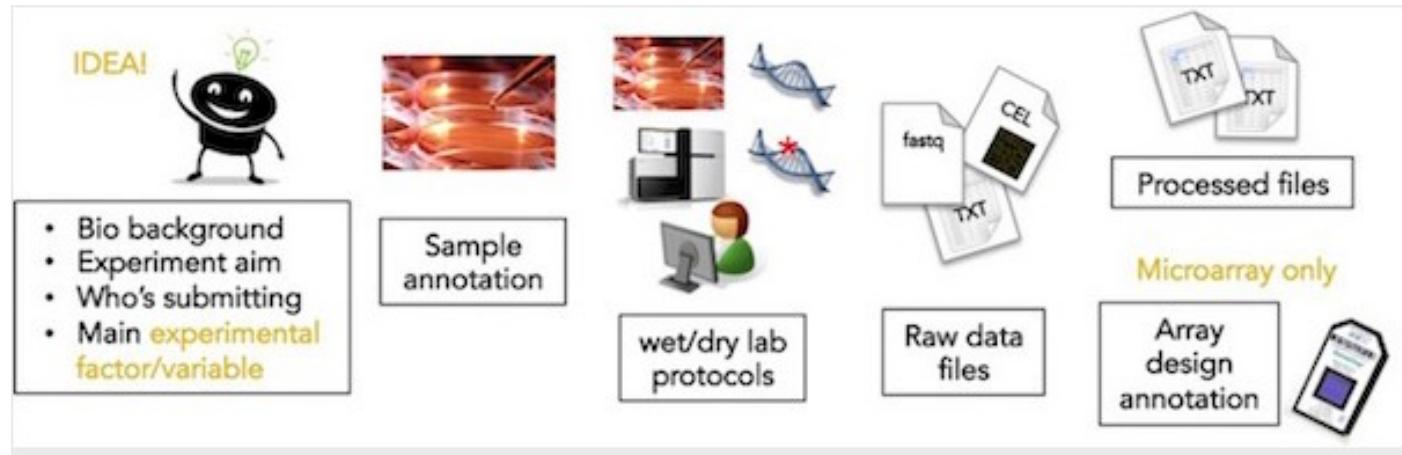
Entry	Entry name		Protein names		Gene names
P28332	ADH6_HUMAN		Alcohol dehydrogenase 6		ADH6

## 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)

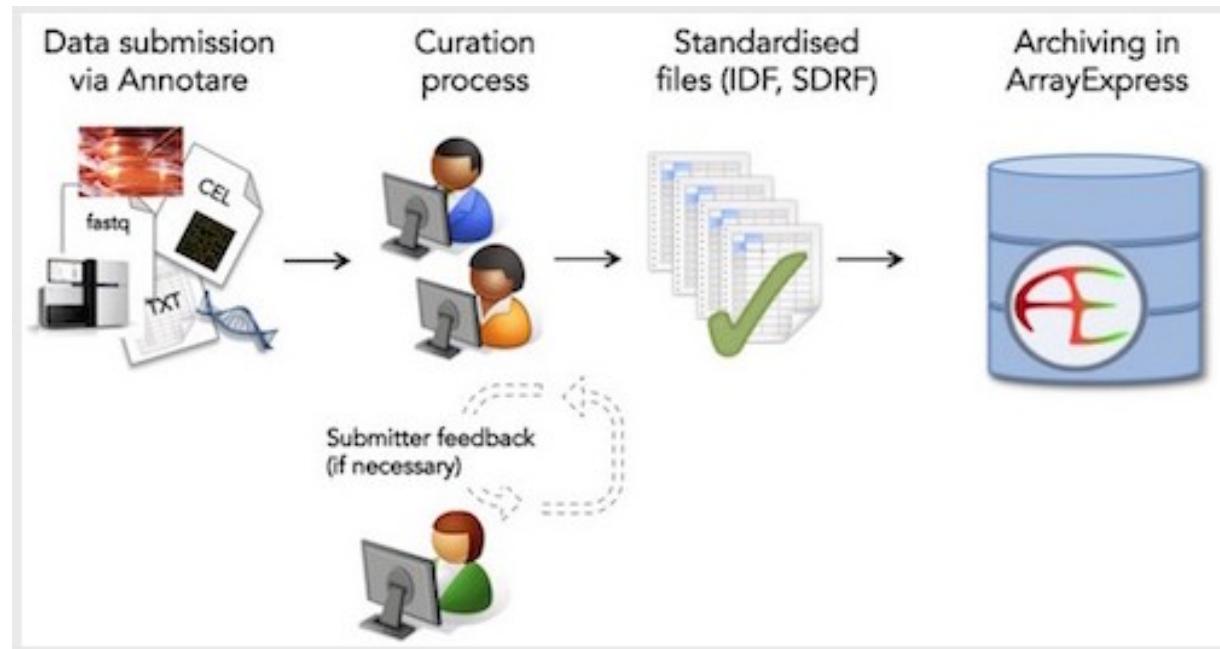


Source: <https://www.ebi.ac.uk/arrayexpress/submit/overview.html>

- 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.

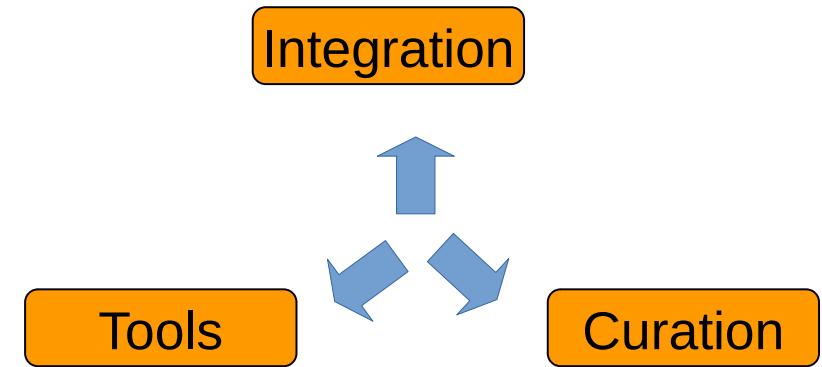


Source: <https://www.ebi.ac.uk/arrayexpress/submit/overview.html>

### III. Centralizing data management

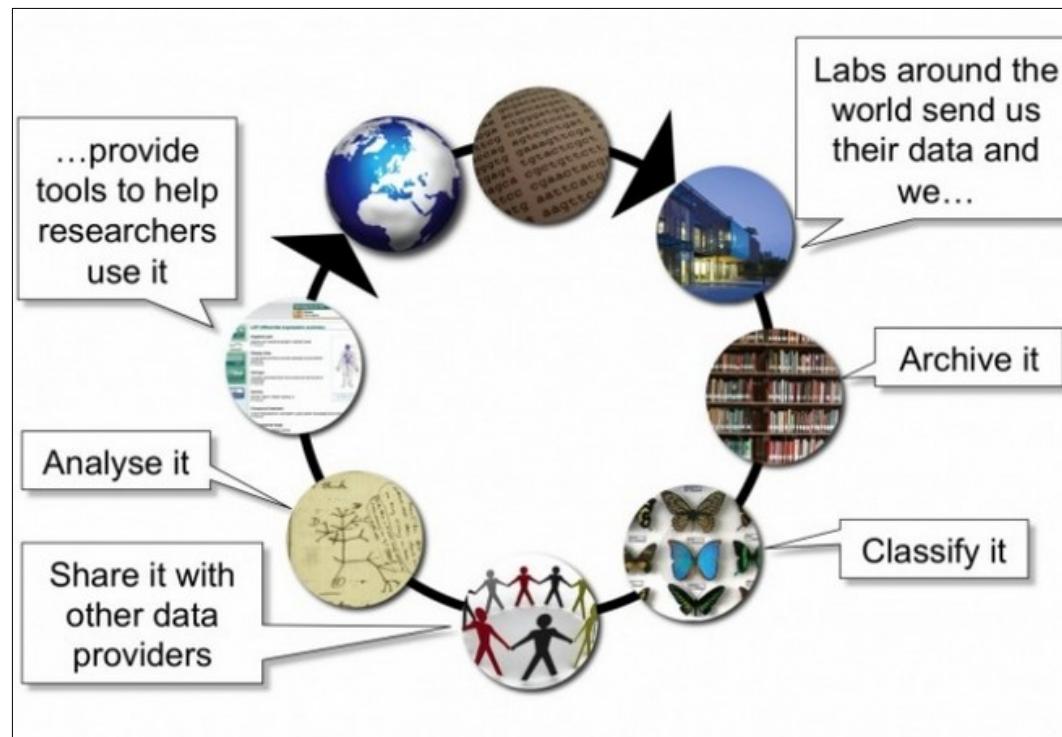
# Centralizing data management

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



# Resource providers

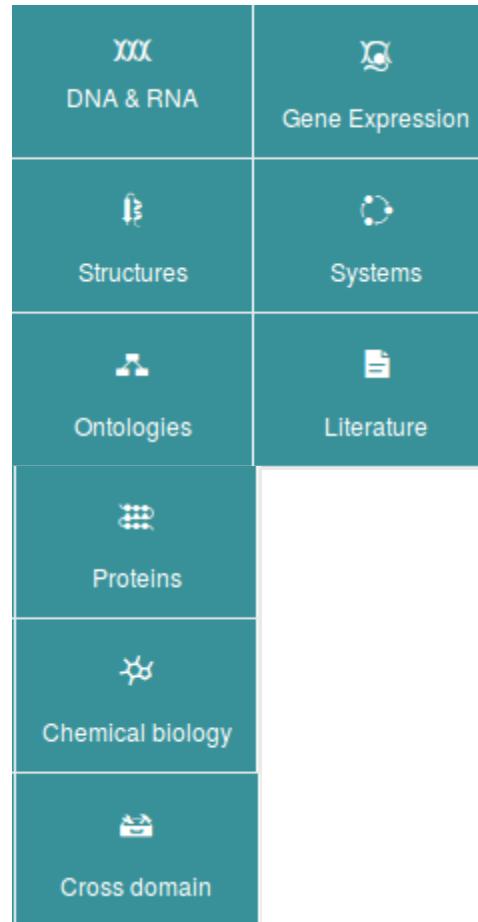
- Big organizations that act as *hubs* that provide transparent access to data sources.



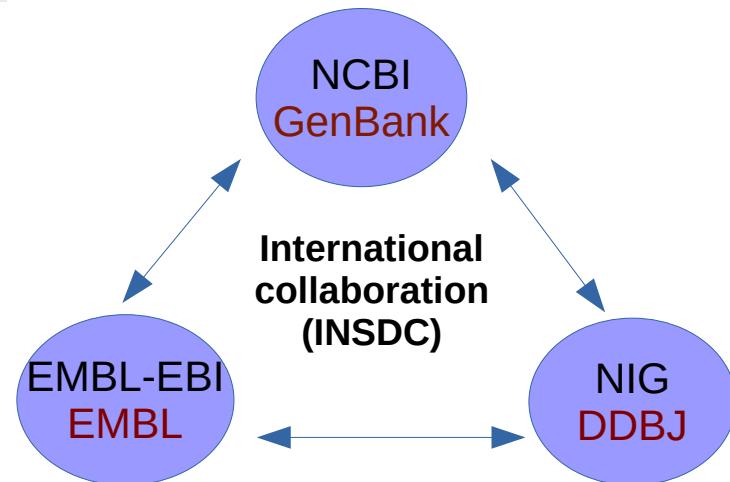
# Resource providers



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



- 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.  
Published in final edited form as:  
[Nature. 2013 Nov 7; 503\(7474\): 59–66.](#)  
doi: [\[10.1038/nature12709\]](#)

PMCID: PMC3983910  
NIHMSID: NIHMS524654  
PMID: [24201279](#)

Cooperation between brain and islet in glucose homeostasis and diabetes

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

► 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: !\[\]\(6a3a10fac78c4674bc151043e1625557\_img.jpg\)](#)

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.

NCBI Resources How To

Nucleotide Nucleotide (brca1) AND "Homo sapiens"[organism:txid9606]  
Create alert Advanced

Learn more about upcoming changes to the Nucleotide, EST, and GSS databases.

Species Animals (507) Summary 20 per page Sort by Default order  
Customize ...

Molecule types mRNA (507) Items: 1 to 20 of 507  
Customize ...  
clear

Source databases RefSeq (507) Filters activated: mRNA, RefSeq. Clear all  
Customize ...  
clear

Sequence length Custom range...  
Release date Custom range...  
Revision date

1. Homo sapiens BRCA1, DNA repair associated (BRCA1), transcript variant 5, mRNA  
Accession: NM\_007299.3 GI: 237681124  
Protein PubMed Taxonomy  
GenBank FASTA Graphics

2. Homo sapiens BRCA1, DNA repair associated (BRCA1), transcript variant 1, mRNA  
Accession: NM\_007294.3 GI: 237757282

miRBase

Home Search Browse Help Download Blog Submit hsa-mir-19a

Stem-loop sequence hsa-mir-19a

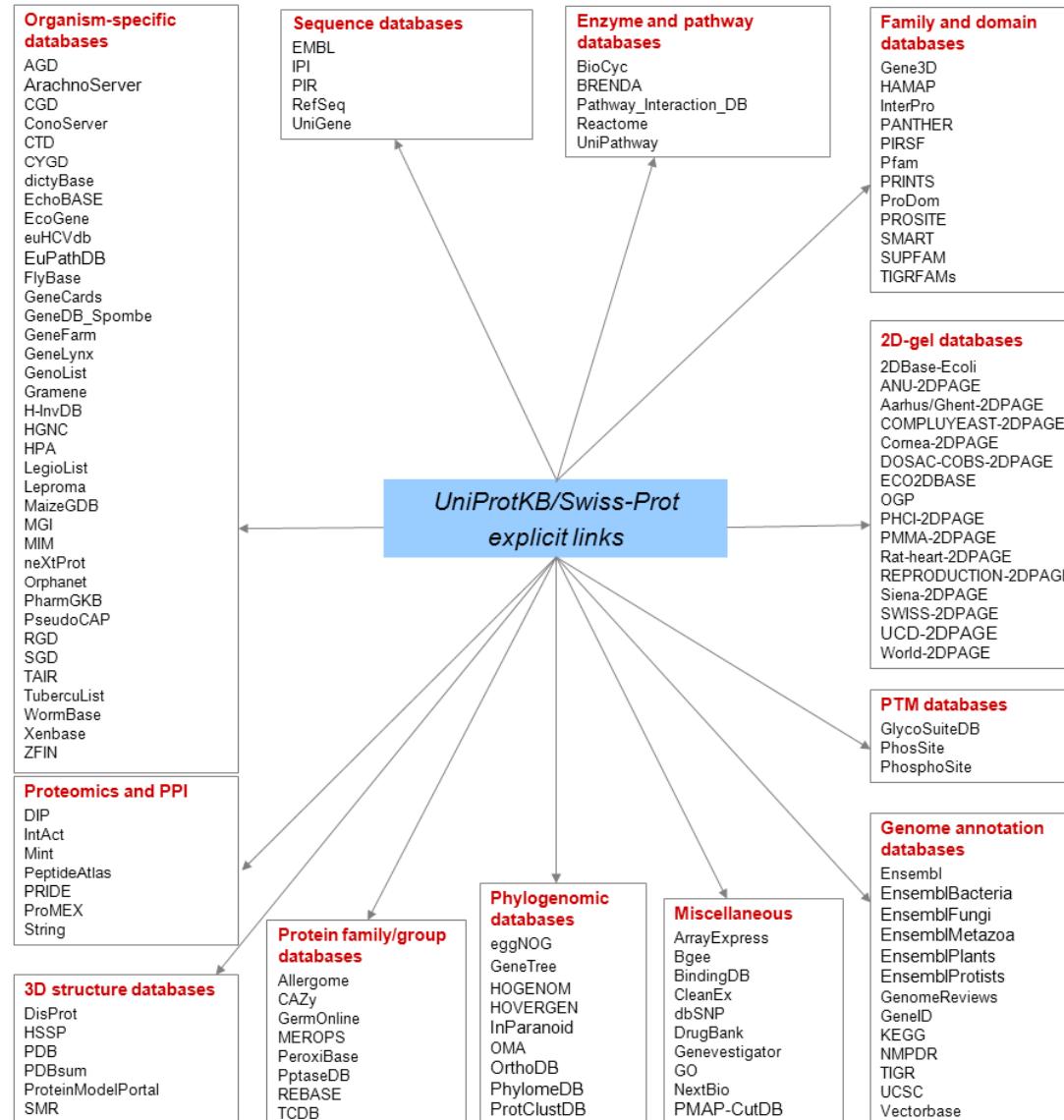
Accession	MI0000073 (change log)
Symbol	HGNC:MIR19A
Description	Homo sapiens miR-19a stem-loop
Gene family	MIPF0000011; mir-19
Literature search	313 open access papers mention hsa-mir-19a (1573 sentences)
Stem-loop	 <pre>       u   u             5' gcag cc cuguuuaguuuugcauag -- --- ag                                   c   g   gg uagucaaaaacguaua aacgug augu g                                   c   u           u   a   lug aa                   u   a                   u   a   </pre> <p>Get sequence</p>

# 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



# 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*

## Builder

All Fields dropdown: colorectal cancer Show index list

AND dropdown: Journal dropdown: Nature Show index list

AND dropdown: Publication Type dropdown: "review"[Publication Type] Show index list Hide index list

Search results (partial list):

- research support, nra, intramural (49210)
- research support, non u s govt (6930275)
- research support, u s govt, non p h s (790770)
- research support, u s govt, p h s (2460270)
- research support, u s government (2902642)
- retracted publication (6332)
- retraction of publication (6645)
- review (2456140)** (highlighted)
- scientific integrity review (243)
- study characteristics (4803808)
- support of research (8501193)

Buttons: Previous 200, Next 200, Refresh index

Bottom row: AND dropdown, All Fields dropdown, Show index list

Bottom buttons: Search (highlighted), or Add to history

# 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

The screenshot shows the NCBI Nucleotide search interface. The search term "mutyh AND \"Homo sapiens\"[orgn:txid9606]" has been entered. The results page displays 20 items out of 38, filtered for mRNA. The first result is for "Homo sapiens mutY DNA glycosylase (MUTYH), transcript variant 13, mRNA". It includes details like accession NM\_001350651.1, GI 1183596751, and links to Protein, PubMed, and Taxonomy. Below this, another result for "Homo sapiens mutY DNA glycosylase (MUTYH), transcript variant 12, mRNA" is shown. A red bracket on the left side groups the filter options under the heading "Using filters".

Species: Summary ▾ 20 per page ▾ Sort by Default order ▾

Items: 1 to 20 of 38

Filters activated: mRNA, RefSeq. [Clear all](#)

[Homo sapiens mutY DNA glycosylase \(MUTYH\), transcript variant 13, mRNA](#)

1. 1,767 bp linear mRNA  
Accession: NM\_001350651.1 GI: 1183596751  
[Protein](#) [PubMed](#) [Taxonomy](#)  
[GenBank](#) [FASTA](#) [Graphics](#)

[Homo sapiens mutY DNA glycosylase \(MUTYH\), transcript variant 12, mRNA](#)

2. 1,831 bp linear mRNA

**Homo sapiens mutY DNA glycosylase (MUTYH), transcript**

NCBI Reference Sequence: NM\_001350651.1  
[GenBank](#) [Graphics](#)

```
>NM_001350651.1 Homo sapiens mutY DNA glycosylase (MUTYH), transcript variant mRNA
CAGCCGGAGCCCGGGTACAACGGAACCTGTAGTCCTCGTGGCTAGTTCAAGCGGAAGGGAGCAGTC
TCTGAAGCTTGAGGAGCCTCTAGAACTATGAGCCGAGGCCCTCCCTCTCCAGAGGCCAGAGGCTT
AAGGCTACTCTGGGAAGCCGCTCACCGCTCGAGCTGCGGGAGCTGAAACTGCGCCATCGTCAGTGTG
GCGGCATGACACCGCTCGTCTCCGCTGAGTCGTCAGTGGGACATGAGGAAGGCCAGGAGCAGCCG
TGGGAAGTGGTACAGGAAGCAGGCCAGGAGCAGAGCATGTAAGAACAAACAGTC
GGCCAAGCCTTCTGCGCTGTAGAGACGTAGCTGAAGTCACAGCCTCCGAGGGAGCCTGCTAAGCTGG
ACGACCAAGAGAACCGGGACCTACCATGGAGAACGGCAGAGATGAGATGGACCTGGACAGGCCGGC
ATATGCTGAAGTGGCCTACACTGAGGACCTGGCCAGTGCTCCCTGGAGGAGGTGAATCAAACCTGGG
```

# Practicum

## 3- Look for MUTYH human protein in UniProtKB

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

### UniProtKB - Q9UIF7 (MUTYH\_HUMAN)

 Basket ▾

#### Display

 BLAST  Align  Format  Add to basket  History

 Feedback  Help video  Other tutorials and videos

Entry

Protein | Adenine DNA glycosylase

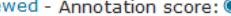
Publications

Gene | MUTYH

Feature viewer

Organism | Homo sapiens (Human)

Feature table

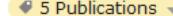
Status |  Reviewed - Annotation score:  - Experimental evidence at protein level<sup>i</sup>

None

Function

#### Function<sup>i</sup>

Names & Taxonomy

Involved in oxidative DNA damage repair. Initiates repair of A\*oxoG to C\*G by removing the inappropriately paired adenine base from the DNA backbone. Possesses both adenine and 2-OH-A DNA glycosylase activities.  5 Publications ▾

Subcellular location

Catalytic activity<sup>i</sup>

# Practicum

## Interaction<sup>i</sup>

### Binary interactions<sup>i</sup>

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

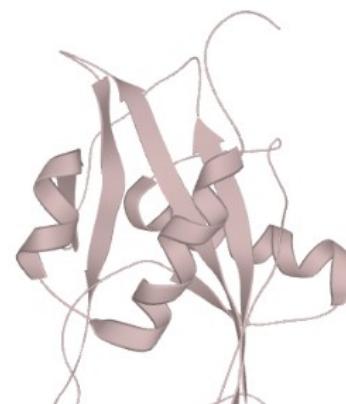
### Protein-protein interaction databases

BioGrid <sup>i</sup>	110681, 11 interactors
DIP <sup>i</sup>	DIP-41972N
IntAct <sup>i</sup>	Q9UIF7, 15 interactors
MINT <sup>i</sup>	Q9UIF7
STRING <sup>i</sup>	9606.ENSP00000361170

## Structure<sup>i</sup>

### Family and domain databases

CDD <sup>i</sup>	cd03431 DNA_Glycosylase_C, 1 hit cd00056 ENDO3c, 1 hit
Gene3D <sup>i</sup>	1.10.1670.10, 1 hit
InterPro <sup>i</sup>	<a href="#">View protein in InterPro</a> IPR011257 DNA_glycosylase IPR004036 Endonuclease-III-like_CS2 IPR003651 Endonuclease3_FeS-loop_motif IPR004035 Endonuclease-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 <sup>i</sup>	<a href="#">View protein in Pfam</a> PF00633 HHH, 1 hit PF00730 HhH-GPD, 1 hit PF14815 NUDIX_4, 1 hit
SMART <sup>i</sup>	<a href="#">View protein in SMART</a> SM00478 ENDO3c, 1 hit SM00525 FES, 1 hit
SUPFAM <sup>i</sup>	SSF48150 SSF48150, 1 hit SSF55811 SSF55811, 1 hit
PROSITE <sup>i</sup>	<a href="#">View protein in PROSITE</a> PS00764 ENDONUCLEASE_III_1, 1 hit PS01155 ENDONUCLEASE_III_2, 1 hit PS51462 NUDIX, 1 hit



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

1 notificació

# Examples of Databases

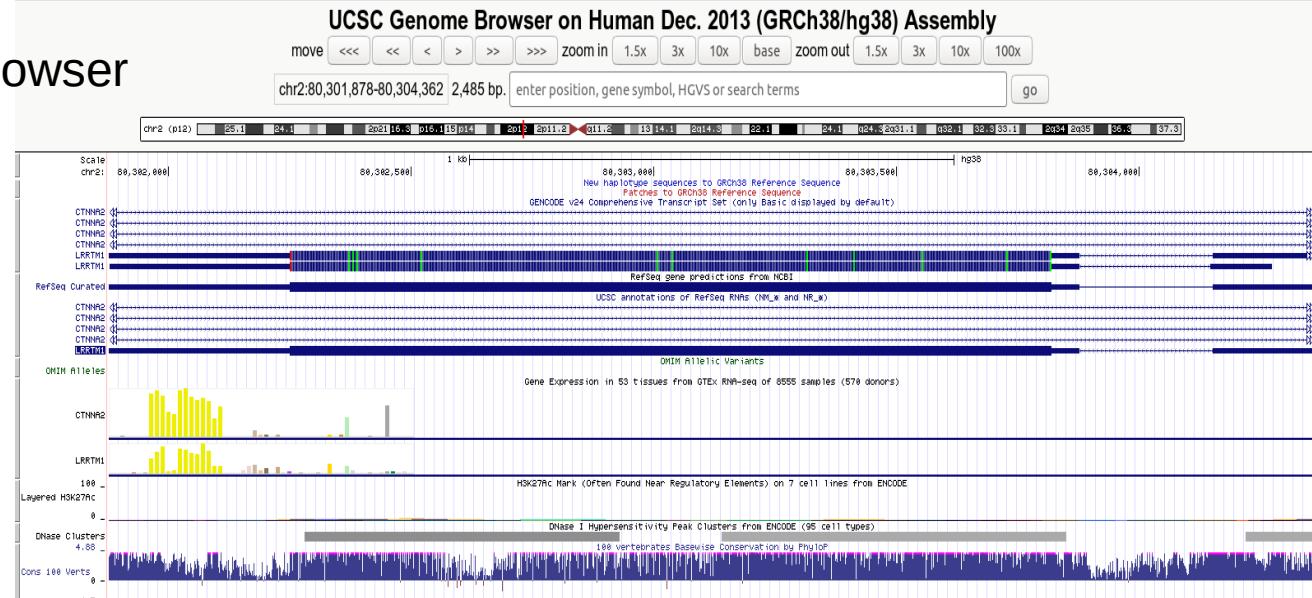
## Genomic databases

- 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

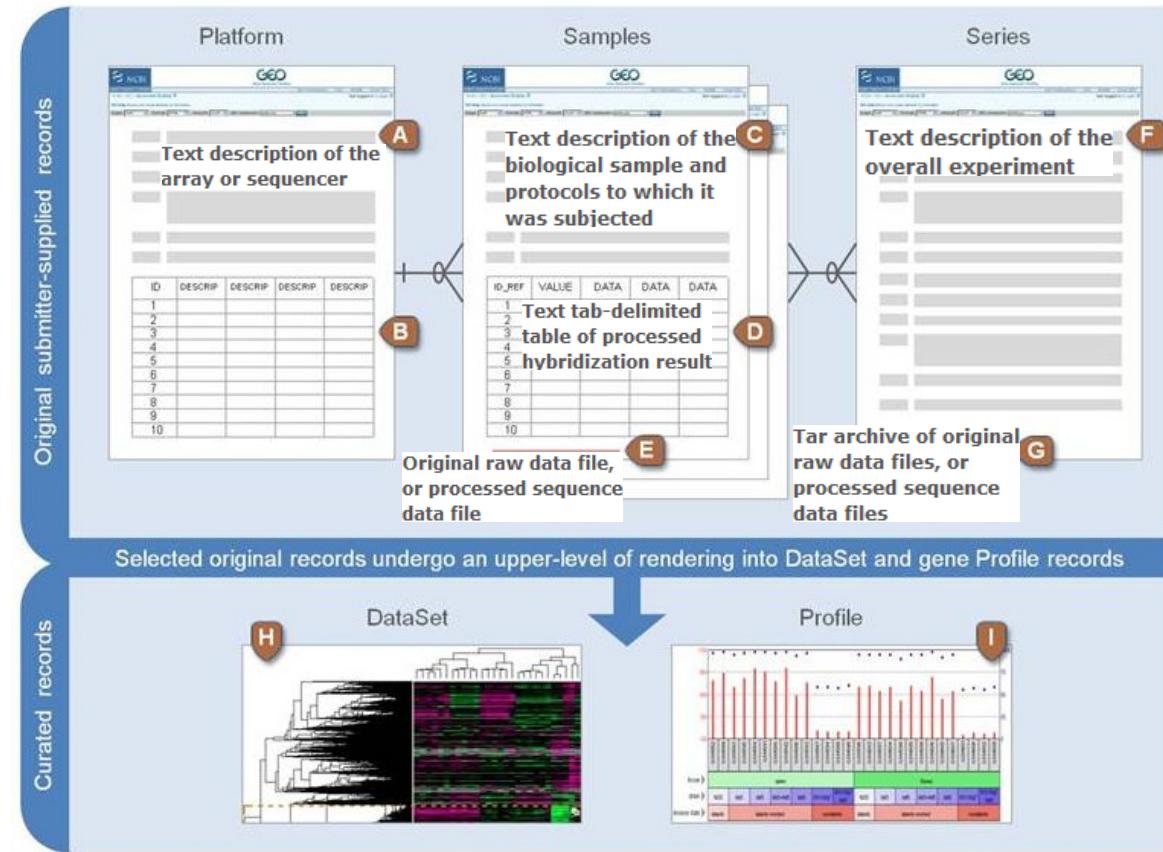
## Gene Expression Databases

- 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	<i>Mus musculus</i>	GPL6887	GSE39549	51
GDS6247	Diet-induced obesity model: white adipose tissue	<i>Mus musculus</i>	GPL6887	GSE39549	40
GDS6177	Acute alcohol consumption effect on whole blood (control...)	<i>Homo sapiens</i>	GPL570	GSE20489	25

# Practicum

## 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	<a href="#">Mus musculus</a>	
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 24 weeks	
Contributor(s)	Kwon E. Choi M	
Platforms (1)	<a href="#">GPL6887 Illumina MouseWG-6 v2.0 expression beadchip</a>	
Samples (91) <a href="#">+ More...</a>	<a href="#">GSM971546</a> Mice fed Normal diet for 2weeks rep1 <a href="#">GSM971547</a> Mice fed Normal diet for 2weeks rep2 <a href="#">GSM971548</a> Mice fed Normal diet for 2weeks rep3	
Relations		
BioProject	<a href="#">PRJNA171109</a>	
<a href="#">Analyze with GEO2R</a>		

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

Download family		Format												
SOFT formatted family file(s)		<a href="#">SOFT</a> 												
MINiML formatted family file(s)		<a href="#">MINiML</a> 												
Series Matrix File(s)		<a href="#">TXT</a> 												
<b>Supplementary file</b> <table border="1"> <thead> <tr> <th>Size</th> <th>Download</th> <th>File type/resource</th> </tr> </thead> <tbody> <tr> <td>8.4 Mb</td> <td>(ftp)(http)</td> <td>TXT</td> </tr> <tr> <td>2.9 Mb</td> <td>(ftp)(http)</td> <td>TXT</td> </tr> <tr> <td>15.8 Mb</td> <td>(http)(custom)</td> <td>TAR</td> </tr> </tbody> </table>			Size	Download	File type/resource	8.4 Mb	(ftp)(http)	TXT	2.9 Mb	(ftp)(http)	TXT	15.8 Mb	(http)(custom)	TAR
Size	Download	File type/resource												
8.4 Mb	(ftp)(http)	TXT												
2.9 Mb	(ftp)(http)	TXT												
15.8 Mb	(http)(custom)	TAR												
Raw data is available on Series record Processed data included within Sample table														

# Practicum

## Retrieving data from GEO

Source name	Adipose tissue of mice
Organism	<a href="#">Mus musculus</a>
Characteristics	strain: C57BL/6J treatment protocol: Normal diet time: 2 weeks 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 Bioanalyzer.
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_2458927	6.5777017

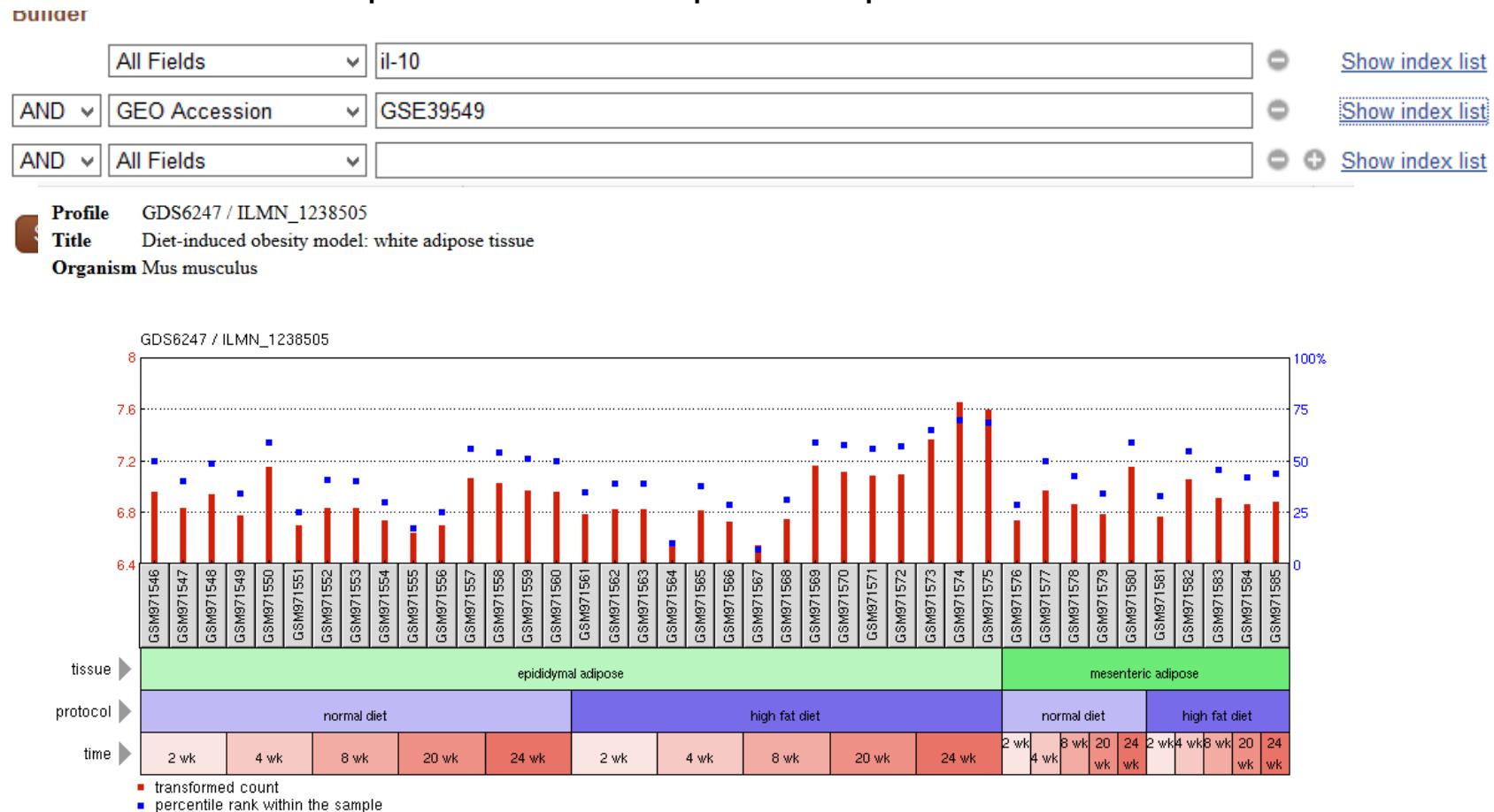
Data table with normalized expression values

# 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](#)
- **Terapeutic 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

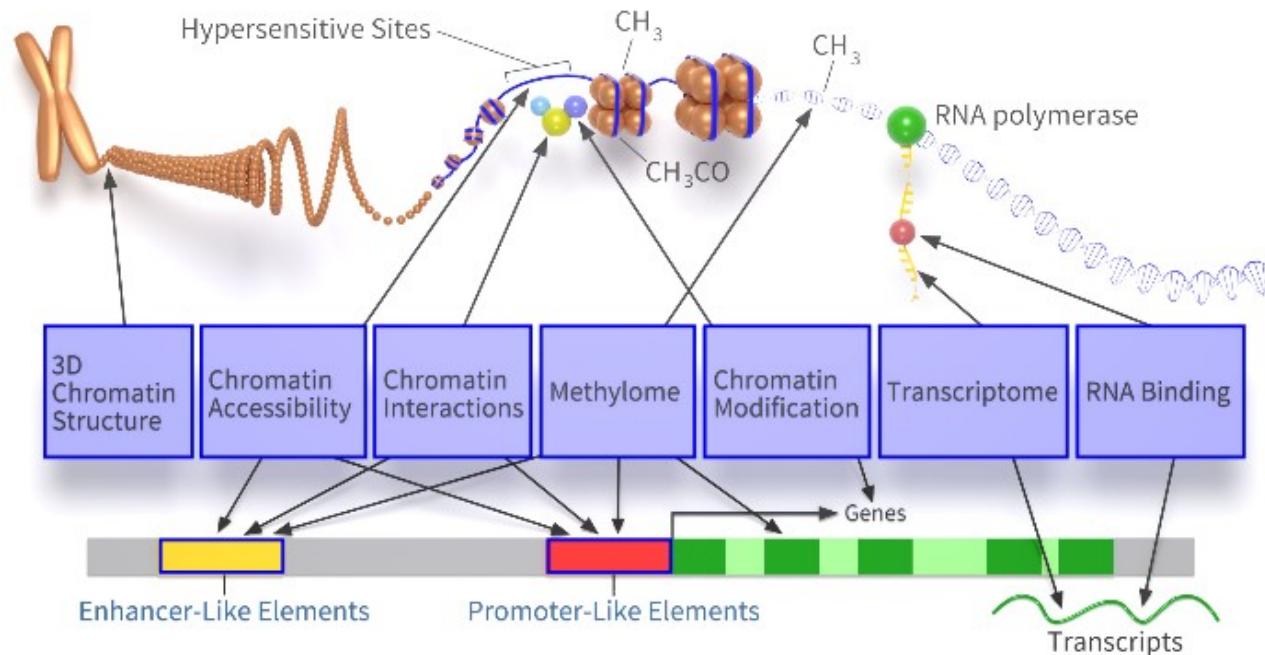
- Collective initiatives: summing efforts from different researchers from different sites of the world.
- **Integrate** different types of data (clinical, multi-omics,...) related to a **specific subject** (or sub-subject)
  - A disease
  - An organism
  - A biological question, tissue, ...
- Provide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

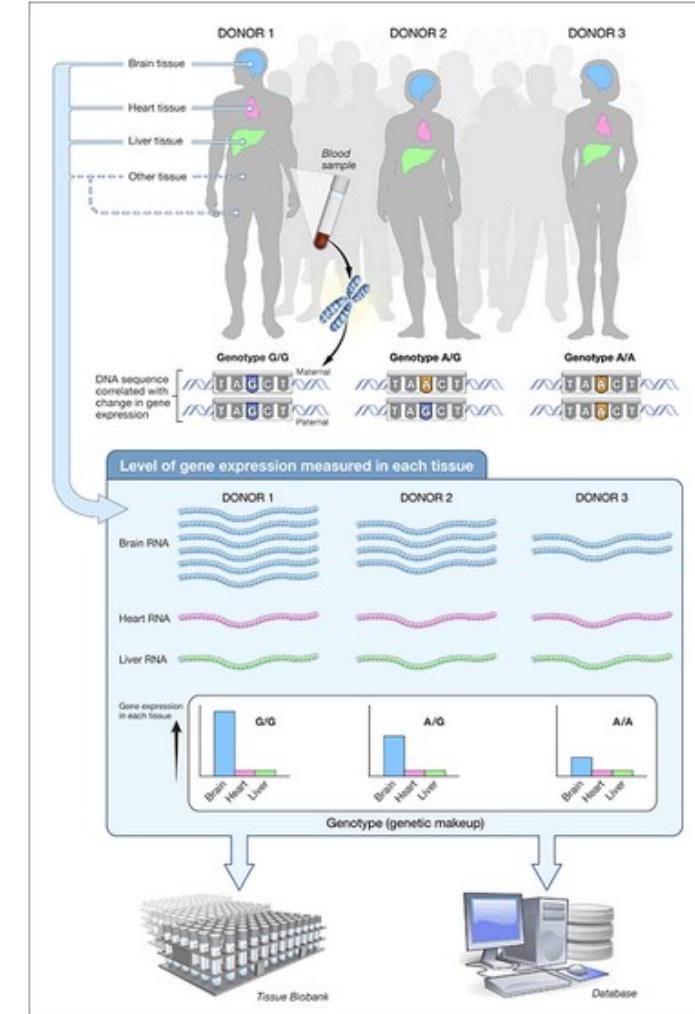
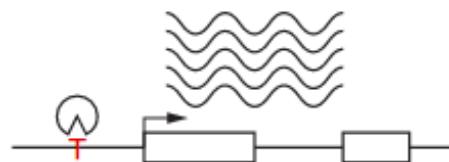


# Subject-specific repositories

## 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.

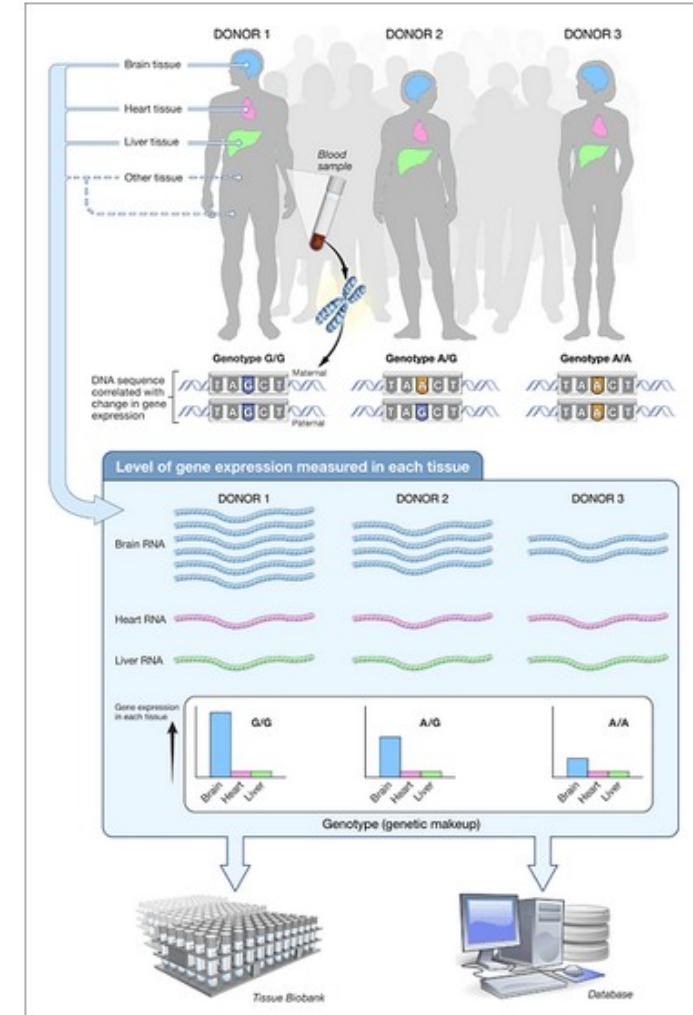


# Subject-specific repositories

## Genotype-Tissue Expression (GTEx) Project

<https://gtexportal.org/home/>

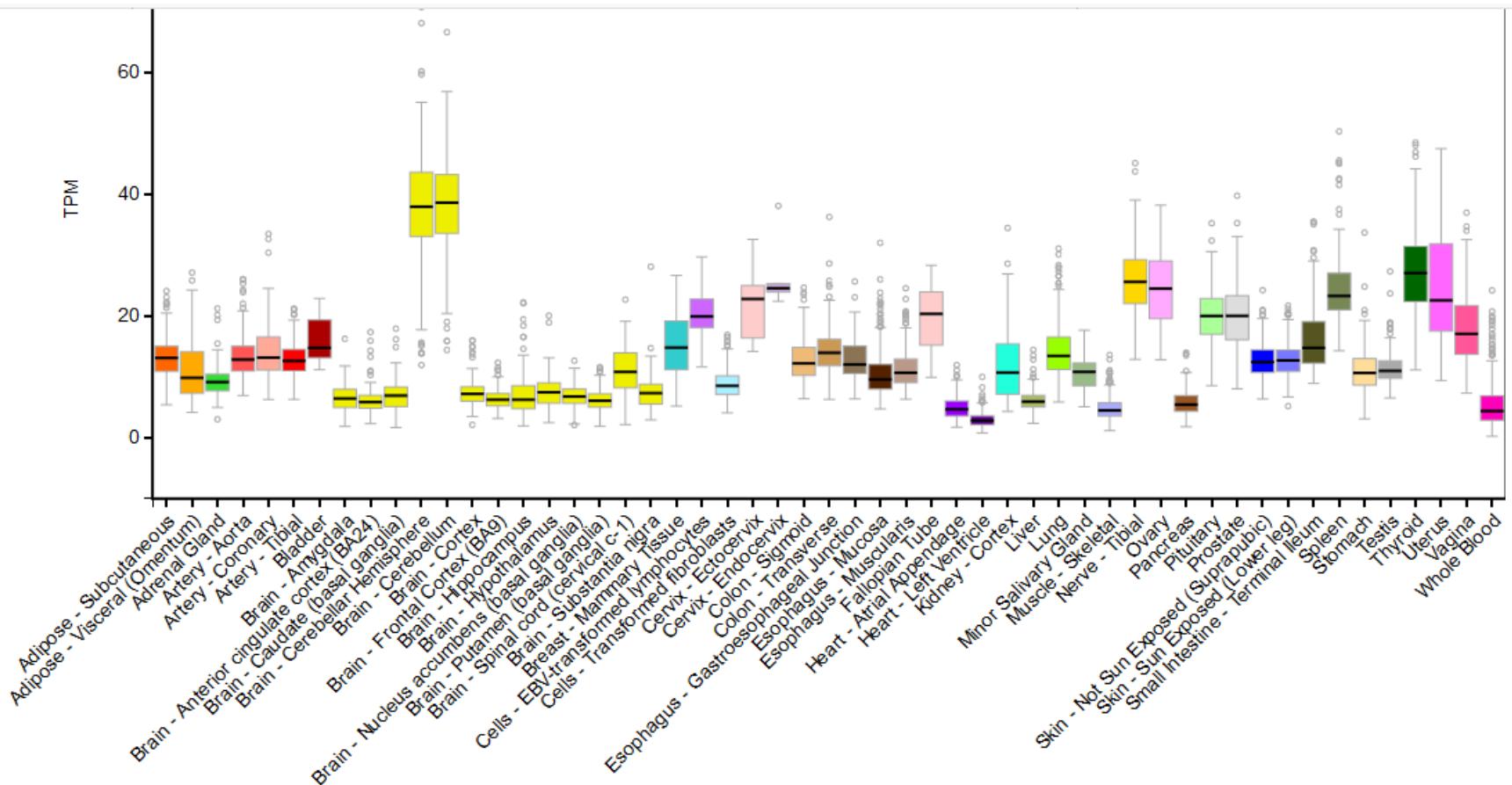
- Types of data provided:
  - Gene/transcript expression in tissues
  - variants associated to gene expression (eQTLs and sQTLs)
  - histology images
  - Patient/sample metadata
- Available datasets:  
<https://www.gtexportal.org/home/datasets>
- Summary statistics:  
<https://gtexportal.org/home/tissueSummaryPage>



# Subject-specific repositories

## Genotype-Tissue Expression (GTEx) Project

Example: Normal tissue expression profile of *mutyh* gene

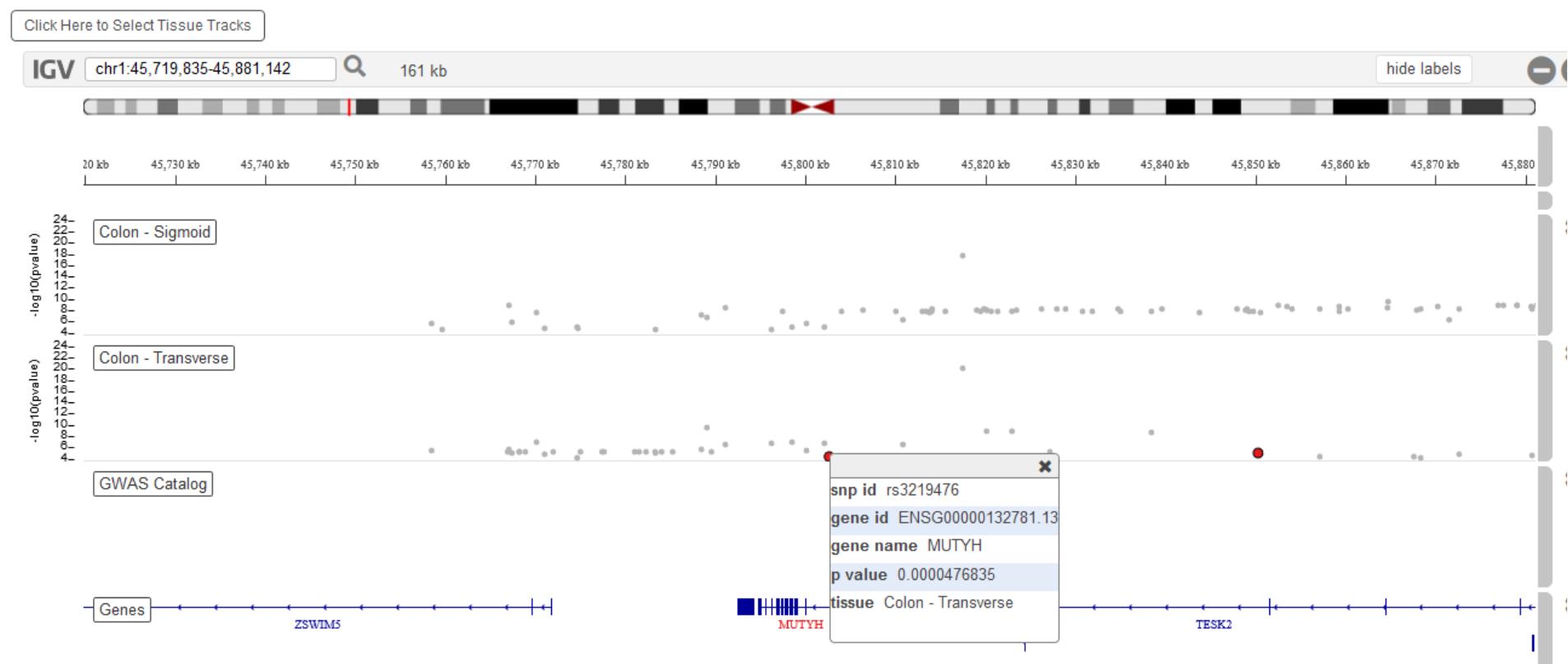


# Subject-specific repositories

## Genotype-Tissue Expression (GTEx) Project

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

### GTEx IGV eQTL Browser



On the selected tissue eQTL tracks:

- Red dots are significant cis-eQTLs for the queried gene or SNP (at FDR<5%).
- Gray dots are significant cis-eQTLs for all other SNP-gene pairs within the genomic region.

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

# 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

NATIONAL CANCER INSTITUTE  
GDC Data Portal

Home Projects Exploration Analysis Repository Quick Search Manage Sets Login Cart 0 GDC Apps

Files Cases Add a File Filter

Start searching by selecting a facet

Add All Files to Cart Manifest View 33,096 Cases in Exploration View Images Browse Analytics

File e.g. 142682.bam, 4f6e2e7a-b...

Data Category

- Simple Nucleotide Variation 127,390
- Transcriptome Profiling 57,685
- Biospecimen 55,223
- Raw Sequencing Data 47,248
- Copy Number Variation 45,256

3 More...

Data Type

- Annotated Somatic Mutation 63,536
- Raw Simple Somatic Mutation 63,536

Start searching by selecting a facet

Add All Files to Cart Manifest View 33,096 Cases in Exploration View Images Browse Analytics

File e.g. 142682.bam, 4f6e2e7a-b...

Data Category

- Simple Nucleotide Variation 127,390
- Transcriptome Profiling 57,685
- Biospecimen 55,223
- Raw Sequencing Data 47,248
- Copy Number Variation 45,256

3 More...

Data Type

- Annotated Somatic Mutation 63,536
- Raw Simple Somatic Mutation 63,536

Showing 1 - 20 of 33,096 cases

Primary Site Project Disease Type Gender Vital Status

Available Files per Data Category

Gender	Files	Seq	Exp	SNV	CNV	Meth	Clinical	Bio	Annotations
Male	20	0	0	0	0	0	0	0	8
Female	12	0	0	0	0	0	0	0	4

Cart Case ID Project Primary Site Gender Files Available Files per Data Category Annotations

TCGA-AF-3912 TCGA-READ Rectosigmoid junction -- Seq Exp SNV CNV Meth Clinical Bio Annotations

# 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.

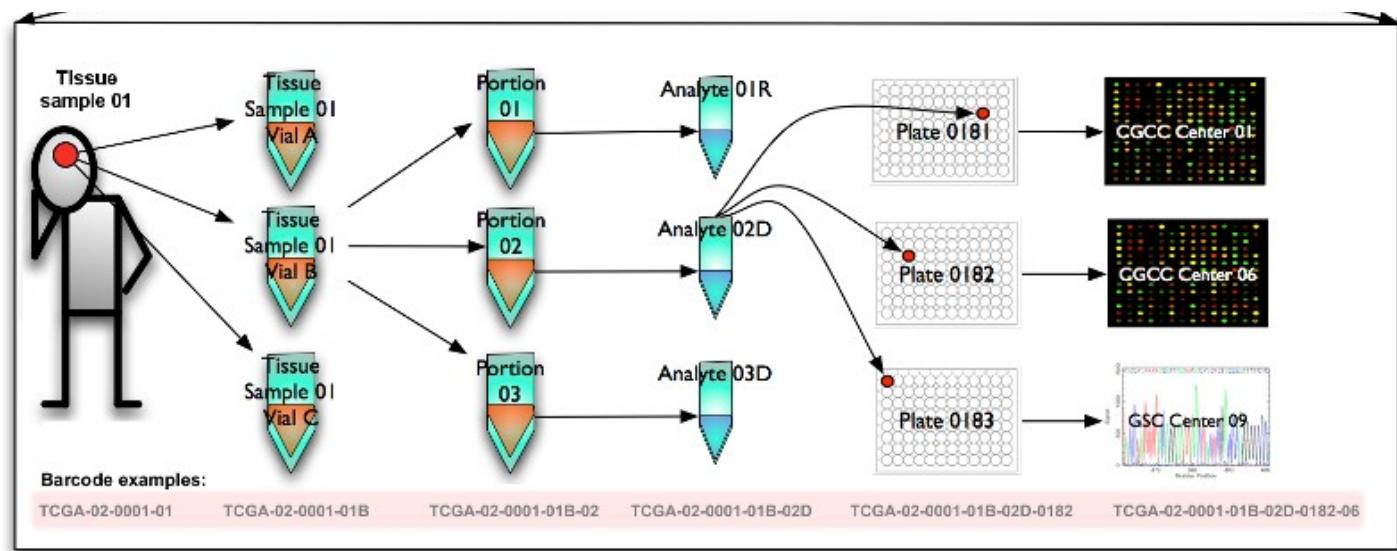
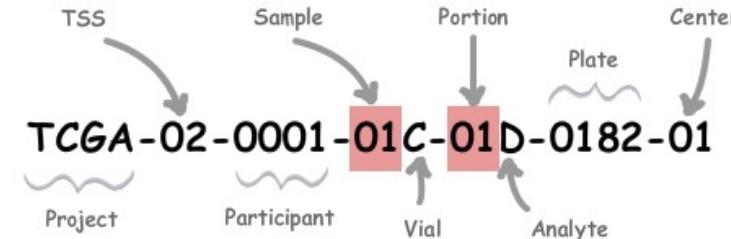
# Subject-specific repositories

## 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)

NATIONAL CANCER INSTITUTE GDC Data Portal

Home Projects Exploration Analysis Repository Quick Search Manage Sets Login Cart 0 GDC Apps

e.g. TCGA-A5-A0G2, 432fe4a9-2...

Upload Case Set

Primary Site

- Ovary 163
- Breast 108
- Bronchus and lung 101
- Corpus uteri 64
- Skin 39

33 More...

Program

- TCGA 810

Project

- TCGA-OV 163
- TCGA-BRCA 108
- TCGA-UCEC 62

Cases (810) Genes (1) Mutations (124) OncoGrid

Primary Site Project Disease Type Gender Vital Status

Showing 1 - 20 of 810 cases

Case ID Project Primary Site Gender Files Available Files per Data Category # Mutations # Genes Slides

Case ID	Project	Primary Site	Gender	Files	Seq	Exp	SNV	CNV	Meth	Clinical	Bio	# Mutations	# Genes	Slides
<a href="#">TCGA-BK-A6W3</a>	<a href="#">TCGA-UCEC</a>	Corpus uteri	Female	56	4	5	16	4	1	10	16	1	1	2
<a href="#">TCGA-GN-A8LK</a>	<a href="#">TCGA-SKCM</a>	Skin	Male	51	4	5	16	4	1	7	14	1	1	2
<a href="#">TCGA-A5-A1OF</a>	<a href="#">TCGA-UCEC</a>	Corpus uteri	Female	56	4	5	16	4	1	10	16	2	1	2
<a href="#">TCGA-L5-A8NM</a>	<a href="#">TCGA-ESCA</a>	Esophagus	Female	54	4	5	16	4	1	8	16	1	1	2
<a href="#">TCGA-BR-8591</a>	<a href="#">TCGA-STAD</a>	Stomach	Male	54	4	5	16	4	1	7	17	1	1	3
<a href="#">TCGA-II7-A9P7</a>	<a href="#">TCGA-KIRP</a>	Kidney	Male	52	4	5	16	4	1	8	14	1	1	2

Biospecimen Clinical JSON TSV Save/Edit Case Set

# Subject-specific repositories



Omics DI

Browse

Submit Data

Databases

API

Help ▾

Login

Organism, repository, gene, tissue, accession

Examples: Cancer, Homo sapiens, Orbitrap, Q9HAU5, Phospho, Hela, PXD001416

differentially further generated  
potential pathways more derived known  
extracted sequencing number  
experiments expression  
studies related overall tumor  
revealed including sample  
series novel analysis molecular  
obtained regulation samples  
important transcriptome patients  
following through disease target  
keywords mechanisms effects

Description  Sample

Data



# 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



[About ▾](#) [Partners](#) [Related resources](#) [Bulk downloads](#) [Submit data](#)

[Viral Sequences](#) [Host Sequences](#) [Expression](#) [Proteins](#) [Biochemistry](#) [Literature](#)

Accelerating research through data sharing

- 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	<i>Saccharomyces cerevisiae</i>	<a href="#">Saccharomyces Genome Database</a>
Fission yeast	<i>Schizosaccharomyces pombe</i>	<a href="#">PomBase</a>
Clawed frog	<i>Xenopus</i>	<a href="#">Xenbase</a>
Fruitfly	<i>Drosophila melanogaster</i>	<a href="#">FlyBase</a>
Mouse	<i>Mus musculus</i>	<a href="#">Mouse Genome Informatics</a>
Nematode	<i>Caenorhabditis elegans</i>	<a href="#">WormBase</a>
Rat	<i>Rattus norvegicus</i>	<a href="#">Rat Genome Database</a>
Social amoeba	<i>Dictyostelium discoideum</i>	<a href="#">DictyBase</a>
Thale cress	<i>Arabidopsis thaliana</i>	<a href="#">The Arabidopsis Information Resource</a>
Maize	<i>Zea mays ssp. mays</i>	<a href="#">MaizeGDB</a>
Zebrafish	<i>Danio rerio</i>	<a href="#">Zebrafish Information Network</a>
Yeast	<i>Candida albicans</i>	<a href="#">CGD</a>
Bacteria	<i>Escherichia coli</i>	<a href="#">EcoCyc</a>

# 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*)

# Tools



[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](#)

# 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

 National Library of Medicine  
*National Center for Biotechnology Information*

Log in

COVID-19 is an emerging, rapidly evolving situation.  
Get the latest public health information from CDC: <https://www.coronavirus.gov>.  
Get the latest research from NIH: <https://www.nih.gov/coronavirus>.  
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: <https://www.ncbi.nlm.nih.gov/sars-cov-2/>.

Search NCBI colon cancer X Search

Results found in 31 databases

Literature	Genes	Proteins
Bookshelf 8,260	Gene 4,087	Conserved Domains 48
MeSH 26	GEO DataSets 12,615	Identical Protein Groups 3,244
NLM Catalog 654	GEO Profiles 779,994	Protein 10,835
PubMed 142,473	HomoloGene 14	Protein Clusters 2
PubMed Central 311,396	PopSet 5	Sparcle 297
		Structure 421

# Tools



BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Using this website

Annotation and prediction

Data access

API & software

About us



Help & Documentation

API & Software

Ensembl Tools

## In this section

### Ensembl Variant Effect Predictor

- VEP web interface
- VEP command line
- Data formats
- Variant Recoder
- Haplosaurus
- VEP FAQ
- Variant Simulator
- VCF to PED Converter

Search documentation...

Go

## Ensembl Tools

We provide a number of ready-made tools for processing both our data and yours. We routinely delete results from our servers after 10 days, so you will need to save the results indefinitely.

### Processing your data

Name	Description	Online tool
<a href="#">Variant Effect Predictor</a> 	Analyse your own variants and predict the functional consequences of known and unknown variants via our Variant Effect Predictor (VEP) tool.	
<a href="#">BLAST/BLAT</a>	Search our genomes for your DNA or protein sequence.	
<a href="#">File Chameleon</a>	Convert Ensembl files for use with other analysis tools	
<a href="#">Assembly Converter</a>	Map (liftover) your data's coordinates to the current assembly.	
<a href="#">ID History Converter</a>	Convert a set of Ensembl IDs from a previous release into their current equivalents.	
<a href="#">Linkage Disequilibrium Calculator</a>	Calculate LD between variants using genotypes from a selected population.	
<a href="#">VCF to PED converter</a>	Parse a vcf file to create a linkage pedigree file (ped) and a marker information file, which together may be loaded into ld visualization tools like Haploview.	

# Tools

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

# Tools

## Galaxy Europe

Analyze Data Workflow Visualize ▾ Shared Data ▾ Help ▾ Login or Register  

### Tools

search tools  

- Get Data
- Send Data
- Collection Operations

**GENERAL TEXT TOOLS**

- Text Manipulation
- Filter and Sort
- Join, Subtract and Group

**GENOMIC FILE MANIPULATION**

- Convert Formats
- FASTA/FASTQ
- FASTQ Quality Control
- Quality Control
- SAM/BAM
- BED

### COVID-19 research!

Want to learn the best practices for the analysis of SARS-CoV-2 data using Galaxy? Visit the Galaxy SARS-CoV-2 portal at [covid19.galaxyproject.org](https://covid19.galaxyproject.org). We mirror **all public** SARS-CoV-2 data from ENA in a [Galaxy data library](#) for your convenience. The Galaxy community also created [COVID-19 related trainings](#) and we also maintain a [running document](#) with recent news. Our new preprint about [The landscape of SARS-CoV-2 RNA modifications](#) is out!

"Anyone, anywhere in the world should have free, unhindered access to not just my research, but to the research of every great and enquiring mind across the spectrum of human understanding." – Prof. Stephen Hawking

### News

- Nov 14, 2020  **UseGalaxy.eu Tool Updates for 2020-11-14**
- Nov 7, 2020  **UseGalaxy.eu Tool Updates for 2020-11-07**
- Nov 3, 2020  **November Galactic News!**
- Oct 31, 2020  **UseGalaxy.eu Tool Updates for**

### Events

- Jan 25, 2021 - Jan 29, 2021   **2021 Galaxy Admin Training**
- Dec 10, 2020   **Galaxy Developer Roundtable: Developer Training**
- Dec 7, 2020 - Dec 10, 2020  **Hackathon sur les outils interactifs de Galaxy (GxIT)**
- Dec 3, 2020   **DNA and DTA**

<https://usegalaxy.eu/>

# Tools

 Cytoscape App Store

Submit an App ▾ Search the App Store Sign In

All Apps

## 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

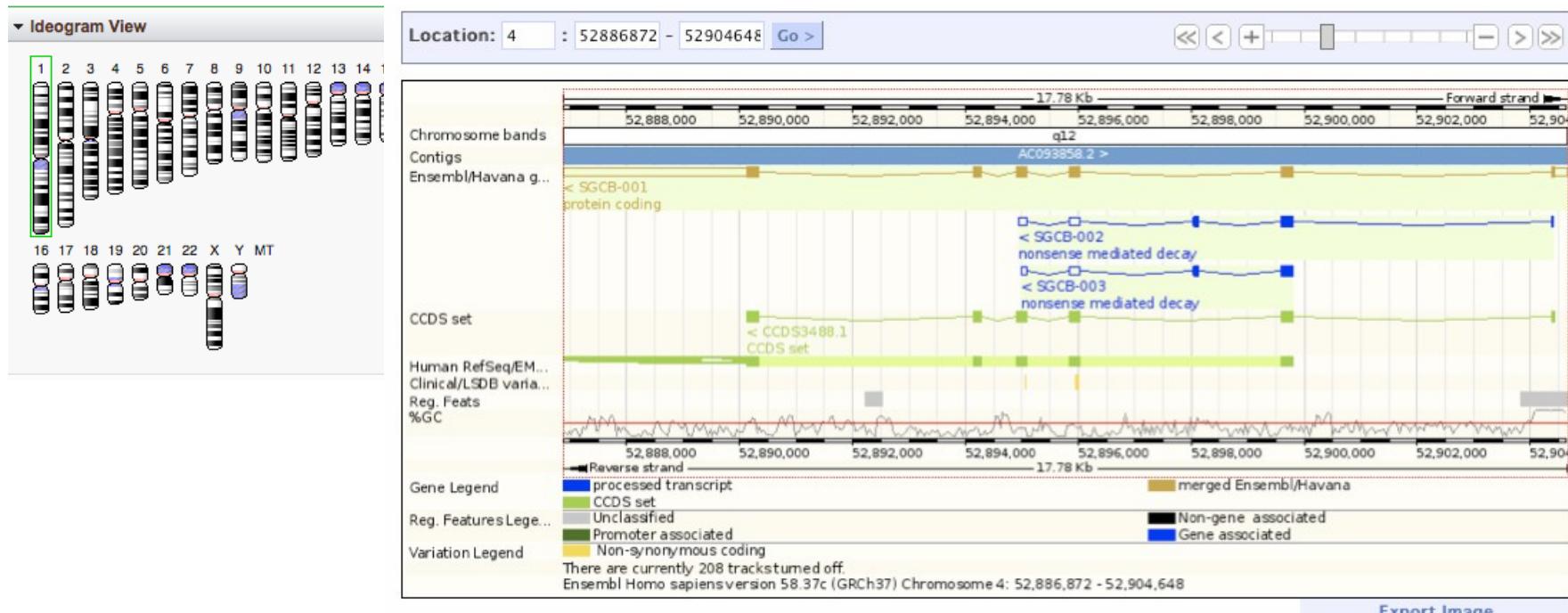
 **OmniPath** 3.0+  
OmniPath: literature curated human signaling pathways

[more newest releases »](#)

# 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

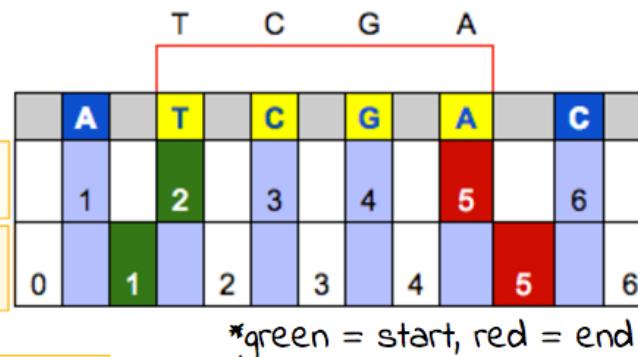
For mouse:

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



1-START, FULLY CLOSED (UCSC GB WEB)

0-START, HALF-OPEN (UCSC GB TABLES)

→ 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: chr1:12714000**1**-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 12714000**0** 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>

The screenshot shows the Ensembl homepage with several callout boxes highlighting specific features:

- Search for a gene, region of interest, disease, variant etc**: Points to the main search bar at the top.
- Select species of interest**: Points to the "All genomes" section where users can select a species.
- Search for a sequence with BLAST**: Points to the "Search for a sequence with BLAST" button.
- Get help**: Points to the "Get help" link in the top right.
- Search here too**: Points to the search bar in the top right corner.
- See the current release number and what's new**: Points to the "Ensembl Release 100 (April 2020)" section which lists updates.
- Export data with BioMart**: Points to the "Export data with BioMart" link.

**Ensembl Release 100 (April 2020)**

- Update to GENCODE 34 (human) and GENCODE M25 (mouse)
- Update of gnomAD genomic allele frequencies to version 3
- New genomes: 3 mammals, 7 fish, 6 birds, 4 reptiles
- Updated genomes: Platypus and Northern Pike
- New interface for configuration of multidimensional track hubs

[More release news](#) (if on our blog)

**Other news from our blog**

- 29 May 2020: [Ensembl under lockdown – Part 3 \(P\)](#)
- 28 May 2020: [Normalising variants to standardise Ensembl VEP output \(P\)](#)
- 21 May 2020: [Ensembl under lockdown – Part 2 \(P\)](#)

# 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.



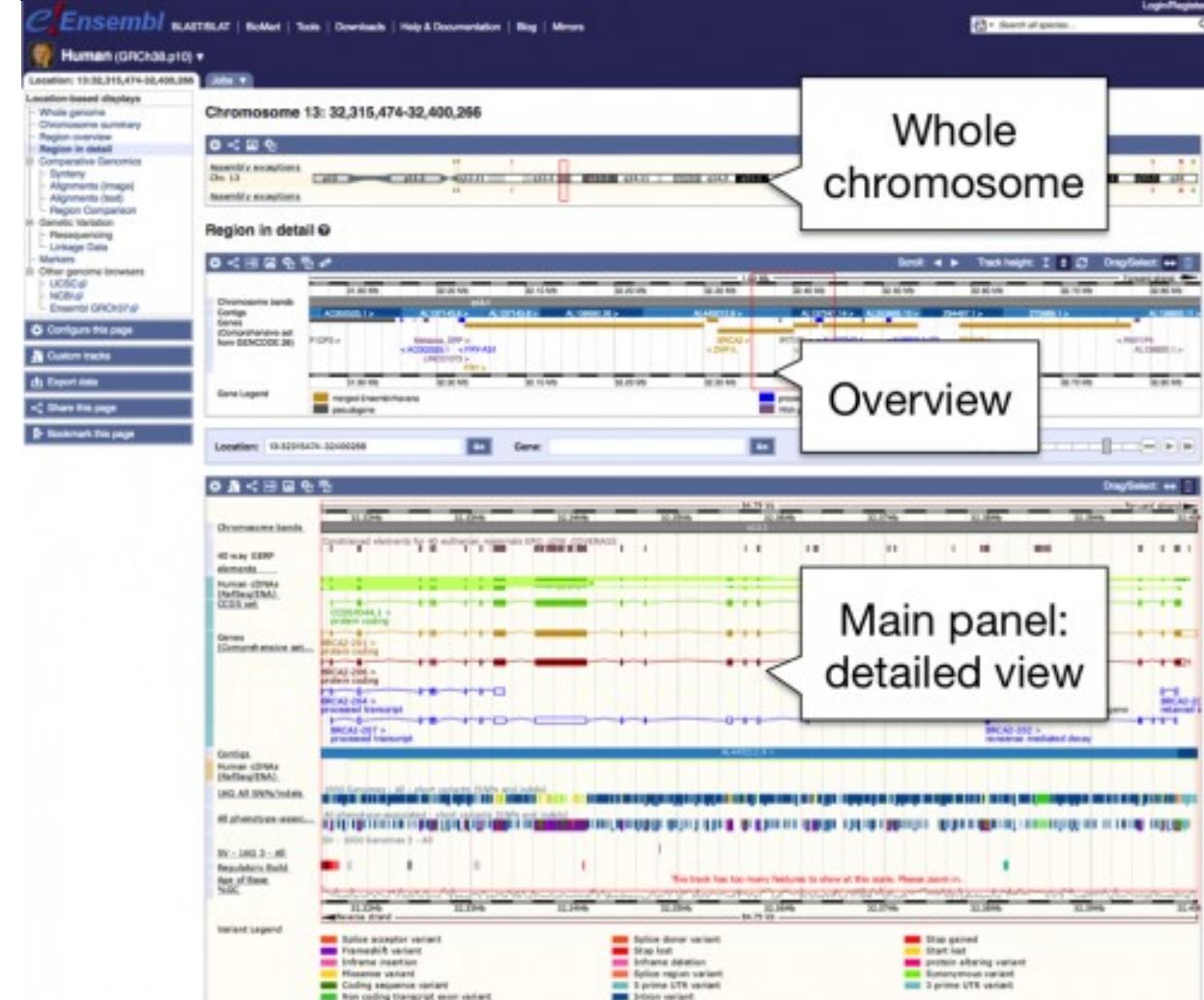
- The available annotations (tracks) to be displayed are configured through the 'Configure this page' button

A screenshot of the 'Configure this page' interface. At the top, there are four tabs: Configure Region Image, Configure Overview Image, Configure Chromosome Image, and Personal Data. The Configure Region Image tab is selected. Below the tabs, there is a section titled "Active tracks" with a list of available configurations. The "Sequence and assembly" section is expanded, showing options like Sequence, Markers, GRC alignments, Simple features, and Clones & misc. regions. Other sections include "Genes and transcripts", "mRNA and protein alignments", and "Variation". Each section has a list of available configurations with their counts in parentheses. On the right side, there is a "Select from available configurations:" dropdown menu with "Current unsaved configuration" selected.

# Genome Browsers

## Ensembl

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



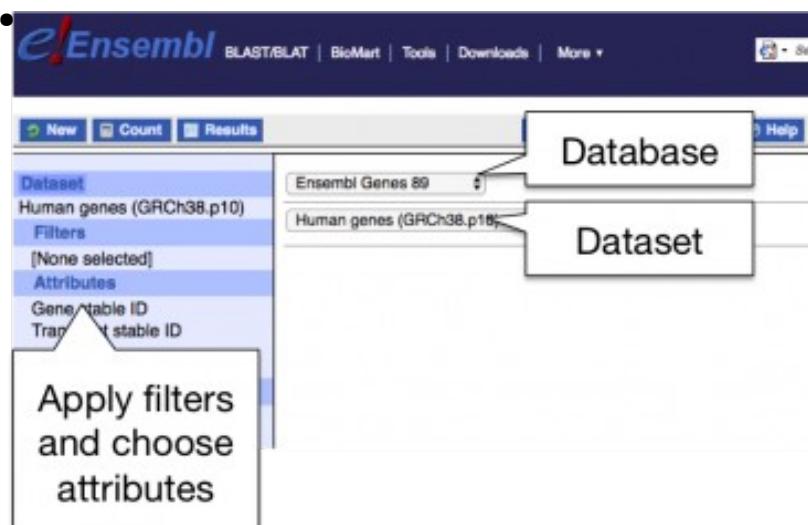
- The location tab's main display is the 'region in detail', which shows a region of the genome up to 1 Mb long in a highly customisable view

# Genome Browsers

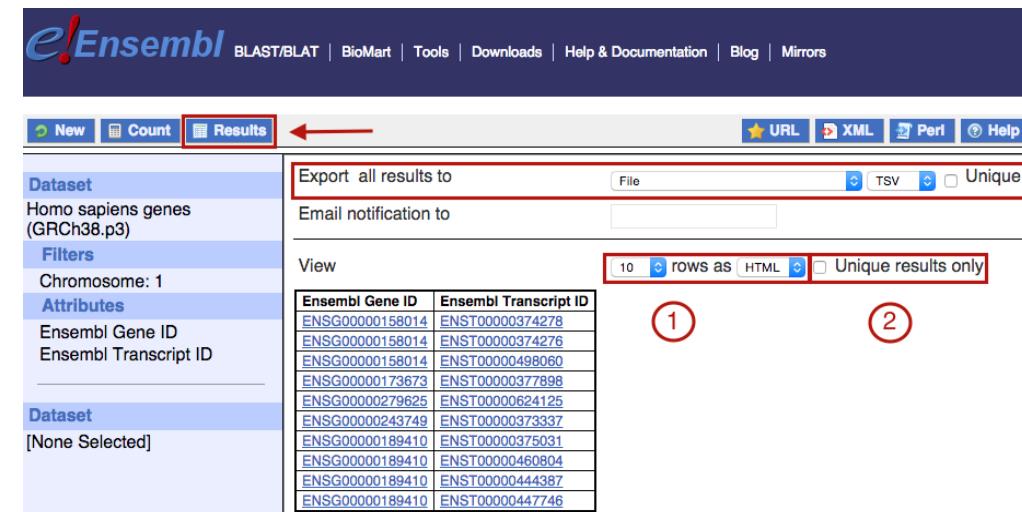
## Ensembl

<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 )



The screenshot shows the Ensembl BioMart search interface. At the top, there are tabs for 'New', 'Count', and 'Results'. The 'Results' tab is highlighted. Below the tabs, there are two main sections: 'Database' and 'Dataset'. The 'Database' section contains a dropdown menu set to 'Ensembl Genes 89'. The 'Dataset' section contains a dropdown menu set to 'Human genes (GRCh38.p10)'. On the left side, there is a sidebar with 'Dataset' and 'Filters' sections. The 'Dataset' section lists 'Human genes (GRCh38.p10)' and 'Human genes (GRCh38.p18)'. The 'Filters' section has a dropdown set to '[None selected]'. A large callout box labeled 'Database' points to the 'Database' section, and another callout box labeled 'Dataset' points to the 'Dataset' section. A text box at the bottom left says 'Apply filters and choose attributes'.



The screenshot shows the Ensembl BioMart search interface with the 'Results' tab selected. The results table displays a list of Ensembl Gene IDs and their corresponding Ensembl Transcript IDs. The table includes columns for 'Ensembl Gene ID' and 'Ensembl Transcript ID'. The first few rows of the table are:

Ensembl Gene ID	Ensembl Transcript ID
ENSG00000158014	ENST00000374278
ENSG00000158014	ENST00000374276
ENSG00000158014	ENST00000498060
ENSG00000173673	ENST00000377898
ENSG00000279625	ENST00000624125
ENSG00000243749	ENST00000373337
ENSG00000189410	ENST00000375031
ENSG00000189410	ENST00000460804
ENSG00000189410	ENST00000444387
ENSG00000189410	ENST00000447746

At the top right, there are export options: 'URL', 'XML', 'Perl', and 'Help'. Below the table, there are buttons for '10 rows as HTML' and 'Unique results only'. Red boxes highlight the 'Results' tab, the 'Export' section, and the 'Unique results only' button.

# Genome Browsers

## Ensembl

- Other tools: <https://www.ensembl.org/index.html>



[\*\*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/>



The image shows the homepage of the UCSC Genome Browser. At the top left is the logo for the University of California Santa Cruz Genomics Institute. To its right is the UCSC logo. The main title "Genome Browser" is prominently displayed. Below the title is a navigation bar with links for Home, Genomes, Genome Browser, Tools, Mirrors, Downloads, My Data, Help, and About Us. On the right side, there is a section titled "Our tools" which lists various genomic analysis tools: Genome Browser, BLAT, Table Browser, Variant Annotation Integrator, Data Integrator, Gene Sorter, Genome Browser in a Box (GBiB), In-Silico PCR, LiftOver, and Track Hubs, each with a brief description. A link "More tools..." is also present.

## Our tools

- **Genome Browser**  
interactively visualize genomic data
- **BLAT**  
rapidly align sequences to the genome
- **Table Browser**  
download data from the Genome Browser database
- **Variant Annotation Integrator**  
get functional effect predictions for variant calls
- **Data Integrator**  
combine data sources from the Genome Browser database
- **Gene Sorter**  
find genes that are similar by expression and other metrics
- **Genome Browser in a Box (GBiB)**  
run the Genome Browser on your laptop or server
- **In-Silico PCR**  
rapidly align PCR primer pairs to the genome
- **LiftOver**  
convert genome coordinates between assemblies
- **Track Hubs**  
import and view external data tracks

[More tools...](#)

# Genome Browsers

## UCSC Genome Browser

- 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)

**Find Position**

**Human Assembly**  
Dec. 2013 (GRCh38/hg38)

**Position/Search Term**  
Irrtm1  
Current position: chr3:52,221,080-52,226,163

**BLAT Search Genome**

Genome:  Search ALL      Assembly:      Query type:      Sort output:      Output type:

Human      Dec. 2013 (GRCh38/hg38)      BLAT's guess      query,score      hyperlink

submit      I'm feeling lucky      clear

Paste in a query sequence to find its location in the genome. Multiple sequences may be searched if separated by lines starting with '>' followed by the sequence name.

**File Upload:** Rather than pasting a sequence, you can choose to upload a text file containing the sequence.  
Upload sequence:  No file selected.

# 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

[NRXN1 \(ENST00000404971.5\) at chr2:49920350-51032399](#) - Homo sapiens neurexin 1 (NRXN1), transcript variant alpha2,  
[NRXN1 \(ENST00000625672.2\) at chr2:49918505-51028456](#) - Cell surface protein involved in cell-cell-interactions, e

### NCBI RefSeq genes, curated subset (NM\_\*, NR\_\*, and YP\_\*)

[NM\\_178839.4 at chr2:80301878-80304362](#)

### NCBI RefSeq genes, predicted subset (XM\_\* and XR\_\*)

[XM\\_017003986.1 at chr2:80302014-80304738](#)

[XM\\_017003987.1 at chr2:80302014-80304738](#)

### RefSeq Genes

[LRRTM1 at chr2:80301878-80304362](#) - (NM\_178839) leucine-rich repeat transmembrane neuronal protein 1 precursor

### Non-Human RefSeq Genes

[LRRTM1 at chr2:80301917-80304282](#) - (NM\_001257467) leucine-rich repeat transmembrane neuronal protein 1 precursor  
[LRRTM1 at chr2:80302082-80304737](#) - (NM\_001080304) leucine-rich repeat transmembrane neuronal protein 1 precursor  
[LRRTM1 at chr2:80288477-80304427](#) - (NM\_001133111) leucine-rich repeat transmembrane neuronal protein 1 precursor  
[LRRTM1 at chr2:80288876-80304610](#) - (NM\_001109374) leucine-rich repeat transmembrane neuronal protein 1 precursor  
[Lrrtm1 at chr2:80301870-80304896](#) - (NM\_028880) leucine-rich repeat transmembrane neuronal protein 1 precursor  
[Lrrtm1 at chr2:80301870-80304896](#) - (NM\_001362109) leucine-rich repeat transmembrane neuronal protein 1 precursor  
[Lrrtm1 at chr2:80287776-80304896](#) - (NR\_155300)  
[Lrrtm1 at chr2:80287776-80304896](#) - (NR\_155299)

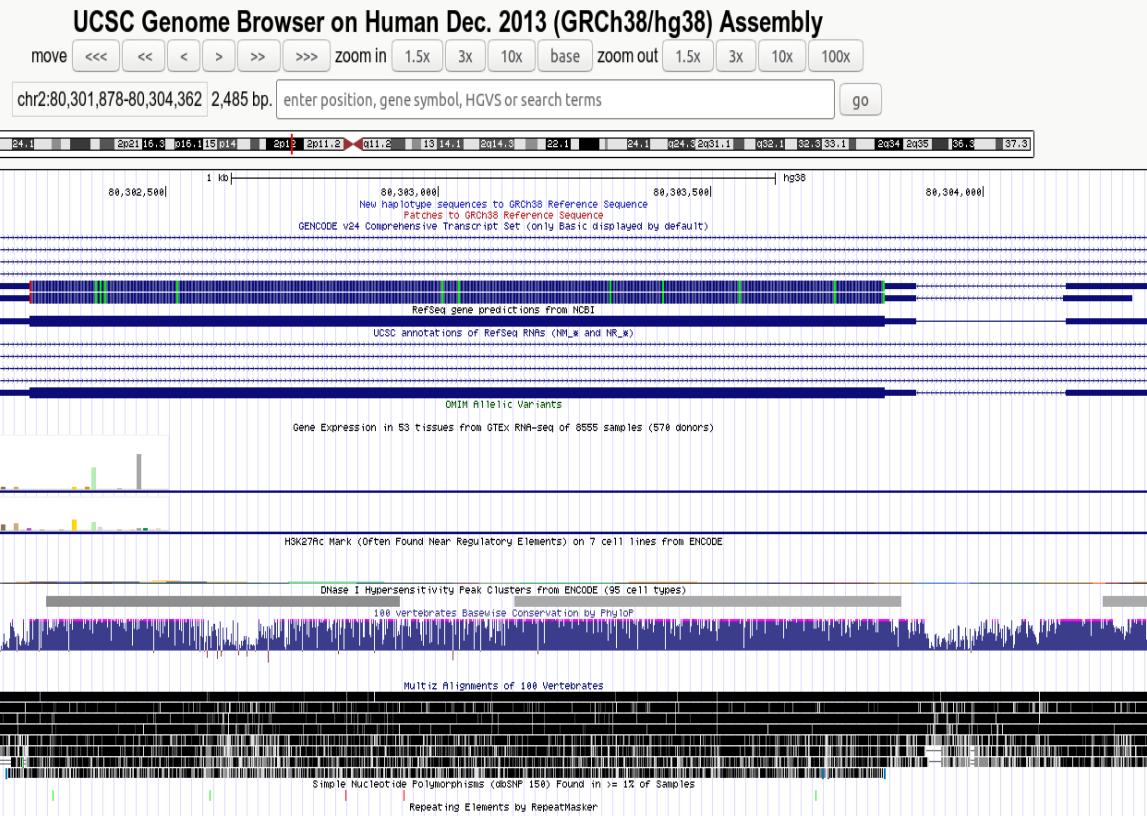
### Basic Gene Annotation Set from GENCODE Version 28 (Ensembl 92)

[LRRTM1 at chr2:80301878-80304749](#)

[LRRTM1 at chr2:80301878-80304737](#)

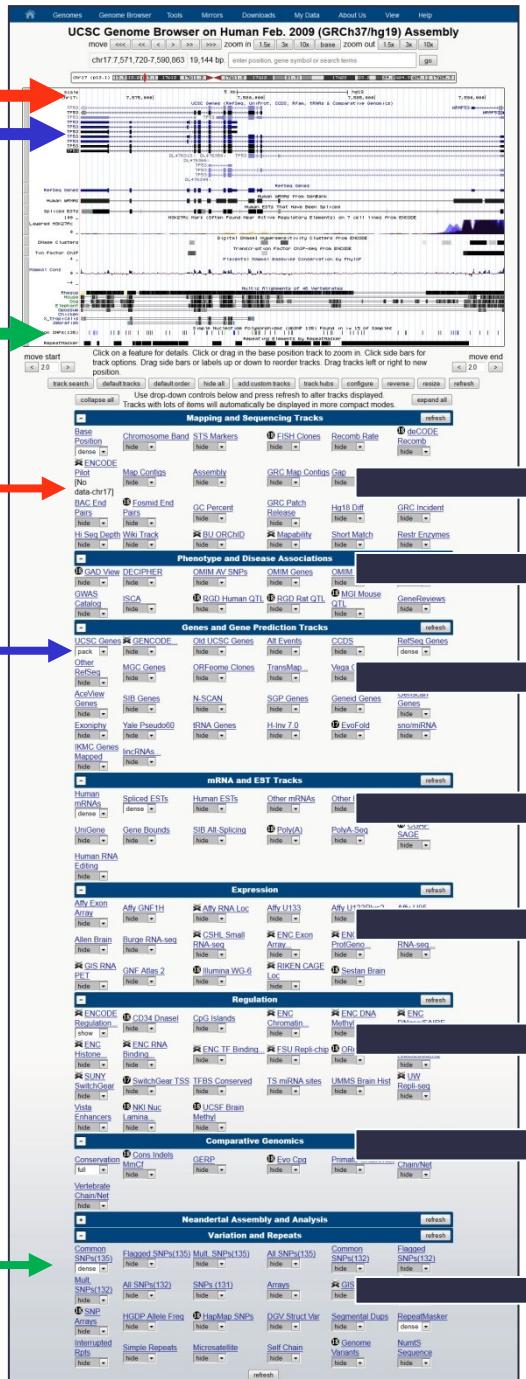
# Genome Browsers

- 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)**

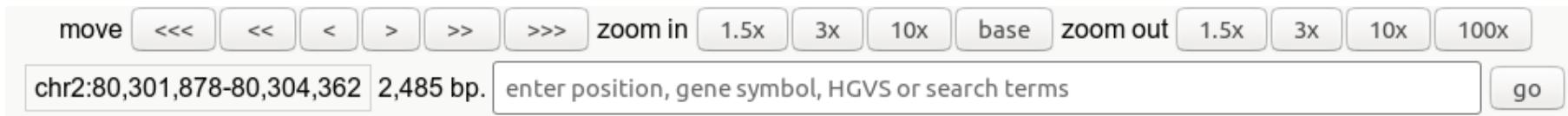
→ **Regulation (including TFBS)**

→ **Comparative Genomics**  
*As a group*  
*Individual species*

→ **Variation and Repeats**  
*(including SNPs, copy number variation)*

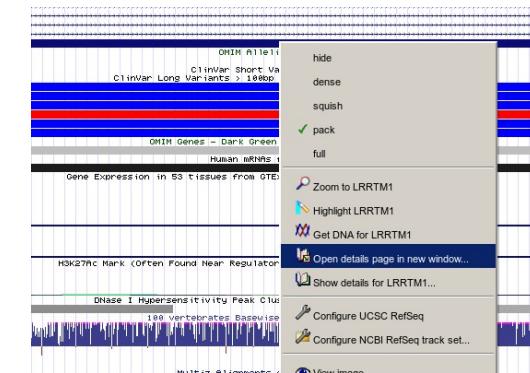
# Genome Browsers

- Change your view or location with controls at the top

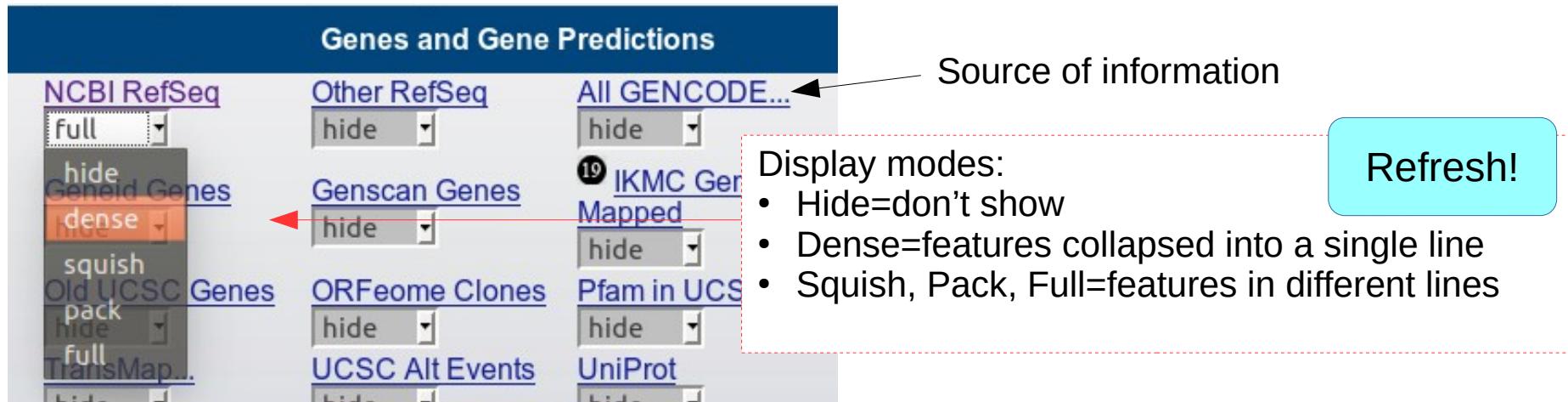


A screenshot of a genome browser interface. At the top, there is a set of buttons for navigating the genome: move (<<<, <<, <, >, >>, >>>), zoom in (1.5x, 3x, 10x, base), and zoom out (1.5x, 3x, 10x, 100x). Below these are two input fields: one containing "chr2:80,301,878-80,304,362 2,485 bp." and another for "enter position, gene symbol, HGVS or search terms". A "go" button is located to the right of the search field.

- 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



A screenshot of a genome browser interface showing a list of tracks and their display modes. The tracks include:

- NCBI RefSeq: full, hide, Geneid Genes, dense, squish, Old UCSC Genes, pack, full, TransMap.
- Other RefSeq: hide, Genscan Genes, ORFeome Clones, UCSC Alt Events.
- All GENCODE...: hide, 19 IKMC Genes Mapped, Pfam in UCSB, UniProt.

An arrow points from the "Source of information" text to the "All GENCODE..." track. Another arrow points from the "Display modes:" text to the "Geneid Genes" entry in the NCBI RefSeq section. A red dashed box highlights the "Display modes:" section, which contains the following text:

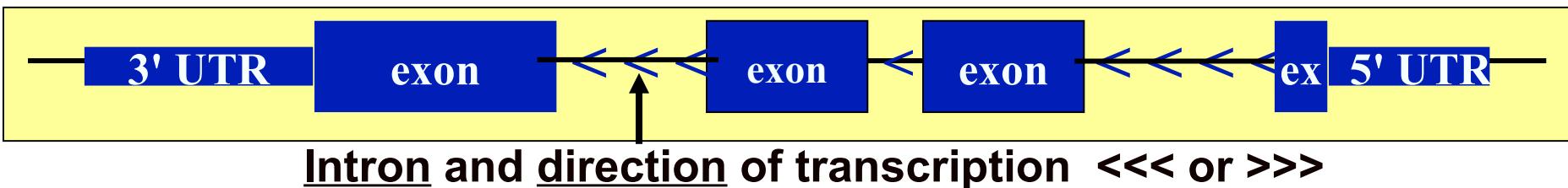
Display modes:

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

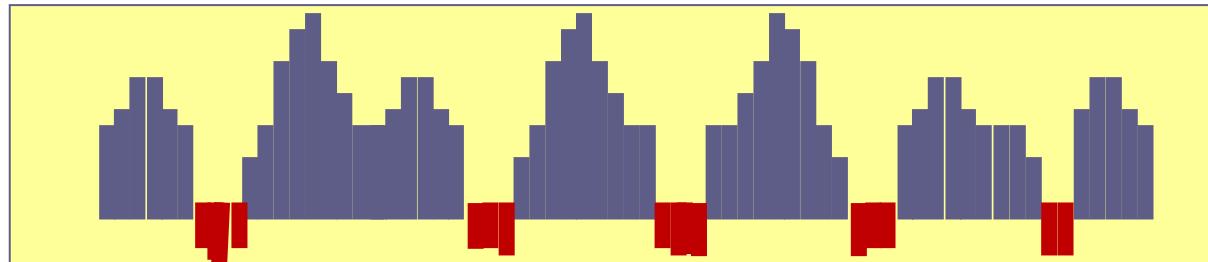
A blue callout bubble with the text "Refresh!" is located in the bottom right corner.

# Genome Browsers

- Some visual clues:



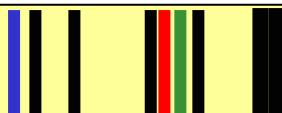
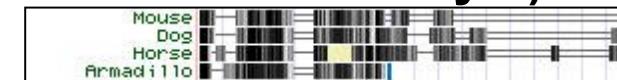
Sequence conservation



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



Tick marks; a single location (STS, SNP)

# Practicum

## Retrieving information with the UCSC Genome Browser

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

1. What is the genomic localization of human *Irrtm1* 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 *Irrtm1* gene to localize this gene in mouse genome. (Use BLAT)

# Practicum

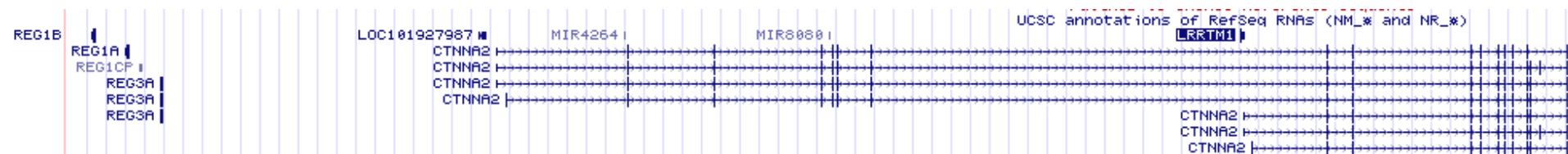
1. What is the genomic localization of human *Irrtm1* gene?

- 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



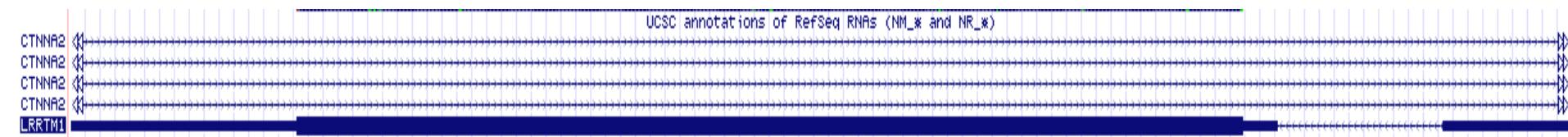
3. How many exons has the gene? 2

- 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



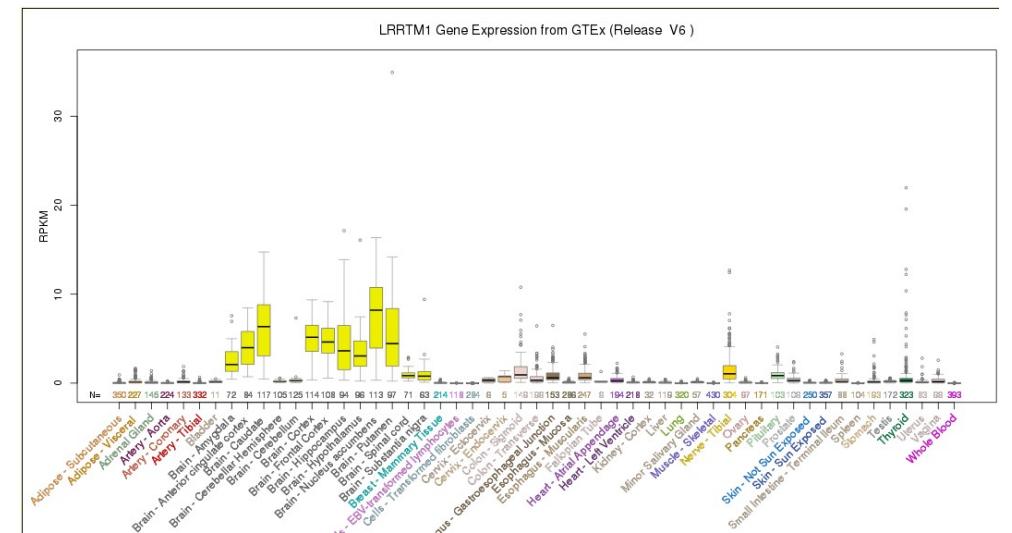
5. Can you find SNPs in this gene? yes

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



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

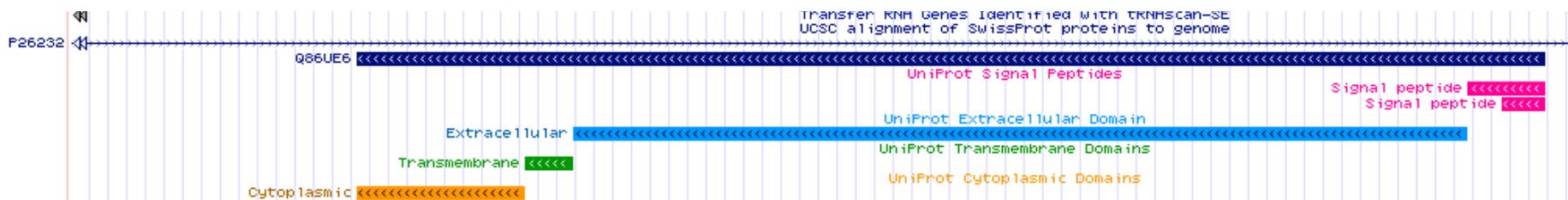
- Info in track “Expression” > “GTEx”



# Practicum

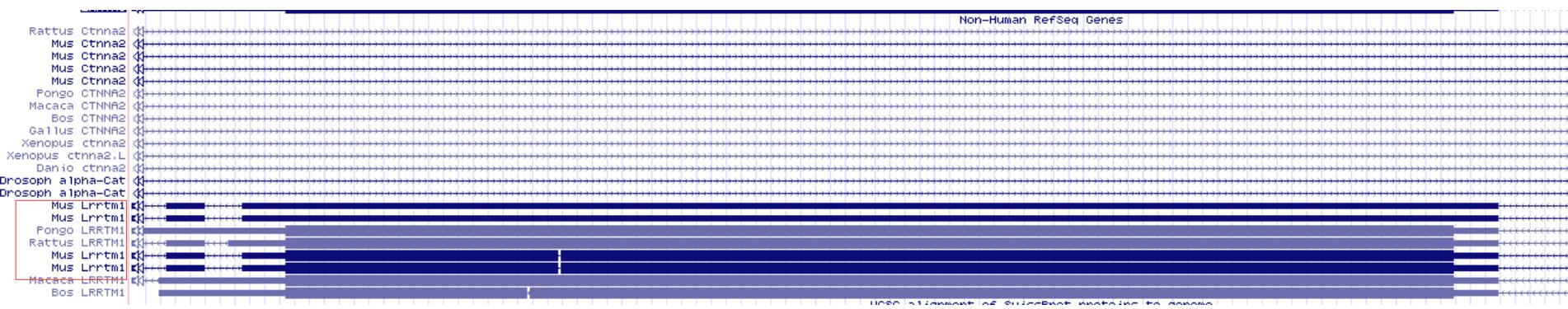
7. Does the protein encoded by this gene have a transmembrane domain?

- 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 *Irrtm1* gene to localize this gene in mouse genome.  
 (Use BLAT)

- Get the CDS sequence of human gene:

## Get Genomic Sequence Near Gene

Note: if you would prefer to get DNA for more than one feature of the

### Sequence Retrieval Region Options:

- Promoter/Upstream by  bases
- 5' UTR Exons
- CDS Exons
- 3' UTR Exons
- Introns
- Downstream by  bases
- One FASTA record per gene.
- One FASTA record per region (exon, intron, etc.) with
- Split UTR and CDS parts of an exon into separate FASTA records

Note: if a feature is close to the beginning or end of a chromosome it

- Use BLAT to align this sequence on Mouse genome:

## BLAT Search Genome

Genome:  Search ALL

Assembly:

Query type:

Sort output:

Mouse

Dec. 2011 (GRCm38/mm10)

DNA

query,score

```

ATGGATTTCTCTGCTCGTCTCTGTCTATACTGGCTGCTGAGGGAGGC
CTCGGGGGTGGCTTGTGCTGCTGGGGGCTGCTTCAGATGCTGCCG
CCGCCCCCAGCGGGTGCAGCTGTGCCGTGCGAGGGGGCGCTGCTG
TACTCGGAGGGCCTCAACCTCACCGAGGCGCCAACACCTGTCCGGCT
GCTGGCTTGTCCCTGCCTACAAACAGCTCTCGAGCTGCCGCCGGCC
AGTTCACGGGTTAATGAGCTCACGGCTCTATCTGGATACAATCAC
ATCTGCTCGTGAGGGGACGCCCTTCAGAAACTGCCGAGTTAAGGA
ACTCACGCTGAGTCCAACCAGATACCCAACCTGCCAACACACCCTCC
GGCCATGCCAACCTGGCAGCGGACCTCTCGTACAACAAGCTGCGAG
GGCTCGGCCGACCTCTCCACGGGCTGCCGAAGCTCACACGCTGCA
TATGCGGGCCAACGCCATTCCAGTTGTGCCGTGCGCATCTTCAGGACT
GCCGAGCTCAAGTTCTGACATGGATACAATCAGCTAAGAGCTG
GCCGCGCAACTTTGCCGGCTTAAAGCTACCGAGCTGCCACCTCGA
GCACAAAGCACTGGTCAAGGTGAACCTCGCCACTCCGCCCTCATCT
CCCTGCACTGCTCTGCCGGAGGAACAGGTGGCATTGGGGTACG

```

submit

I'm feeling lucky

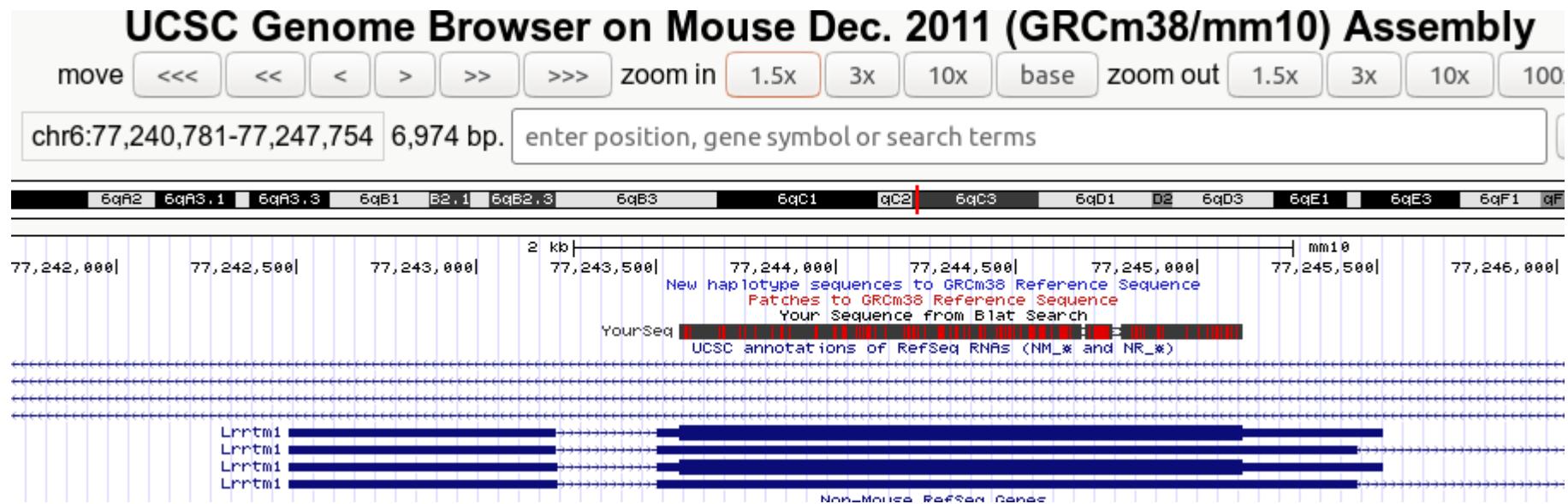
clear

# Practicum

- Visualize entry with highest score

→

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

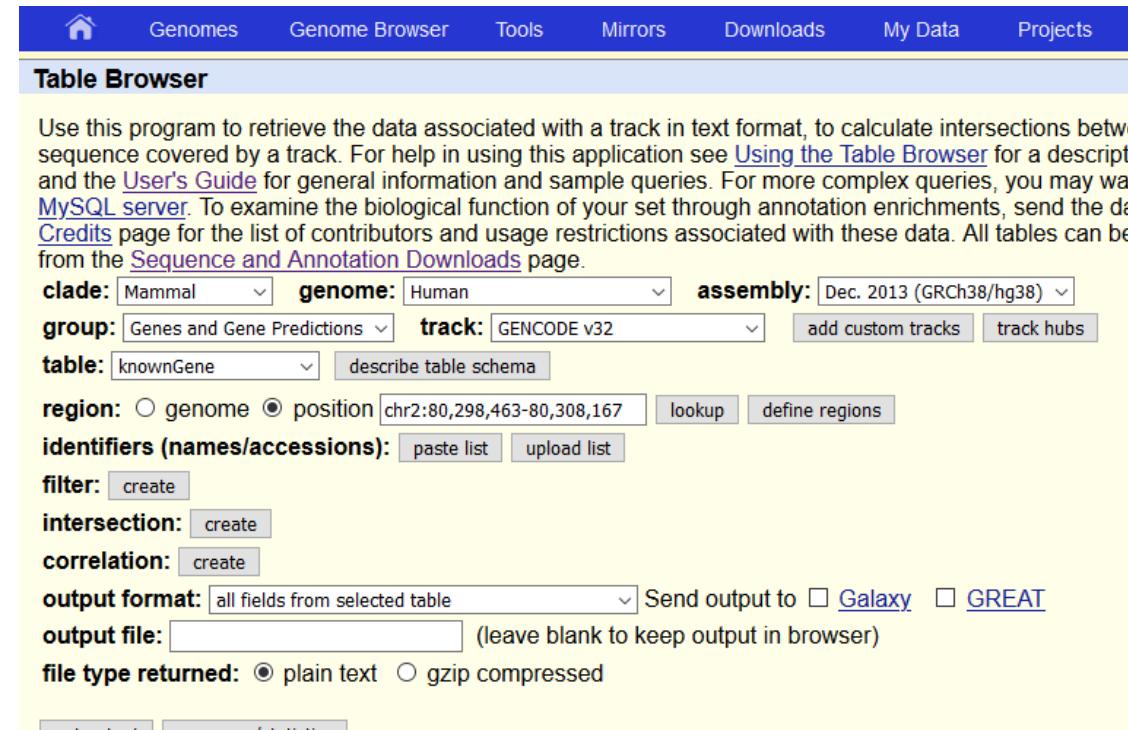


# Genome Browsers

## 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



The screenshot shows the UCSC Table Browser interface. At the top, there is a navigation bar with links for Home, Genomes, Genome Browser, Tools, Mirrors, Downloads, My Data, and Projects. Below the navigation bar, the title "Table Browser" is displayed. The main content area contains a text block explaining the purpose of the browser, followed by a form for specifying search parameters. The form includes fields for clade (Mammal), genome (Human), assembly (Dec. 2013 (GRCh38/hg38)), group (Genes and Gene Predictions), track (GENCODE v32), and table (knownGene). There are also buttons for "add custom tracks" and "track hubs". Below these, there are sections for "region" (with radio buttons for "genome" and "position" and a specific coordinate input), "identifiers (names/accessions)" (with "paste list" and "upload list" buttons), "filter" (with a "create" button), "intersection" (with a "create" button), "correlation" (with a "create" button), "output format" (set to "all fields from selected table"), "Send output to" (checkboxes for Galaxy and GREAT), "output file" (input field with a note about leaving it blank), and "file type returned" (radio buttons for "plain text" and "gzip compressed"). At the bottom, there are "get output" and "summary/statistics" buttons.

## Examples

To reset **all** user cart settings (including custom tracks), [click here](#).

# Practicum

## 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

#### Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, sequence covered by a track. For help in using this application see [Using the Table Browser](#) for a description of the c and the [User's Guide](#) for general information and sample queries. For more complex queries, you may want to use [G MySQL server](#). To examine the biological function of your set through annotation enrichments, send the data to [GRE Credits](#) page for the list of contributors and usage restrictions associated with these data. All tables can be download from the [Sequence and Annotation Downloads](#) page.

clade: Mammal    genome: Human    assembly: Dec. 2013 (GRCh38/hg38)

group: Variation    track: dbSNP 153    add custom tracks    track hubs

table: Common dbSNP(153) (dbSnp153Common)    describe table schema

**Note:** Most dbSNP tables are huge. Trying to download them through the Table Browser usually leads to a timeout. Please see our [Data Access FAQ](#) on how to download dbSNP data.

region:  genome  position chr2:80,301,878-80,304,752    lookup    define regions

identifiers (names/accessions):

filter:

subtrack merge:

intersection:

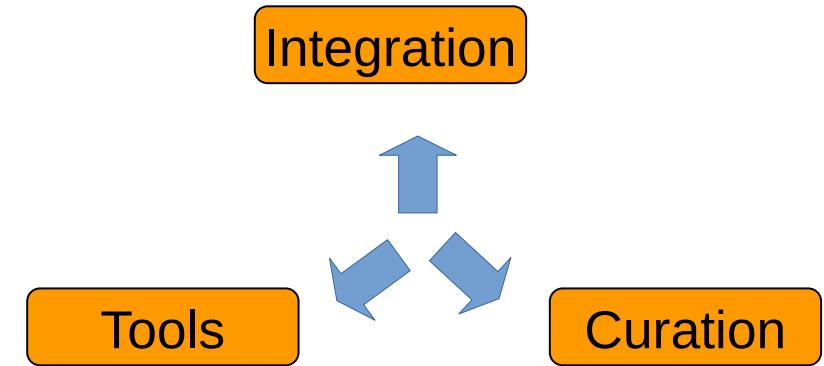
output format:  Send output to  Galaxy  GREAT

# 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

