

Databases in molecular biology (part II)

Bioinformatics Course UEB-VHIR
November 2020

Ricardo Gonzalo¹, Mireia Ferrer¹, Álex Sánchez^{1,2}
Berta Miró¹, 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...

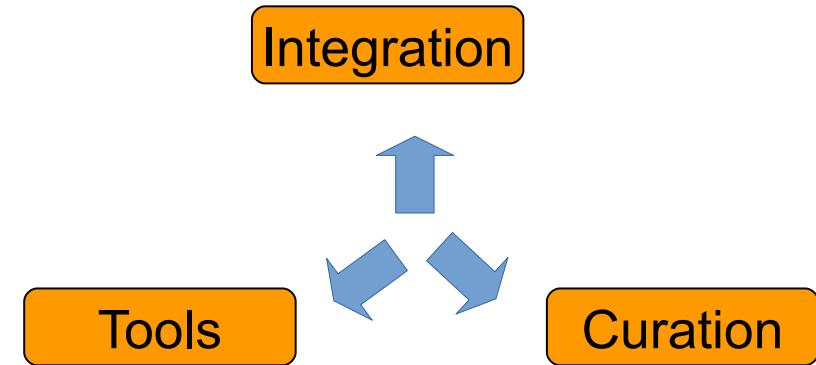
- **Resource Providers**



NCBI 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



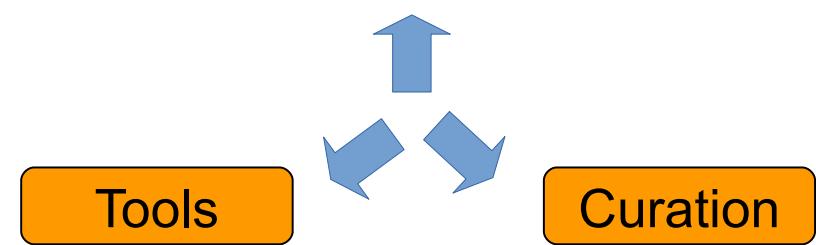
DNA & RNA	Gene Expression
Structures	Systems
Ontologies	Literature
Proteins	
Chemical biology	
Cross domain	



Summary

In this session

- General
 - Resource providers Integration
- 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

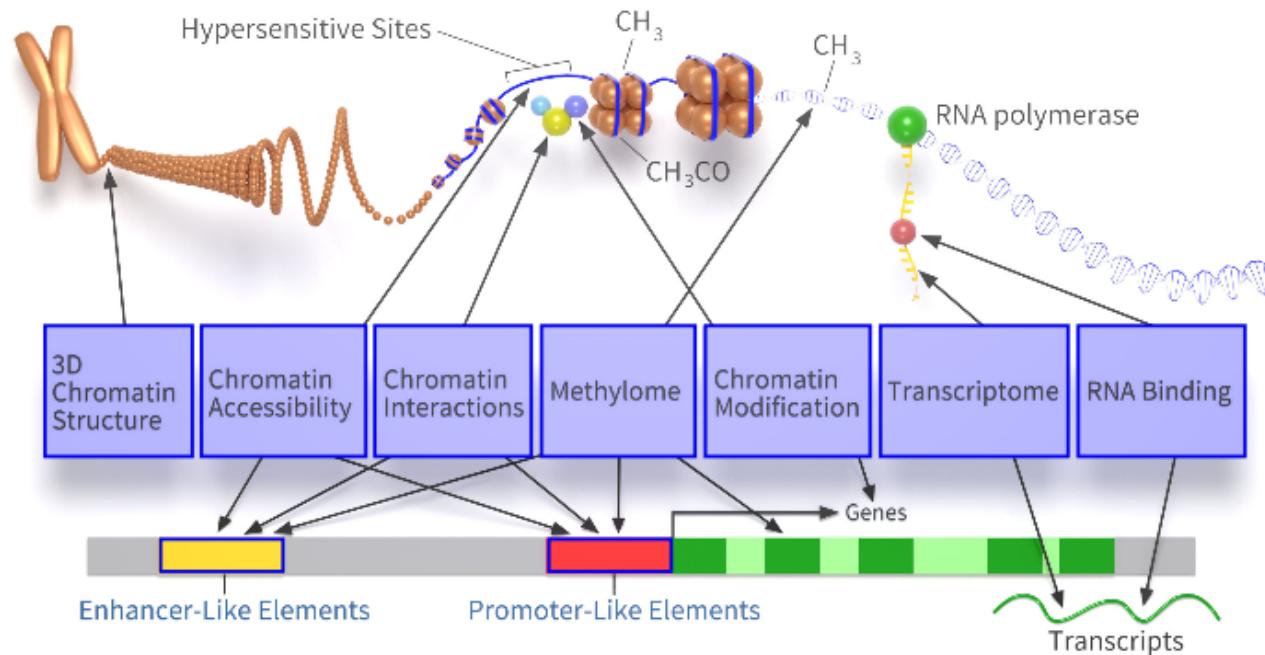
- Collective initiatives: summing efforts from different researchers from different sites of the world.
- **Integrate** different types of data (clinical, multi-omics,...) related to a **specific subject** (or sub-subject)
 - A disease
 - An organism
 - A biological question, tissue, ...
- Provide curated,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

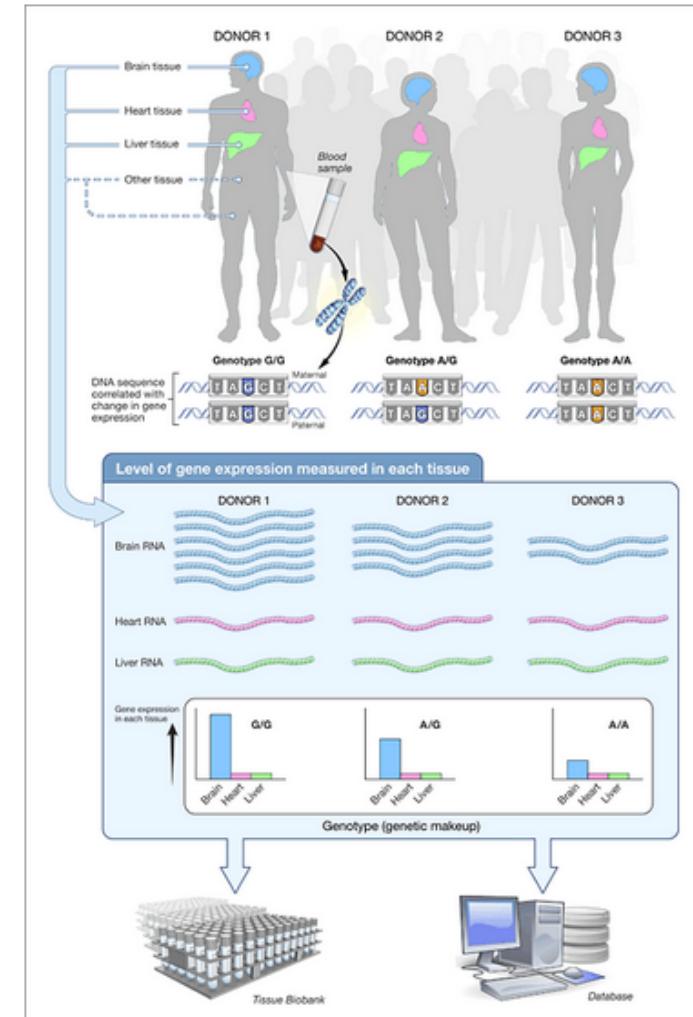
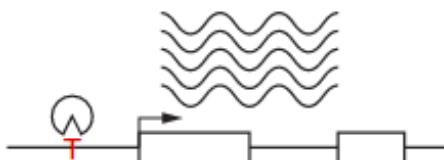


Subject-specific repositories

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.

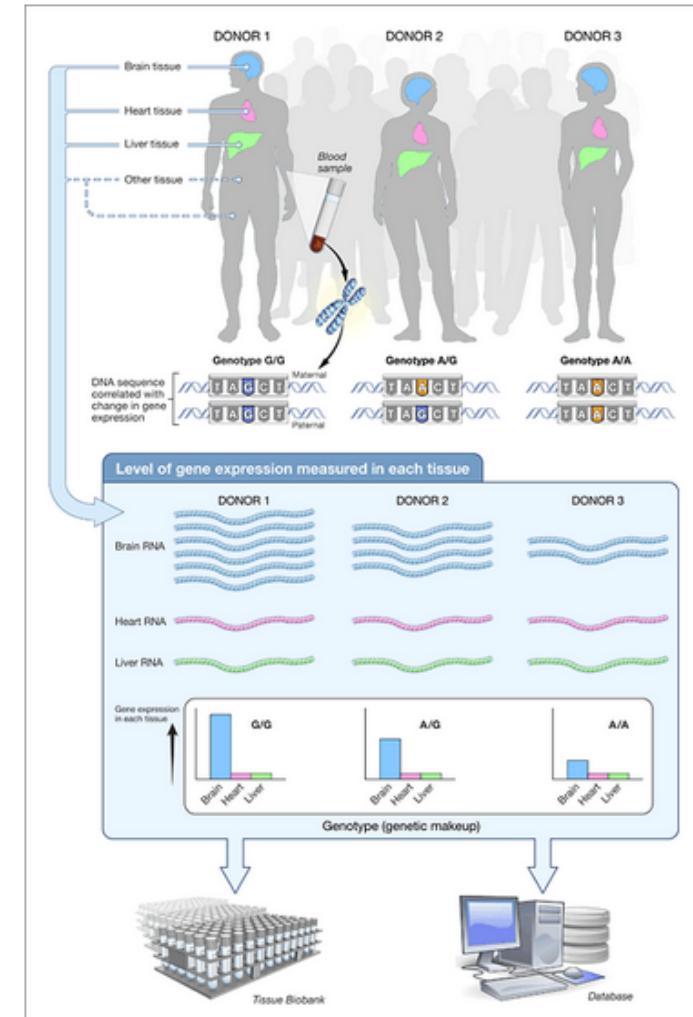


Subject-specific repositories

Genotype-Tissue Expression (GTEx) Project

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

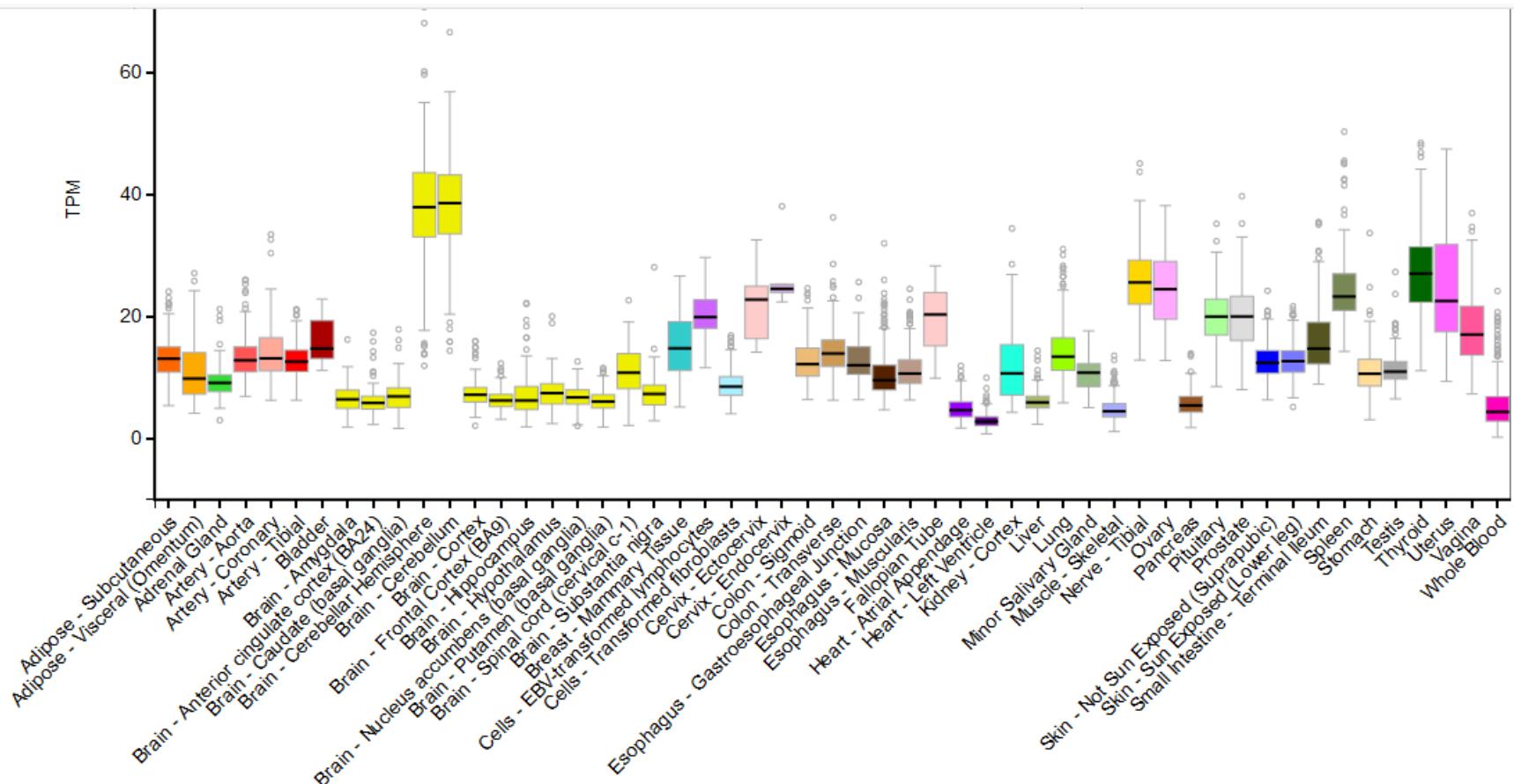
- Types of data provided:
 - Gene/transcript expression in tissues
 - variants associated to gene expression (eQTLs and sQTLs)
 - histology images
 - Patient/sample metadata
- Available datasets:
<https://www.gtexportal.org/home/datasets>
- Summary statistics:
<https://gtexportal.org/home/tissueSummaryPage>



Practicum

Genotype-Tissue Expression (GTEx) Project

Example: Normal tissue expression profile of *mutyh* gene

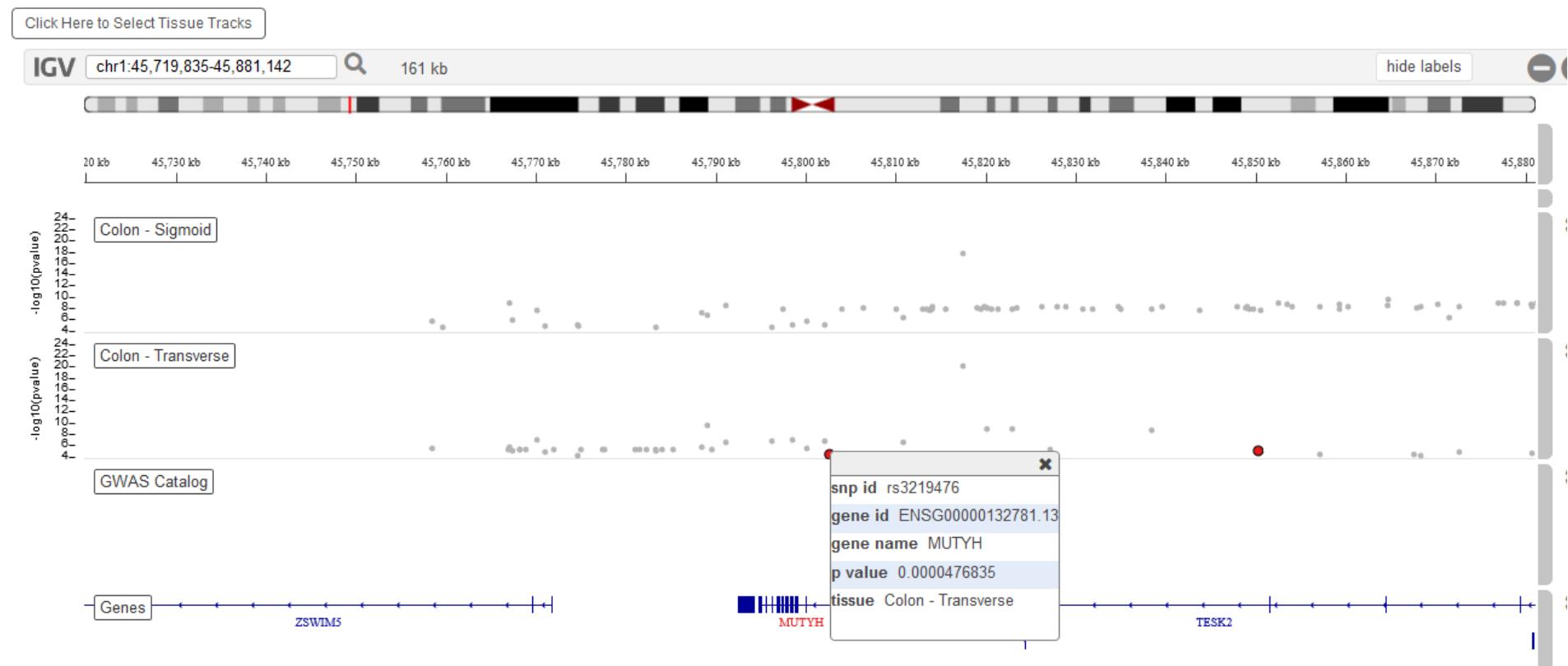


Practicum

Genotype-Tissue Expression (GTEx) Project

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

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)	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

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

NIH 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 Advanced Search

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...

358,092 33,096

Primary Site Project Disease Type Gender Vital Status

Showing 1 - 20 of 33,096 cases

Cart Case ID Project Primary Site Gender Files Available Files per Data Category Annotations Seq Exp SNV CNV Meth Clinical Bio

TCGA-AF-3912 TCGA-READ Rectosigmoid junction -- 20 0 0 0 0 0 0 8 12 4

This screenshot shows the GDC Data Portal interface. At the top, there's a navigation bar with links for Home, Projects, Exploration, Analysis, Repository (which is currently selected), Quick Search, Manage Sets, Login, Cart (containing 0 items), and GDC Apps. Below the navigation is a search bar with the placeholder 'Start searching by selecting a facet'. To the left, there's a sidebar with sections for 'File' (containing a search input and a file example 'e.g. 142682.bam, 4f6e2e7a-b...') and 'Data Category' (listing various types of genomic data with their counts: Simple Nucleotide Variation (127,390), Transcriptome Profiling (57,685), Biospecimen (55,223), Raw Sequencing Data (47,248), Copy Number Variation (45,256), and three more categories). To the right, there are several visualizations: four pie charts showing the distribution of Primary Site, Project, Disease Type, and Gender across the dataset; and a table at the bottom showing the count of available files per data category (Seq, Exp, SNV, CNV, Meth, Clinical, Bio) for the TCGA-AF-3912 project. The table also includes annotations for each category.

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.

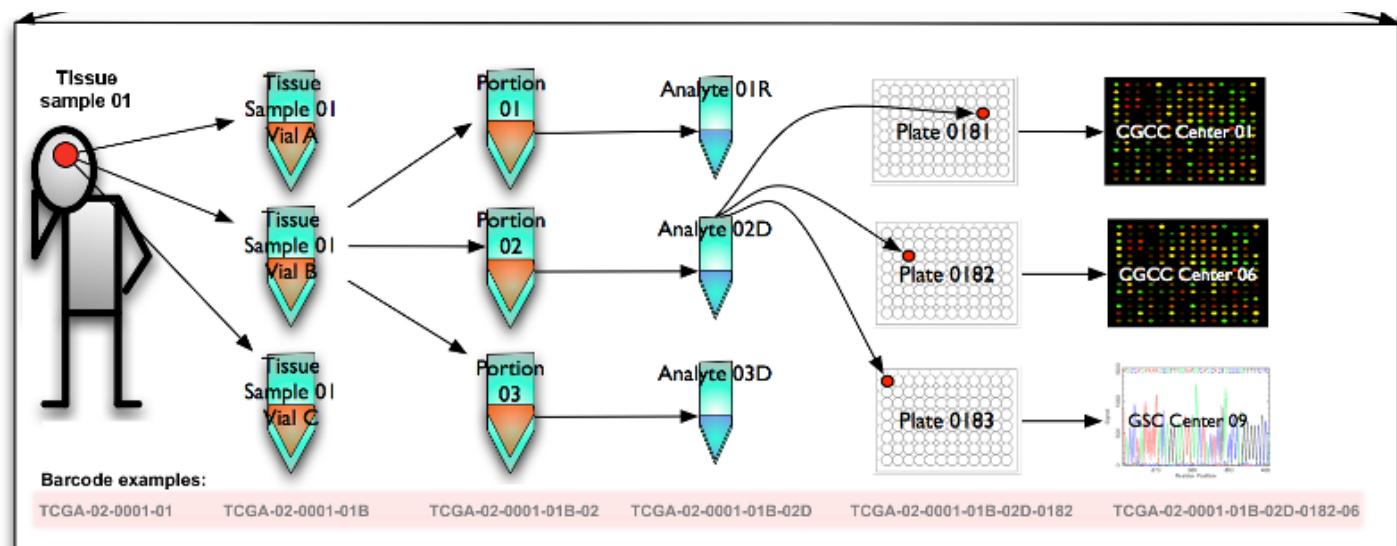
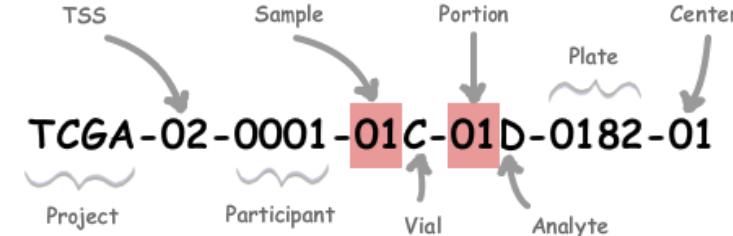
Subject-specific repositories

The Cancer Genome Atlas (TCGA)

<https://cancergenome.nih.gov/>

Data in <https://portal.gdc.cancer.gov/>

- To take into account when downloadina 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

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

Biospecimen Clinical JSON TSV Save/Edit Case Set

Case ID	Project	Primary Site	Gender	Available Files per Data Category										# Mutations	# Genes	Slides
				Seq	Exp	SNV	CNV	Meth	Clinical	Bio						
TCGA-BK-A6W3	TCGA-UCEC	Corpus uteri	Female	56	4	5	16	4	1	10	16			1	1	(2)
TCGA-GN-A8LK	TCGA-SKCM	Skin	Male	51	4	5	16	4	1	7	14			1	1	(2)
TCGA-A5-A1OF	TCGA-UCEC	Corpus uteri	Female	56	4	5	16	4	1	10	16			2	1	(2)
TCGA-L5-A8NM	TCGA-ESCA	Esophagus	Female	54	4	5	16	4	1	8	16			1	1	(2)
TCGA-BR-8591	TCGA-STAD	Stomach	Male	54	4	5	16	4	1	7	17			1	1	(3)
TCGA-UZ-A9PZ	TCGA-KIRP	Kidney	Male	52	4	5	16	4	1	8	14			1	1	(2)

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

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

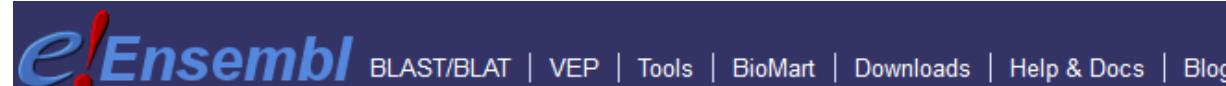


Fast sensitive protein homology searches using profile hidden Markov models (HMMs) for querying against both sequence and HMM target databases.

[Sequence similarity search](#)

[Protein function analysis](#)

Tools



Using this website Annotation and prediction Data access API & software About us

[Home](#) > [Help & Documentation](#) > [API & Software](#) > [Ensembl Tools](#)

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 please be able to save the results indefinitely.

Processing your data

Name	Description	Online tool
Variant Effect Predictor 	Analyse your own variants and predict the functional consequences of known and unknown variants via our Variant Effect Predictor (VEP) tool.	
BLAST/BLAT	Search our genomes for your DNA or protein sequence.	
File Chameleon	Convert Ensembl files for use with other analysis tools	
Assembly Converter	Map (liftover) your data's coordinates to the current assembly.	
ID History Converter	Convert a set of Ensembl IDs from a previous release into their current equivalents.	
Linkage Disequilibrium Calculator	Calculate LD between variants using genotypes from a selected population.	
VCF to PED converter	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

Search:

[Home](#) [Install](#) [Help](#) [Developers](#) [About](#)

[Home](#) » BiocViews

All Packages

Bioconductor version 3.12 (Release)

Autocomplete biocViews search:

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

Packages found under FunctionalGenomics:

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

Show [All](#) [▼](#) entries

Search table:

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

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. 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 »

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](#)

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

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

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

PathLinker 3.0+
Reconstructs signaling pathways from protein interaction networks

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

OmniPath 3.0+
OmniPath: literature curated human signaling pathways

more newest releases »

Tools



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



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*

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] Hide index list

Result list (partial):

- 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

Buttons: Search, 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 in Homo sapiens. The first result is for "Homo sapiens mutY DNA glycosylase (MUTYH), transcript variant 13, mRNA". It provides 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. To the left, a sidebar lists various filters: Species (Animals, Customize...), Molecule types (mRNA, RefSeq), Source databases (INSDC, GenBank), Sequence length (Custom range...), and Release date (Custom range...). A red curly brace on the left groups the "Species" and "Molecule types" sections under the heading "Using filters".

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 13, mRNA
CAGCCGGAGCCCGGGTACAACGGAACCTGTAGTCTCTCGTGGCTAGTTCAAGCGGAAGGGAGCAGTC
TCTGAAGCTTGAGGAGCCTCTAGAACTATGAGCCGAGGCCCTCCCTCTCCAGAGGCCAGAGGCTT
AAGGCTACTCTGGGAAGCCGCTCACCGCTCGAGCTGCGGGAGCTGAAACTGCGCCATCGTCAGTGTG
GCGGCATGACACCGCTCGTCTCCGCGTGAAGCTGCTGTGGGCATCATGAGGAAGGCCAGGAGCAGCC
TGGGAAGTGGTACAGGAAGCAGGCCAGGAGCAGAGCATGTAAGAACAAACAGTC
GGCCAAGCCTTCTGCGTGTAGAGACGTAGCTGAAGTCACAGCCTCCGAGGGAGCCTGCTAAGCTGG
ACGACCAAGAGAACCGGGACCTACCATGGAGAACGGCAGAGATGAGATGGACCTGGACAGGCCGGC
ATATGCTGAAGTGGCTACACTGAGGACCTGGCCAGTGCTTCCCTGGAGGAGGTGAATCAAACCTGGG
```

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

Publications

Feature viewer

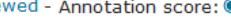
Feature table

None

Protein | Adenine DNA glycosylase

Gene | MUTYH

Organism | Homo sapiens (Human)

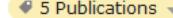
Status |  Reviewed - Annotation score:  - Experimental evidence at protein levelⁱ

Function

Names & Taxonomy

Subcellular location

Functionⁱ

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 ▾

Catalytic activityⁱ

Practicum

Interactionⁱ

Binary interactionsⁱ

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

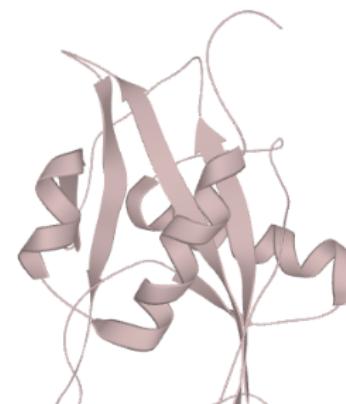
Protein-protein interaction databases

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

Structureⁱ

Family and domain databases

CDD ⁱ	cd03431 DNA_Glycosylase_C , 1 hit cd00056 ENDO3c , 1 hit
Gene3D ⁱ	1.10.1670.10 , 1 hit
InterPro ⁱ	View protein in InterPro 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 ⁱ	View protein in Pfam PF00633 HHH , 1 hit PF00730 HhH-GPD , 1 hit PF14815 NUDIX_4 , 1 hit
SMART ⁱ	View protein in SMART SM00478 ENDO3c , 1 hit SM00525 FES , 1 hit
SUPERFAMILY ⁱ	SSF48150 SSF48150 , 1 hit SSF55811 SSF55811 , 1 hit
PROSITE ⁱ	View protein in PROSITE PS00764 ENDONUCLEASE_III_1 , 1 hit PS01155 ENDONUCLEASE_III_2 , 1 hit PS51462 NUDIX , 1 hit



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

1 notificación

Practicum

Example of database cross-search with Entrez

 **National Library of Medicine**
National Center for Biotechnology Information

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**

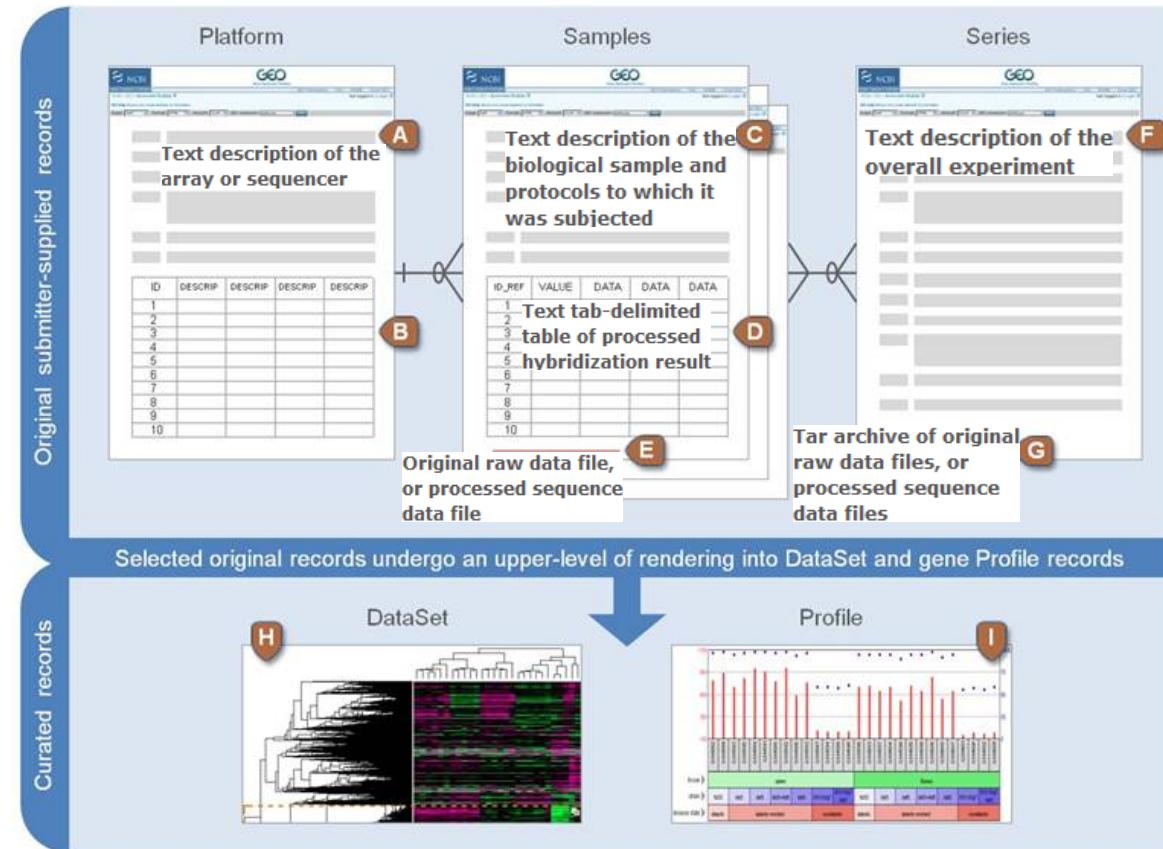
Results found in 31 databases

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

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

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)		SOFT 																
MINiML formatted family file(s)		MINiML 																
Series Matrix File(s)		TXT 																
<table border="1"> <thead> <tr> <th>Supplementary file</th> <th>Size</th> <th>Download</th> <th>File type/resource</th> </tr> </thead> <tbody> <tr> <td>GSE39549_Matrix_non-normalized_EPI.txt.gz</td> <td>8.4 Mb</td> <td>(ftp)(http)</td> <td>TXT</td> </tr> <tr> <td>GSE39549_Matrix_non-normalized_MES.txt.gz</td> <td>2.9 Mb</td> <td>(ftp)(http)</td> <td>TXT</td> </tr> <tr> <td>GSE39549_RAW.tar</td> <td>15.8 Mb</td> <td>(http)(custom)</td> <td>TAR</td> </tr> </tbody> </table> <p>Raw data is available on Series record Processed data included within Sample table</p>			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
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															

Practicum

Retrieving data from GEO

Source name	Adipose tissue of mice
Organism	Mus musculus
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	6.5777917

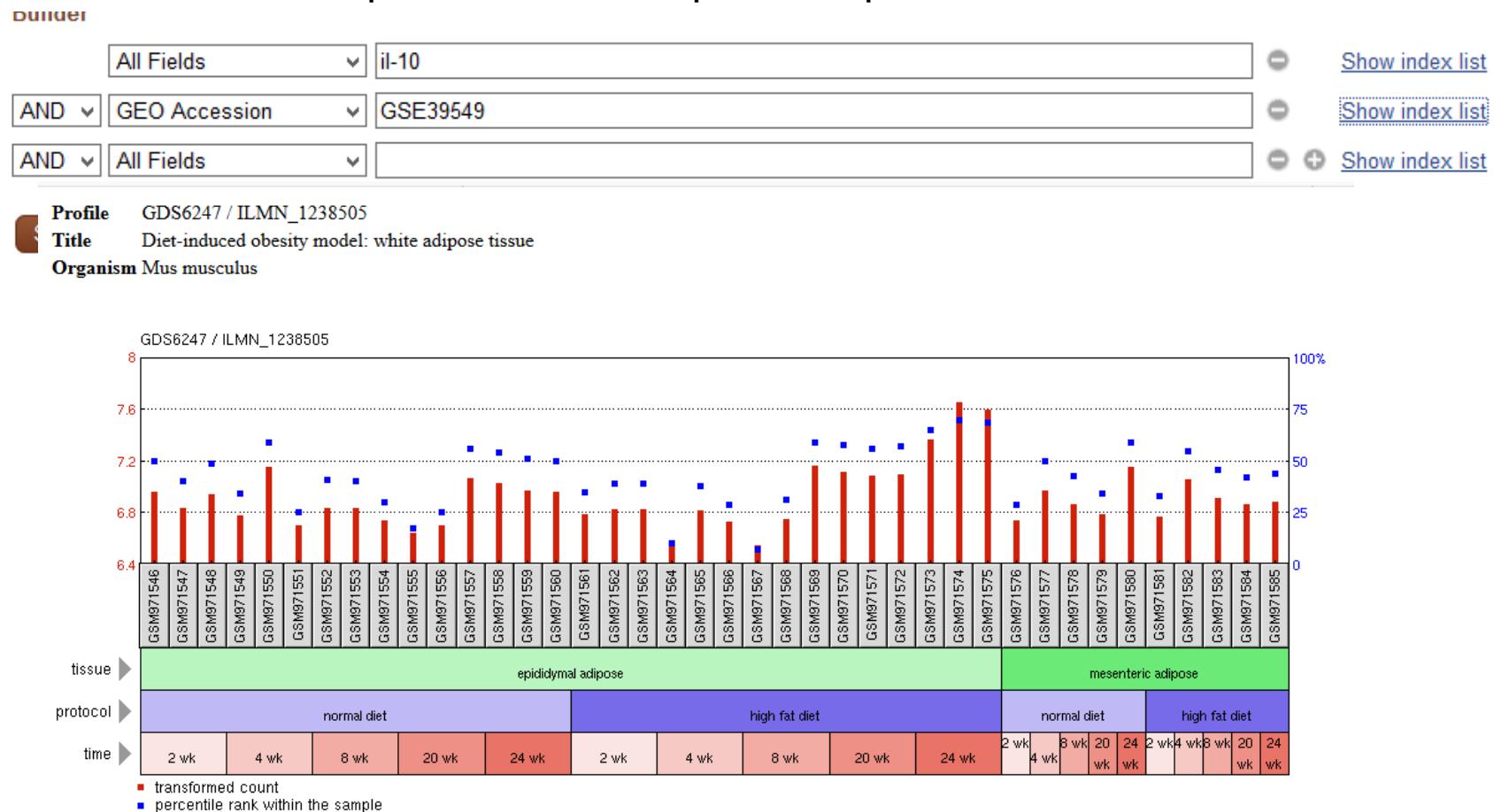
Data table with normalized expression values

Practicum

Retrieving data from GEO

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

Example: search for expression profile of IL-10



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)

Practicum

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.

Practicum

Downloading data from TCGA using R

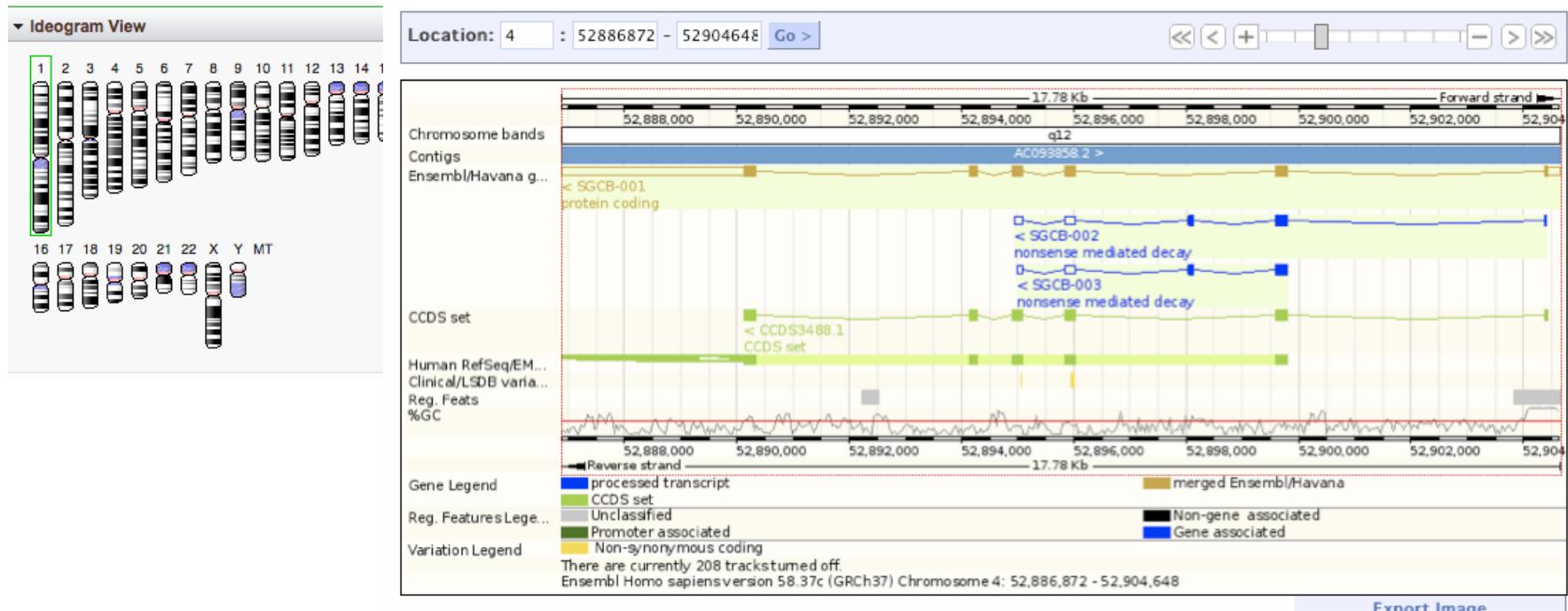


6. Prepare a `SummarizedExperiment` (SE) object from the query using `GDCprepare(query)` function, where `query` will be the query performed above.
7. Examine the data. A `SummarizedExperiment` object has three main matrices that can be accessed using the `SummarizedExperiment` package:
 - Sample matrix: 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

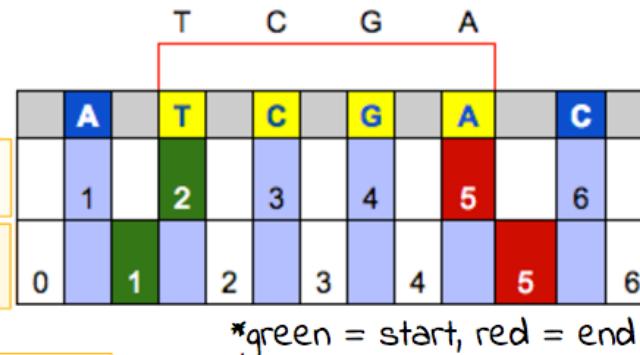
For mouse:

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 ($\text{length}=\text{end}-\text{start}$)
 - 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: chr1:12714000**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](#)
- 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>

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

- Search for a sequence with BLAST**: Points to the BLAST/BLAST+ search bar at the top.
- Search for a gene, region of interest, disease, variant etc**: Points to the main search bar where "BRCA2" is entered.
- Select species of interest**: Points to the "All genomes" section where "Pig breeds" is selected.
- 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 and links to release news.

Key sections visible on the page:

- Ensembl logo and navigation bar**: Includes links for BLAST/BLAST+, VEP, Tools, BioMart, Downloads, Help & Docs, and Blog.
- BioMart**: A tool for exporting custom datasets from Ensembl.
- Export data with BioMart**: A tool for searching genomes based on DNA or protein sequences.
- Variant Effect Predictor**: A tool for analyzing variants and predicting their functional consequences.
- Search bar**: Allows users to search across all species.
- All genomes**: A list of species, with "Pig breeds" currently selected.
- Favourite genomes**: Lists Human (GRCh38.p13), Mouse (GRCm38.p6), and Zebrafish (GRCz11).
- Other news from our blog**: Lists recent posts including "Ensembl under lockdown – Part 3" and "Normalising variants to standardise Ensembl VEP output".

Genome Browsers

Ensembl

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

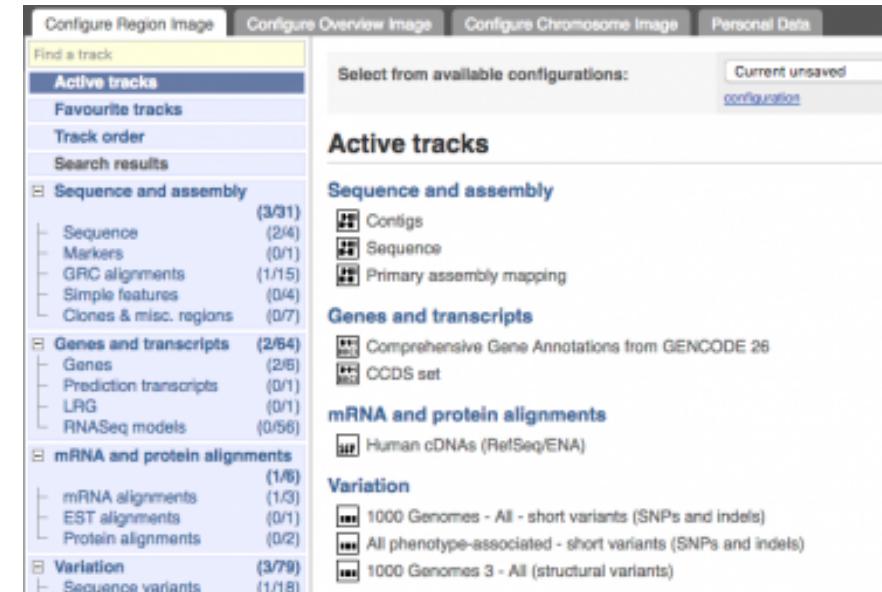
- The navigation of the Ensembl website is organised into tabs, or main pages. The location, gene, transcript, variant and regulation tabs allow data browsing at that level.



Human (GRCh38.p13) ▾

Location: 13:32,315,086-32,400,266 Gene: BRCA2 Transcript: BRCA2-201 Variant: rs55880202 Regulation: ENSR00000060894 Jobs ▾

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



Configure Region Image Configure Overview Image Configure Chromosome Image Personal Data

Find a track

Active tracks

- Favourite tracks
- Track order
- Search results

Sequence and assembly (3/31)

- Sequence (2/4)
- Markers (0/1)
- GRC alignments (1/15)
- Simple features (0/4)
- Clones & misc. regions (0/7)

Genes and transcripts (2/64)

- Genes (2/6)
- Prediction transcripts (0/1)
- LRG (0/1)
- RNASeq models (0/56)

mRNA and protein alignments (1/6)

- mRNA alignments (1/3)
- EST alignments (0/1)
- Protein alignments (0/2)

Variation (3/79)

- Sequence variants (1/18)

Select from available configurations:

Current unsaved configuration

Active tracks

Sequence and assembly

- Configs
- Sequence
- Primary assembly mapping

Genes and transcripts

- Comprehensive Gene Annotations from GENCODE 26
- CCDS set

mRNA and protein alignments

- Human cDNAs (RefSeq/ENA)

Variation

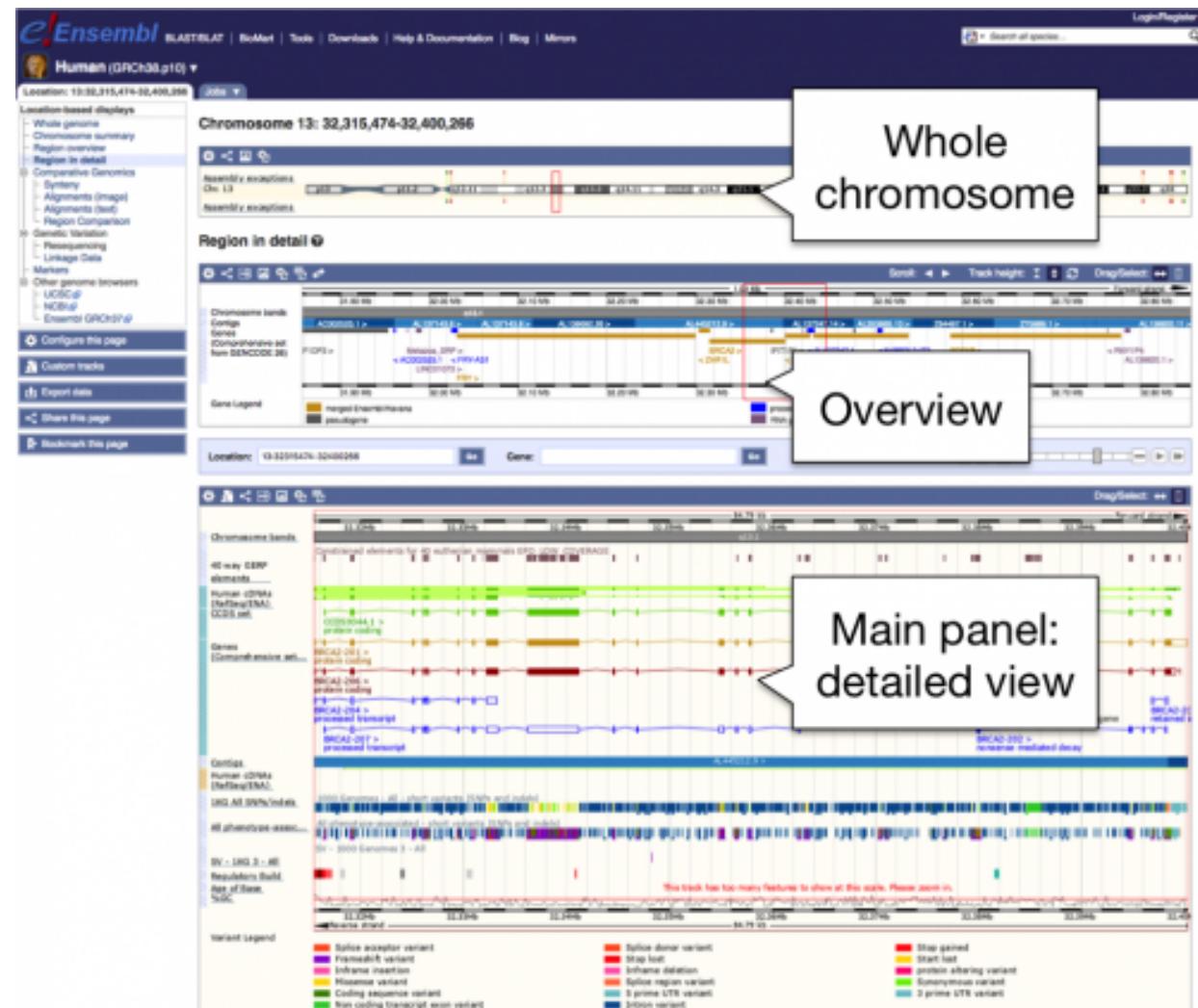
- 1000 Genomes - All - short variants (SNPs and indels)
- All phenotype-associated - short variants (SNPs and indels)
- 1000 Genomes 3 - All (structural variants)

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

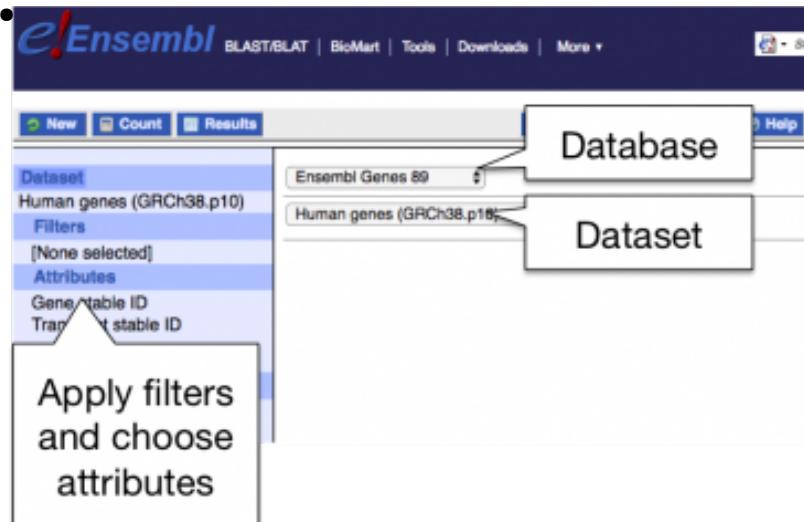


Genome Browsers

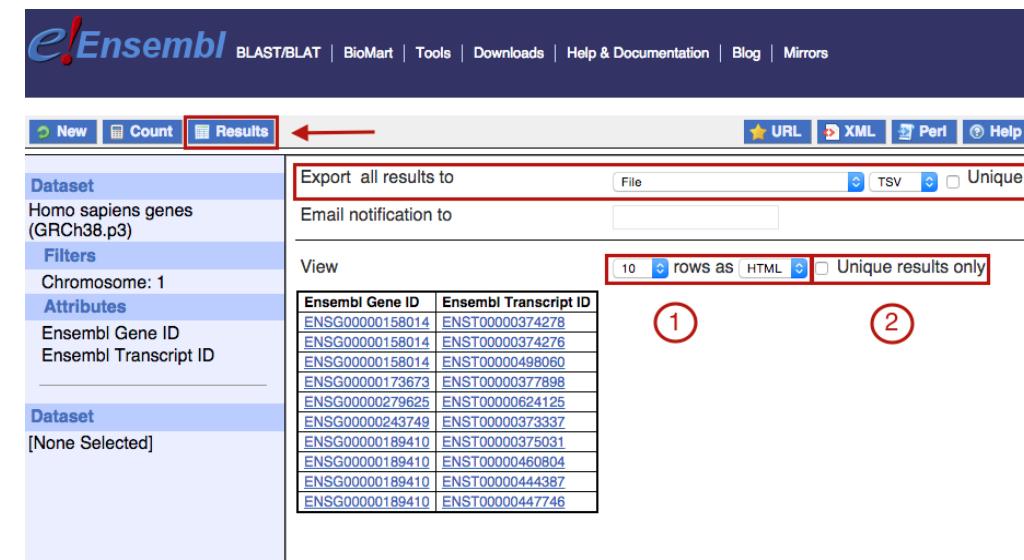
Ensembl

<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'. Below these, a 'Database' section is selected, showing 'Ensembl Genes 89' and 'Human genes (GRCh38.p10)'. A 'Dataset' section below it shows 'Human genes (GRCh38.p10)'. On the left, there are 'Dataset' and 'Filters' sections, and a large box at the bottom says 'Apply filters and choose attributes'.



The screenshot shows the Ensembl BioMart search interface with the 'Results' tab selected. At the top, there are buttons for 'New', 'Count', and 'Results'. Below these, there are sections for 'Dataset' (set to 'Homo sapiens genes (GRCh38.p3)'), 'Filters' (set to 'Chromosome: 1'), and 'Attributes' (set to 'Ensembl Gene ID' and 'Ensembl Transcript ID'). To the right, there are options to 'Export all results to' (File, XML, Perl, Help), 'Email notification to' (input field), and 'View' (10 rows as HTML, Unique results only). A table below lists Ensembl Gene IDs and Ensembl Transcript IDs. Red circles numbered 1 and 2 point to the 'Unique results only' checkbox and the 'View' dropdown respectively.

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

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/>



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 displayed prominently. A navigation bar below the header includes links for Home, Genomes, Genome Browser, Tools, Mirrors, Downloads, My Data, Help, and About Us. The main content area features a large blue DNA helix graphic. To the right of the graphic 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, Track Hubs, and More tools... Each tool entry includes a brief description.

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

<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)

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: Dec. 2013 (GRCh38/hg38) Query type: BLAT's guess Sort output: query,score Output type: hyperlink



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)

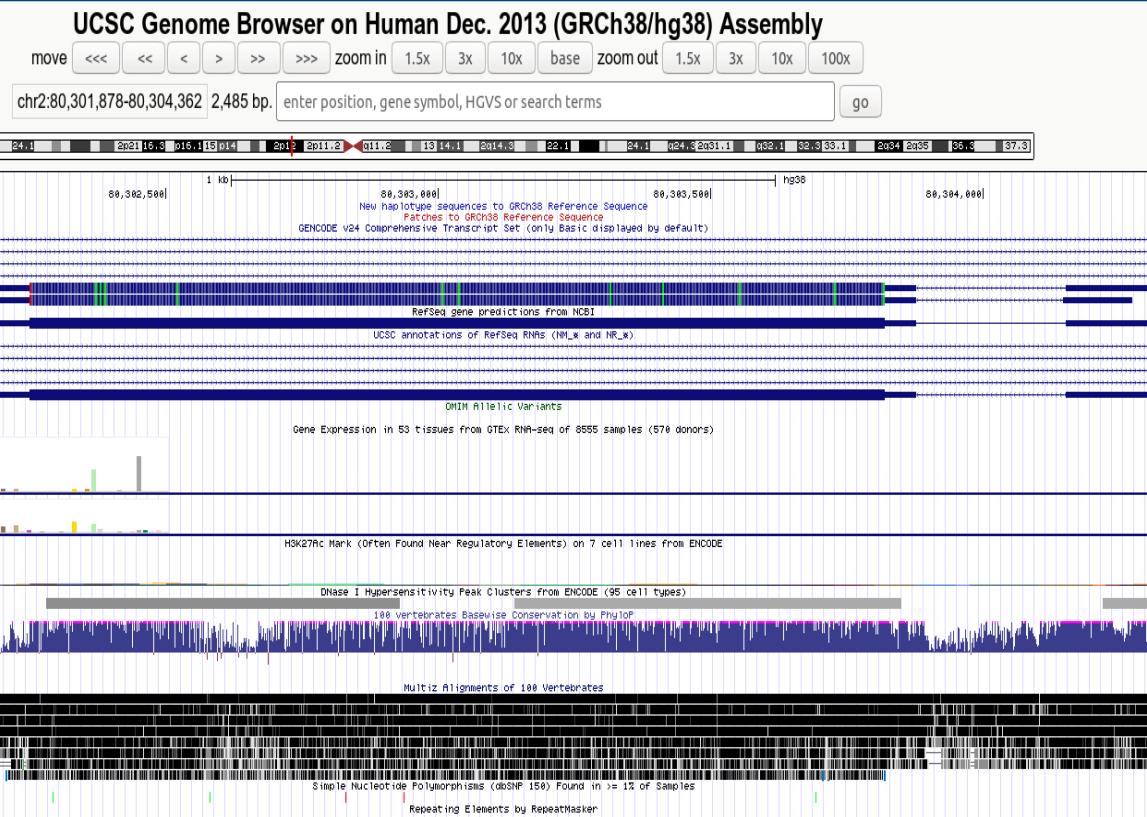
Basic Gene Annotation Set from GENCODE Version 28 (Ensembl 92)

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

[LRRTM1 at chr2:80301878-80304738](#)

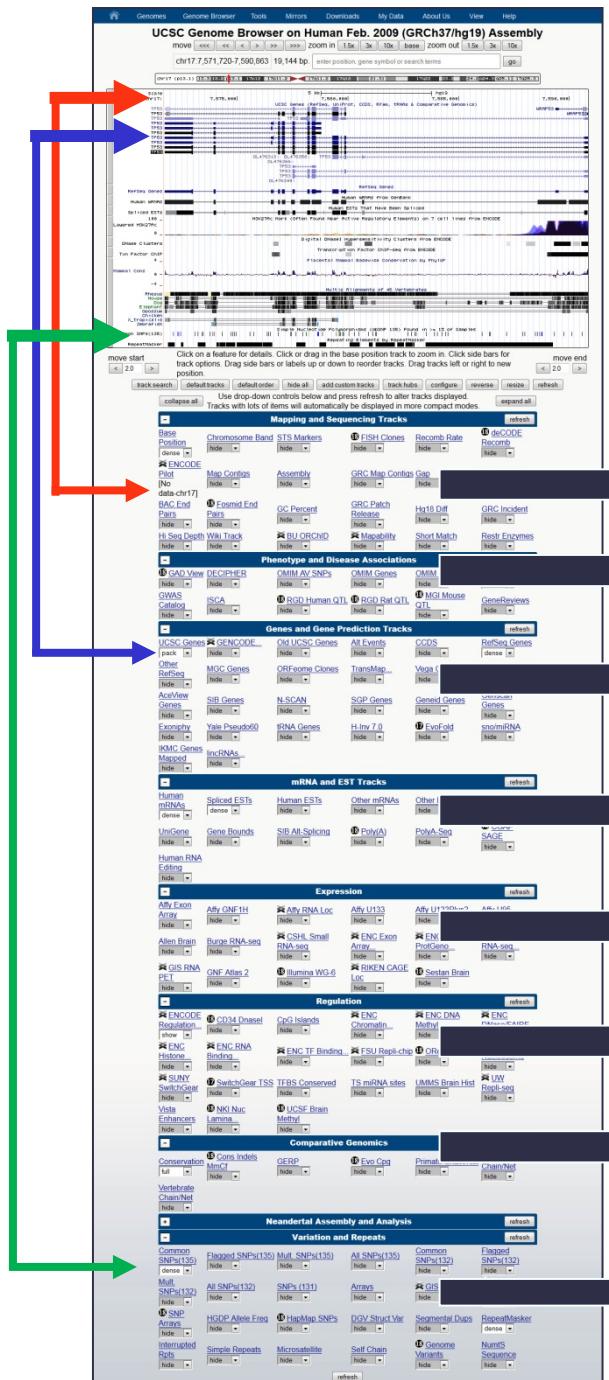
Genome Browsers

- 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

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)

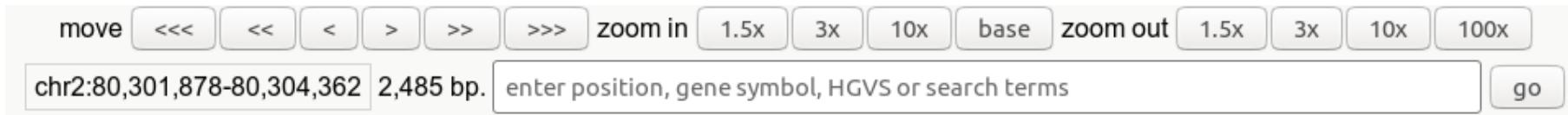
Comparative Genomics
As a group
Individual species

Variation and Repeats

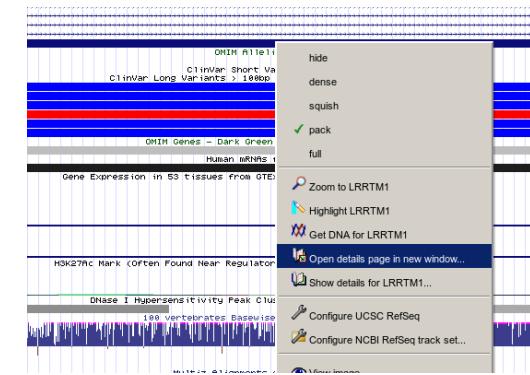
(including SNPs, copy number variation)

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

Display modes:

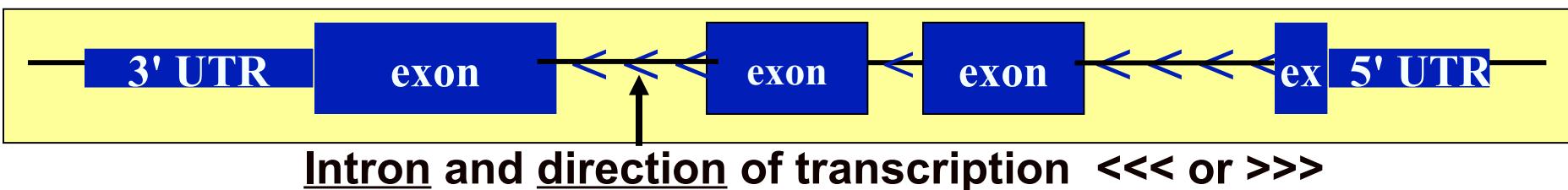
- Hide=don't show
- Dense=features collapsed into a single line
- Squish, Pack, Full=features in different lines

Refresh!

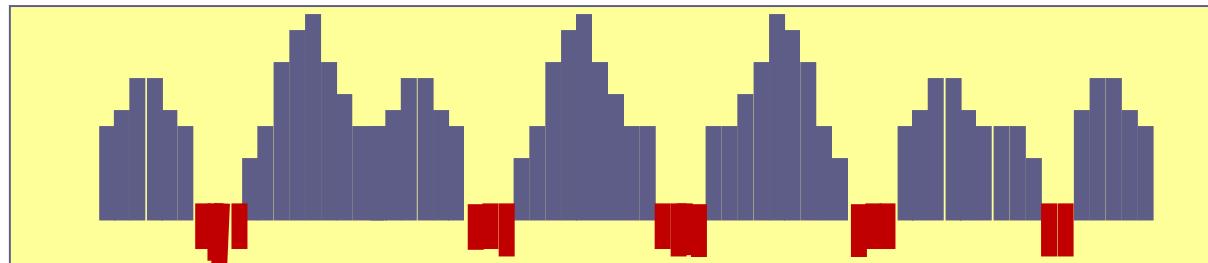
Genes and Gene Predictions		
NCBI RefSeq	Other RefSeq	All GENCODE...
full	hide	hide
hide	Genscan Genes	19 IKMC Ger Mapped
Geneid Genes	hide	hide
dense	ORFeome Clones	Pfam in UCS
squish	hide	hide
Old UCSC Genes	UCSC Alt Events	UniProt
pack	hide	hide
full	TransMap...	hide
TransMap...	hide	hide

Genome Browsers

- Some visual clues:



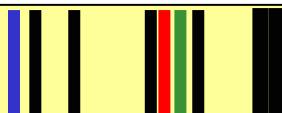
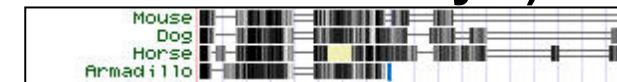
Sequence conservation



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

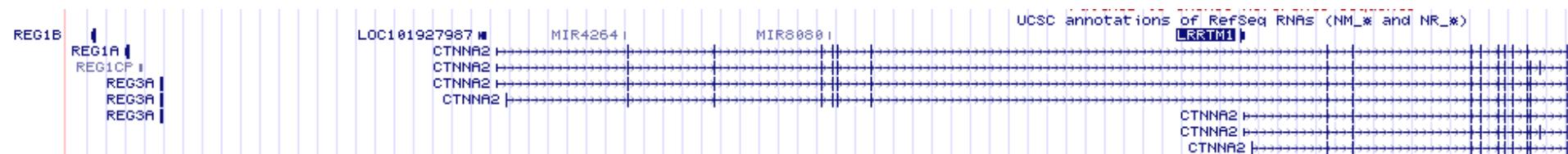
1. What is the genomic localization of human *Irrtm1* gene?

- 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



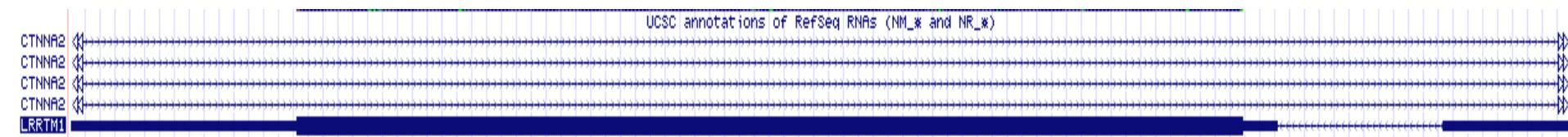
3. How many exons has the gene? 2

- 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



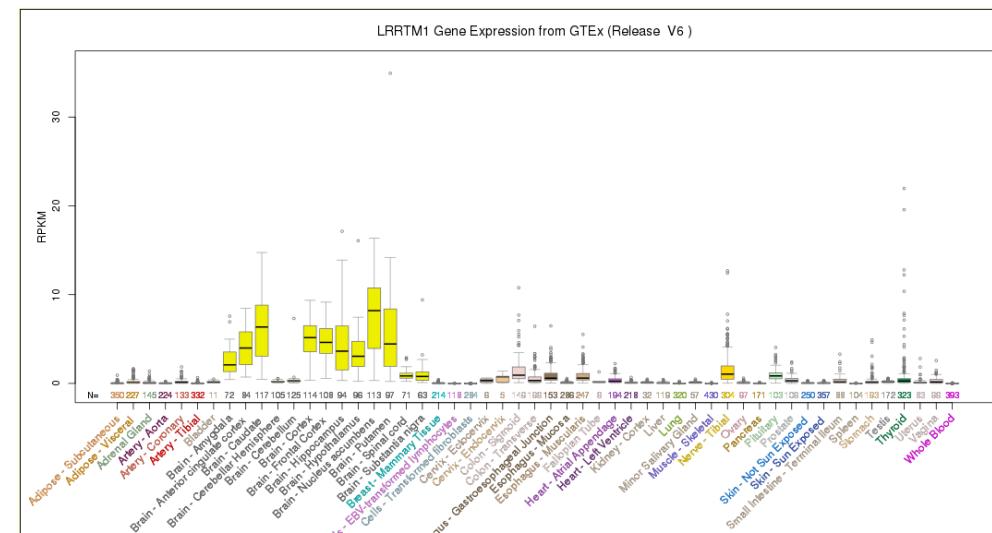
5. Can you find SNPs in this gene? yes

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



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

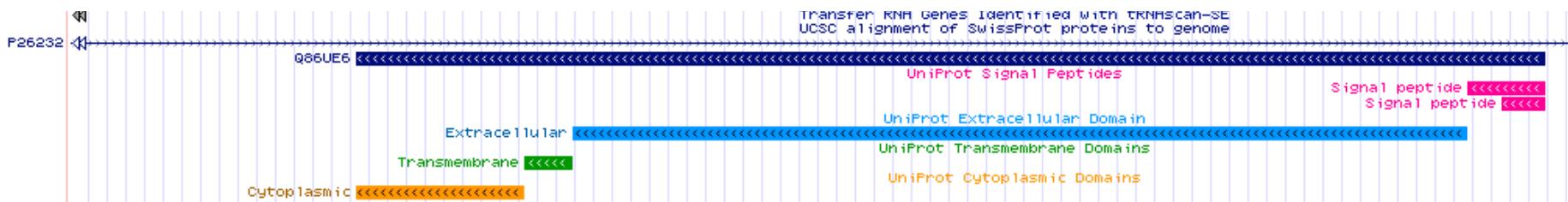
- Info in track “Expression” > “GTEx”



Practicum

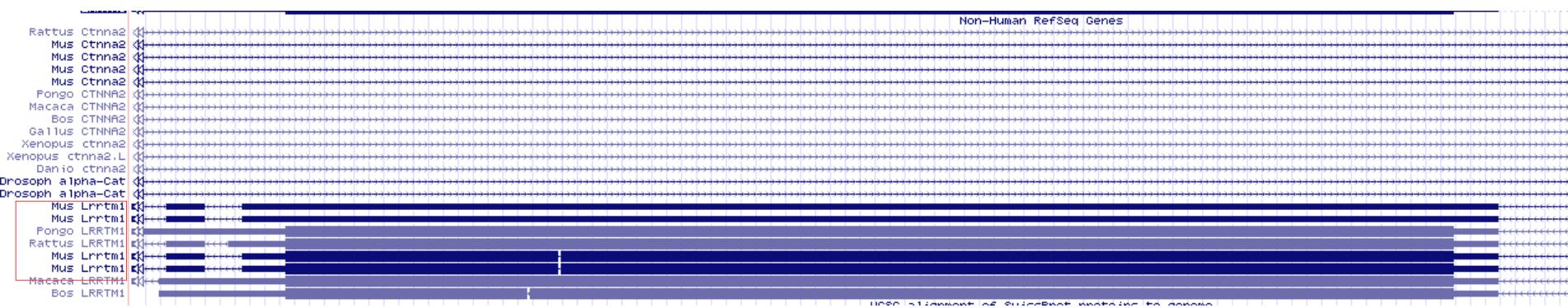
7. Does the protein encoded by this gene have a transmembrane domain?

- 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

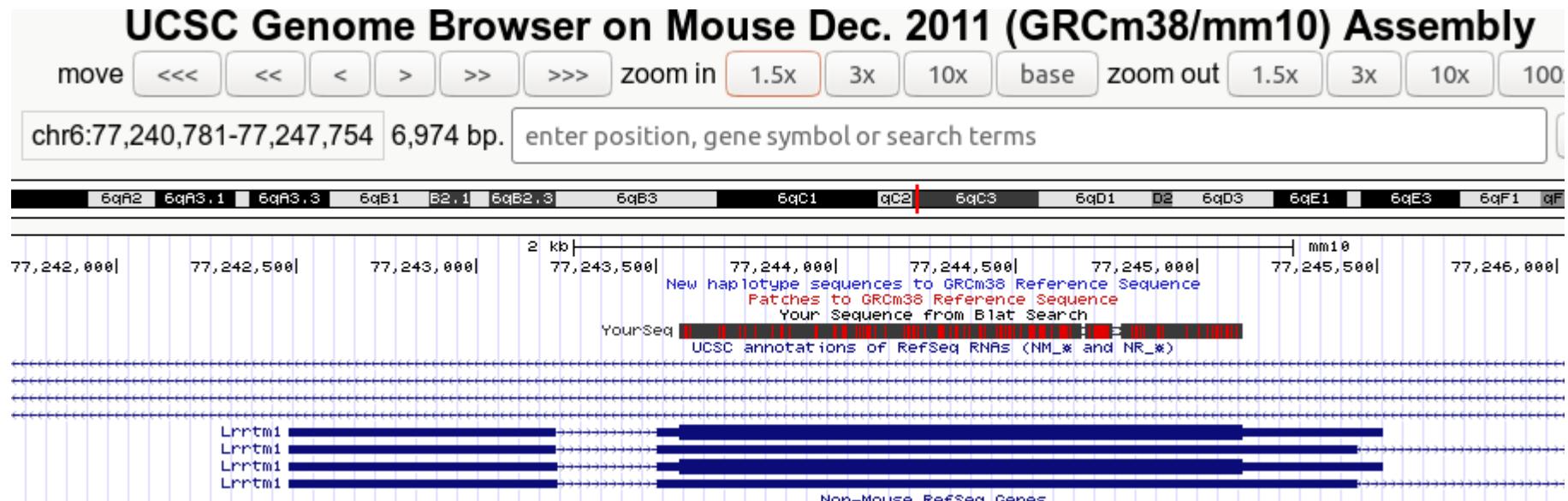
Genome: <input type="checkbox"/> Search ALL	Assembly: <input type="text" value="Dec. 2011 (GRCm38/mm10)"/>	Query type: <input type="text" value="DNA"/>	Sort output: <input type="text" value="query,score"/>
<input type="text" value="ATGGATTTCTCTGCTCGGTCTCTGCTATACTGGCTGCTGAGGGAGGC"/> CTCGGGGGTGGCTTGTGCTGCTGGGGGCTGCTTCAGATGCTGGCG CCGCCCCCAGCGGGTGCAGCTGTGCCGTGCGAGGGCGGCTGCTG TACTCGGAGGGCCTCAACCTCACCGAGGCGCCAACACCTGTCCGGCT GCTGGCTTGTCCCTGCCTACAAACAGCTCTCGAGCTGCCGCCGCC AGTTCACGGGTTAATGAGCTCACGGCTCTATCTGGATACAATCAC ATCTGCTCGTGAGGGGACGCCCTTCAGAAACTGCCGAGTTAAGGA ACTCACGCTGAGTCCAACCAGATACCCAACCTGCCAACACCCCTTC GGCCATGCCAACCTGGCAGCGGACCTCTCGTACAACAAAGCTGAG GCGCTGGCGGGACCTCTCCACGGGCTGCCGAAGCTCACACGGCTGCA TATGCGGGCCAACGCCATTCCAGTTGTGCCGTGCGCATCTTCAAGGACT GCGCAGCTCAAGTTCTGACATGGATACAATCAGCTAAGAGCTG GCGCGCAACTTTGCCGGCTTAAAGCTACCGAGCTGCACCTCGA GCACAAAGCACTGGTCAAGGTGAACCTCGCCACTCCGCGCCTCATCT CCCTGCACTGCTCTGCGGGAGGAACAGGTGGCATTGGGGTAGC			
<input type="button" value="submit"/> <input type="button" value="I'm feeling lucky"/> <input type="button" value="clear"/>			

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

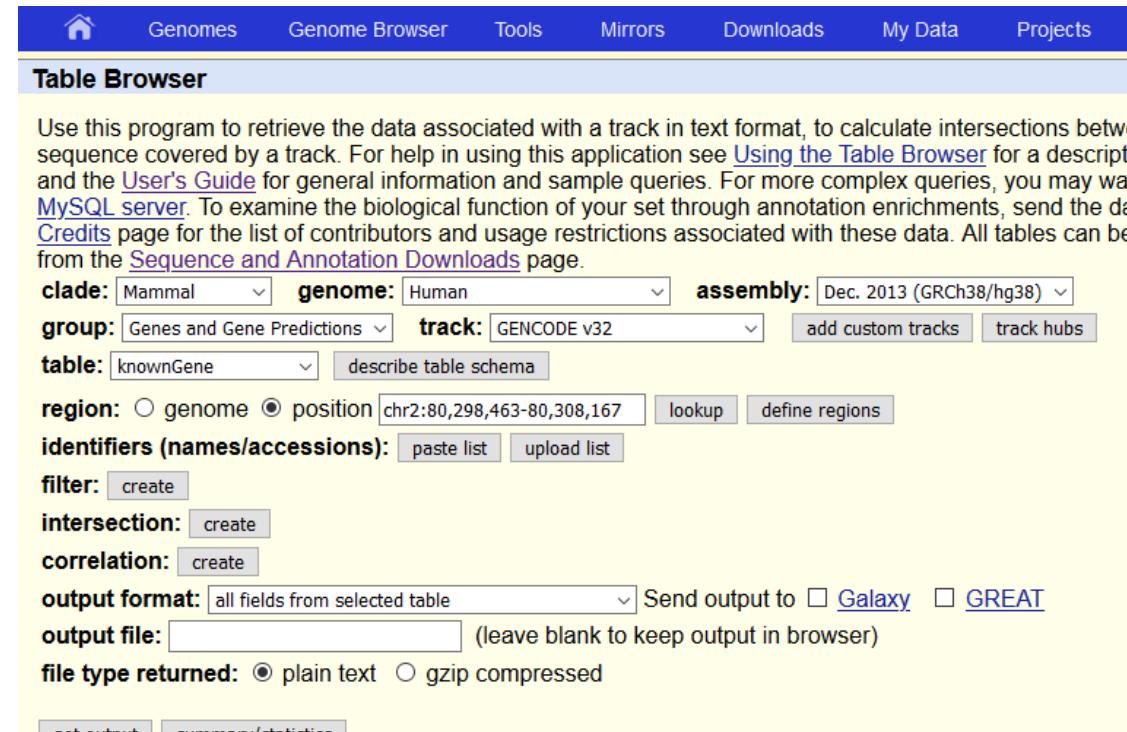


Genome Browsers

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



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. A descriptive text block explains the purpose of the browser, mentioning its ability to retrieve data associated with a track, calculate intersections, and provide help through the User's Guide and MySQL server. It also mentions annotation enrichments and credits. Below this text, there are several input fields and buttons for specifying search parameters: "clade" (Mammal), "genome" (Human), "assembly" (Dec. 2013 (GRCh38/hg38)), "group" (Genes and Gene Predictions), "track" (GENCODE v32), "table" (knownGene), "region" (radio buttons for genome or position, with a specific position example: chr2:80,298,463-80,308,167), "identifiers (names/accessions)" (paste list, upload list), "filter" (create), "intersection" (create), "correlation" (create), "output format" (all fields from selected table), "Send output to" (checkboxes for Galaxy and GREAT), "output file" (input field with a note: leave blank to keep output in browser), "file type returned" (radio buttons for plain text or gzip compressed), and two buttons at the bottom: "get output" and "summary/statistics".

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

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

