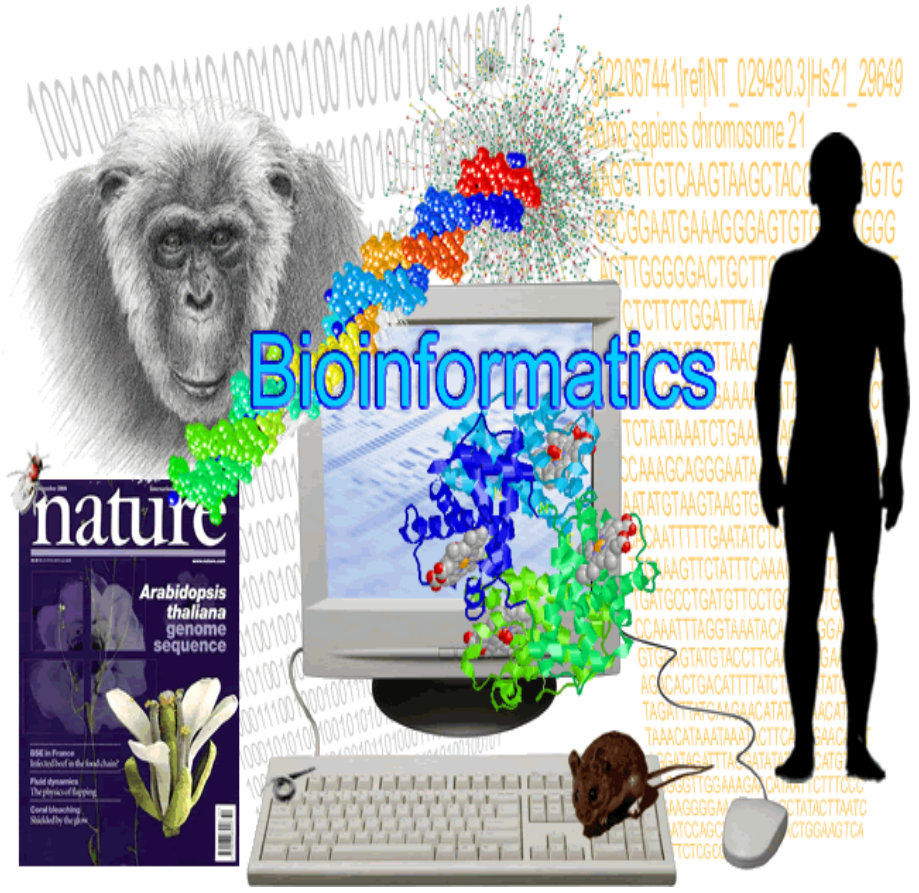
Databases in molecular biology

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* : 1737 bio-databases in 2018!

http://www.oxfordjournals.org/nar/database/cap/

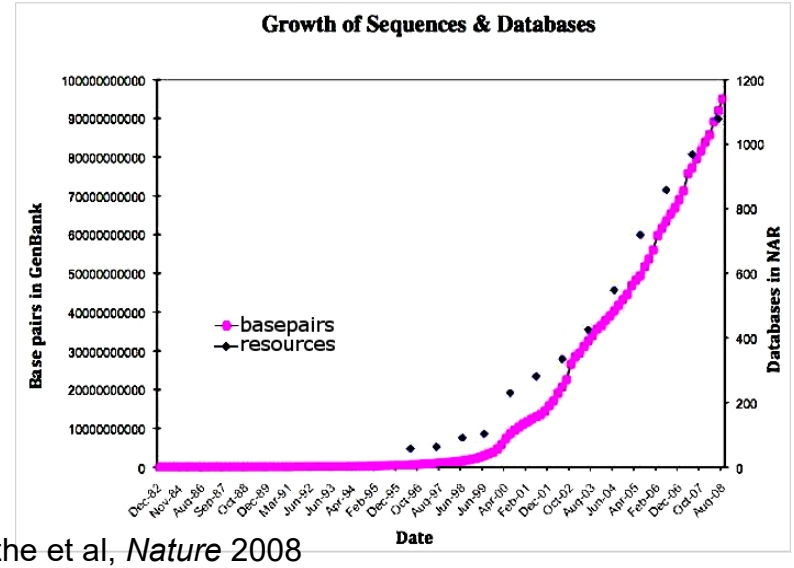
Biological Databases

NAR Database Summary Paper Category List

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

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

Challenges



Lathe et al, *Nature* 2008

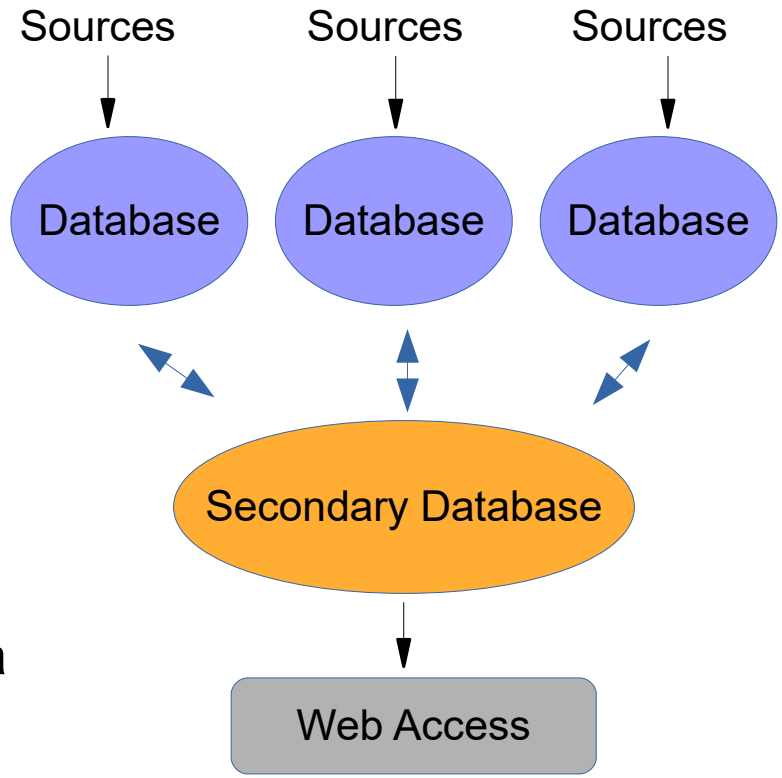
This large number of databases, though extremely useful, can lead to its own issues of redundancy and lack of integration.

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

I. Integrating the information

Integrating the information

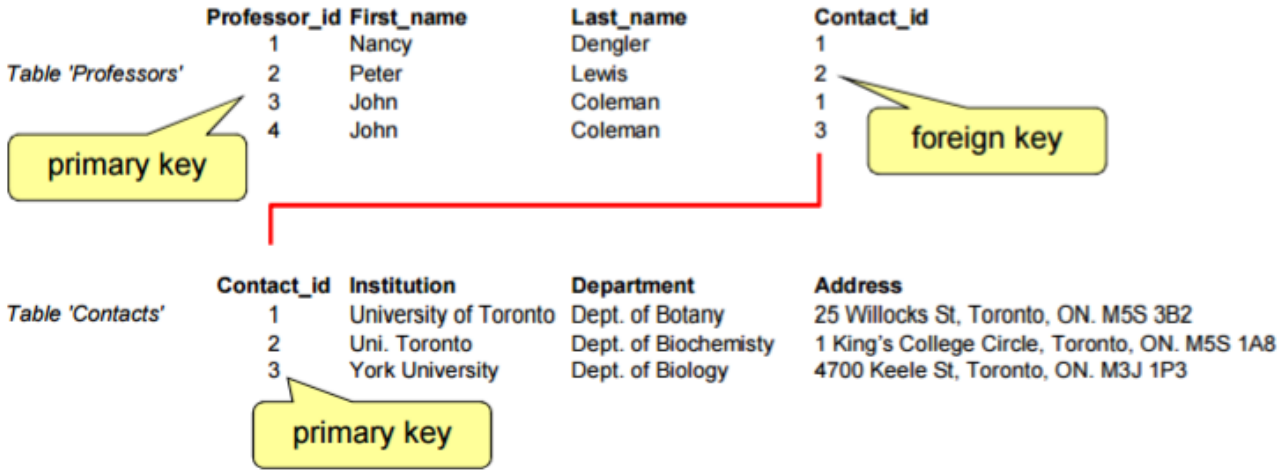
- **Primary databases:**
  - often hold only one type of specific data which is stored in their own archive.
  - upload new data from experiments and update entries
- **Secondary databases:**
  - use other databases as their source of information.
  - often already process or analyze the data to get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



Relational databases

Integrating the information

- Relational databases:
  - cross-reference to other databases through a common field (unique identificator)
  - flat-file format



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.

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**

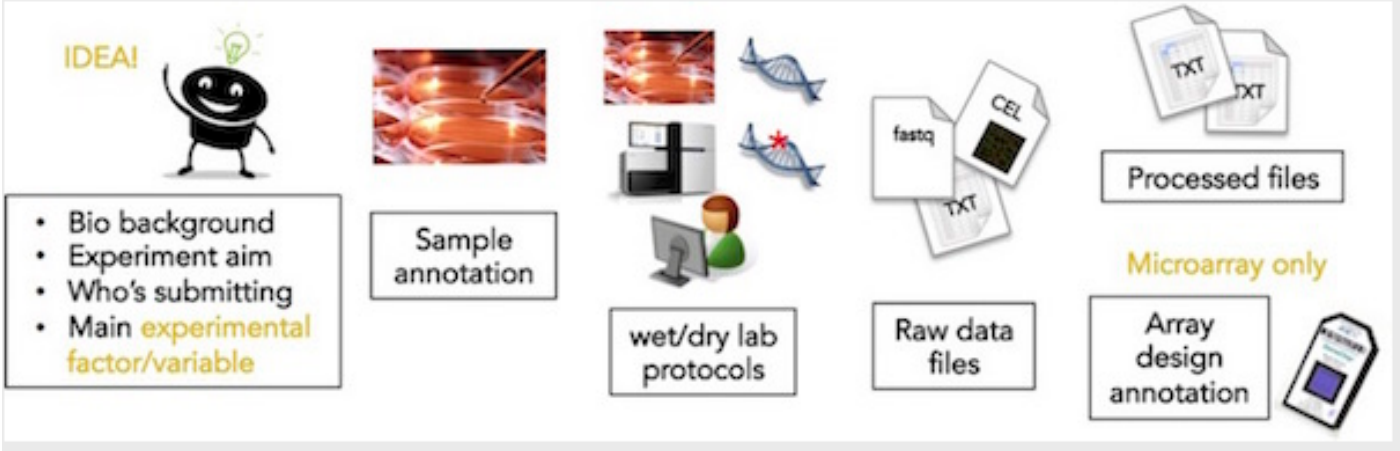
II. Data Annotation and Curation

Data Annotation

- Collaborative efforts to provide as much information about the data
  - Annotation of sequences/genomes: chromosome position, gene function, …
  - Metadata: information for an experiment, identification of samples, …
  - The **minimum information standard** is a set of guidelines for reporting data
- Benefits:
  - Ensures the verification, interpretation and reproducibility of data
  - Facilitates the creation of structured databases and development of analysis tools

---

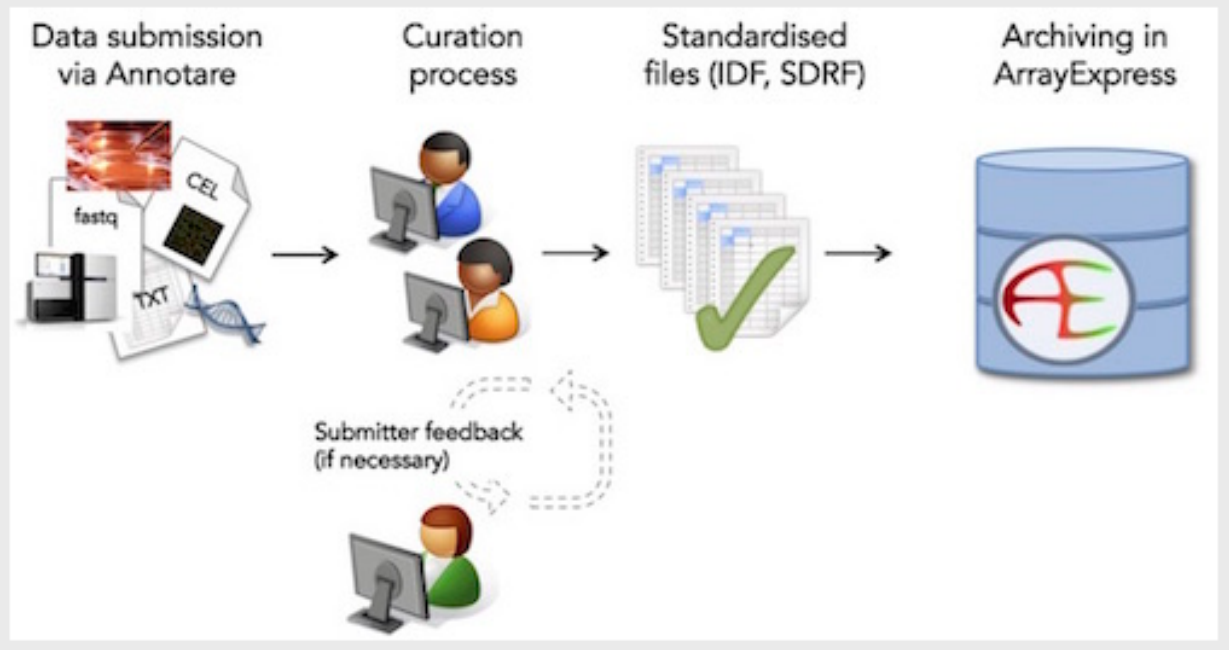
Example: MIAME The minimum information about a microarray experiment



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

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.

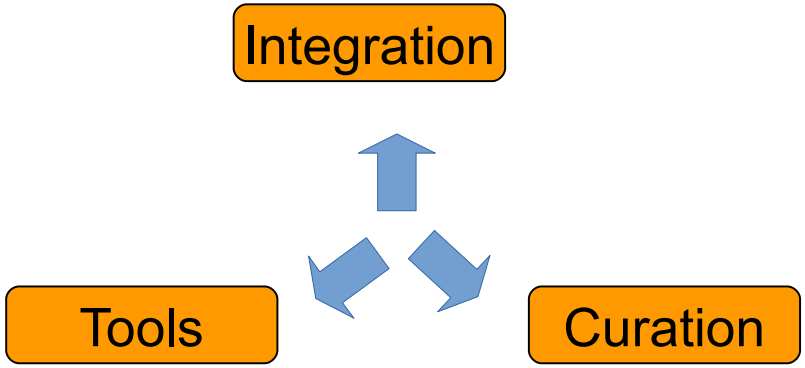


Source: https://www.ebi.ac.uk/arrayexpress/submit/overview.html

III. Centralizing data management

Centralizing data management

- General
  - **Resource providers**
- Subject-specific
  - **Collaborative projects**
  - **Genomic Browsers**



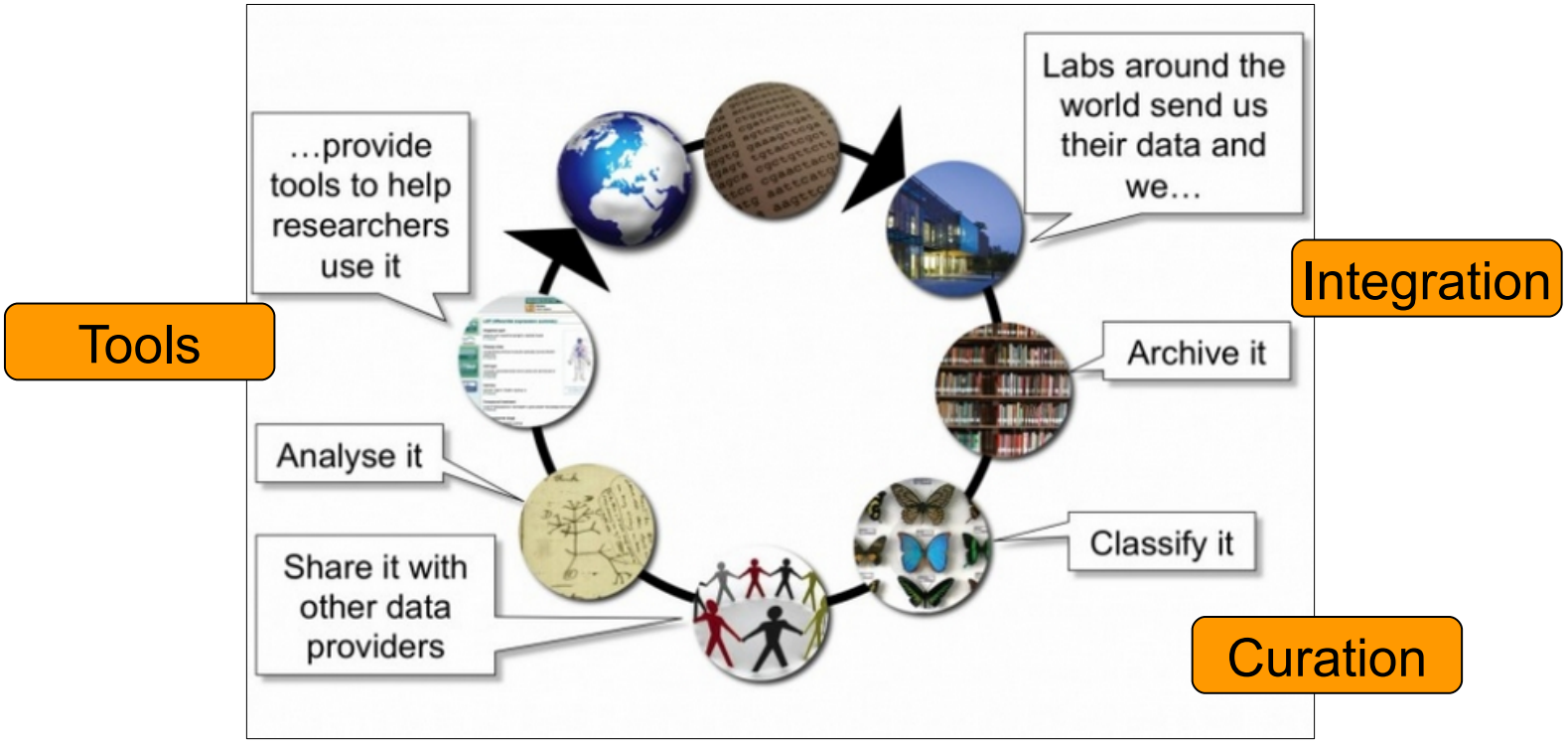
Resource providers

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

NCBI

EMBL-EBI

NATIONAL INSTITUTE OF GENETICS - JAPAN - NIG



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

Resource providers

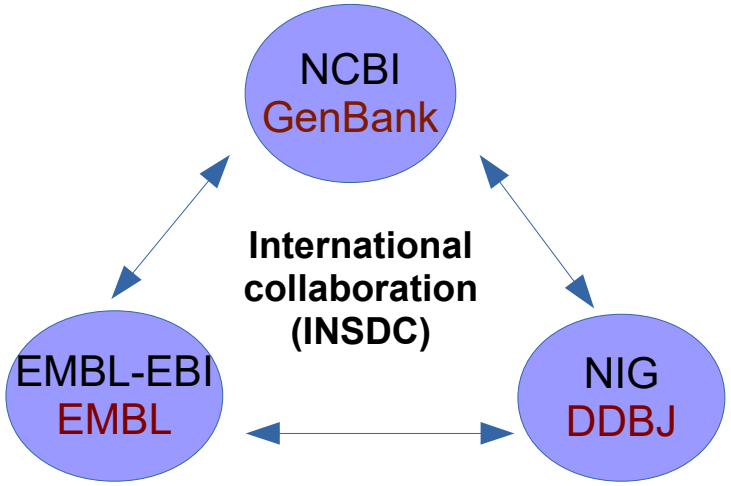
NCBI

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

EMBL-EBI

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

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



Resource providers

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

Clustal Omega

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

Multiple sequence alignment

InterProScan

InterProScan searches sequences against InterPro's predictive protein signatures.

Protein feature detection | Sequence motif recognition

BLAST [protein]

Fast local similarity search tool for protein sequence databases.

Sequence similarity search

BLAST [nucleotide]

Fast local similarity search tool for nucleotide sequence databases.

Sequence similarity search

HMMER

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

Sequence similarity search | Protein function analysis

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

Examples of Databases and guided practicum

Examples of Databases

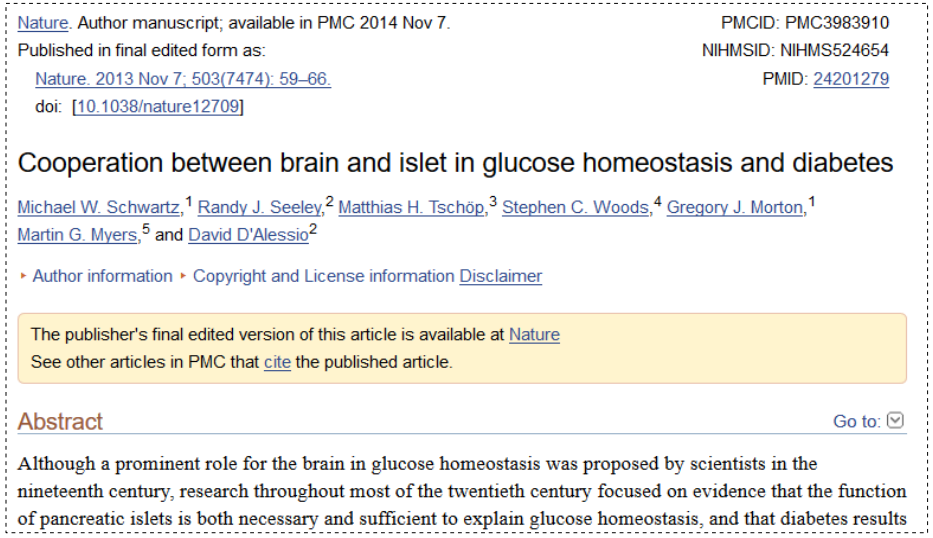
Taxonomic DB

- 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

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

Author information · Copyright and License information Disclaimer

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

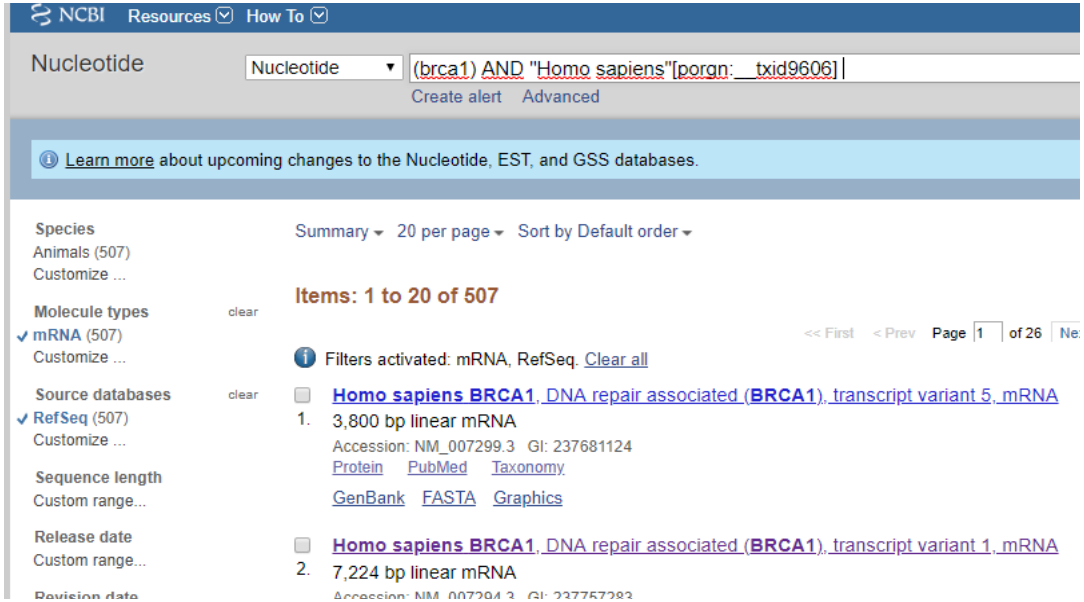
Abstract Go to:

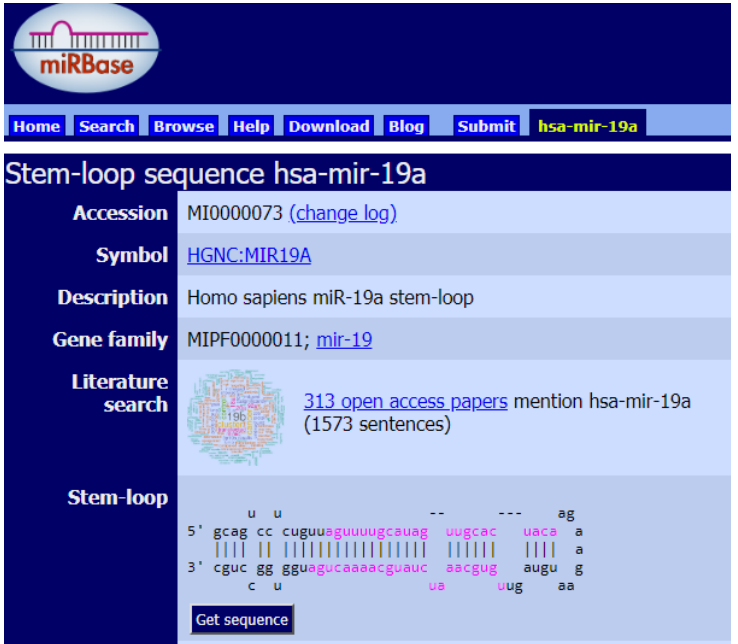
Although a prominent role for the brain in glucose homeostasis was proposed by scientists in the nineteenth century, research throughout most of the twentieth century focused on evidence that the function of pancreatic islets is both necessary and sufficient to explain glucose homeostasis, and that diabetes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

Examples of Databases

Nucleotide DB

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





Examples of Databases

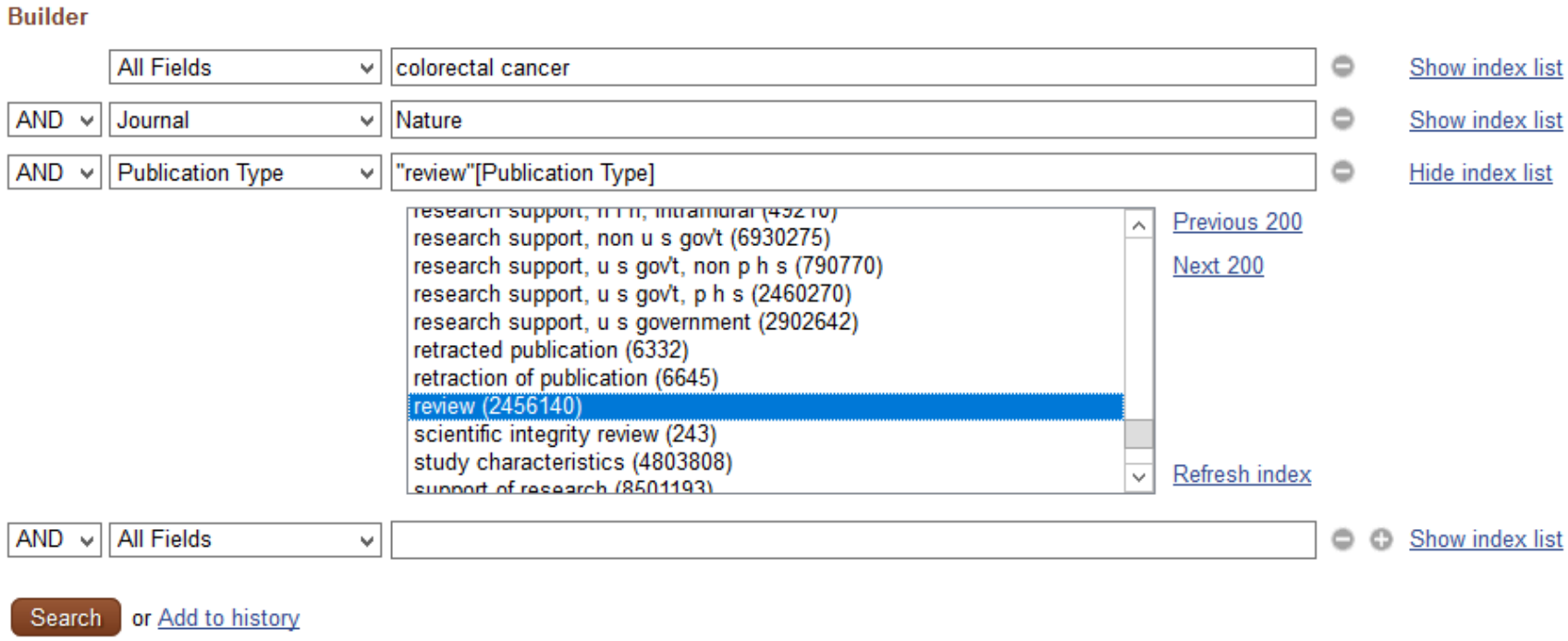
Protein DB

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

Practicum

Formulating specific queries and retrieving nucleotide sequences

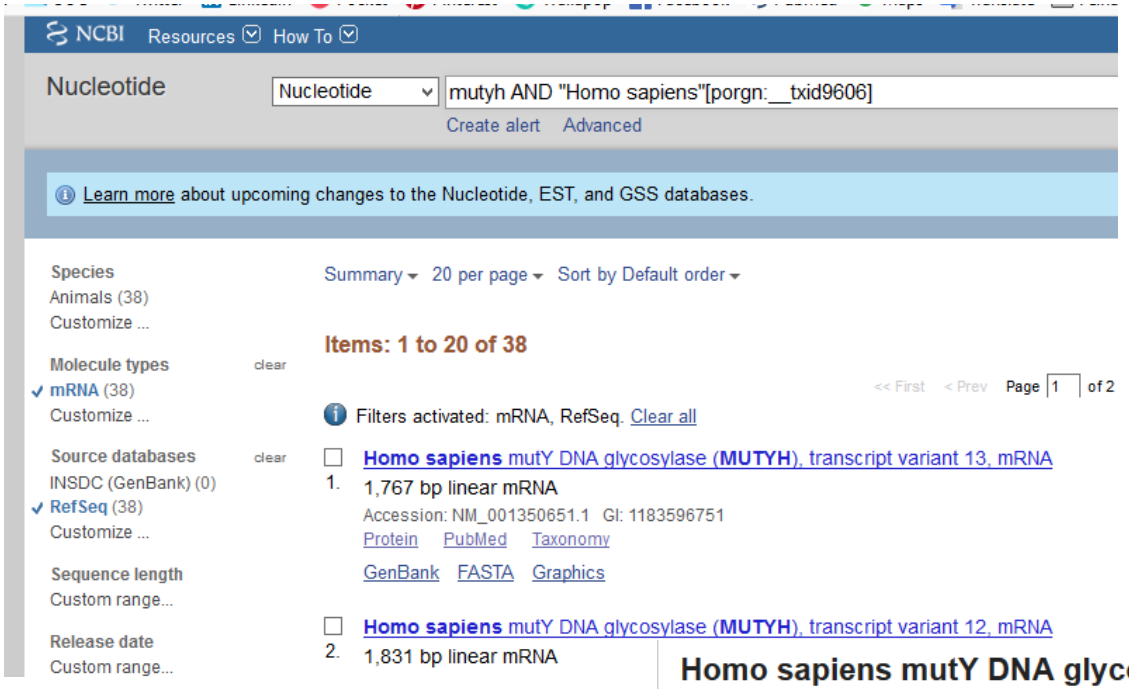
1- Using PubMed Advanced Search, look for a *review* paper published in *Nature* on *colorectal cancer* and authored by *David SS*

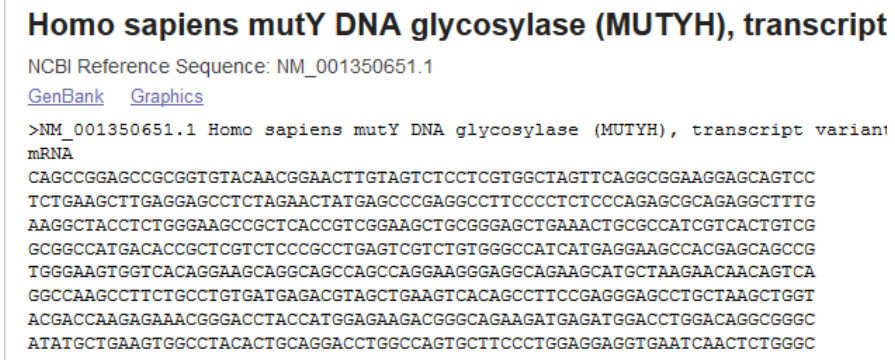


Practicum

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

Using filters



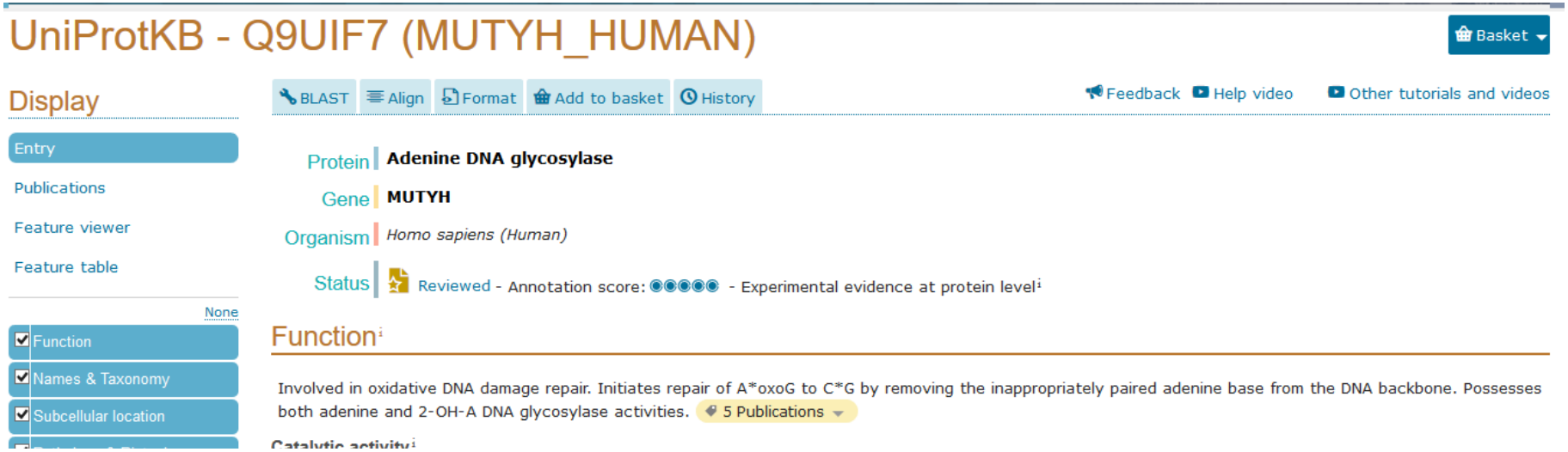


Practicum

Retrieving protein information

3- Look for MUTYH human protein in UniProtKB

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



Practicum

Interaction

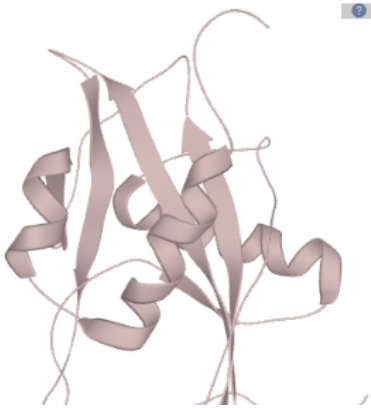
Binary interactions

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

Protein-protein interaction databases

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

Structure



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

Family and domain databases

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

Practicum

Can we learn more about this gene?

4- What is the genomic context of this gene?

5- Are there known mutations associated to cancer disease?

6- In which tissues is the gene normally expressed?

7- Are there SNPs associated to changes in *mutyh* expression?

Examples of Databases

Genomic databases

- Organize information on genomes including sequences, maps, chromosomes, assemblies, and annotations
- **ENCODE** (Encyclopedia of DNA Elements) Project: 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)
- **Genome Browsers:** provide tools for visualization and integrative genomic analysis:
  - NCBI Genome Data Viewer
  - EBI's Ensembl [Biomart]
  - University of California, Santa Cruz (UCSC) Genome Browser [Table Browser]
- Species-specific genome databases (eg. Mouse Genome Informatics)

Examples of Databases

Genomic databases

- Basic and advanced genome annotations:
  - Genes
    - Genomic location
    - Gene model structures
      - Exons
      - Introns
      - UTRs
    - Transcript(s)
      - Pseudogenes
      - Non-coding RNA
    - Protein(s)
    - Links to other sources of information

Examples of Databases

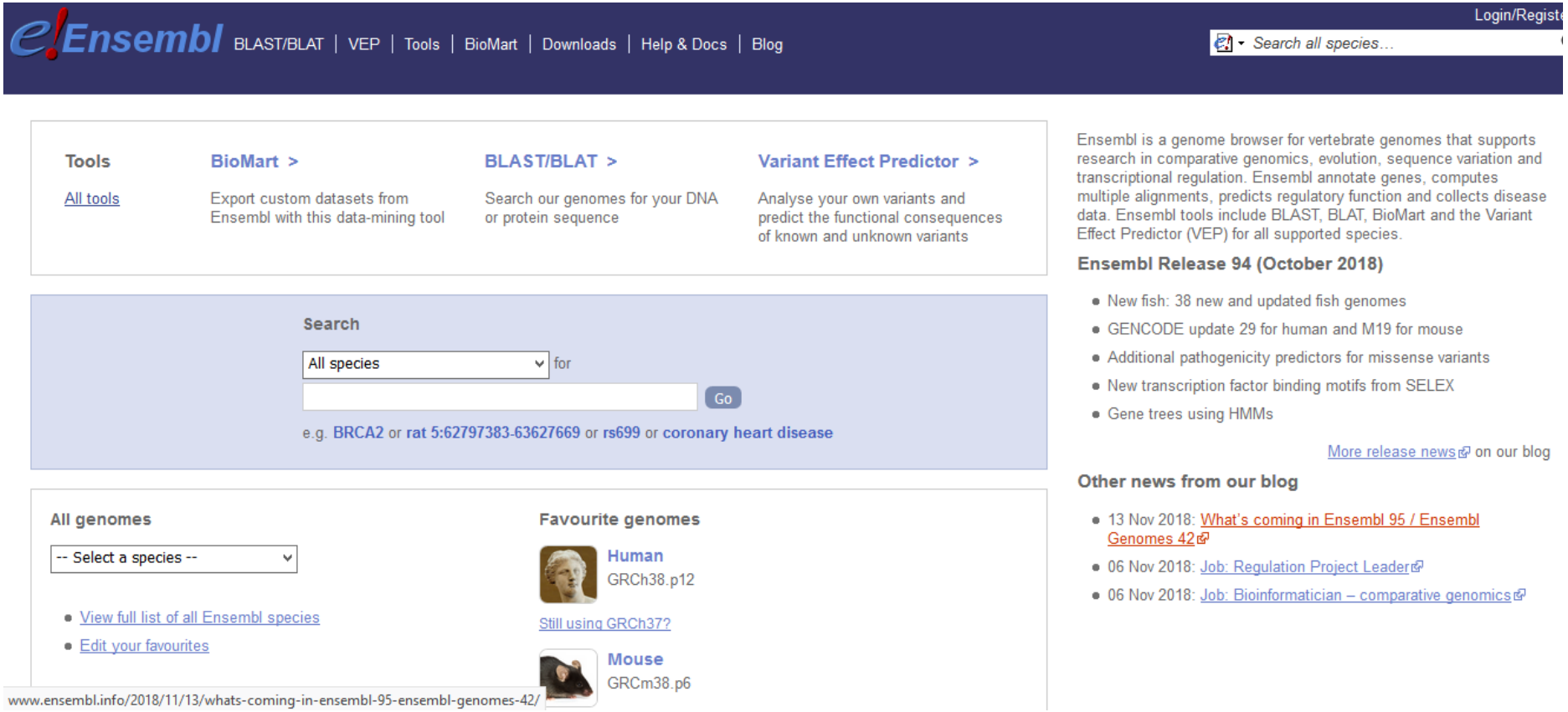
Genomic databases

- Basic and advanced genome annotations:
  - Cytogenetic bands
  - Polymorphic markers
    - Sequence Tagged Sites (STS)
  - Genetic variation
    - Single Nucleotide Polymorphisms (SNPs)
    - Deletion-Insertion Polymorphisms (DIPs)
    - Short Tandem Repeats (STRs)
  - Repetitive sequences
  - Expressed Sequence Tags (ESTs)
  - cDNAs or mRNAs from related species
  - Regions of sequence homology

Genomic Browsers

Ensembl

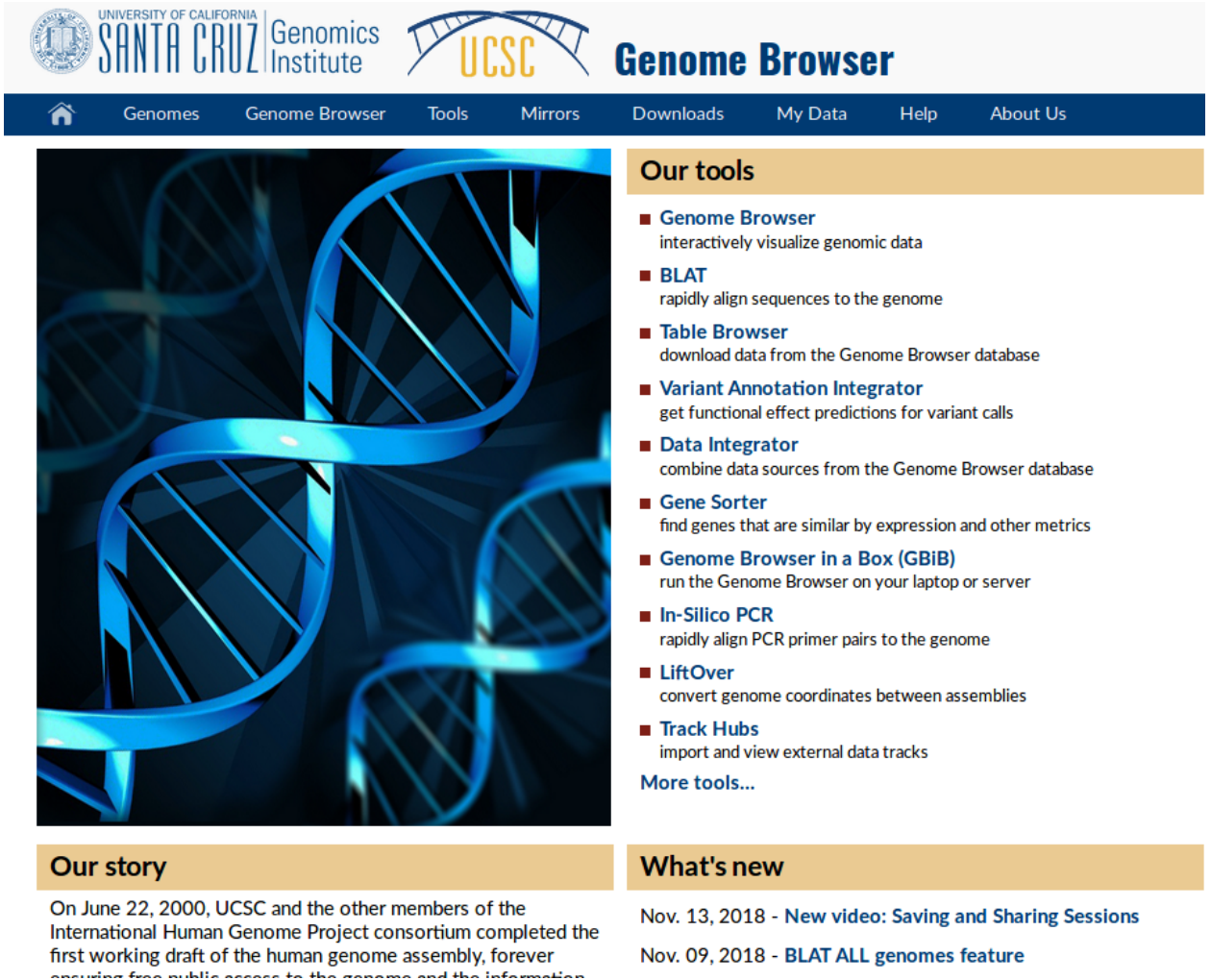
https://www.ensembl.org/index.html



Genomic Browsers

UCSC Genome Browser

https://genome.ucsc.edu/

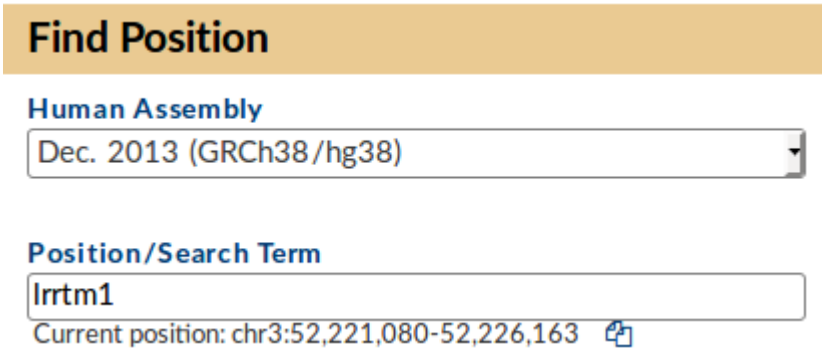


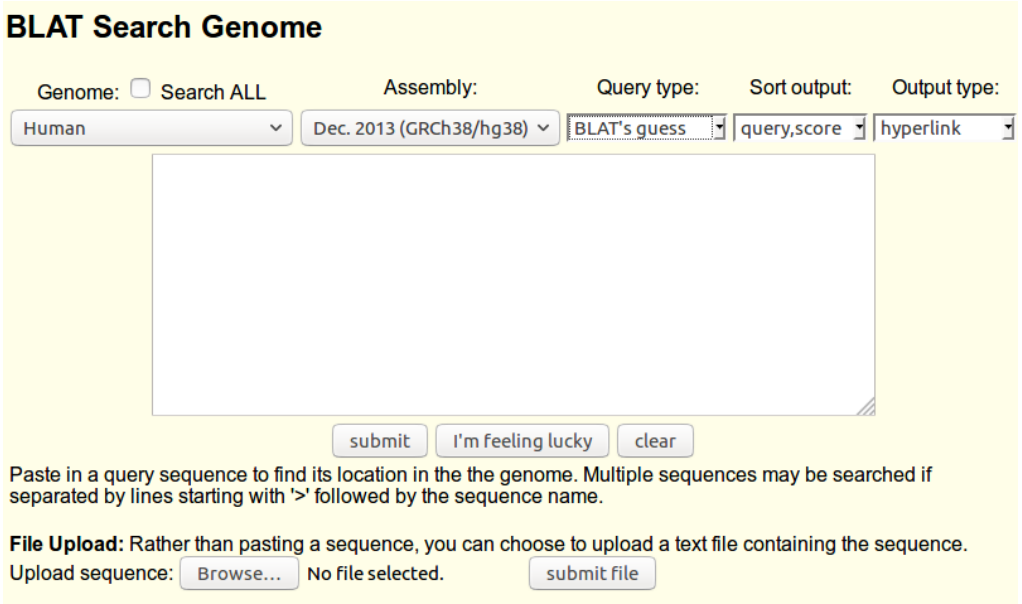
Practicum

UCSC Genome Browser

https://genome.ucsc.edu/

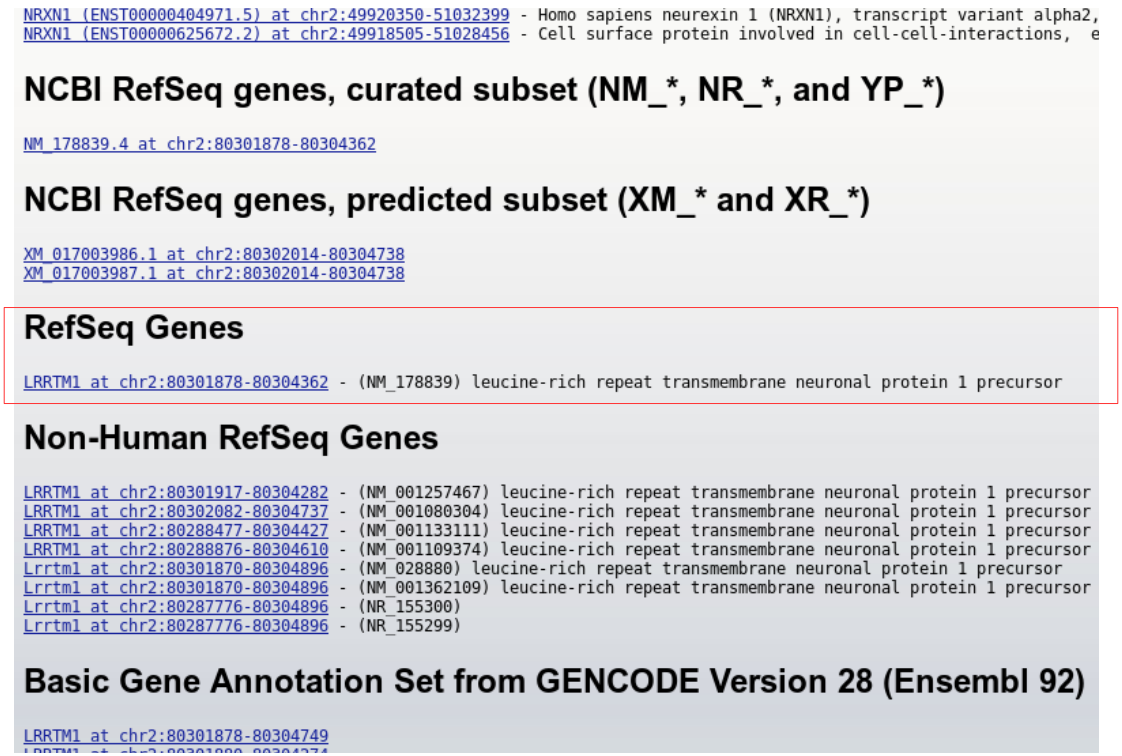
- Different search options:
  a) By gene/transcript/protein name, symbol or ID: LRRTM1
  b) By Chromosome number or region: chr11:1038475-1075482
  c) By Keywords: kinase, receptor
  d) By sequence (BLAT tool)
  e) By track type (Track search)





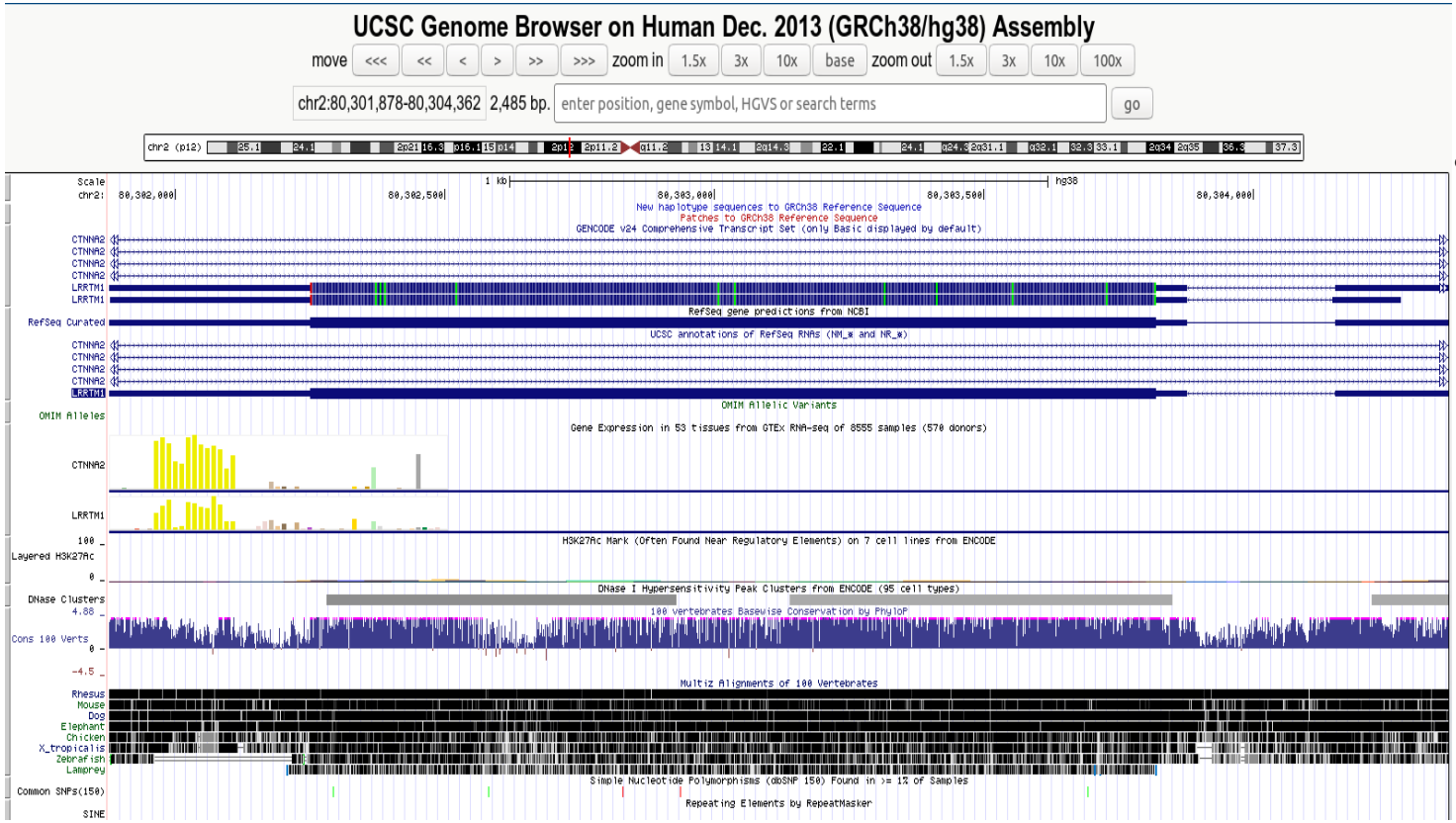
Practicum

- 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



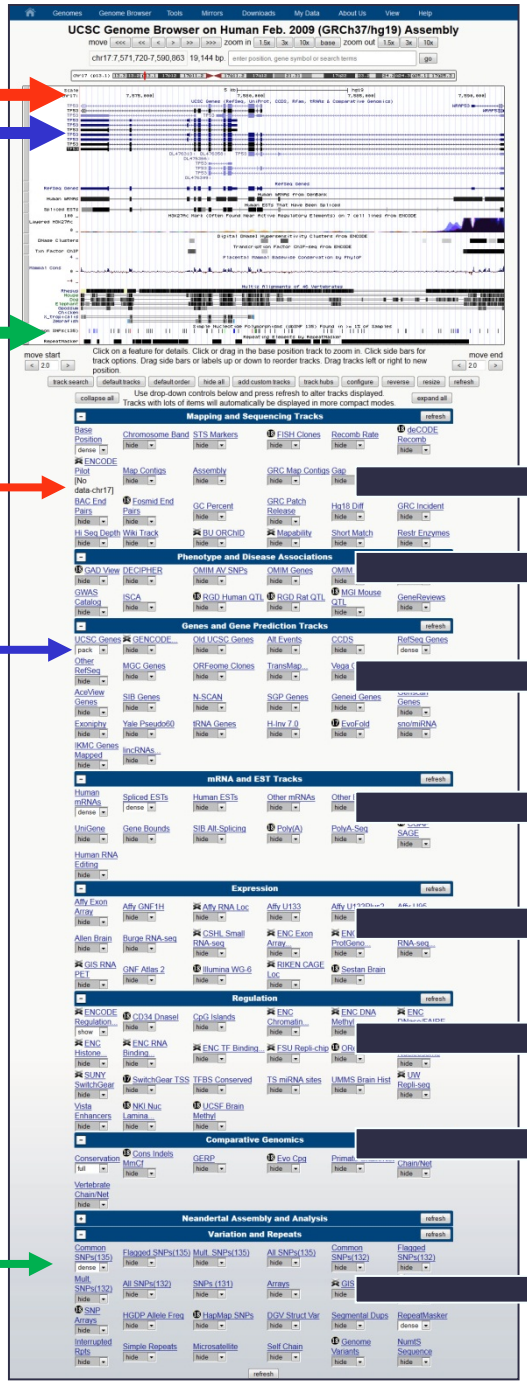
Practicum

- 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

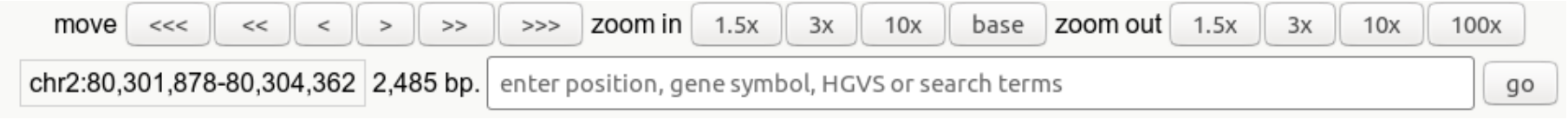
Individual species

Variation and Repeats

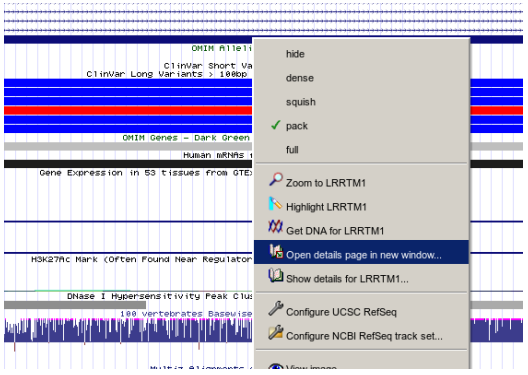
(including SNPs, copy number variation)

Practicum

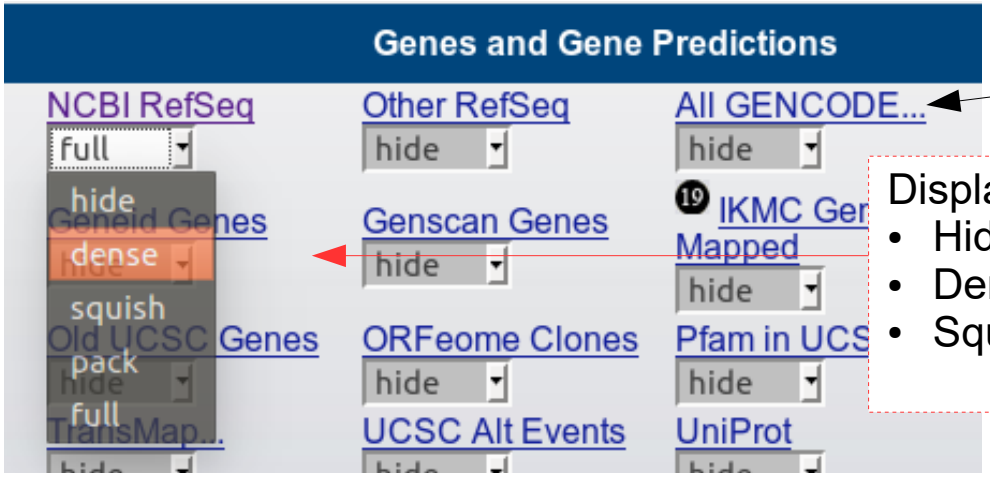
- Change your view or location with controls at the top



- 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



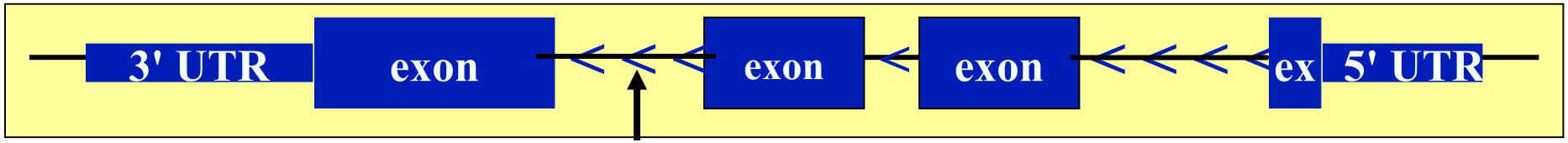
Source of information

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

Refresh!

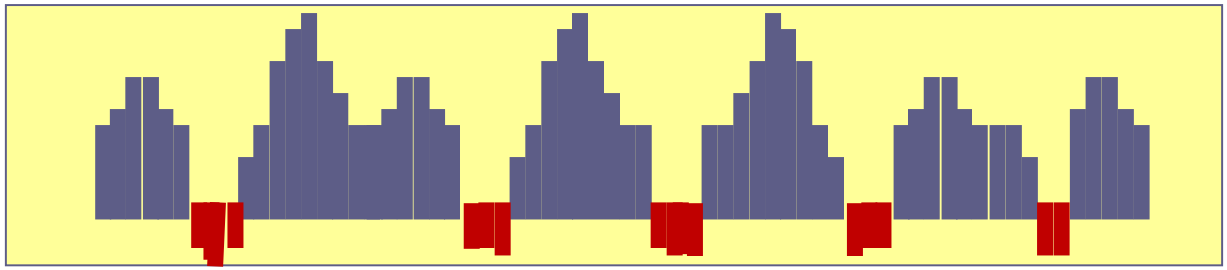
Practicum

- Some visual clues:



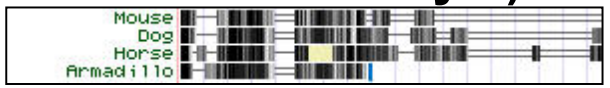
Intron and direction of transcription <<< or >>>

Sequence conservation



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

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



Practicum

Retrieving information with the UCSC Genome Browser

https://genome.ucsc.edu/

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

-chromosome:

-position:

-strand:

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

3. How many exons has the gene?

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

5. Can you find SNPs in this gene?

6. In which tissue is this gene mainly expressed?

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

8. Has this gene an ortholog in mouse?

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

Practicum

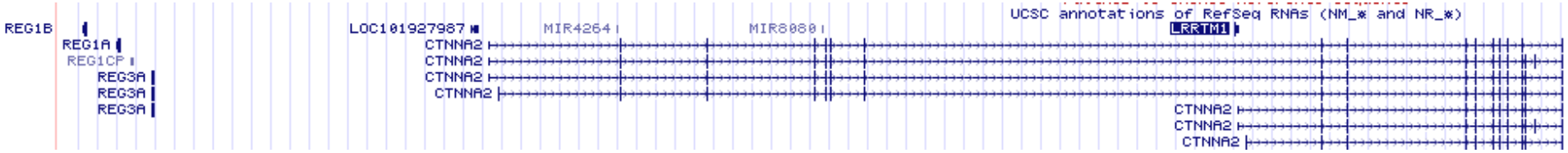
1. What is the genomic localization of human *lrrtm1* gene?

   - Click on the gene to see the information

Position: chr2:80301878-80304362
Band: 2p12
Genomic Size: 2485
Strand: -
Gene Symbol: LRRTM1
CDS Start: complete
CDS End: complete

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

   - Zoom out in genome viewer



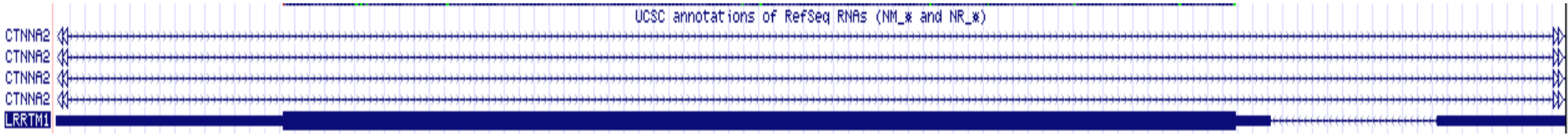
3. How many exons has the gene? 2

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

Practicum

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

   - Use Genome Viewer or Table Browser for more detailed information

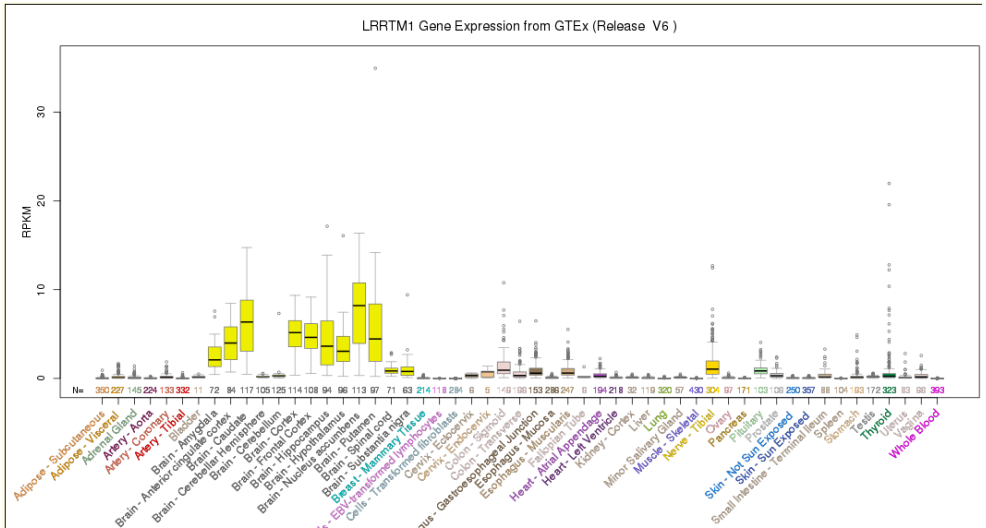


5. Can you find SNPs in this gene? yes

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

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

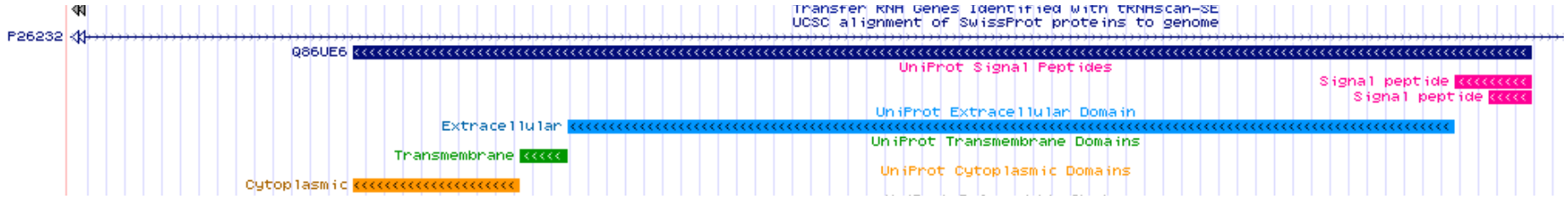
   - Info in track “Expression” > “GTEx”



Practicum

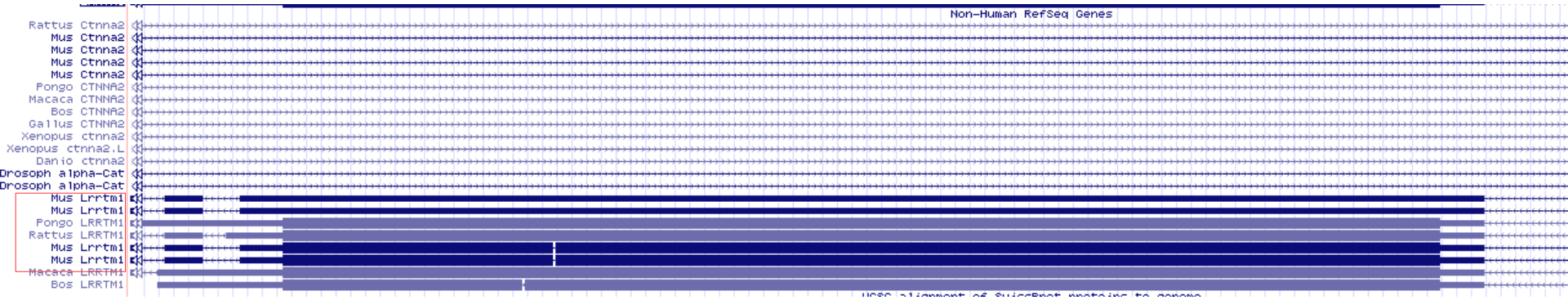
7. Does the protein encoded by this gene have a transmembrane domain?

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



8. Has this gene an ortholog in mouse? yes

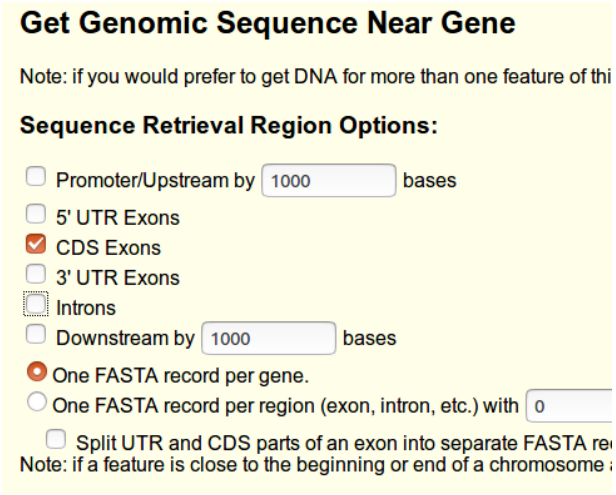
- Set display mode of track “Gene and Gene Predictions” > “Other RefSeq” to “pack”



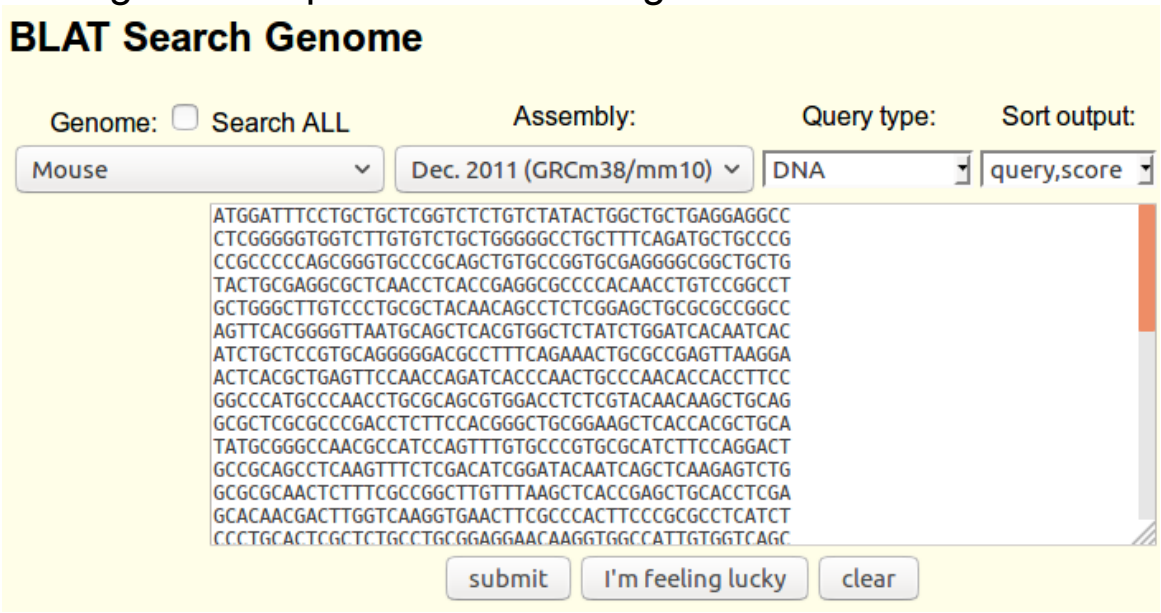
Practicum

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

- Get the CDS sequence of human gene:



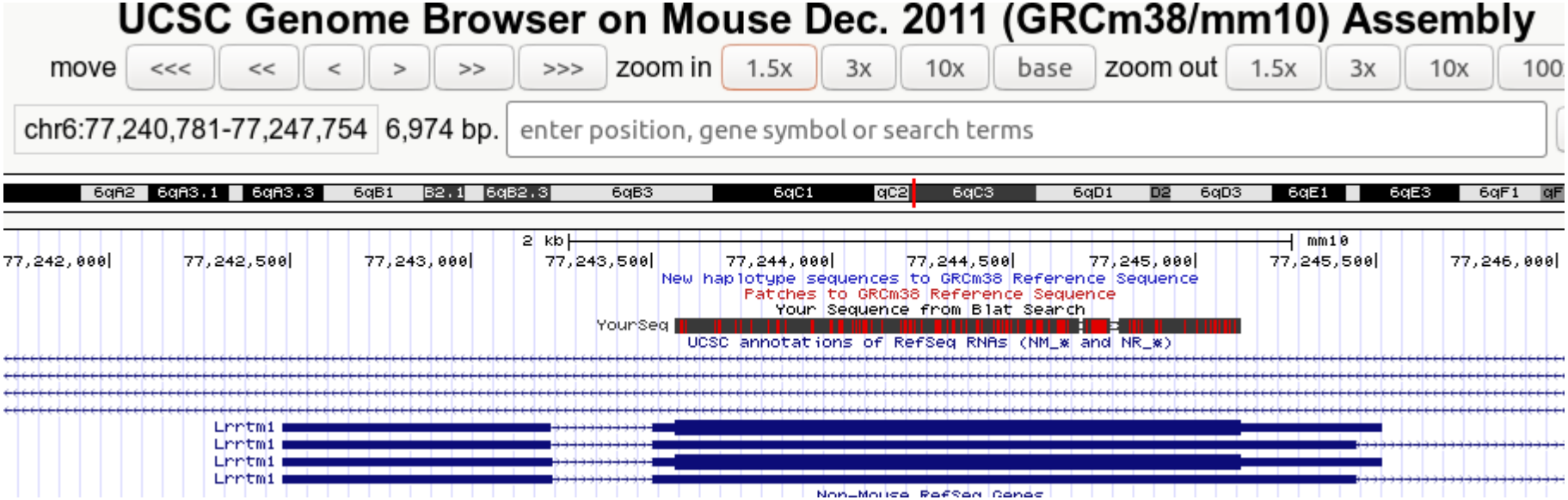
- Use BLAT to align this sequence on Mouse genome:



Practicum

- Visualize entry with highest score

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



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
- 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.
- GTEx

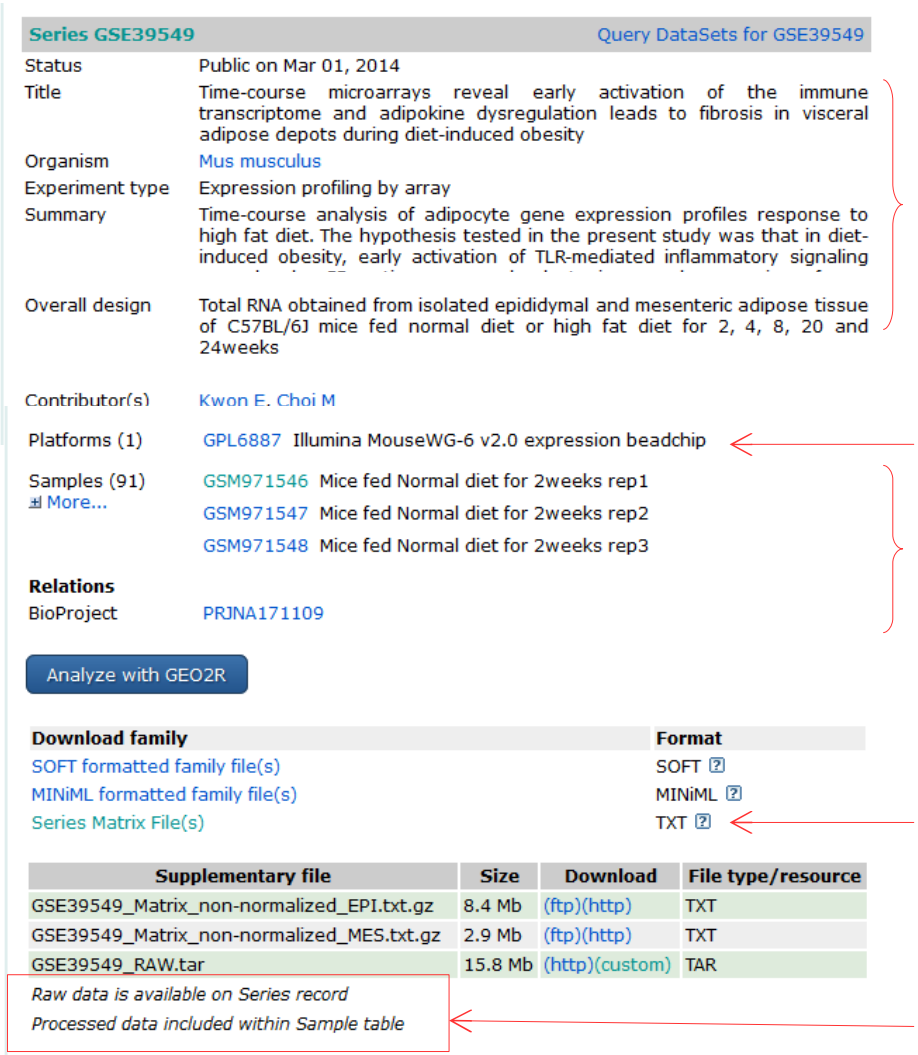
Retrieving data from GEO

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

Example with GDS browser:

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

Retrieving data from GEO



Series GSE39549 Query DataSets for GSE39549

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

Analyze with GEO2R

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

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

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

Study information

Platform used (data table with annotation of probes)

Samples

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

Info on data files

Practicum

Retrieving data from GEO

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

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

Data table header descriptions	
ID_REF	
VALUE	normalized signal

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

Data table with normalized expression values

Practicum

Retrieving data from GEO

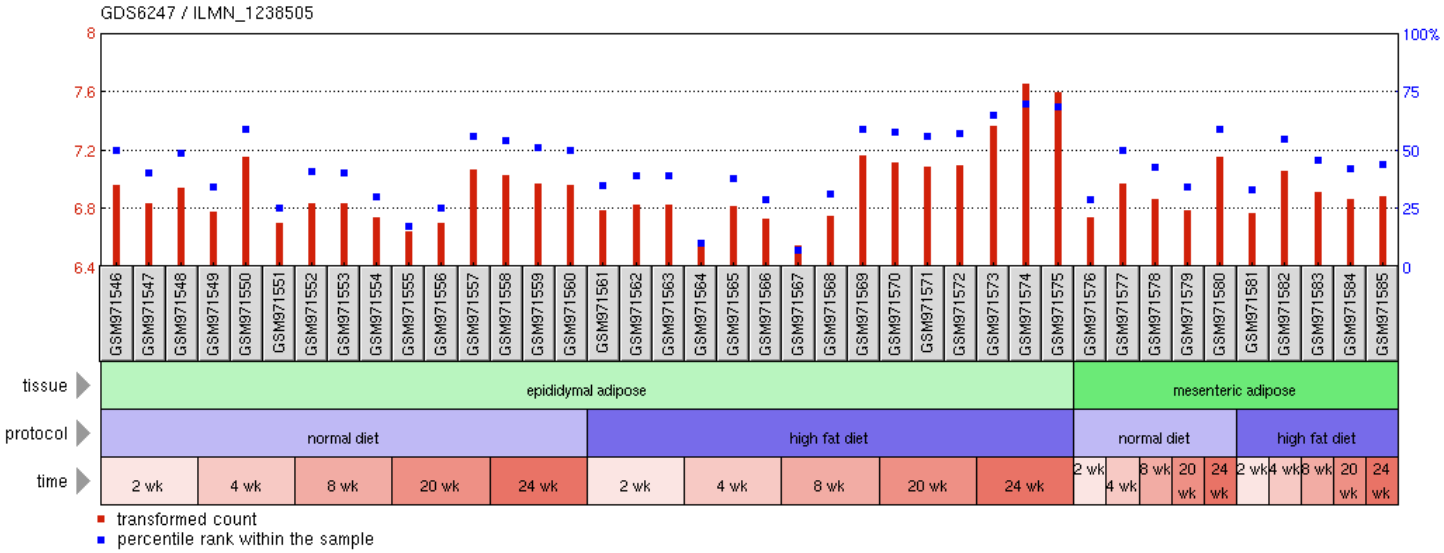
- GEO Profiles: stores individual gene expression profiles from curated DataSets.

Example: search for expression profile of IL-10

Builder

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

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



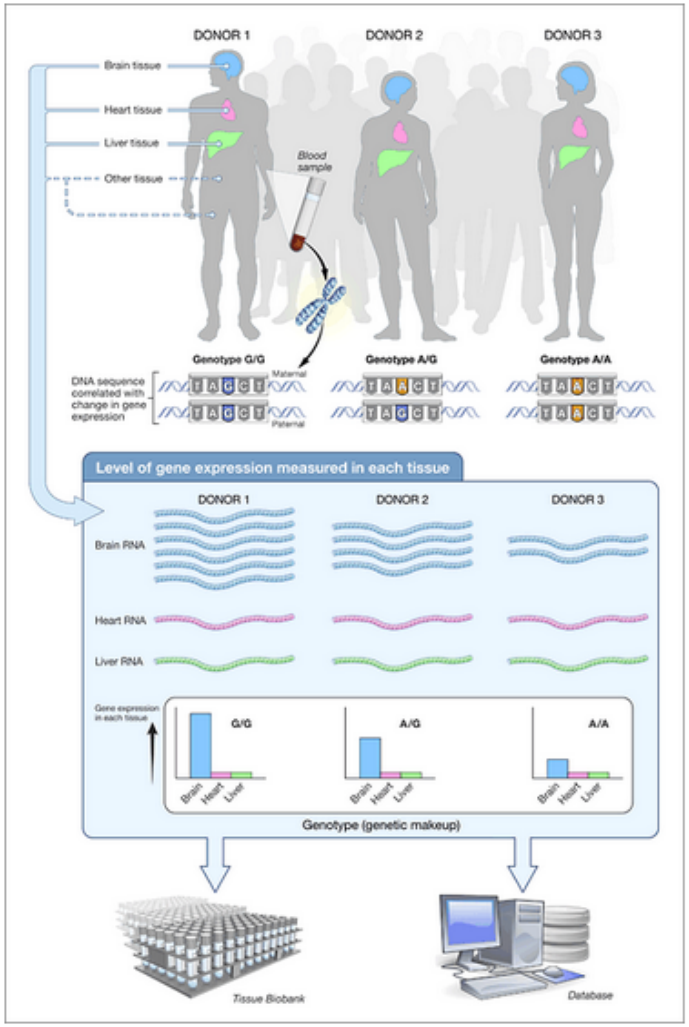
Examples of Databases

Genotype-Tissue Expression (GTEx) Project

https://gtexportal.org/home/

- Public resource to study tissue-specific gene expression and regulation and its relationship to genetic variation across individuals.
- Samples from 53 non-diseased tissues across nearly 1000 individuals
- Types of data provided:
  - Gene/transcript expression in tissues
  - variants associated to gene expression (expression quantitative trait loci, or eQTLs)

  - histology images
  - Patient/sample metadata



Practicum

Genotype-Tissue Expression (GTEx) Project

https://gtexportal.org/home/

GTEx Portal

GTEx | Datasets | Expression | QTLs & Browsers | Sample Data | Biobank | Documentation | Publications | Contact | FAQs

Search Gene or SNP ID... Sign In

2018-04-04
New eQTL Dashboard Visualization
Read More >>

Current Release

Latest Version: V7

Download | Summary Statistics | How to cite GTEx?

The Genotype-Tissue Expression (GTEx) project is an ongoing effort to build a comprehensive public resource to study tissue-specific gene expression and regulation. Samples were collected from 53 non-diseased tissue sites across nearly 1000 individuals, primarily for molecular assays including WGS, WES, and RNA-Seq. Remaining samples are available from the GTEx Biobank. The GTEx Portal provides open access to data including gene expression, QTLs, and histology images.

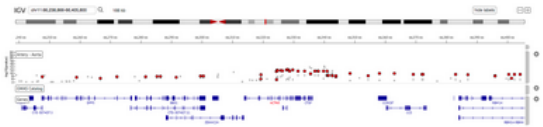
The current release is V7 including 11,688 samples, 53 tissues and 714 donors.

Genetic Association

Single-Tissue eQTLs

Search eQTL by gene or SNP ID

GTEx IGV eQTL Browser



GTEx Gene-eQTL Visualizer

Transcriptome

Search expression by gene ID...

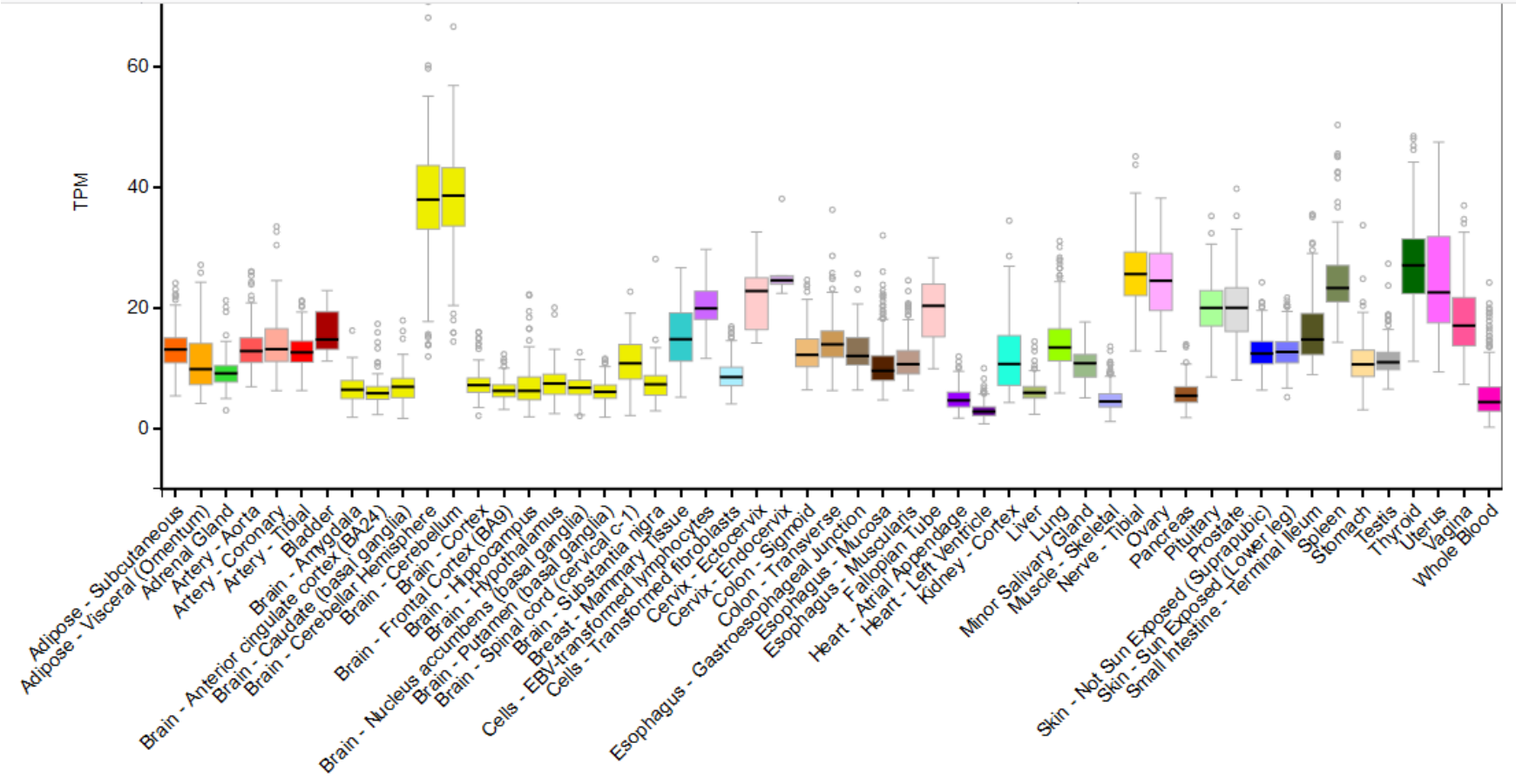
Top Expressed Genes in a Tissue

Gene Expression in Tissues

Exon and Transcript Expression

Genotype-Tissue Expression (GTEx) Project

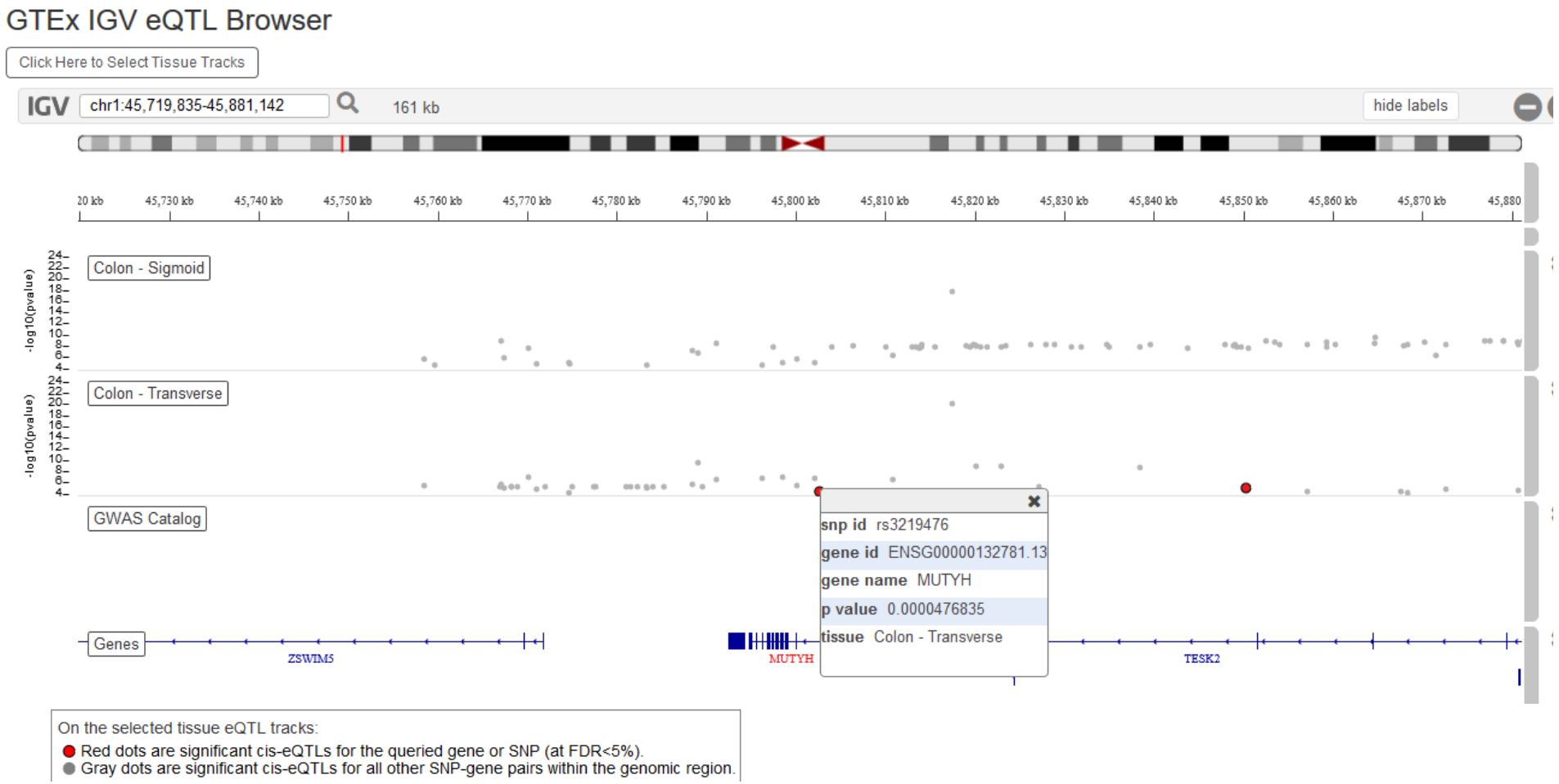
Example: Normal tissue expression profile of *mutyh* gene



Practicum

Genotype-Tissue Expression (GTEx) Project

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



Other Databases

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

Other Databases

The Cancer Genome Atlas (TCGA)

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

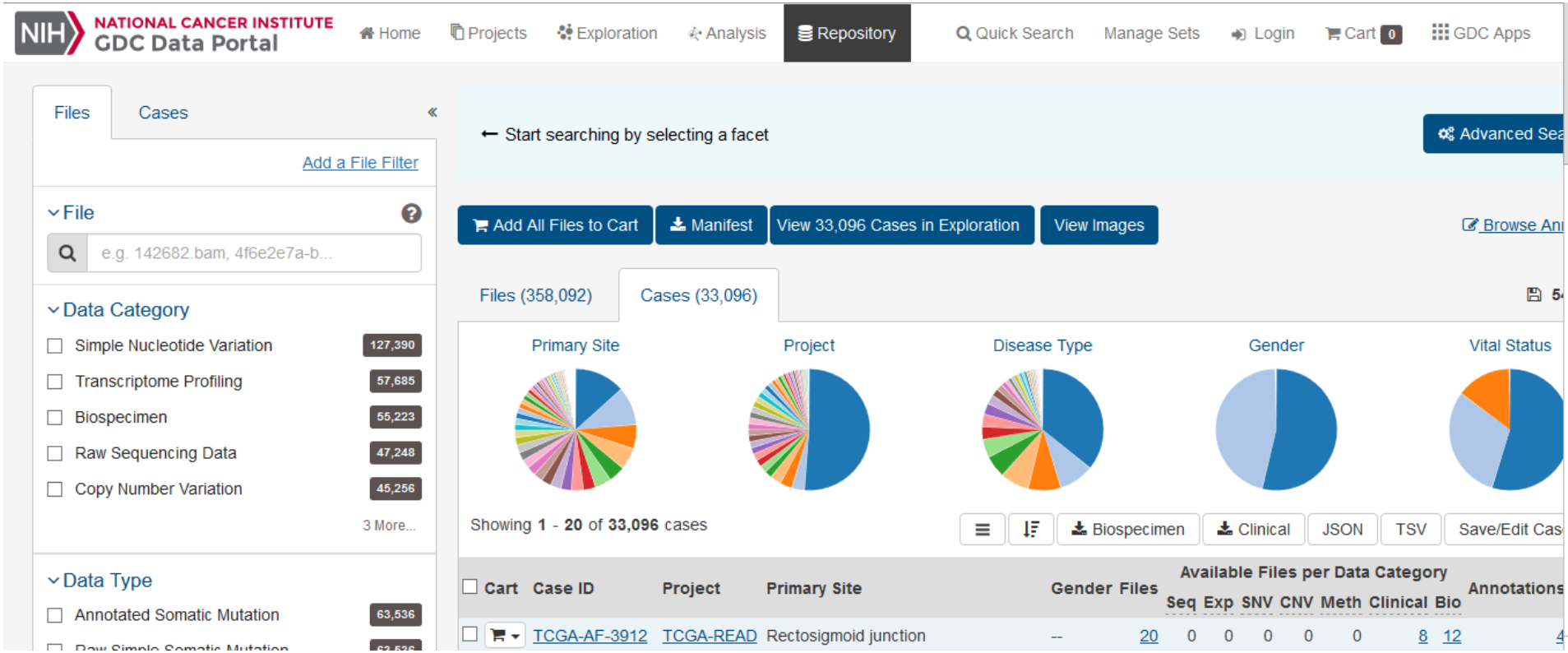
- Collaboration between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) that has generated comprehensive, multi-dimensional maps of the key genomic changes in 33 types of cancer.
- High-quality tumor and matched normal samples from over 11,000 patients. Dataset,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

Practicum

The Cancer Genome Atlas (TCGA)

https://cancergenome.nih.gov/

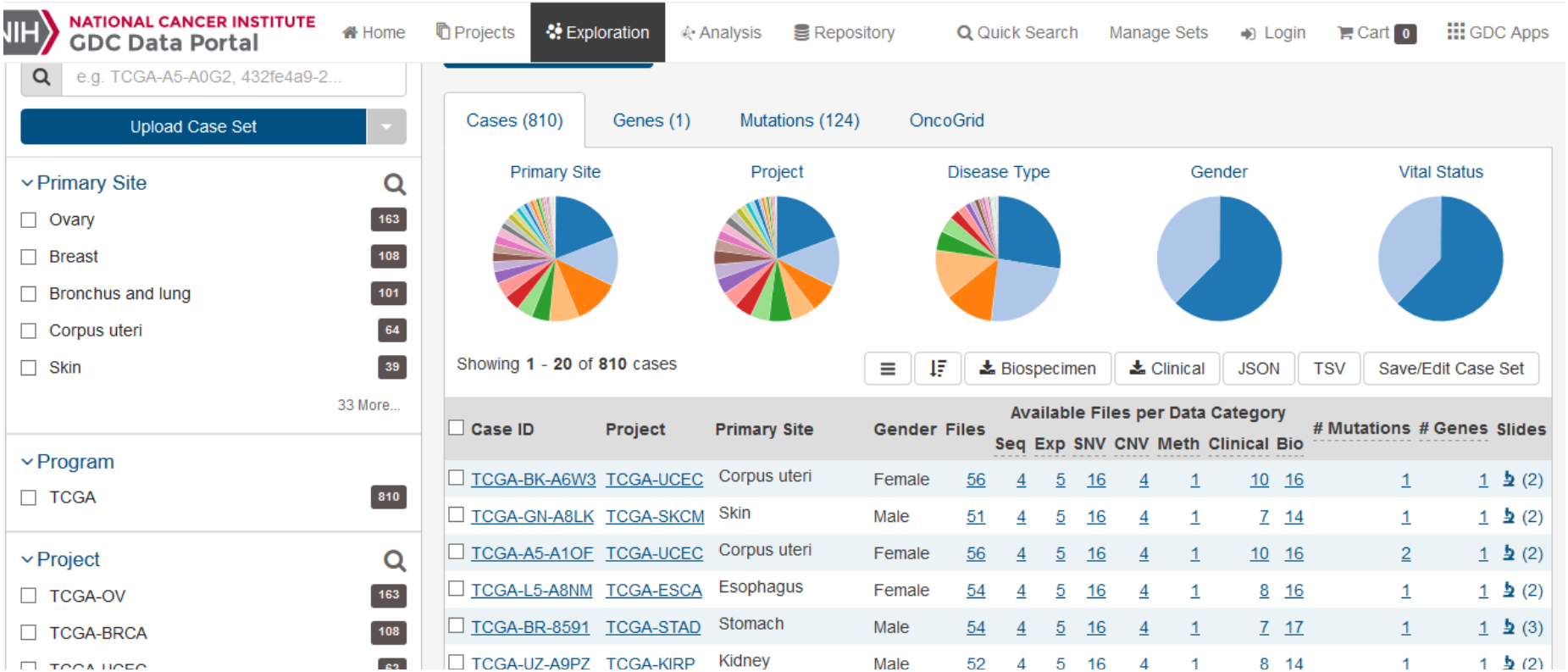
Data in https://portal.gdc.cancer.gov/



Practicum

The Cancer Genome Atlas (TCGA)

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



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