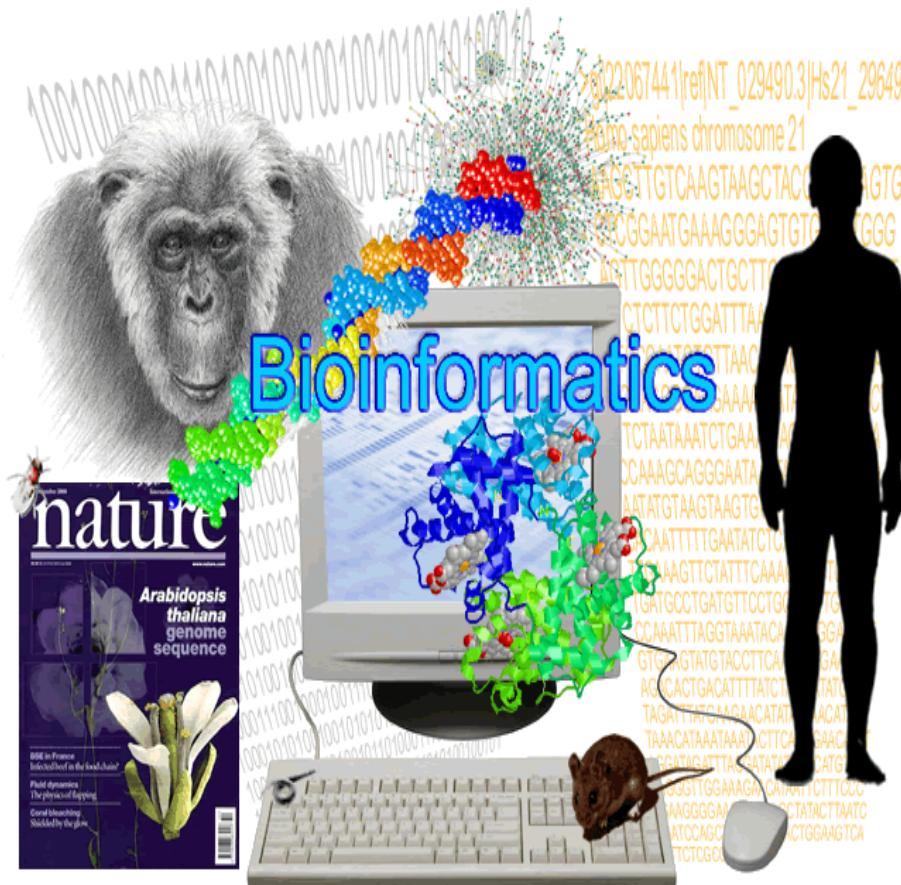


BIOINFORMATICS COURSE

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

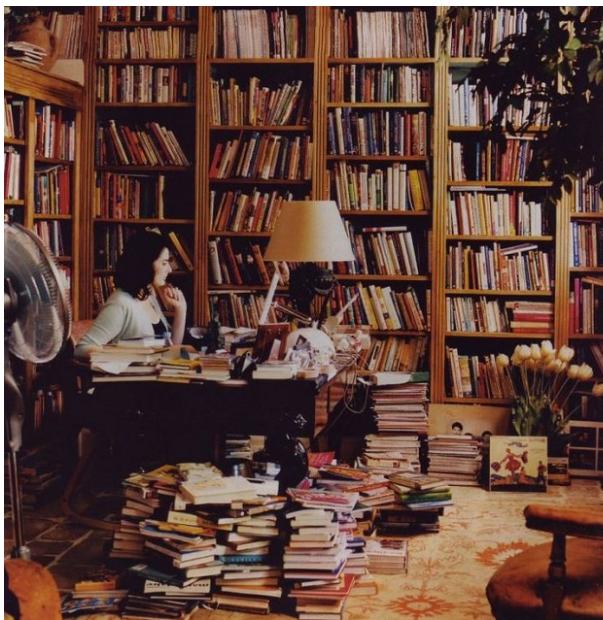
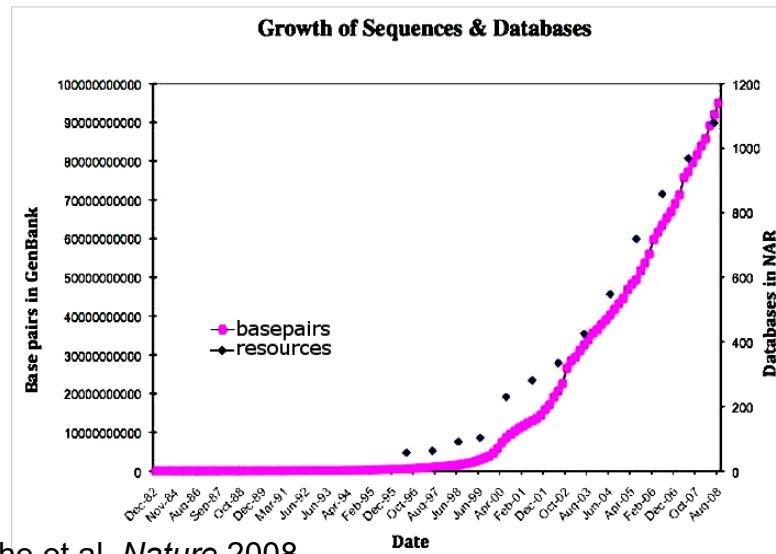
• Microarray Databases

- ArrayExpress
- MIAME

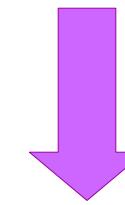
• Literature Databases

- MEDLINE
- OMIM
- Patent Abstracts
- more...

Challenges



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

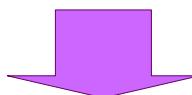
Integrating the information

- **Primary databases:**

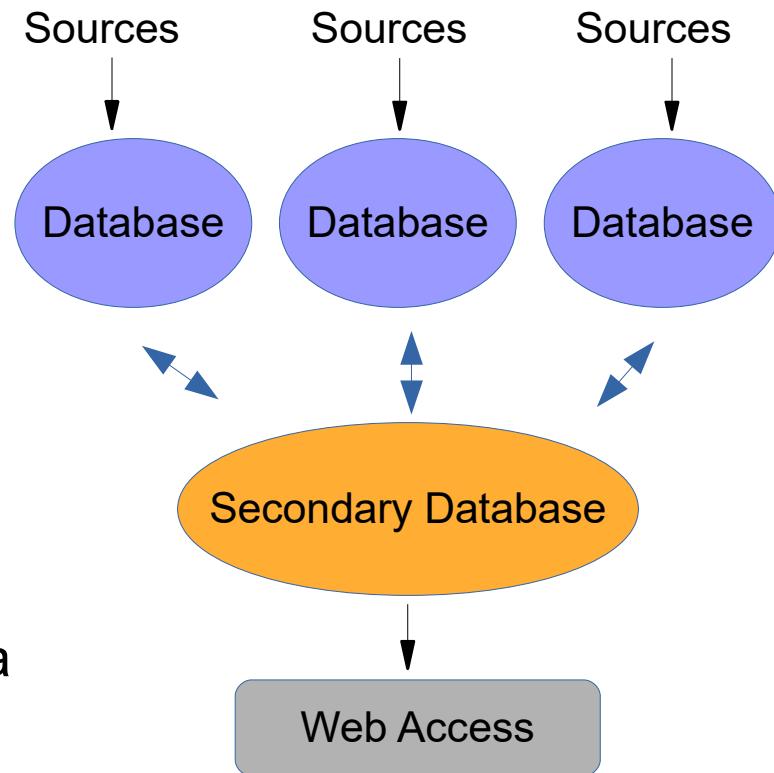
- often hold only one type of specific data which is stored in their own archive.
- upload new data from experiments and update entries

- **Secondary databases:**

- use other databases as their source of information.
- often already process or analyze the data to get new results.



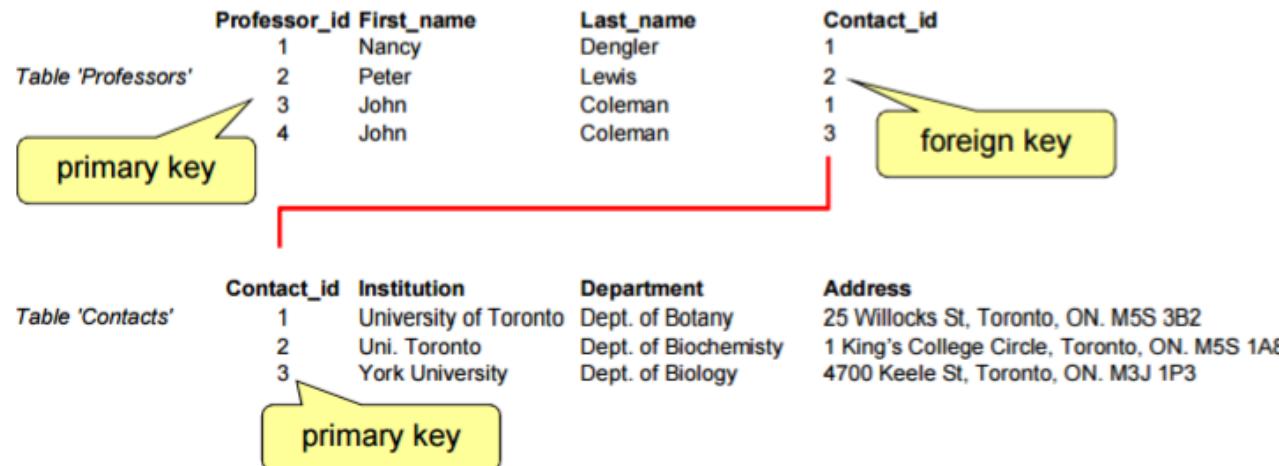
Relational databases



Integrating the information

- Relational databases:

- cross-reference to other databases through a common field (unique identifier)
- 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

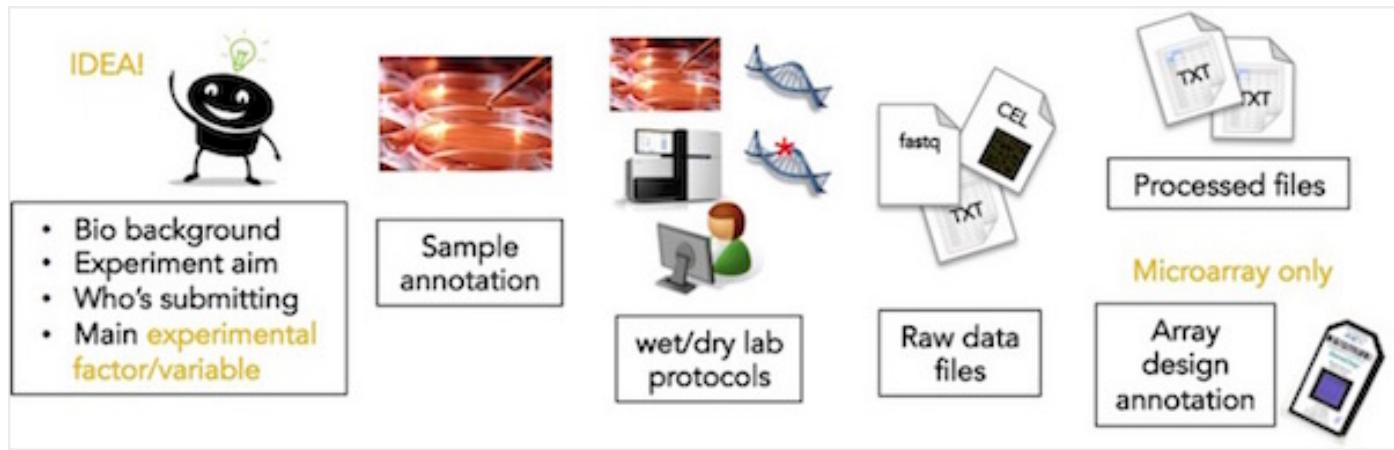
Integrating the information

II. Data Annotation and Curation

Data Annotation

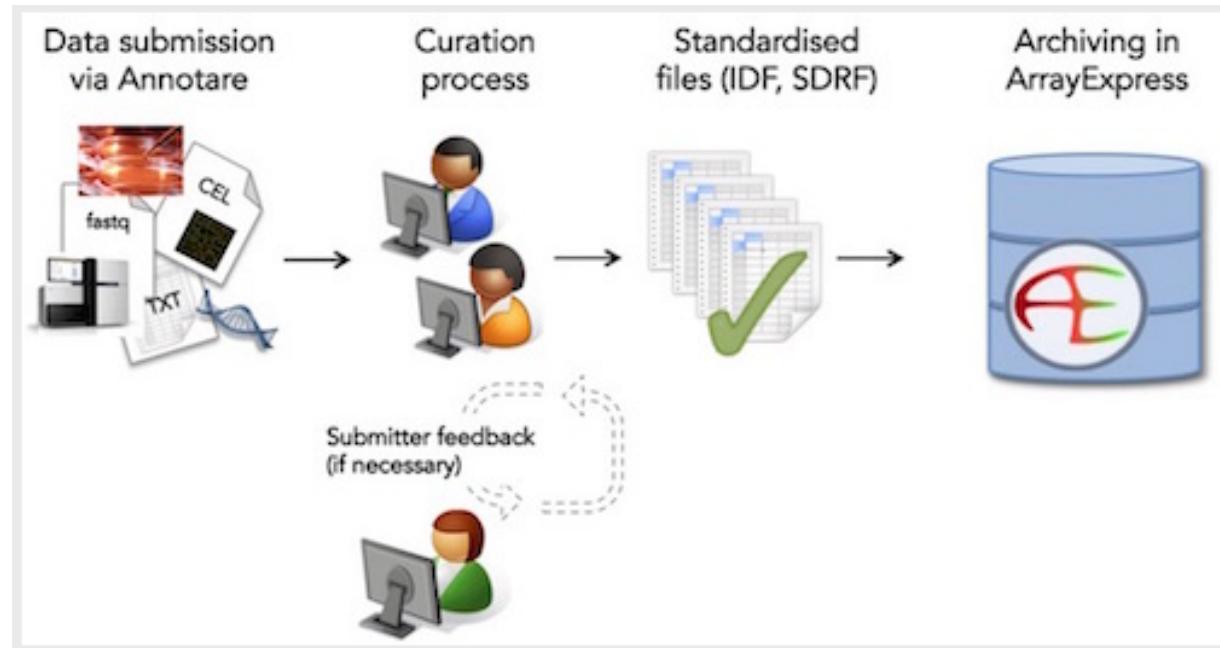
- Collaborative efforts to provide as much information about the data
 - Annotation of sequences/genomes: chromosome position, gene function, ...
 - Metadata: information for an experiment, identification of samples, ...
 - ◆ The **minimum information standard** is a set of guidelines for reporting data
- 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



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.

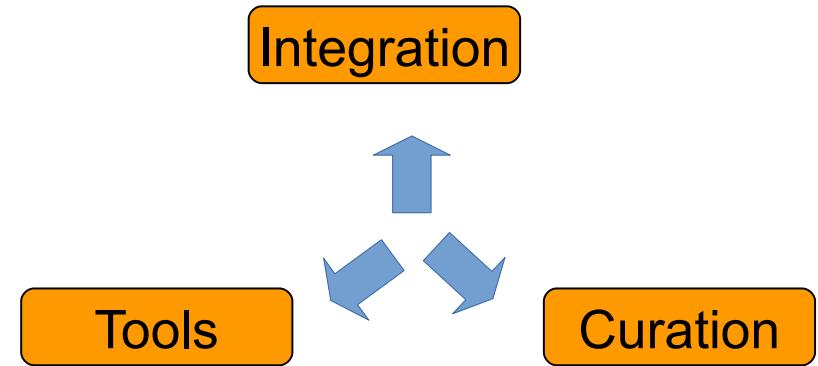


Source: <https://www.ebi.ac.uk/arrayexpress/submit/overview.html>

III. Centralizing data management

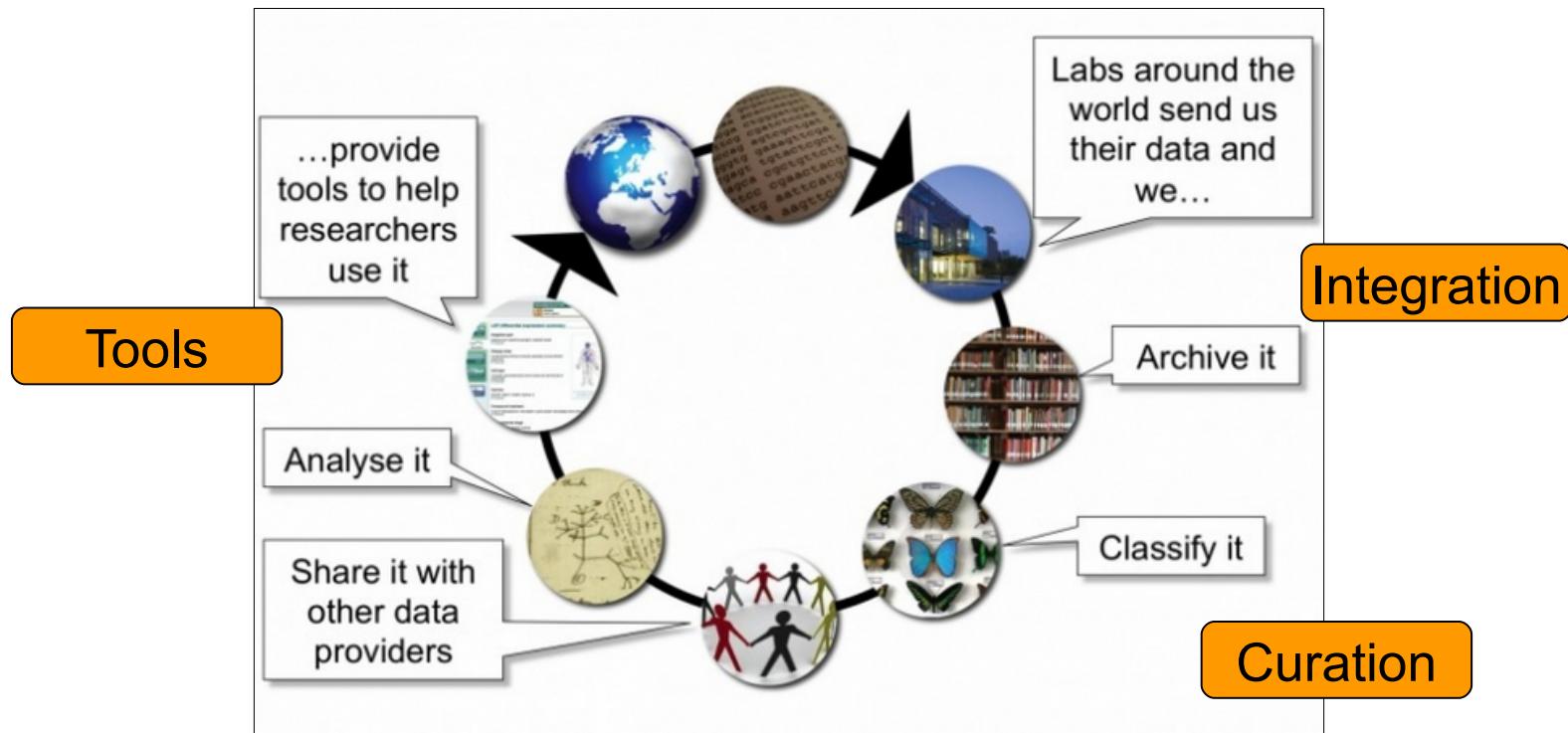
Centralizing data management

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



Resource providers

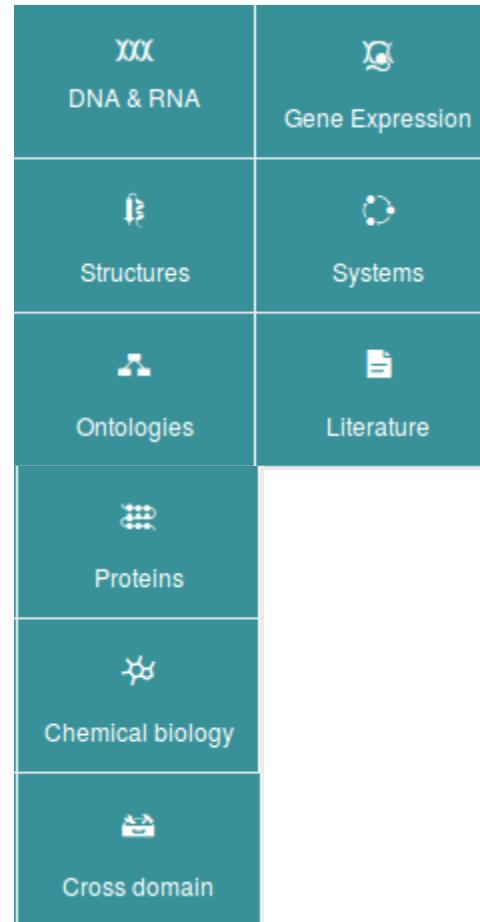
- Big organizations that act as *hubs* that provide transparent access to data sources.



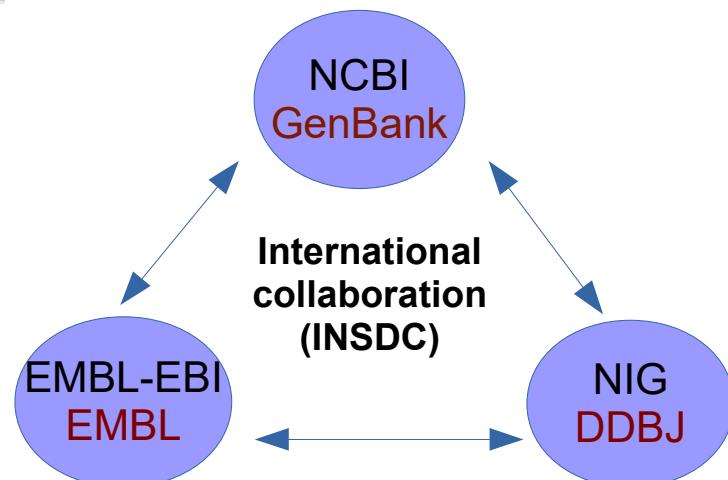
Resource providers



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



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



Resource providers



[Amino Acid Explorer](#)

[Assembly Archive](#)

[Basic Local Alignment Search Tool \(BLAST\)](#)

[Batch Entrez](#)

[BioAssay Services](#)

[BLAST Link \(BLink\)](#)

[BLAST Microbial Genomes](#)

[BLAST RefSeqGene](#)

[CDTree](#)

[Cn3D](#)

[COBALT](#)

[Concise Microbial Protein BLAST](#)

[Conserved Domain Architecture Retrieval Tool \(CDART\)](#)

[Conserved Domain Search Service \(CD Search\)](#)

[Digital Differential Display \(DDD\)](#)

[Electronic PCR \(e-PCR\)](#)

[Frequency-weighted Link \(FLink\)](#)

Clustal Omega



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

[Multiple sequence alignment](#)

InterProScan



InterProScan searches sequences against InterPro's predictive protein signatures.

[Protein feature detection](#) [Sequence motif recognition](#)

BLAST [protein]



Fast local similarity search tool for protein sequence databases.

[Sequence similarity search](#)

BLAST [nucleotide]



Fast local similarity search tool for nucleotide sequence databases.

[Sequence similarity search](#)

HMMER



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

[Sequence similarity search](#) [Protein function analysis](#)

- Provide a wide variety of data analysis tools that allow users to 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'Alessio²

► Author information ► Copyright and License information [Disclaimer](#)

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

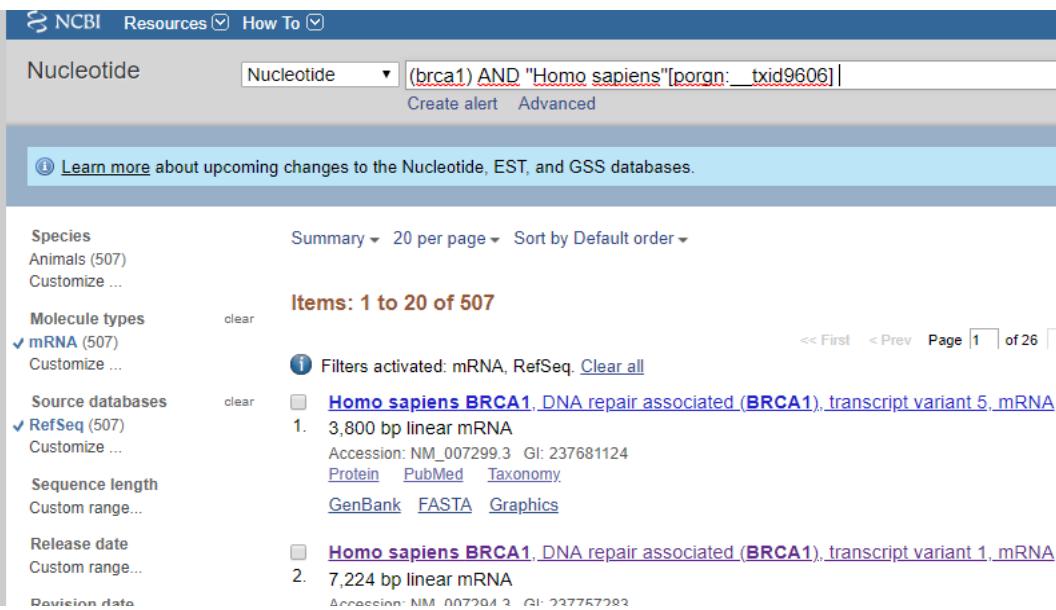
[Abstract](#) [Go to: !\[\]\(6a3a10fac78c4674bc151043e1625557_img.jpg\)](#)

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

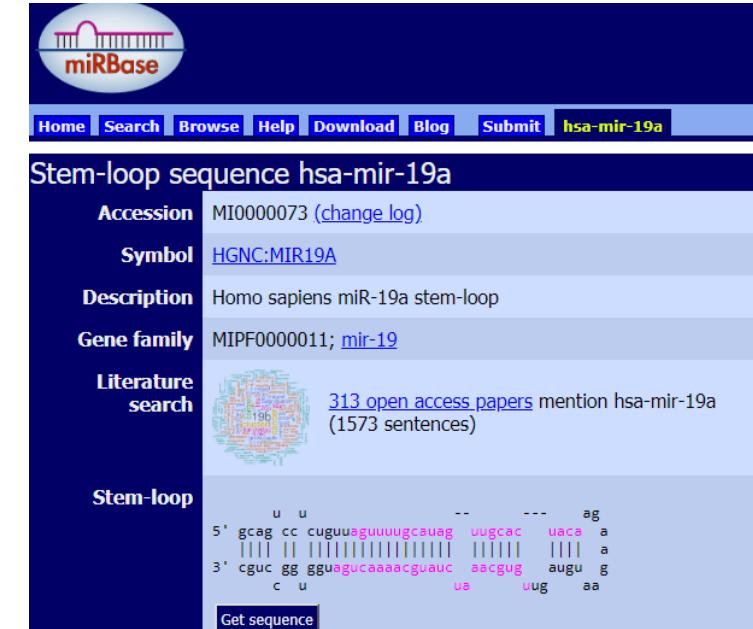
Examples of Databases

Nucleotide DB

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


 NCBI Resources How To
 Nucleotide Nucleotide (brca1) AND "Homo sapiens"[organism:txid9606]
 Create alert Advanced
 Learn more about upcoming changes to the Nucleotide, EST, and GSS databases.
 Species Animals (507) Summary 20 per page Sort by Default order
 Molecule types mRNA (507) Items: 1 to 20 of 507
 Source databases RefSeq (507)
 Release date Custom range...
 Revision date

Items: 1 to 20 of 507
 Filters activated: mRNA, RefSeq. [Clear all](#)
 Homo sapiens BRCA1, DNA repair associated (BRCA1), transcript variant 5, mRNA
 1. 3,800 bp linear mRNA
 Accession: NM_007299.3 GI: 237681124
[Protein](#) [PubMed](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Graphics](#)
 Homo sapiens BRCA1, DNA repair associated (BRCA1), transcript variant 1, mRNA
 2. 7,224 bp linear mRNA
 Accession: NM_007299.3 GI: 237767292


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

Stem-loop sequence hsa-mir-19a
Accession MI0000073 ([change log](#))
Symbol HGNC:MIR19A
Description Homo sapiens miR-19a stem-loop
Gene family MIPF0000011; mir-19
Literature search 313 open access papers mention hsa-mir-19a (1573 sentences)
Stem-loop


```

      u   u          --   ---   ag
5' gcag cc cuguuaguuuugcauag uugcac uaca a
            |   |   |   |   |   |
3' cguc gg gguaguacaaacguauc aacgug augu g
            c   u           ua   lug aa
  
```

[Get sequence](#)

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*

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

Search results (partial list):

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

Buttons: Previous 200, Next 200, Refresh index

Bottom controls: AND dropdown, All Fields dropdown, Show index list, Search button, Add to history link

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 the "Homo sapiens mutY DNA glycosylase (MUTYH), transcript variant 13, mRNA". It provides details such as accession number NM_001350651.1, GI number 1183596751, and links to Protein, PubMed, and Taxonomy. Below this, another result for "Homo sapiens mutY DNA glycosylase (MUTYH), transcript variant 12, mRNA" is listed. A red bracket on the left side groups the "Species" and "Molecule types" filter sections.

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

Molecule types clear

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
ATATGCTGAAGTGGCTACACTGAGGACCTGGCCAGTGCTCCCTGGAGGAGGTGAATCAAACCTGGG
```

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?

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 level¹

Function

Names & Taxonomy

Subcellular location

Function¹

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 activity¹

Practicum

Interactionⁱ

Binary interactionsⁱ

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

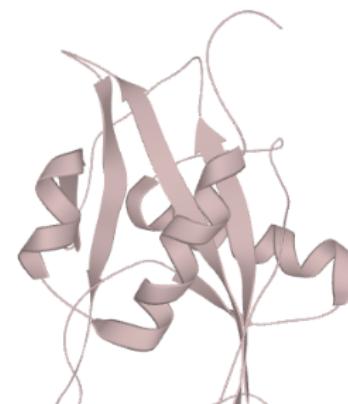
Protein-protein interaction databases

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

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>


BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog
Login/Register
 Search all species...

Tools	BioMart >	BLAST/BLAT >	Variant Effect Predictor >	
All tools	Export custom datasets from Ensembl with this data-mining tool	Search our genomes for your DNA or protein sequence	Analyse your own variants and predict the functional consequences of known and unknown variants	Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

Search

All species for

e.g. [BRCA2](#) or [rat 5:62797383-63627669](#) or [rs699](#) or [coronary heart disease](#)

All genomes	Favourite genomes
<input type="button" value="-- Select a species --"/> <ul style="list-style-type: none"> View full list of all Ensembl species Edit your favourites 	 Human GRCh38.p12 Still using GRCh37?  Mouse GRCm38.p6

www.ensembl.info/2018/11/13/whats-coming-in-ensembl-95-ensembl-genomes-42/

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

Ensembl Release 94 (October 2018)

- New fish: 38 new and updated fish genomes
- GENCODE update 29 for human and M19 for mouse
- Additional pathogenicity predictors for missense variants
- New transcription factor binding motifs from SELEX
- Gene trees using HMMs

[More release news](#) on our blog

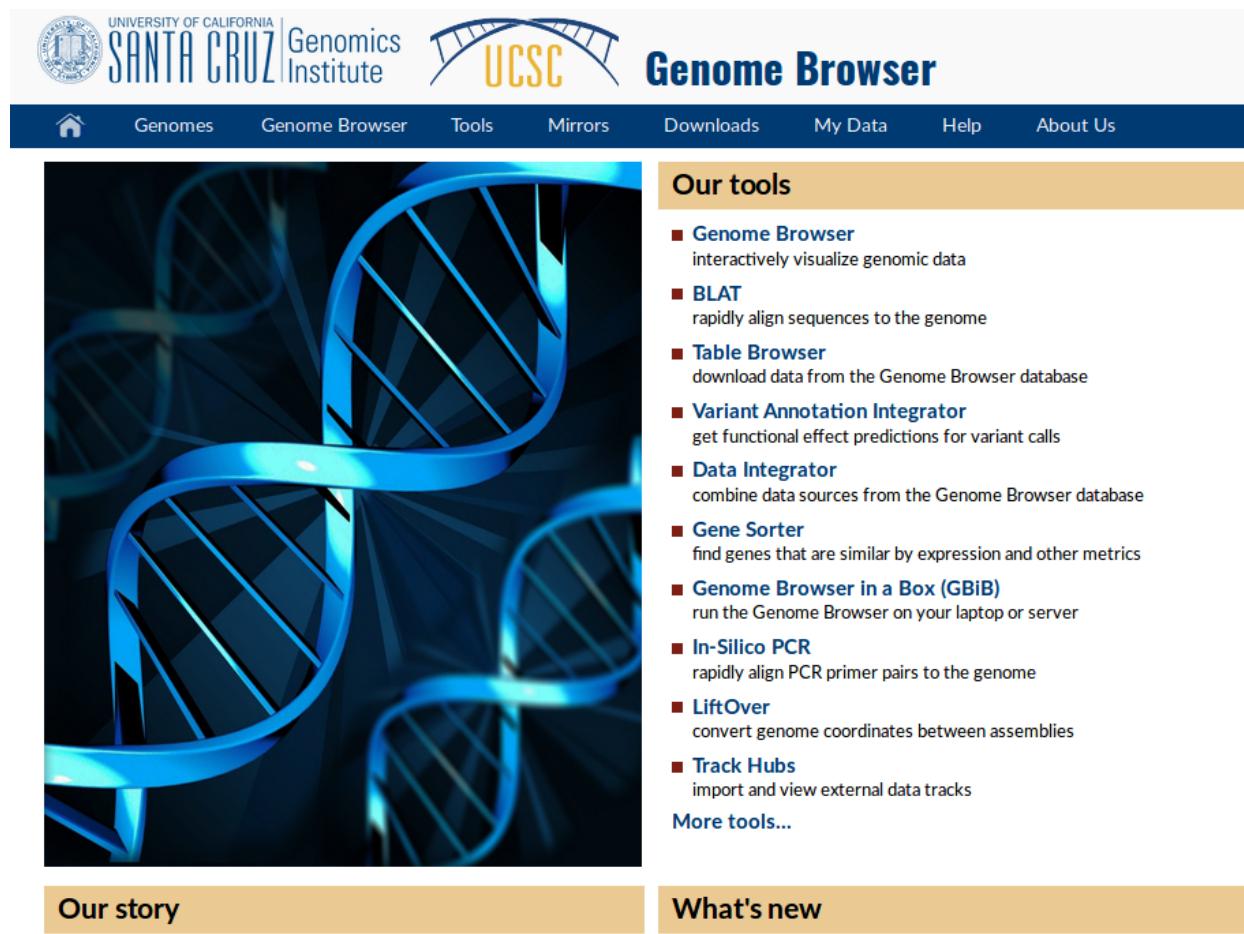
Other news from our blog

- 13 Nov 2018: [What's coming in Ensembl 95 / Ensembl Genomes 42](#)
- 06 Nov 2018: [Job: Regulation Project Leader](#)
- 06 Nov 2018: [Job: Bioinformatician – comparative genomics](#)

Genomic Browsers

UCSC Genome Browser

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



The screenshot shows the UCSC Genome Browser homepage. At the top, there's a header with the University of California Santa Cruz Genomics Institute logo and the text "UCSC Genome Browser". Below the header is a navigation bar with 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 sidebar titled "Our tools" which lists various genomic tools: Genome Browser, BLAT, Table Browser, Variant Annotation Integrator, Data Integrator, Gene Sorter, Genome Browser in a Box (GBiB), In-Silico PCR, LiftOver, and Track Hubs, along with a "More tools..." link. At the bottom left is a "Our story" section, and at the bottom right is a "What's new" section with links to recent news items.

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...](#)

Our story

On June 22, 2000, UCSC and the other members of the International Human Genome Project consortium completed the first working draft of the human genome assembly, forever ensuring free public access to the genome and the information it contains.

What's new

Nov. 13, 2018 - New video: Saving and Sharing Sessions

Nov. 09, 2018 - BLAT ALL genomes feature

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)

Find Position

Human Assembly
Dec. 2013 (GRCh38/hg38)

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

BLAT Search Genome

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

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

submit I'm feeling lucky clear

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

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

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

[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

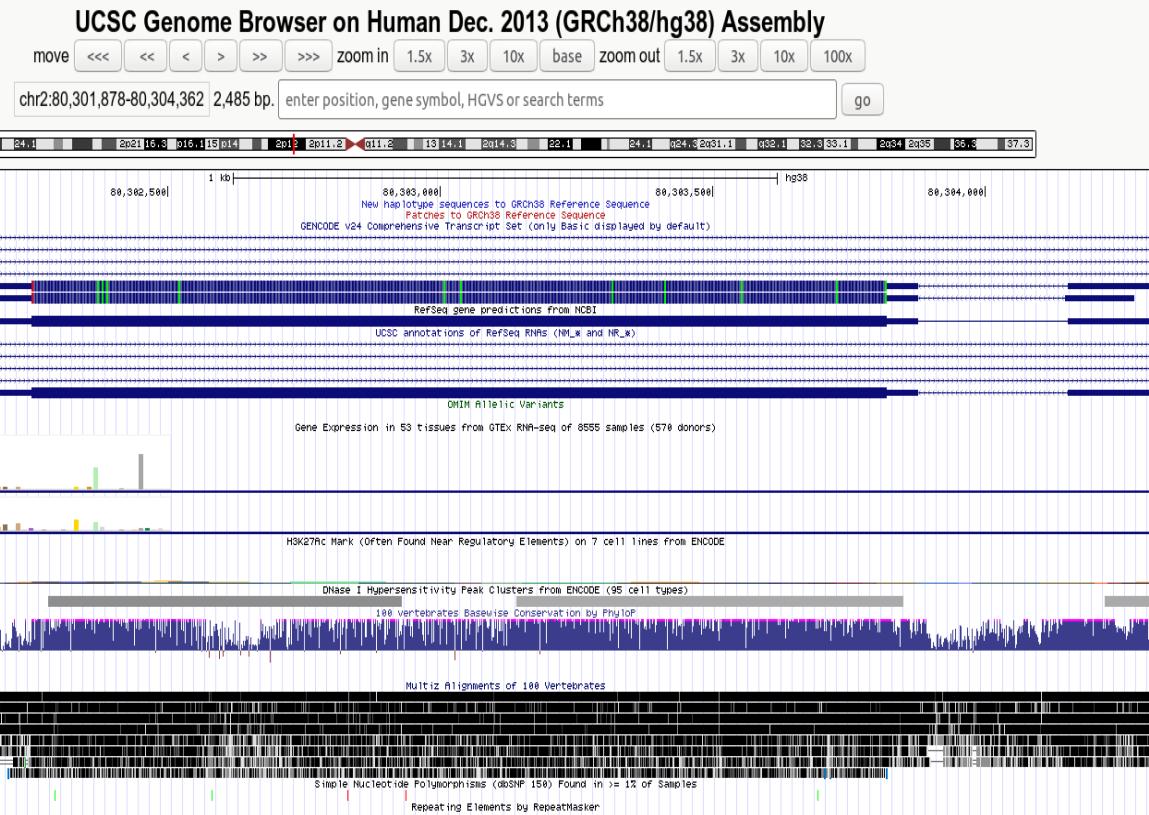
[LRRTM1 at chr2:80301917-80304282](#) - (NM_001257467) leucine-rich repeat transmembrane neuronal protein 1 precursor
[LRRTM1 at chr2:80302082-80304737](#) - (NM_001080304) leucine-rich repeat transmembrane neuronal protein 1 precursor
[LRRTM1 at chr2:80288477-80304427](#) - (NM_001133111) leucine-rich 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](#)

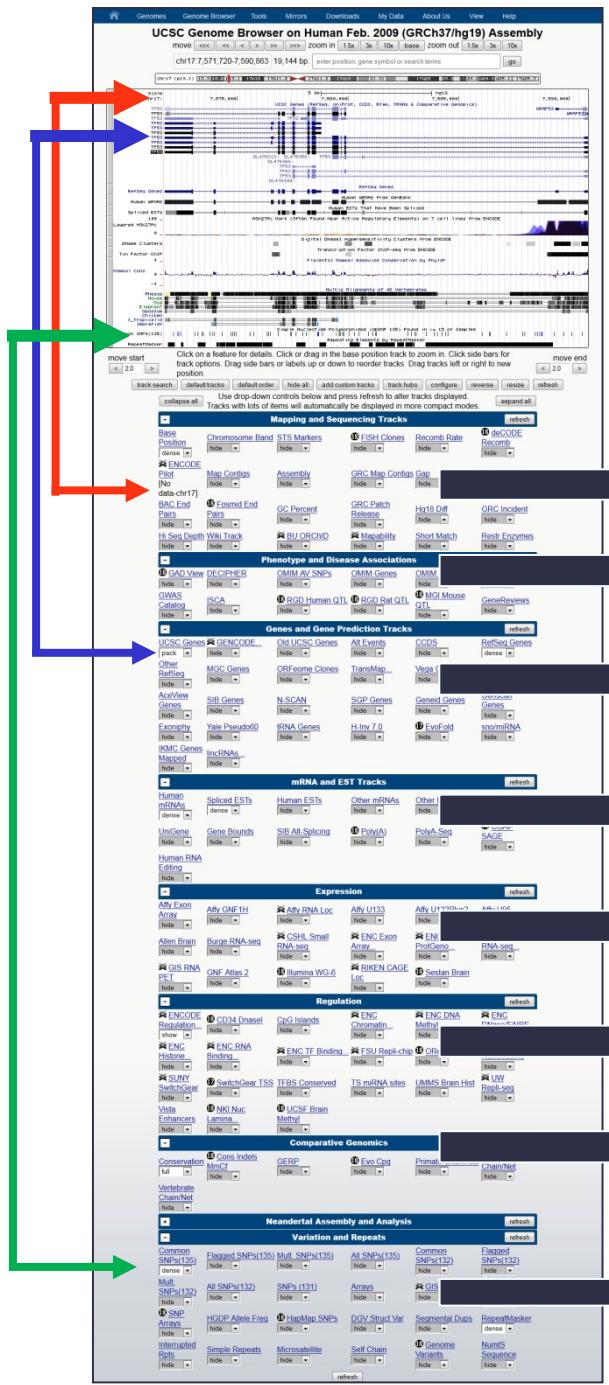
Practicum

- 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

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)

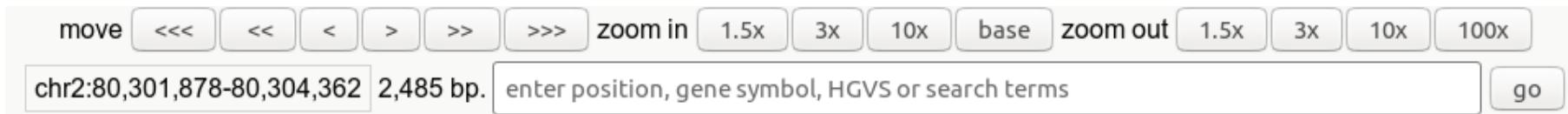
Regulation (including TFBS)

Comparative Genomics
As a group
Individual species

Variation and Repeats
(including SNPs, copy number variation)

Practicum

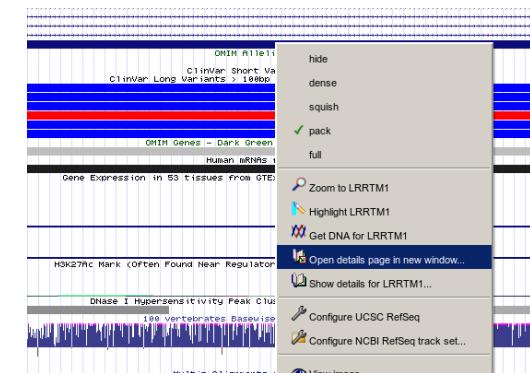
- Change your view or location with controls at the top



move <<< << < > >> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x

chr2:80,301,878-80,304,362 2,485 bp. enter position, gene symbol, HGVS or search terms go

- 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

Genes and Gene Predictions

Source of information

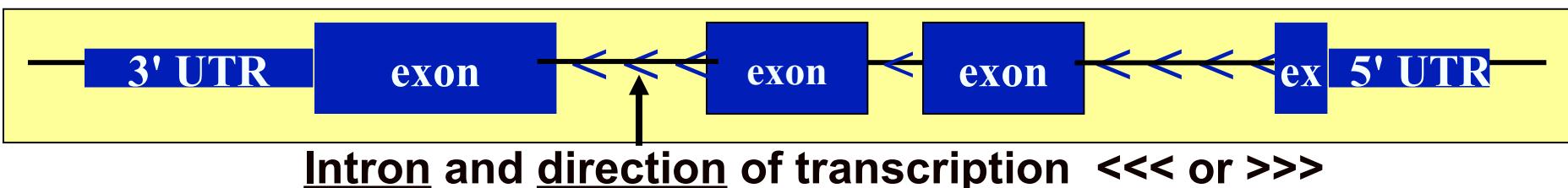
NCBI RefSeq Other RefSeq All GENCODE...
 full hide 19 IKMC Ger Mapped
 hide Genscan Genes hide
 hide ORFeome Clones hide
 hide UCSC Alt Events hide
 hide UniProt hide

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

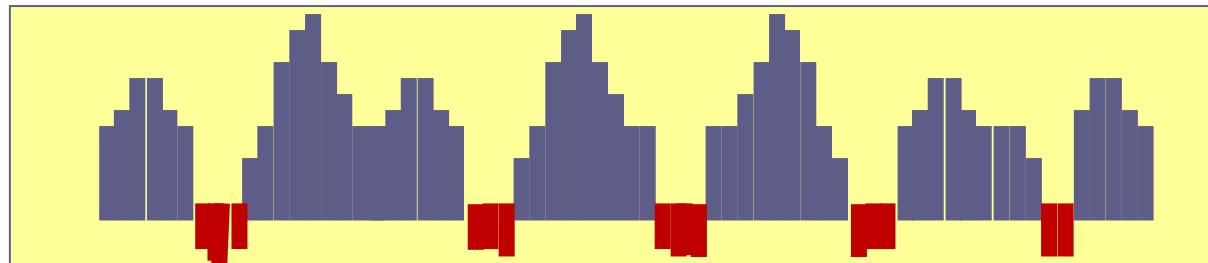
Refresh!

Practicum

- Some visual clues:



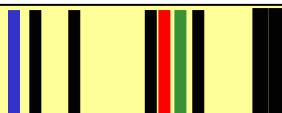
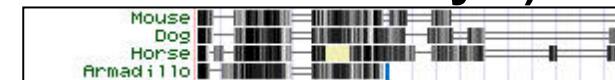
Sequence conservation



**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

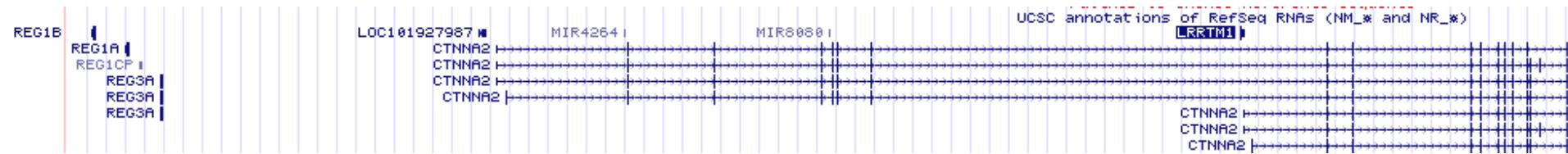
1. What is the genomic localization of human *Irrtm1* gene?

- Click on the gene to see the information

Position: [chr2:80301878-80304362](#)
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



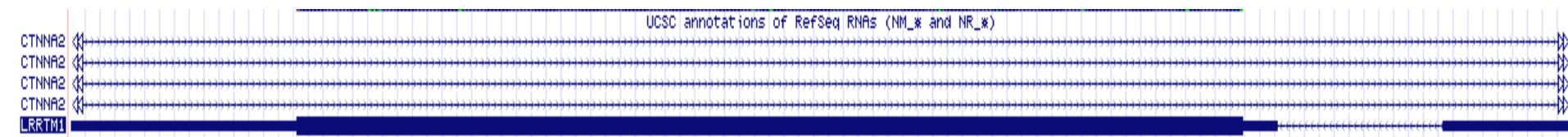
3. How many exons has the gene? 2

- 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



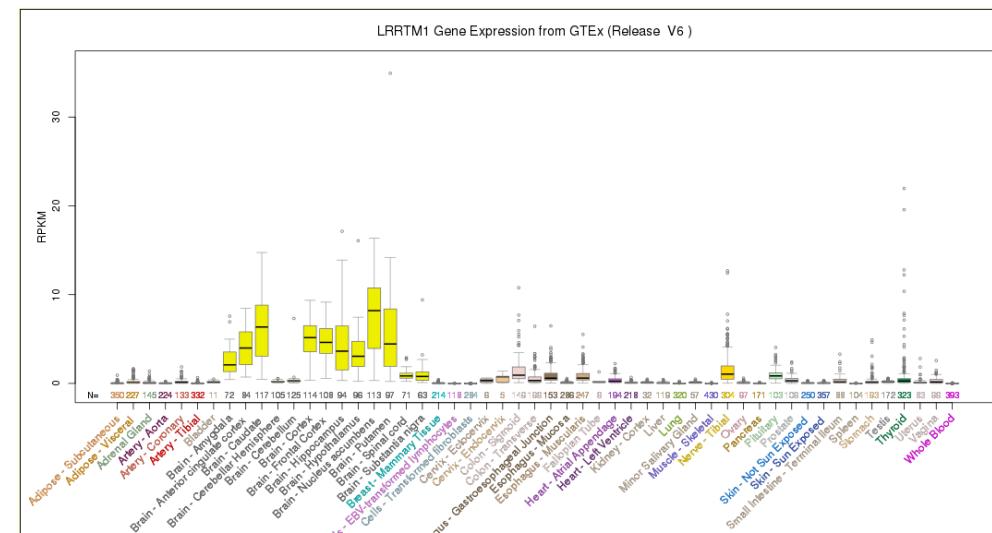
5. Can you find SNPs in this gene? yes

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



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

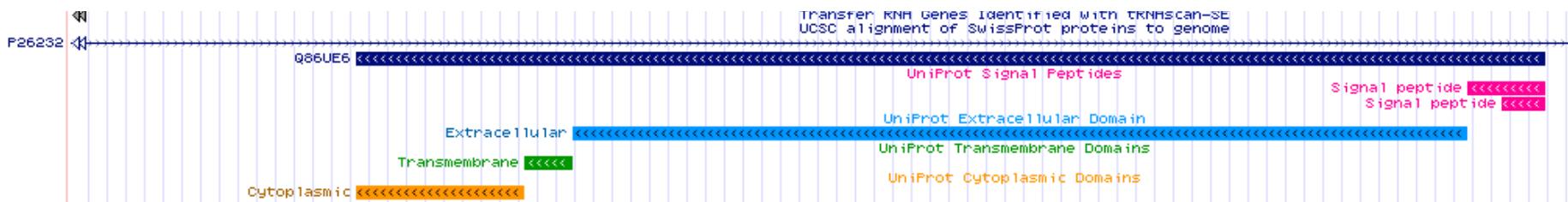
- Info in track “Expression” > “GTEx”



Practicum

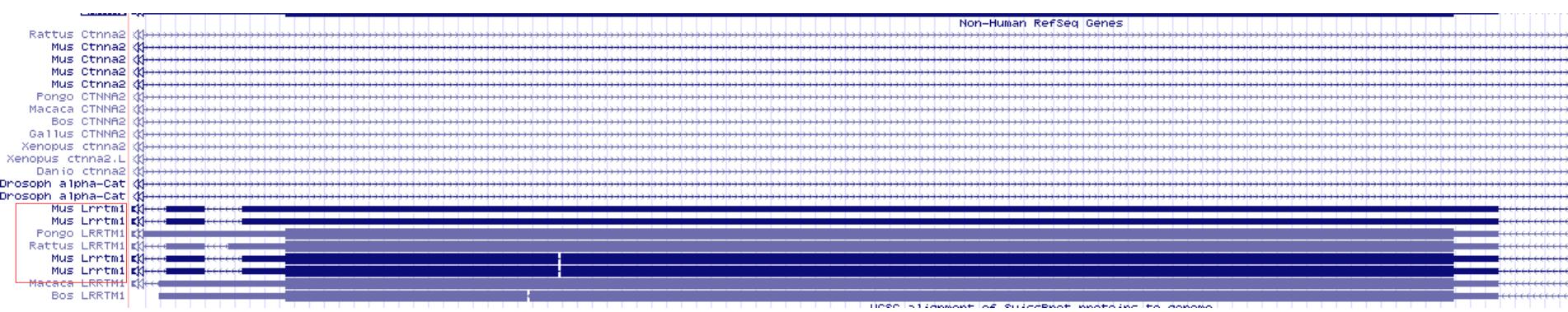
7. Does the protein encoded by this gene have a transmembrane domain?

- 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 *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

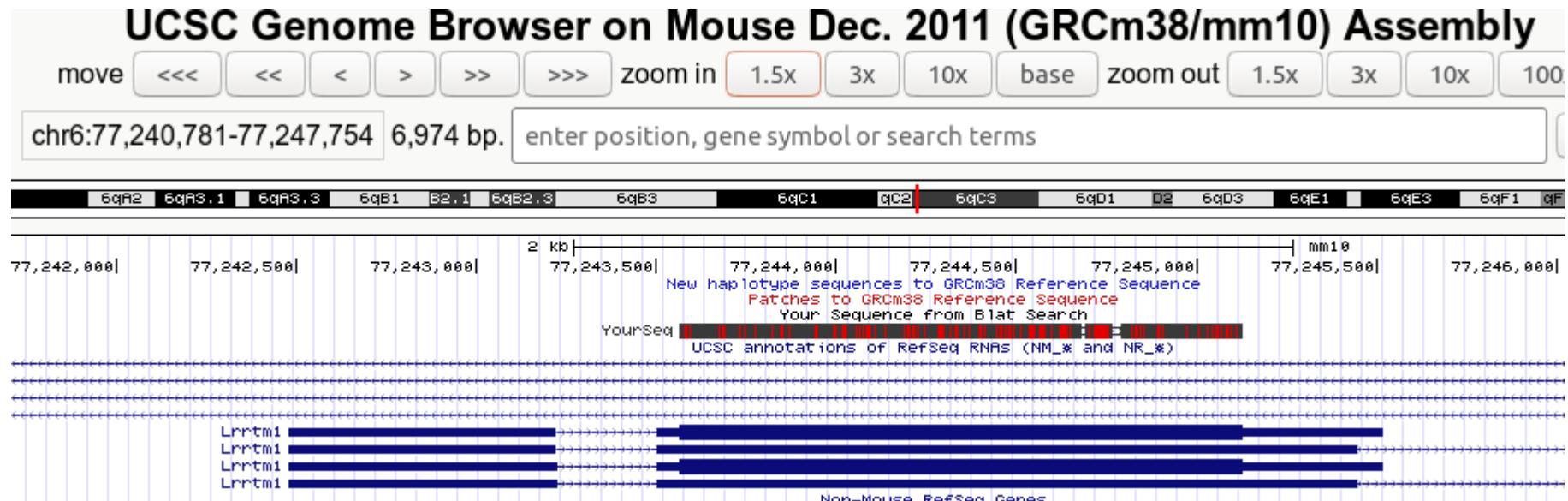
Genome: <input type="checkbox"/> Search ALL	Assembly: <input type="text" value="Dec. 2011 (GRCm38/mm10)"/>	Query type: <input type="text" value="DNA"/>	Sort output: <input type="text" value="query,score"/>
<input type="text" value="ATGGATTTCTCTGCTCGGTCTCTGCTATACTGGCTGCTGAGGGAGGC"/> CTCGGGGGTGGCTTGTGCTGCTGGGGGCTGCTTCAGATGCTGCCG CCGCCCCCAGCGGGTGCAGCTGTGCCGTGCGAGGGCGGCTGCTG TACTCGGAGGGCCTCAACCTCACCGAGGCGCCAACACCTGTCCGGCT GCTGGCTTGTCCCTGCCTACAAACAGCTCTCGAGCTGCCGCCGCC AGTTCACGGGTTAATGAGCTCACGGCTCTATCTGGATACAATCAC ATCTGCTCGTGAGGGGACGCCCTTCAGAAACTGCCGAGTTAAGGA ACTCACGCTGAGTCCAACCAGATACCCAACCTGCCAACACCCCTTC GGCCATGCCAACCTGGCAGCGGACCTCTCGTACAACAAGCTGCG GCGCTGGCGGGACCTCTCCACGGGCTGCCGAAGCTCACACGCTGCA TATGCGGGCCAACGCCATTCCAGTTGTGCCGTGCGCATCTTCAGGACT GCGCAGCTCAAGTTCTGACATGGATAACATCAGCTAAGAGCTG GCGCGCAACTTTGCGCCGCTTAAAGCTACCGAGCTGCACCTCGA GCACAAAGCACTGGTCAAGGTGAACCTCGCCACTCCGCGCCTCATCT CCCTGCACTGCTCTGCGGGAGGAACAGGTGGCATTGGGTGAGC			
<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



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](#)

Practicum

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	<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

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
Supplementary file	
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
<i>Raw data is available on Series record Processed data included within Sample table</i>	

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	6.5777017

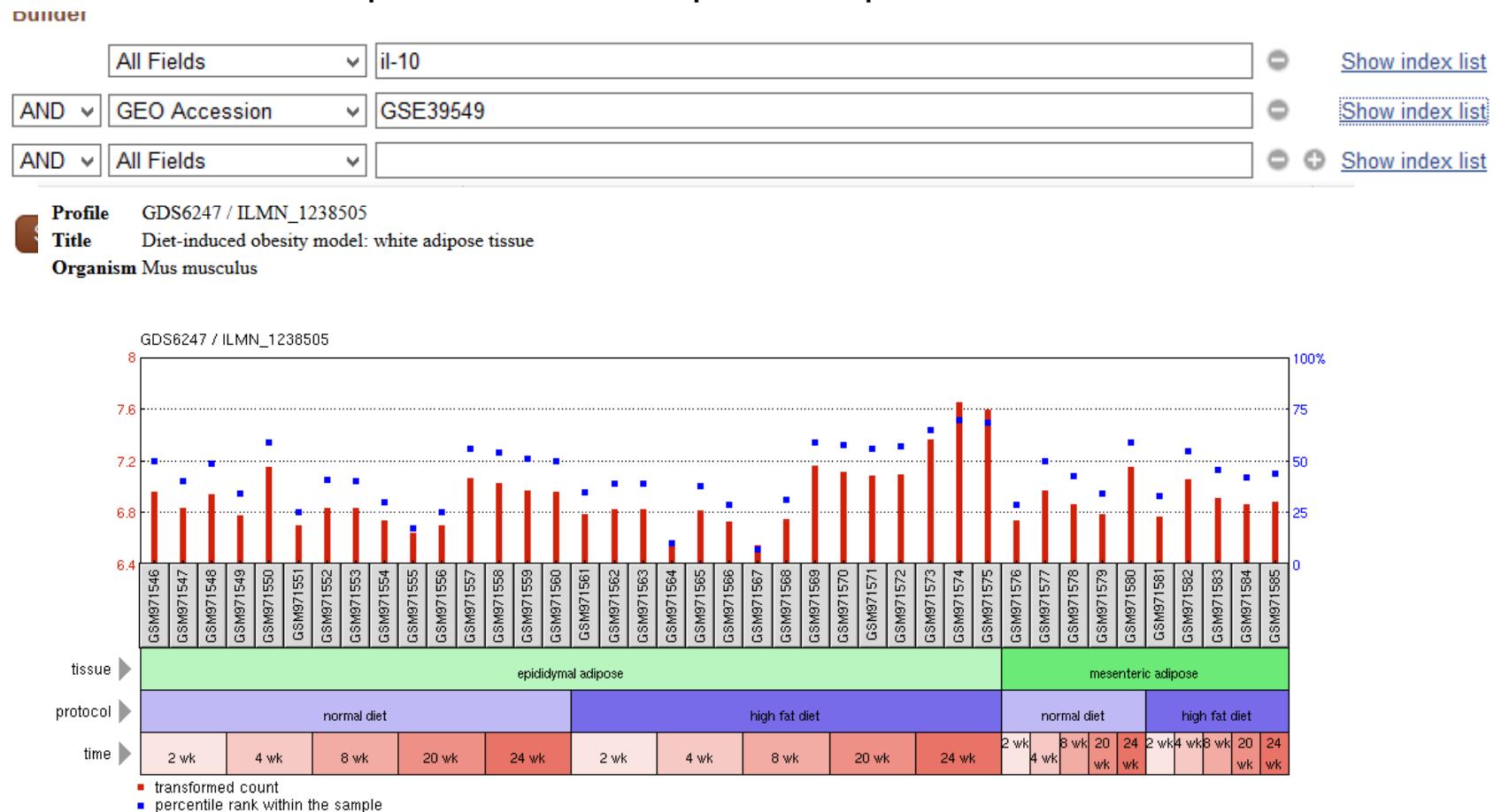
Data table with normalized expression values

Practicum

Retrieving data from GEO

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

Example: search for expression profile of IL-10



Examples of Databases

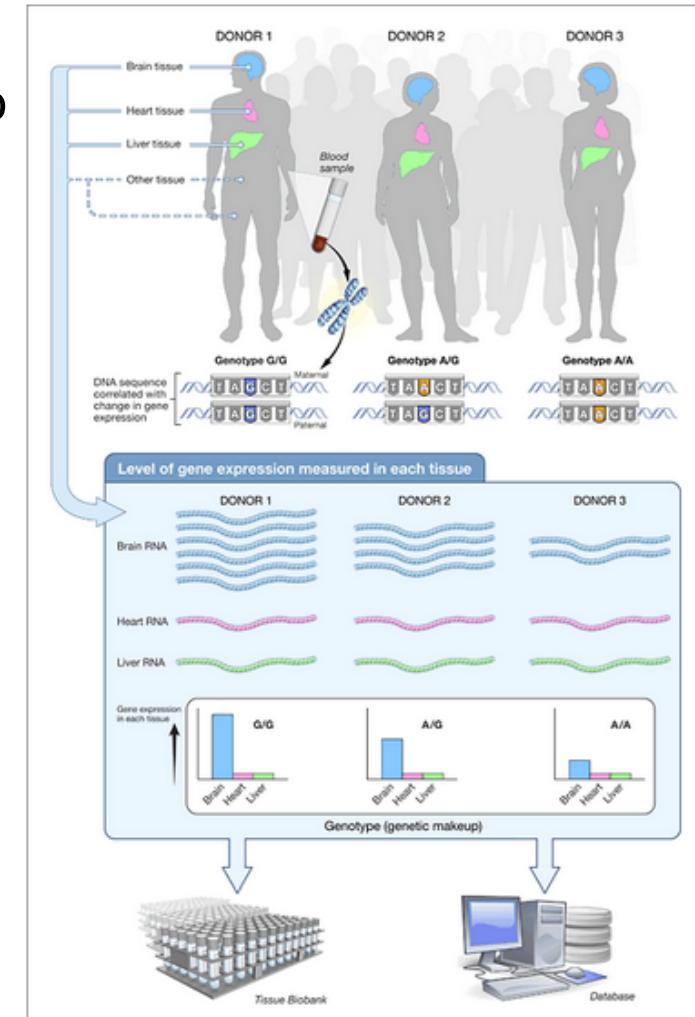
Genotype-Tissue Expression (GTEx) Project

<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/>

The screenshot shows the GTEx Portal homepage. At the top is a navigation bar with links for GTEX, Datasets, Expression, QTLs & Browsers, Sample Data, Biobank, Documentation, Publications, Contact, and FAQs. A search bar for gene or SNP ID is also present. Below the navigation is a banner featuring a stylized image of a tissue sample. To the right of the banner is a date (2018-04-04), a message about a new eQTL Dashboard Visualization, and a "Read More" link. The main content area is divided into three sections: "Current Release" (with information about Version V7 and download links), "Genetic Association" (with a search bar for eQTLs and a "Single-Tissue eQTLs" section), and "Transcriptome" (with a search bar for expression by gene ID and sections for Top Expressed Genes in a Tissue, Gene Expression in Tissues, and Exon and Transcript Expression).

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

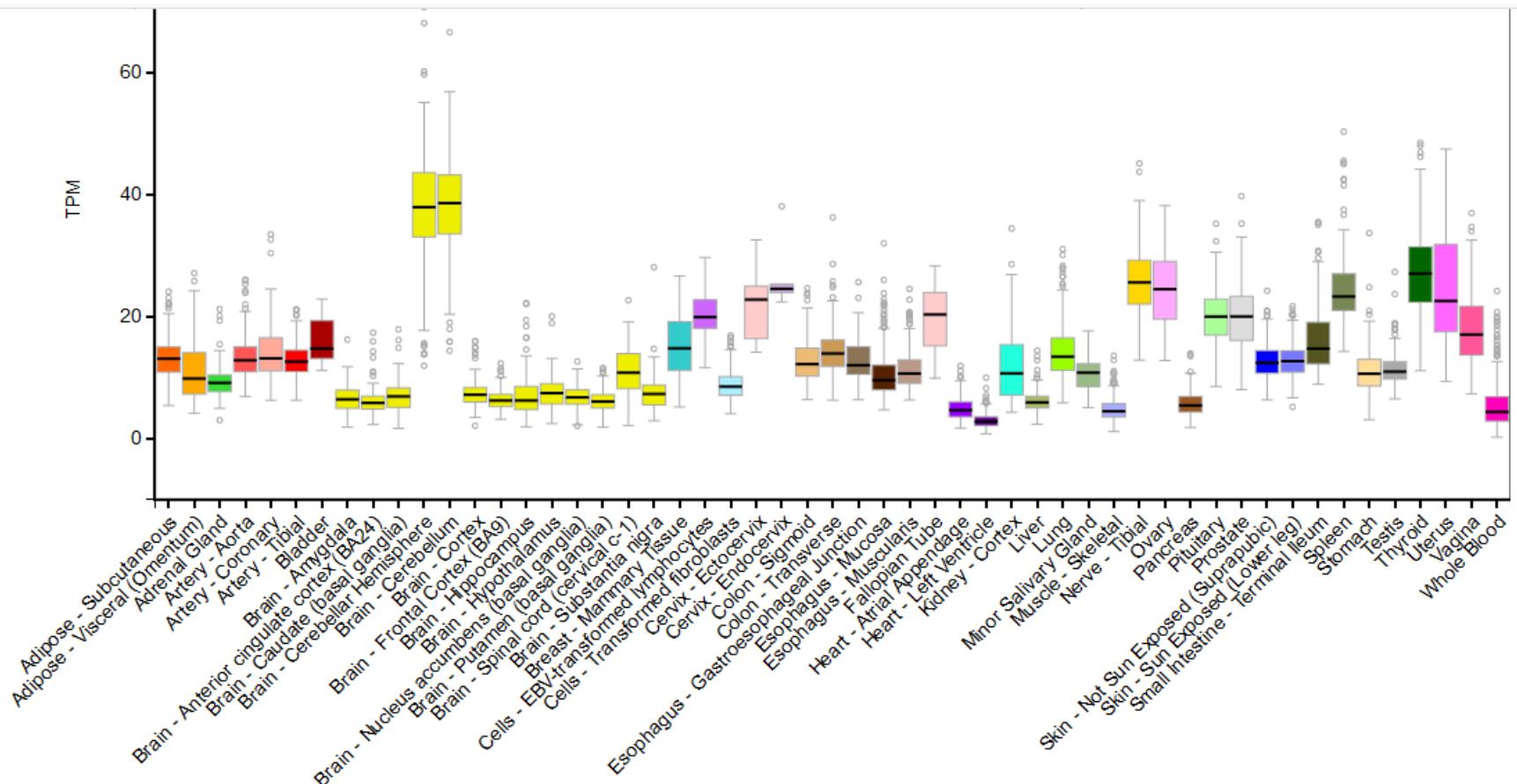
Exon and Transcript Expression

2018-04-04
New eQTL Dashboard Visualization
[Read More >](#)

Practicum

Genotype-Tissue Expression (GTEx) Project

Example: Normal tissue expression profile of *mutyh* gene

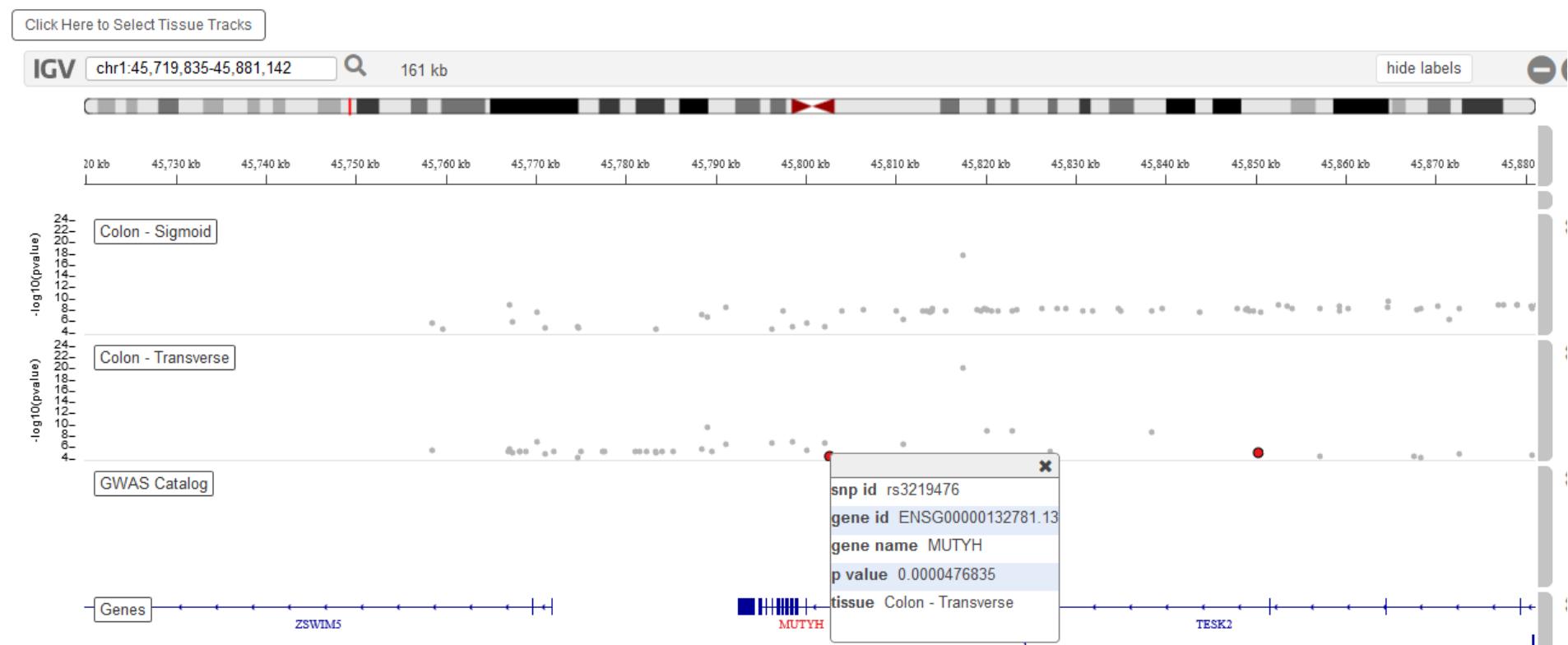


Practicum

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.

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](#)
- **Terapeutic targets**
 - [Therapeutic targets database](#)
 - [PharmGKB](#) : pharmacogenomics (impact of genetics on drug response)
- **Disease-related**
 - [DisGeNet](#): genes and variants associated to human diseases
 - [TCGA, COSMIC](#) (Cancer)

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/>

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 Advanced Search

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

Browse Analytics

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

Data Category

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

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

TCGA-AF-3912 TCGA-READ Rectosigmoid junction -- 20 Seq Exp SNV CNV Meth Clinical Bio

8 12 4

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)

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

