

# Databases in molecular biology

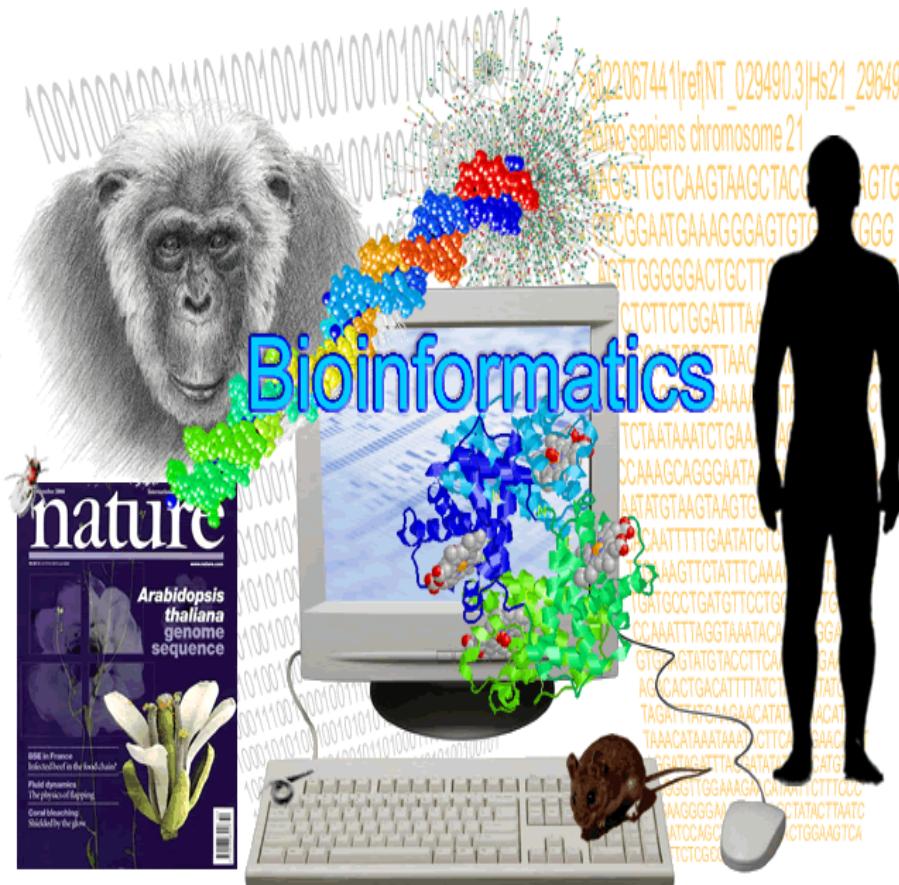
Bioinformatics Course UEB-VHIR  
November 2020

Ricardo Gonzalo<sup>1</sup>, Mireia Ferrer<sup>1</sup>, Álex Sánchez<sup>1,2</sup>  
Berta Miró<sup>1</sup>, Angel Blanco<sup>1,2</sup>

1 Unitat d'Estadística i Bioinformàtica (UEB) VHIR

2 Departament de Genètica Microbiologia i Estadística, UB

# Information in the omics era



- Massive quantities of information (not necessarily “big data”)
- Open-access
- For this information to be accessible it must be properly stored.
- Access to information
  - Must be fast
  - Must be flexible
- This has been made possible
  - Creating databases
  - Distributing them through the web

# Biological Databases

- **Definition:** *libraries* of life sciences information, collected from scientific experiments, published literature, high-throughput experiment technology and computational analysis
- **What are they for?**
  - Storage of information
  - Data organization
  - Access information
  - Knowledge discovery
- There are many different general and specialized databases.
  - Large list published yearly in *NAR* : 1637 in 2020!
    - <https://www.oxfordjournals.org/nar/database/c/>

# Biological Databases

## NAR Database Summary Paper Category List

Nucleotide Sequence Databases

RNA sequence databases

Protein sequence databases

Structure Databases

Genomics Databases (non-vertebrate)

Metabolic and Signaling Pathways

Human and other Vertebrate Genomes

Human Genes and Diseases

Microarray Data and other Gene Expression Databases

Proteomics Resources

Other Molecular Biology Databases

Organelle databases

Plant databases

Immunological databases

Cell biology

• Nucleotide Databases

- ASD
- ATD
- EMBL-Bank
- EMBL CDS
- Ensembl
- Genome Reviews
- IMGT/HLA

• Protein Databases

- CSA
- GOA
- IntAct
- IntEnz
- InterPro

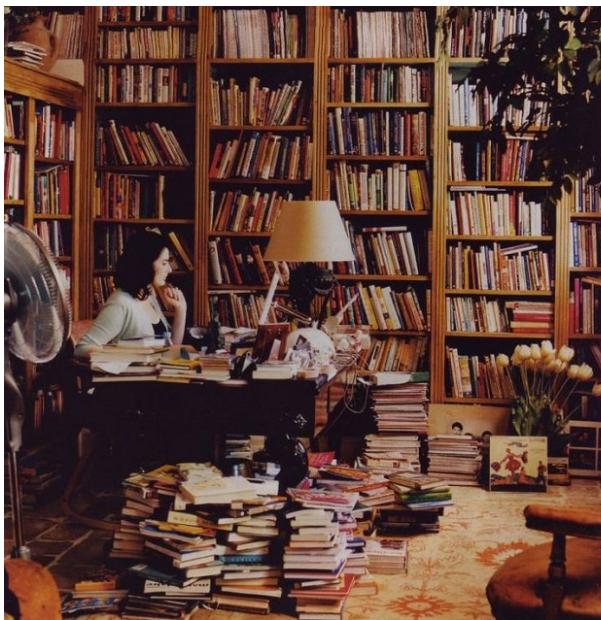
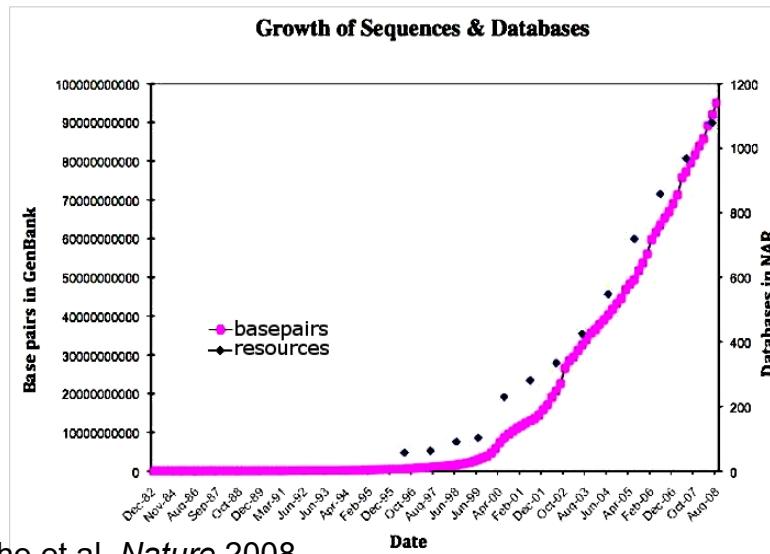
• Microarray Databases

- ArrayExpress
- MIAME

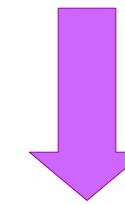
• Literature Databases

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

# Challenges



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



- Structure/Integrate information
- Annotation and Curation
- Centralize data management

# I. Structuring and Integrating the information

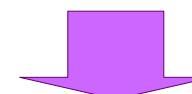
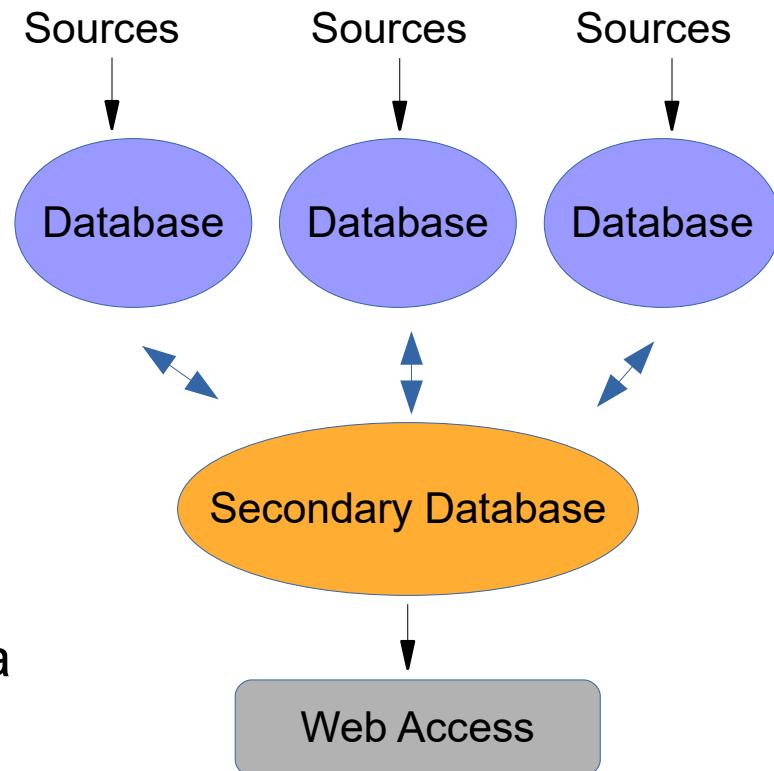
# Integrating the information

- **Primary databases:**

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

- **Secondary databases:**

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



**Different formats and models  
for structuring the data**

# Integrating the information

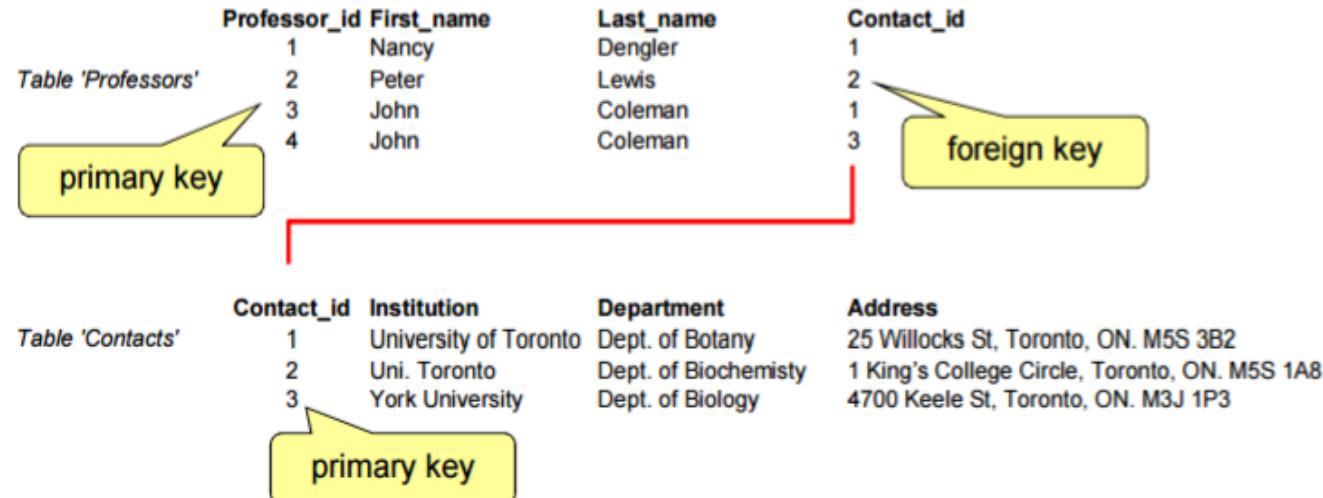
- The quality of the information in a database is closely related to its structure
- This aspect is also crucial for its efficiency and accessibility.
- There are many different types of file formats depending on the type of information they contain / source of data
- Most are based on the flat-file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

## FASTA format (sequences)

Identifier	Description
sequence	>136435 Mus Musculus basic domain/leucine zipper transcription factor mRIIA, complete cds. gccccggccgcgtccccagacaaaaggcttggccggccggccggccggccgtgcgcctcgctccccgcctcccc cggttgcgcgtcttcgcggccggctttggctggcgcgtcccggccggccgaaagtttccccgcggcag cgccggctgagcctcgcttttagcgatggccggagctgagcatggggcaagagactgcccaccagccgct ggccatggagtacgtcaacgacttcgaccttctaagttcgacgtgaagaaggagccctggggcgccgga gcgtccggccggccatgcacacgcctgcagcctgctggctgggtcgccaccccgctcagcactccgt

# Integrating the information

- Different models exist to relate/integrate the information (Relational, Hierarchical, Networks...).
- Most common model: Relational databases
  - flat-file format (text file)
  - Many tables linked to each other: cross-referencing through a key (common) field (unique identifier)



# Integrating the information

- In many databases an entry can be identified in 2 (ore more!) different ways:
  - **Identifier** ("locus" in GenBank, "entry name" in UniProt): is a string of letters and digits. May change if the database curators decide that is no longer appropriate.
  - **Accession code (number)**: is a number (possibly with a few characters in front) that uniquely identifies an entry in its database. It is supposed to be stable.
  - **Versions and Gene Indices**: The same accession number may be associated with a different GI if a newer or corrected sequence is submitted.

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

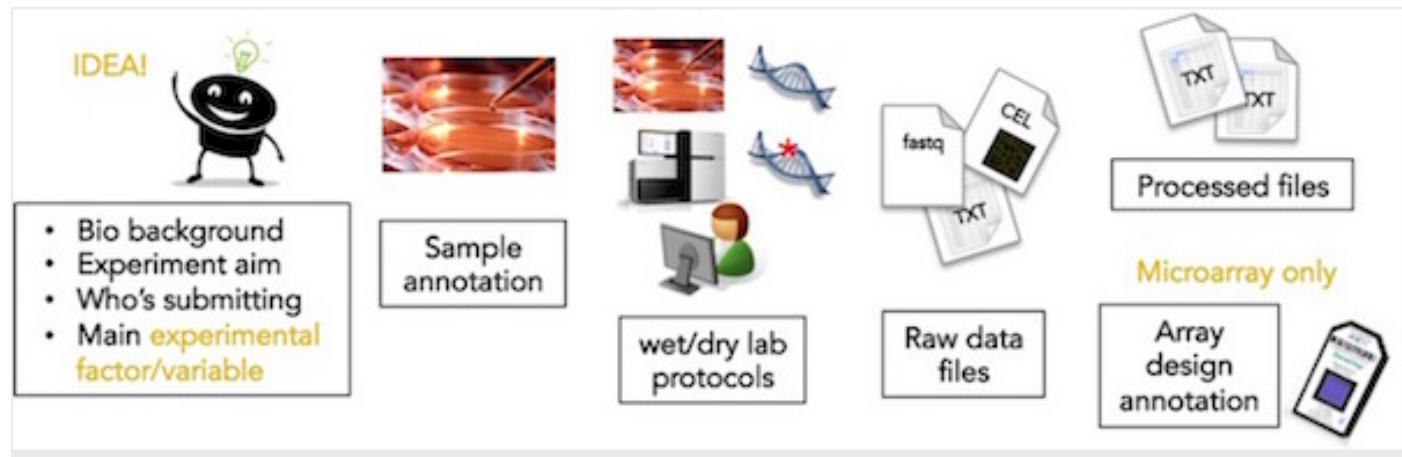
# Integrating the information

## II. Data Annotation and Curation

# Data Annotation

- Different levels of annotation
  - Data: annotation of sequences/genomes (chromosome position, gene function, ...)
  - Metadata: information for an experiment, identification of samples, ...
- Collaborative efforts to provide as much information about the data
- The **Minimum Information Standard** is a set of guidelines for reporting data

MIAME (Minimum Information About a Microarray Experiment)



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

- Benefits:
  - Ensures the verification, interpretation and reproducibility of data
  - Facilitates the creation of structured databases and development of analysis

# Data Curation

- It involves annotation, publication and presentation of the data such that the value of the data is maintained over time, and the data remains available for reuse and preservation.
- May be done by database experts or experts of the scientific community.

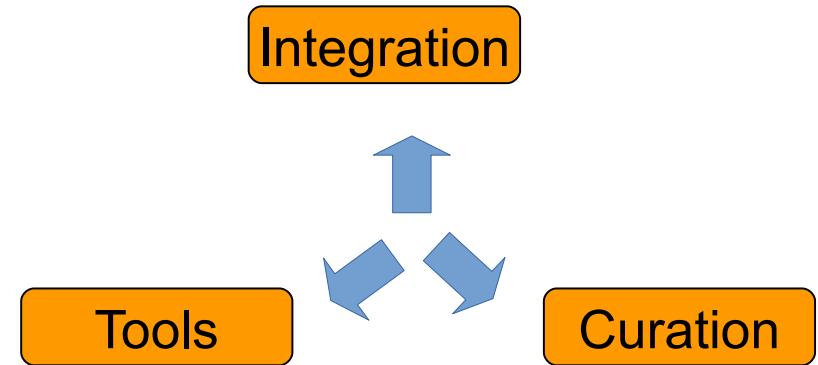


Source: <https://www.ebi.ac.uk/arrayexpress/submit/overview.html>

### **III. Centralizing data management**

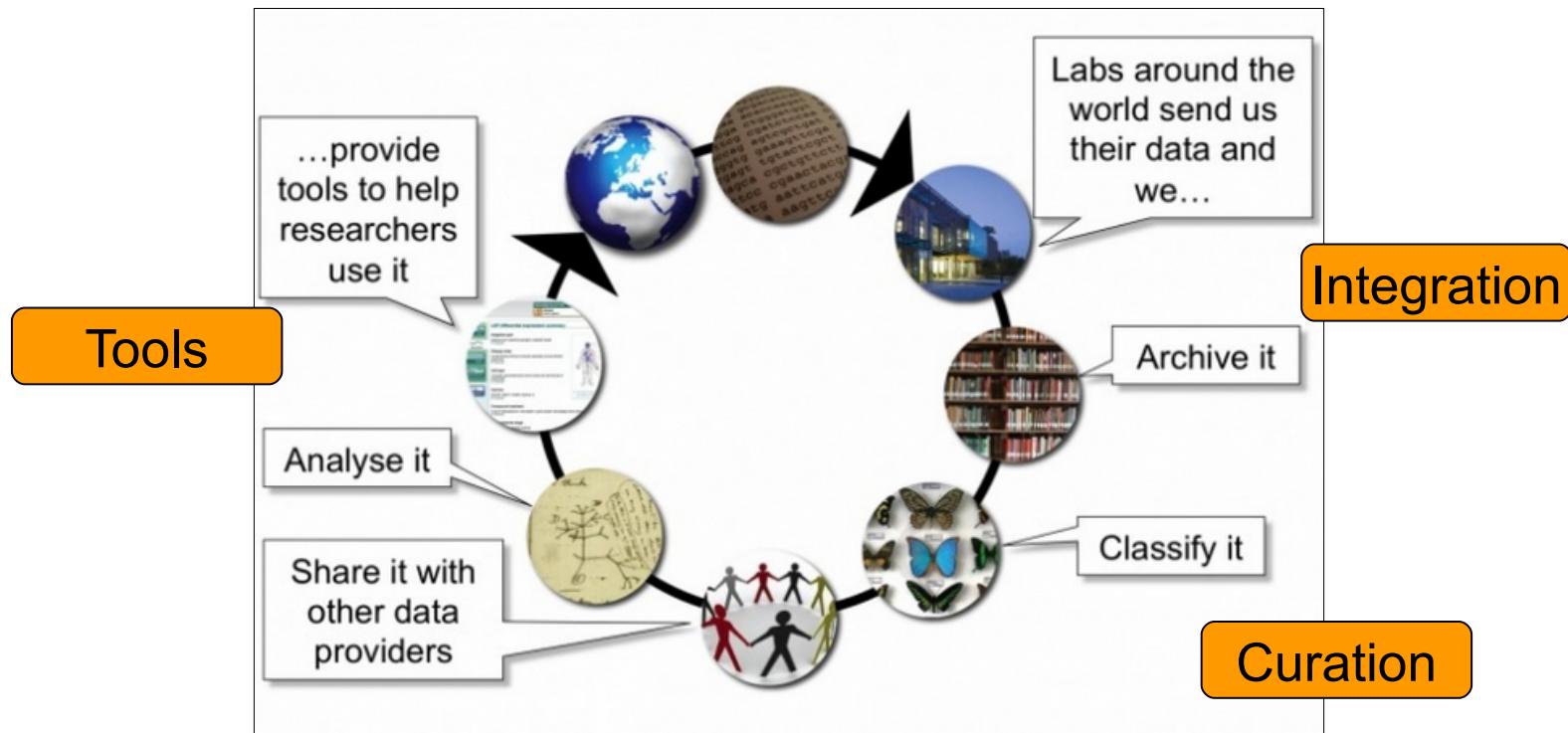
# Centralizing data management

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



# Resource providers

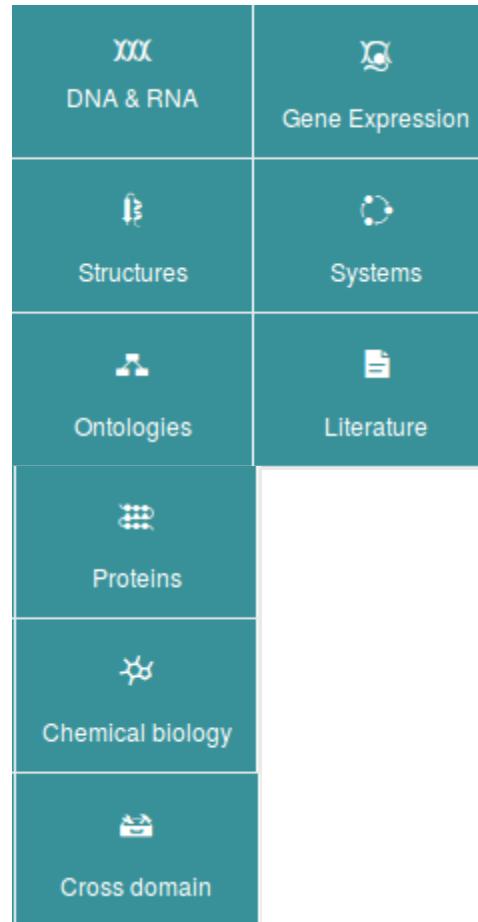
- Big organizations that act as *hubs* that provide transparent access to data sources.



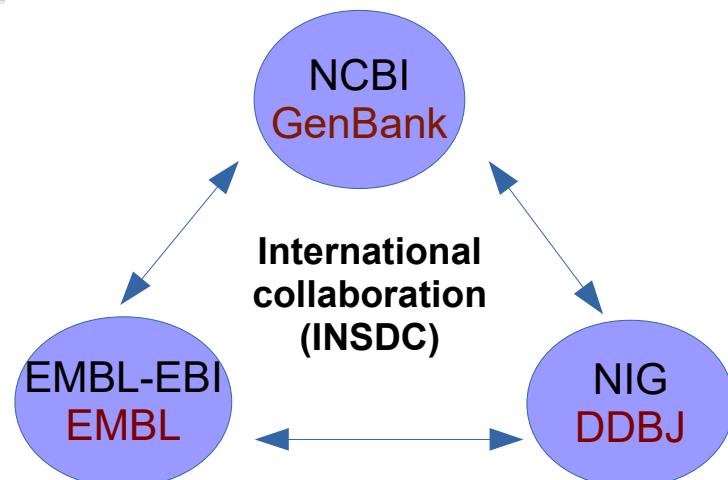
# Resource providers



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



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



# Resource providers



[Amino Acid Explorer](#)

[Assembly Archive](#)

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

[Batch Entrez](#)

[BioAssay Services](#)

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

[BLAST Microbial Genomes](#)

[BLAST RefSeqGene](#)

[CDTree](#)

[Cn3D](#)

[COBALT](#)

[Concise Microbial Protein BLAST](#)

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

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

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

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

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

## Clustal Omega



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

[Multiple sequence alignment](#)

## InterProScan



InterProScan searches sequences against InterPro's predictive protein signatures.

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

## BLAST [protein]



Fast local similarity search tool for protein sequence databases.

[Sequence similarity search](#)

## BLAST [nucleotide]



Fast local similarity search tool for nucleotide sequence databases.

[Sequence similarity search](#)

## HMMER



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

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

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

# Centralizing data management

- General
  - **Resource providers**

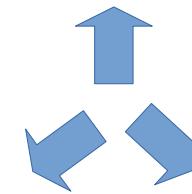
Integration

- Subject-specific

- **Collaborative projects**
- **Multi-omics repositories**
- **Genomic Browsers**

Tools

Curation



Session 3

# Examples of Databases

# Examples of Databases

## Literature DB

- Contain different types of bibliographic information (articles, reviews, books, patents...). Not only peer-reviewed!
- PubMed (NCBI): references and abstracts on life and biomedical sciences
- Europe PMC (EBI-EMBL): a full-text literature database for life sciences
- ArXiv: repository of electronic pre-prints after moderation
- Patent databases (eg. EPO) can be accessed from EBI-search
- Biocatalogue: provides a curated catalog of life-sciences web services

Nature. Author manuscript; available in PMC 2014 Nov 7.  
Published in final edited form as:  
[Nature. 2013 Nov 7; 503\(7474\): 59–66.](#)  
doi: [\[10.1038/nature12709\]](#)

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

Cooperation between brain and islet in glucose homeostasis and diabetes

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

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

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

[Abstract](#) [Go to: !\[\]\(6349ea74862db5fb00756ca97efd2d40\_img.jpg\)](#)

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

# Examples of Databases

## Taxonomic DB

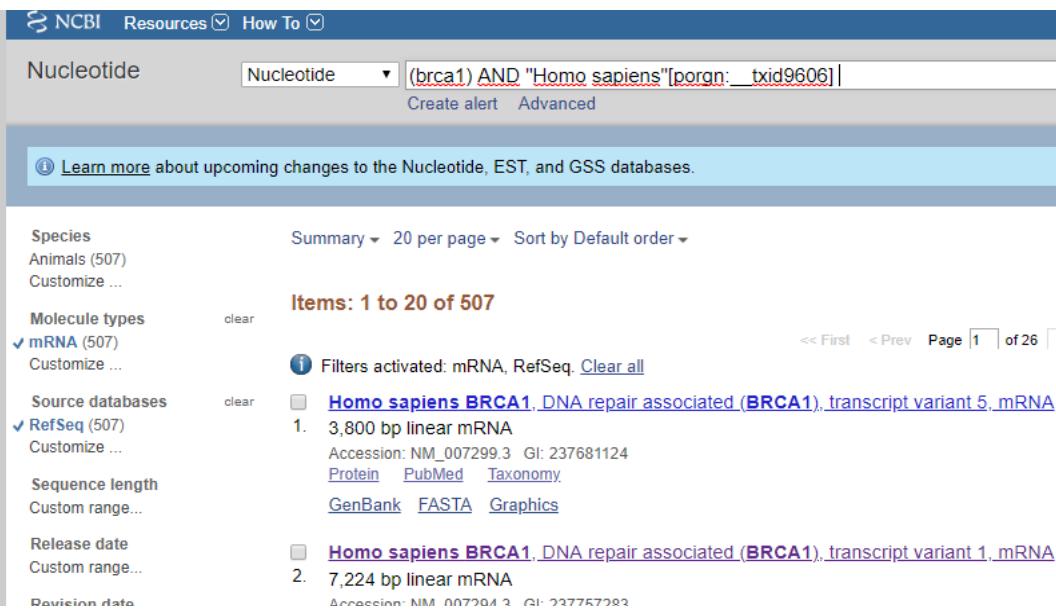
- Contain information about the classification of organisms, mainly from molecular data
- **Taxonomy DB**: curated classification and nomenclature for all of the organisms in the public sequence databases.
- This represents about 10% of described species



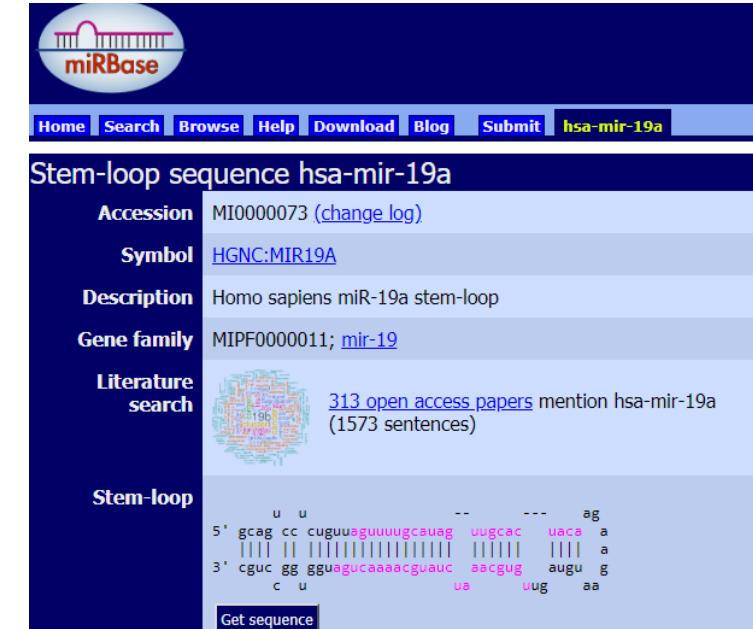
# Examples of Databases

## Nucleotide DB

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


 NCBI Resources How To  
 Nucleotide Nucleotide (brca1) AND "Homo sapiens"[organism:txid9606]  
 Create alert Advanced  
 Learn more about upcoming changes to the Nucleotide, EST, and GSS databases.  
 Species Animals (507) Summary 20 per page Sort by Default order  
 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
  
```

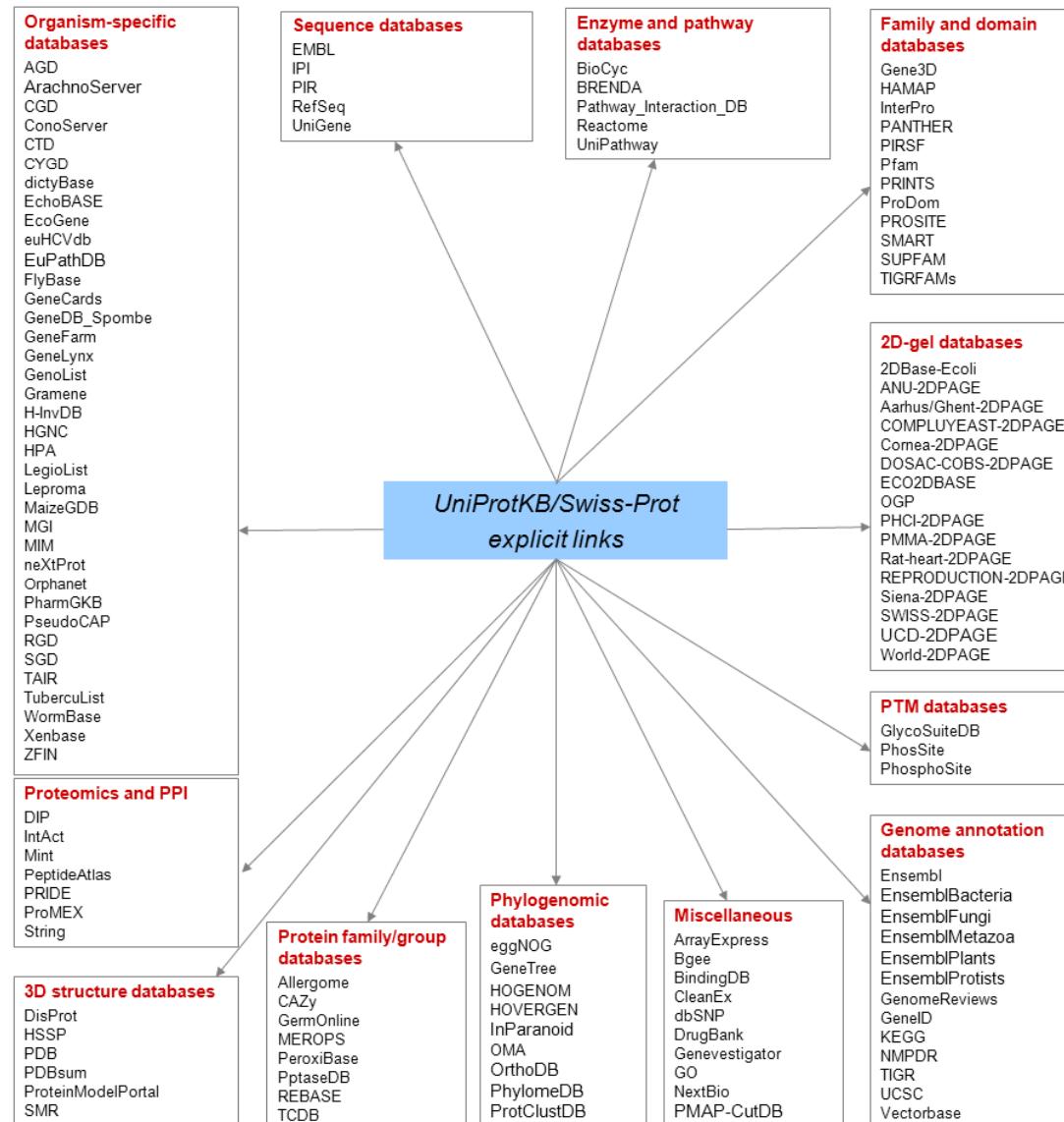
[Get sequence](#)

# 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**

# Examples of Databases



# 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)
- Species-specific genome databases (eg. [Mouse Genome Informatics](#))
- Genome Browsers: provide tools for visualization and integrative genomic analysis
  - NCBI [Genome Data Viewer](#)
  - EBI's [Ensembl](#)
  - [UCSC Genome Browser](#)

# Examples of Databases

## Gene Expression Databases

- Contain gene expression data derived from microarray, next-generation sequencing, and other forms of high-throughput functional genomics data submitted by the research community
- [Gene Expression Omnibus \(GEO\)](#) / [ArrayExpress \(EBI\)](#)
- [Sequence Read Archive \(SRA\)](#): stores raw sequencing data and alignment information from high-throughput sequencing platforms
- [GTEx](#)
- [Expression Atlas](#): provides gene expression results for different organisms, including metazoans and plants. Expression profiles of tissues from Human Protein Atlas, GTEx and FANTOM5, and of cancer cell lines from ENCODE, CCLE and Genentech projects can be explored.

# Examples of Databases

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

# Examples of databases

- And many more that can be found in / accessed from:
  - Large list published yearly in [NAR](#)
  - Journal-recommended repositories for publication ([Nature](#))
  - Resource providers portals ([NCBI](#) / [EBI-EMBL](#))
  - Integrative projects (disease-specific / organism-specific)
  - Genome Browsers (genome-oriented)

# Tools for exploiting database information

# Tools

- What are they for?
  - Search of information (eg. *Entrez*)
  - Finding/comparing sequences (eg. *BLAST*)
  - Data exploration and visualization (eg. *Genome Browsers*)
  - Manipulating and analyzing data
  - Make predictions
  - Knowledge discovery (data mining)
  - Downloading/Exporting data
- Can be accessed through
  - Web interface from resource providers, databases, projects or subject-specific repositories
  - Software (eg. *R*, *Cytoscape*)
  - Platforms (eg. *Galaxy*)

# Tools



[Amino Acid Explorer](#)

[Assembly Archive](#)

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

[Batch Entrez](#)

[BioAssay Services](#)

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

[BLAST Microbial Genomes](#)

[BLAST RefSeqGene](#)

[CDTree](#)

[Cn3D](#)

[COBALT](#)

[Concise Microbial Protein BLAST](#)

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

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

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

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

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

## Clustal Omega



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

[Multiple sequence alignment](#)

## InterProScan



InterProScan searches sequences against InterPro's predictive protein signatures.

[Protein feature detection](#)

[Sequence motif recognition](#)

## BLAST [protein]



Fast local similarity search tool for protein sequence databases.

[Sequence similarity search](#)

## BLAST [nucleotide]



Fast local similarity search tool for nucleotide sequence databases.

[Sequence similarity search](#)

## HMMER

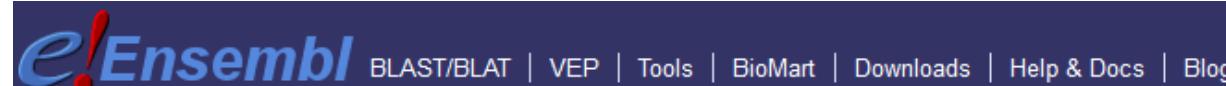


Fast sensitive protein homology searches using profile hidden Markov models (HMMs) for querying against both sequence and HMM target databases.

[Sequence similarity search](#)

[Protein function analysis](#)

# Tools



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

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

## Ensembl Tools

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

### Processing your data

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



# Tools

Search:

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

[Home](#) » BiocViews

## All Packages

### Bioconductor version 3.12 (Release)

Autocomplete biocViews search:

▼ Software (1974)
▶ AssayDomain (791)
▶ BiologicalQuestion (822)
▶ Infrastructure (456)
▼ ResearchField (902)
BiomedicalInformatics (62)
CellBiology (54)
Cheminformatics (13)
ComparativeGenomics (8)
Epigenetics (63)
Epitranscriptomics (1)
FunctionalGenomics (53) <span style="background-color: #e0f2f1;">■</span>
Genetics (200)
ImmunoOncology (447)
Lipidomics (11)
MathematicalBiology (8)
Metabolomics (74)

### Packages found under FunctionalGenomics:

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

Show [All](#)  entries

Search table:

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

# Tools

## Galaxy Europe

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

**Tools**  

search tools

Get Data

Send Data

Collection Operations

**GENERAL TEXT TOOLS**

Text Manipulation

Filter and Sort

Join, Subtract and Group

**GENOMIC FILE MANIPULATION**

Convert Formats

FASTA/FASTQ

FASTQ Quality Control

Quality Control

SAM/BAM

BED

### COVID-19 research!

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

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

### News

Nov 14, 2020  **UseGalaxy.eu Tool Updates for 2020-11-14**

Nov 7, 2020  **UseGalaxy.eu Tool Updates for 2020-11-07**

Nov 3, 2020  **November Galactic News!**

Oct 31, 2020  **UseGalaxy.eu Tool Updates for**

### Events

Jan 25, 2021 - Jan 29, 2021   **2021 Galaxy Admin Training**

Dec 10, 2020   **Galaxy Developer Roundtable: Developer Training**

Dec 7, 2020 - Dec 10, 2020  **Hackathon sur les outils interactifs de Galaxy (GxIT)**

Dec 3, 2020   **DNA and DTA**

<https://usegalaxy.eu/>

# Tools

Cytoscape App Store

Submit an App ▾ Search the App Store Sign In

All Apps

Newest Releases

Get Started with the App Store »

**Categories**

- [collections](#)
- [data visualization](#)
- [network generation](#)
- [network analysis](#)
- [graph analysis](#)
- [online data import](#)
- [automation](#)
- [integrated analysis](#)
- [clustering](#)
- [systems biology](#)
- [utility](#)
- [enrichment analysis](#)
- [visualization](#)
- [data integration](#)

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

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

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

**PathLinker** 3.0+  
Reconstructs signaling pathways from protein interaction networks

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

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

more newest releases »

# Tools



Omics DI

Browse

Submit Data

Databases

API

Help ▾

Login

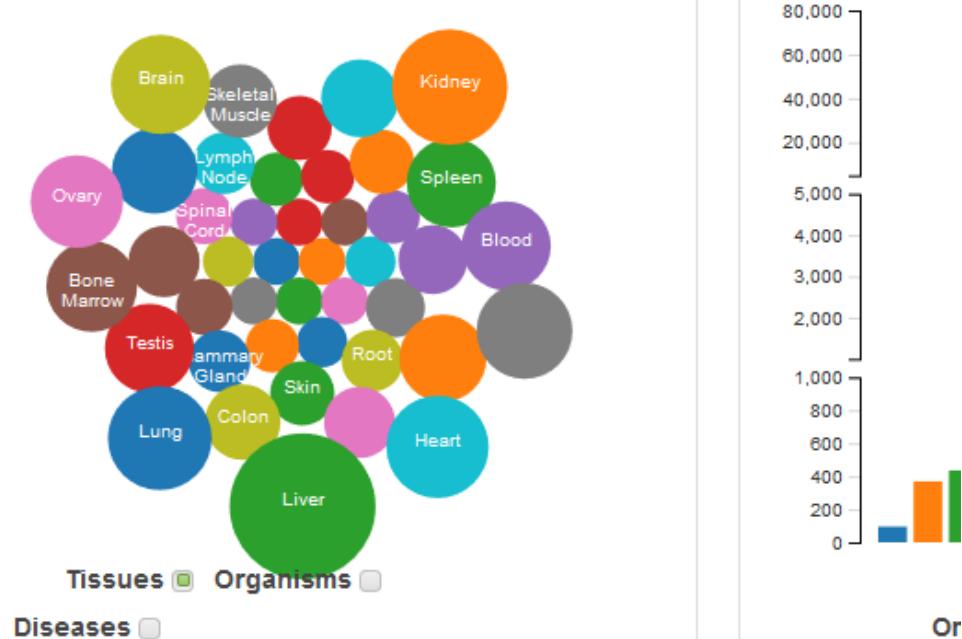
Organism, repository, gene, tissue, accession

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

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

Description  Sample

Data



# Tools

## The Entrez Search and Retrieval System

- Text-based search and retrieval system used at NCBI for all of its major databases
- All databases indexed by Entrez can be searched via a single query string. This returns a unified results page, that shows the number of hits for the search in each of the databases, which are also links to actual search results for that particular database.
- Supports boolean operators (AND, OR, NOT, "", \*)
- Use tags to limit parts of the search statement to particular fields.

```
term [field] OPERATOR term [field]
```

- Start with a general query and refine it progressively using Filters/Limits
- For individual databases, the Advanced Search and Limits pages assist greatly in the construction of complex queries.

# Tools

## The Entrez Search and Retrieval System

Field	Short term	Nucleotide	Available for Database ...			
			Protein	Genome	Structure	PopSet
Accession	ACCN	Yes	Yes	Yes	Yes	Yes
All Fields	ALL	Yes	Yes	Yes	Yes	Yes
Author Name	AUTH	Yes	Yes	Yes	Yes	Yes
EC/RN Number	ECNO	Yes	Yes	Yes	Yes	Yes
Feature Key	FKEY	Yes	No	Yes	No	Yes
Filter	FILT	Yes	Yes	Yes	Yes	Yes
Gene Name	GENE	Yes	Yes	Yes	No	Yes
Issue	ISS	Yes	Yes	Yes	Yes	Yes
Journal Name	JOUR	Yes	Yes	Yes	Yes	Yes
Keyword	KYWD	Yes	Yes	Yes	No	Yes
Modification Date	MDAT	Yes	Yes	Yes	Yes	Yes
Molecular Weight	MOLWT	No	Yes	No	No	No
Organism	ORGN	Yes	Yes	Yes	Yes	Yes
Page Number	PAGE	Yes	Yes	Yes	Yes	Yes
Primary Accession	PACC	Yes	Yes	Yes	No	Yes
Properties	PROP	Yes	Yes	Yes	No	Yes
Protein Name	PROT	Yes	Yes	Yes	No	Yes
Publication Date	PDAT	Yes	Yes	Yes	Yes	Yes
SeqID String	SQID	Yes	Yes	Yes	No	Yes
Sequence Length	SLEN	Yes	Yes	Yes	No	No
Substance Name	SUBS	Yes	Yes	No	Yes	No
Text Word	WORD	Yes	Yes	Yes	Yes	Yes
Title Word	TITL	Yes	Yes	Yes	No	No
Uid	UID	No	No	No	No	No
Volume	VOL	Yes	Yes	Yes	Yes	Yes

# Practicum

## Example of database cross-search with Entrez

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

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

**Search NCBI** colon cancer X **Search**

Results found in 31 databases

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

# Practicum

## Querying databases to answer biological questions

- 1- Using [PubMed](#) Advanced Search, look for a *review* paper published in *Nature* on *colorectal cancer* and authored by *David SS*
- 2- In the abstract, the authors mention a gene associated to the disease. Find a *well annotated mRNA* sequence for this gene using the [Nucleotide DB](#)
- 3- Look for MUTYH human protein in [UniProtKB](#)
  - Identify protein sequence, motifs and 3D structure
  - With which proteins interacts according to *IntAct DB*?

# Practicum

## 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 like accession NM\_001350651.1, GI 1183596751, and links to Protein, PubMed, and Taxonomy. Below this, another result for "Homo sapiens mutY DNA glycosylase (MUTYH), transcript variant 12, mRNA" is listed. A red bracket on the left side groups the filter options under the heading "Using filters".

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

Items: 1 to 20 of 38

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

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

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

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

2. 1,831 bp linear mRNA

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

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

```
>NM_001350651.1 Homo sapiens mutY DNA glycosylase (MUTYH), transcript variant 13, mRNA
CAGCCGGAGCCCGGGTACAACGGAACCTGTAGTCCTCGTGGCTAGTTCAAGCGGAAGGGAGCAGTC
TCTGAAGCTTGAGGAGCCTCTAGAACTATGAGCCGAGGCCCTCCCTCTCCAGAGGCCAGAGGCTT
AAGGCTACTCTGGGAAGCCGCTCACCGCTCGAGCTGCGGGAGCTGAAACTGCGCCATCGTCAGTGTG
GCGGCATGACACCGCTCGTCTCCGCGTGAAGCTGCTGTGGGCATCATGAGGAAGGCCAGGAGCAGCC
TGGGAAGTGGTACAGGAAGCAGGCCAGGAGCAGAGCATGTAAGAACAAACAGTC
GGCCAAGCCTTCTGCGTGTAGAGACGTAGCTGAAGTCACAGCCTCCGAGGGAGCCTGCTAAGCTGG
ACGACCAAGAGAACCGGGACCTACCATGGAGAACGGCAGAGATGAGATGGACCTGGACAGGCCGGC
ATATGCTGAAGTGGCTACACTGAGGACCTGGCCAGTGCTTCCCTGGAGGAGGTGAATCAAACCTGGG
```

# 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

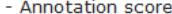
Feature table

None

Protein | Adenine DNA glycosylase

Gene | MUTYH

Organism | *Homo sapiens (Human)*

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

Function

Names & Taxonomy

Subcellular location

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

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 activity<sup>i</sup>

# Practicum

## Interaction<sup>i</sup>

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

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

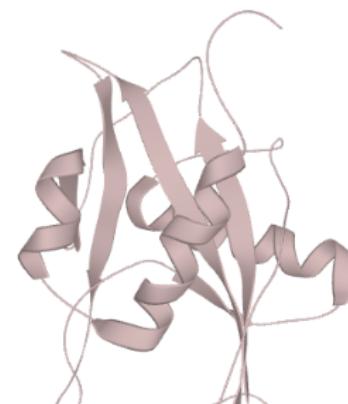
### Protein-protein interaction databases

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

## Structure<sup>i</sup>

### Family and domain databases

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



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

1 notificación

# Practicum

## Gene Expression Omnibus (GEO)

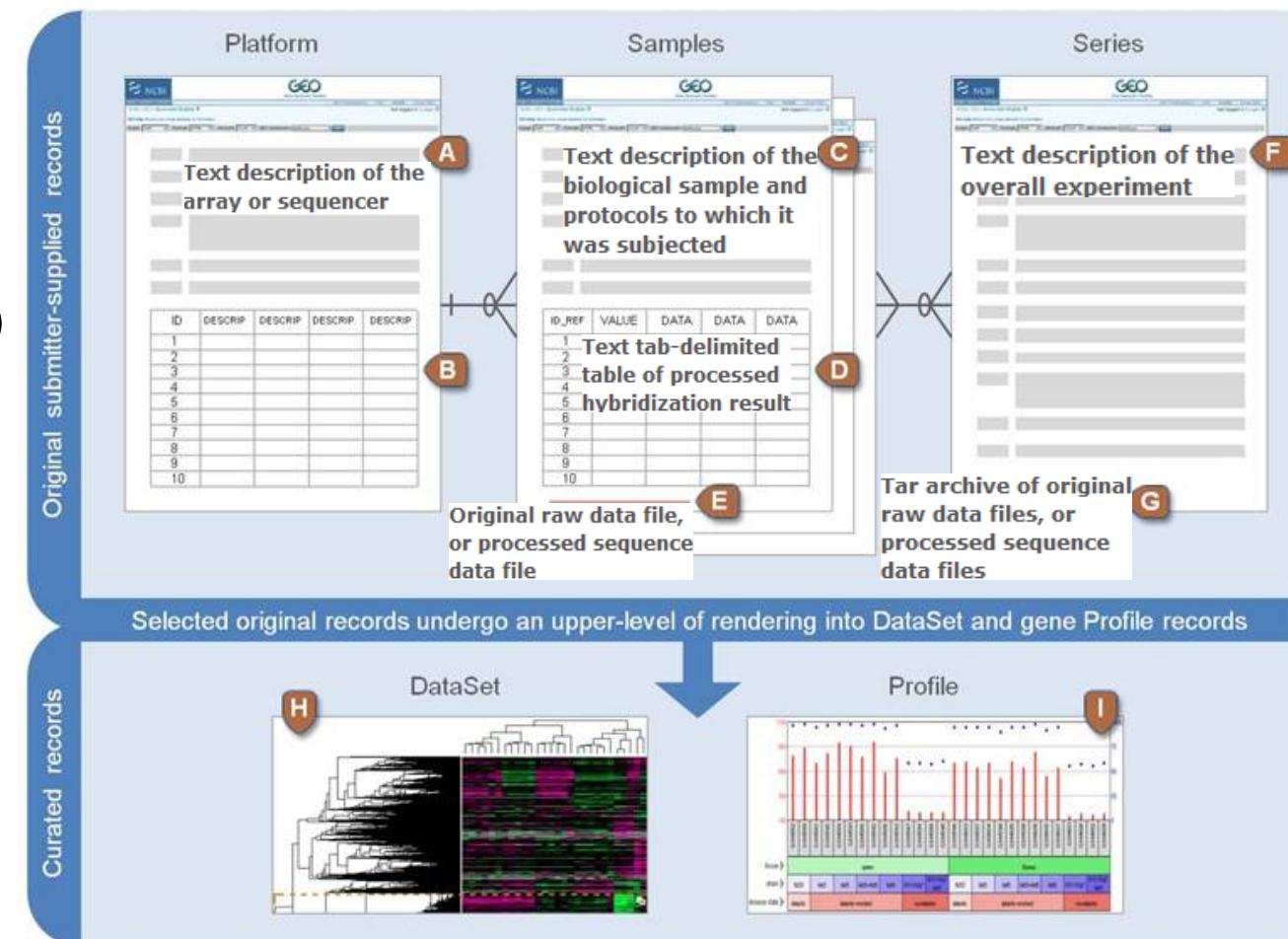
- GEO is an international public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomics data submitted by the research community.
- The three main goals of GEO are:
  - 1) Data organization: provide a robust, versatile database in which to efficiently store high-throughput functional genomic data
  - 2) Data submission: offer simple submission procedures and formats that support complete and well-annotated data deposits from the research community
  - 3) Query and analysis: provide user-friendly mechanisms that allow users to query, locate, review and download studies and gene expression profiles of interest

# Practicum

## Gene Expression Omnibus (GEO)

- Data organization:

- Platform (GPLxxx)
- Samples (GSMxxx)
- Series (GSExxx)



- Datasets (curated) (GSDxxx)
- Profiles (curated)

See some examples

# Practicum

## Gene Expression Omnibus (GEO)

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

Example with GDS browser:

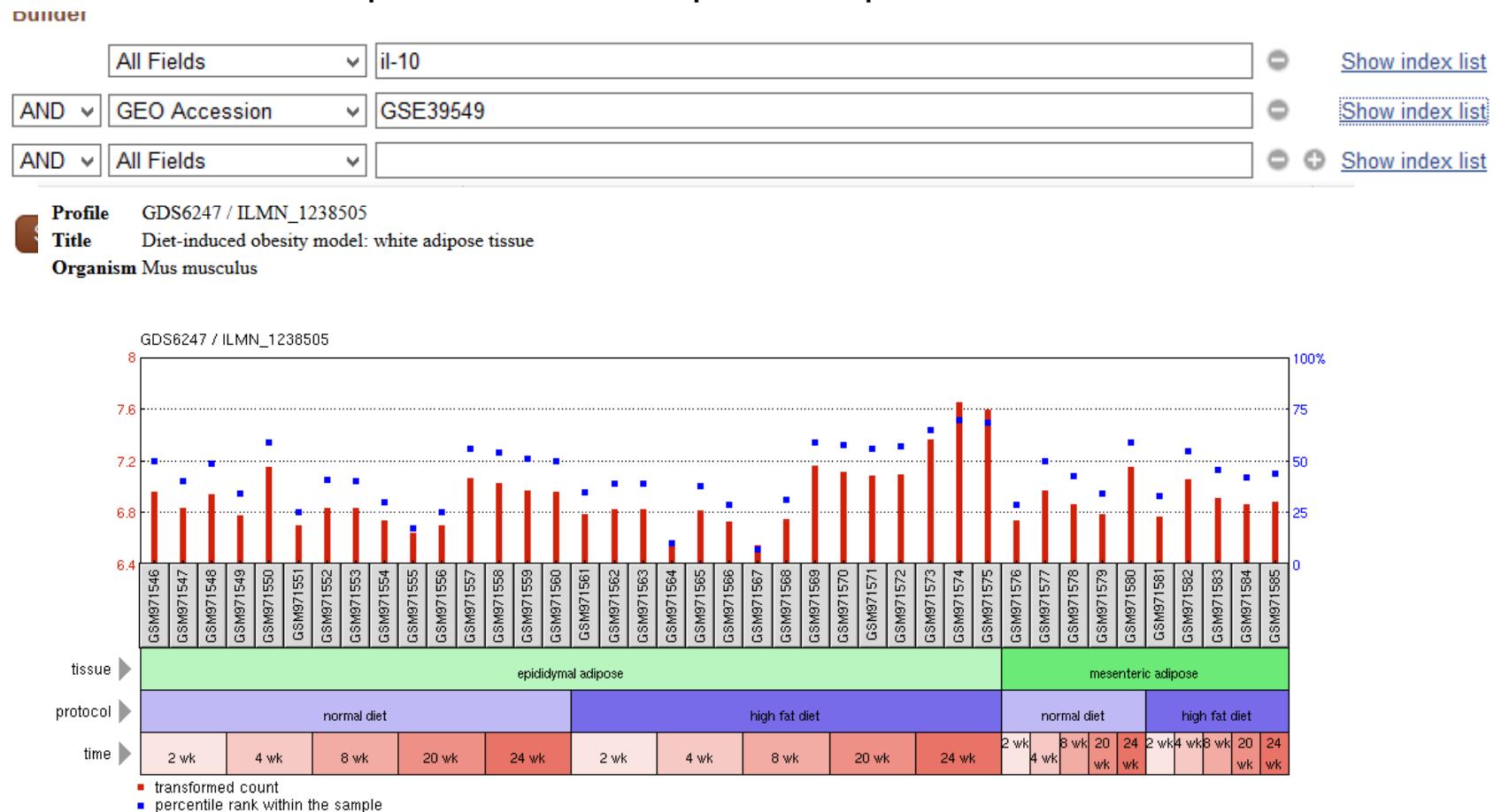
DataSet	Title	Organism(s)	Platform	Series	Samples
GDS6248	Diet-induced obesity model: liver	<i>Mus musculus</i>	GPL6887	GSE39549	51
GDS6247	Diet-induced obesity model: white adipose tissue	<i>Mus musculus</i>	GPL6887	GSE39549	40
GDS6177	Acute alcohol consumption effect on whole blood (control...	<i>Homo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

# Practicum

## Gene Expression Omnibus (GEO)

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

Example: search for expression profile of IL-10



# Practicum

## Gene Expression Omnibus (GEO)

- Formulating queries: GEO DataSets and GEO Profiles are part of NCBI's network of Entrez databases. As with these other databases, data of interest may be located simply by entering keywords into the GEO DataSets or GEO Profiles search boxes. The Advanced Search and Limits pages, linked at the head of the GEO DataSets and GEO Profiles pages, assist greatly in the construction of complex queries.

# Practicum

## Retrieving data from GEO

Series GSE39549		Query DataSets for GSE39549
Status	Public on Mar 01, 2014	
Title	Time-course microarrays reveal early activation of the immune transcriptome and adipokine dysregulation leads to fibrosis in visceral adipose depots during diet-induced obesity	
Organism	<a href="#">Mus musculus</a>	
Experiment type	Expression profiling by array	
Summary	Time-course analysis of adipocyte gene expression profiles response to high fat diet. The hypothesis tested in the present study was that in diet-induced obesity, early activation of TLR-mediated inflammatory signaling	
Overall design	Total RNA obtained from isolated epididymal and mesenteric adipose tissue of C57BL/6J mice fed normal diet or high fat diet for 2, 4, 8, 20 and 24 weeks	
Contributor(s)	Kwon E. Choi M	
Platforms (1)	GPL6887 Illumina MouseWG-6 v2.0 expression beadchip	
Samples (91) <a href="#">+ More...</a>	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	<a href="#">PRJNA171109</a>	
<a href="#">Analyze with GEO2R</a>		

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
<a href="#">Supplementary file</a>	
GSE39549_Matrix_non-normalized_EPI.txt.gz	8.4 Mb <a href="#">(ftp)(http)</a> TXT
GSE39549_Matrix_non-normalized_MES.txt.gz	2.9 Mb <a href="#">(ftp)(http)</a> TXT
GSE39549_RAW.tar	15.8 Mb <a href="#">(http)(custom)</a> 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	<a href="#">Mus musculus</a>
Characteristics	strain: C57BL/6J treatment protocol: Normal diet time: 2 weeks age: 7 weeks tissue: epididymal adipose tissue
Treatment protocol	C57BL/6J mice were fed a high-fat diet (HFD) or normal diet (ND) and sacrificed at 5 time-points (2, 4, 8, 20 and 24 weeks) over 24 weeks.
Extracted molecule	total RNA
Extraction protocol	RNA was extracted with Trizol reagent, followed by clean-up and DNase I treatment with QIAGEN RNeasy mini kit in accordance with the prescribed protocol provided with the kit. Quality control was performed with Agilent Bioanalyzer.
Label	biotin
Label protocol	Biotinylated cRNA were prepared with the Ambion MessageAmp kit for Illumina arrays
Hybridization protocol	Standard Illumina hybridization protocol
Scan protocol	Standard Illumina scanning protocol
Description	Sample name: E2N1 replicate 1
Data processing	Raw data were extracted using the software provided by the manufacturer (Illumina BeadStudio v3.1.3 (Gene Expression Module v3.3.8). The data were normalised by quantile method using ArrayAssist®.

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

### Data table header descriptions

ID_REF	VALUE
	normalized signal

### Data table

ID_REF	VALUE
ILMN_2417611	7.1251793
ILMN_2762289	6.838682
ILMN_2896528	12.505199
ILMN_2721178	11.040463
ILMN_2458927	6.5777917

Data table with normalized expression values

# Practicum

## Retrieving data from GEO in *R*!



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

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

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

Let's try: *Practicum\_geoqueries.Rmd*

# Summary

- **Databases**
  - data collections
  - many types and diverse information
- **Subject-specific repositories**
- **Resource Providers / Collaborative projects**
  - centers or organizations specialized in storing and maintaining databases
  - centralize data management
- **Bioinformatics Tools**
  - computer programs for exploiting the information