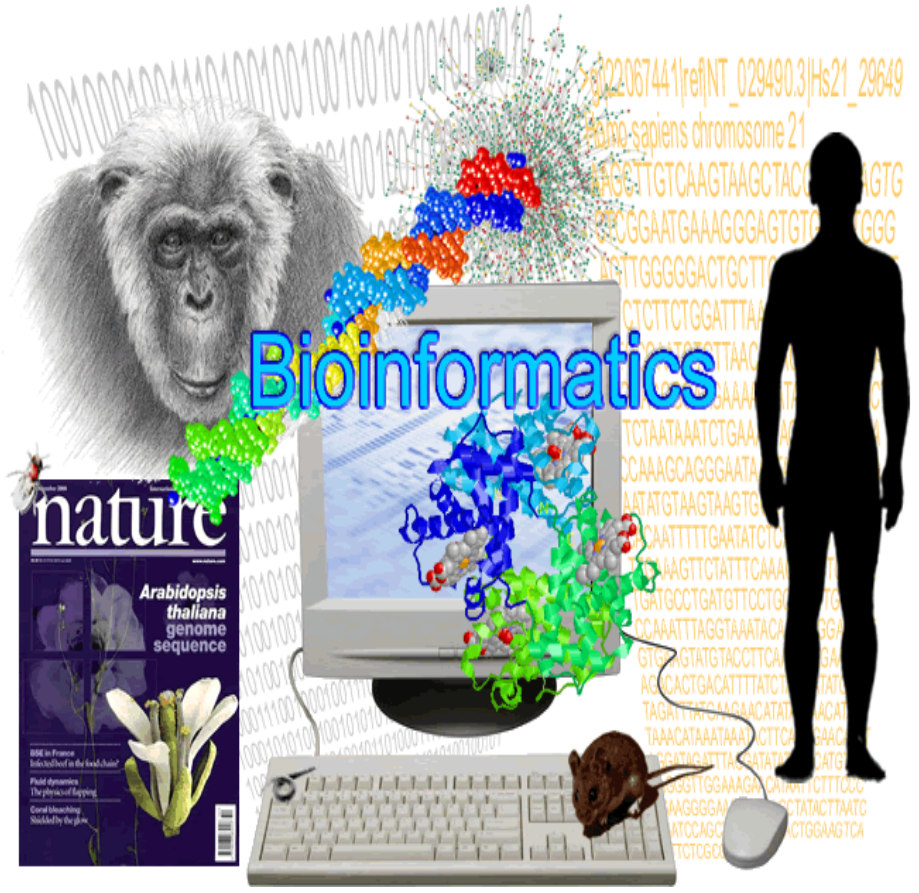
Databases in molecular biology

Bioinformatics Course UEB-VHIR
November 2020

Ricardo Gonzalo[1], Mireia Ferrer[1], Álex Sánchez[1,2]
Berta Miró[1], Angel Blanco[1,2]

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/

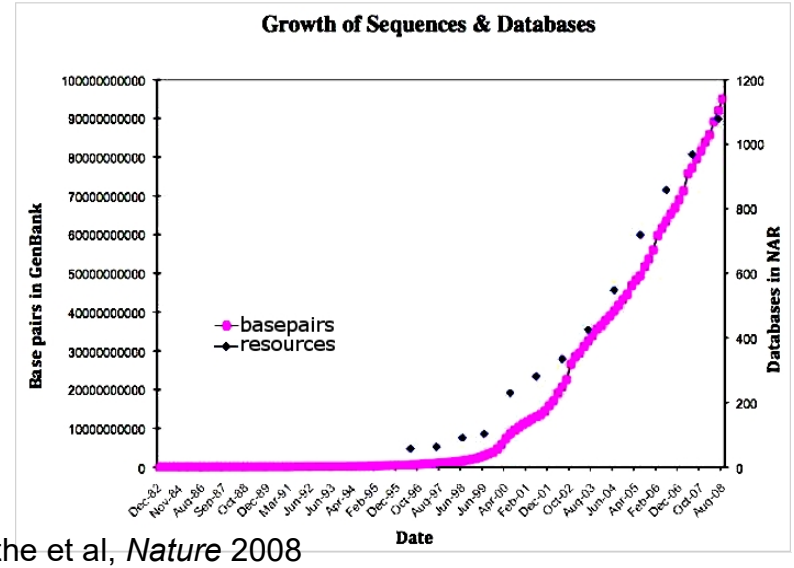
Biological Databases

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



Lathe et al, *Nature* 2008

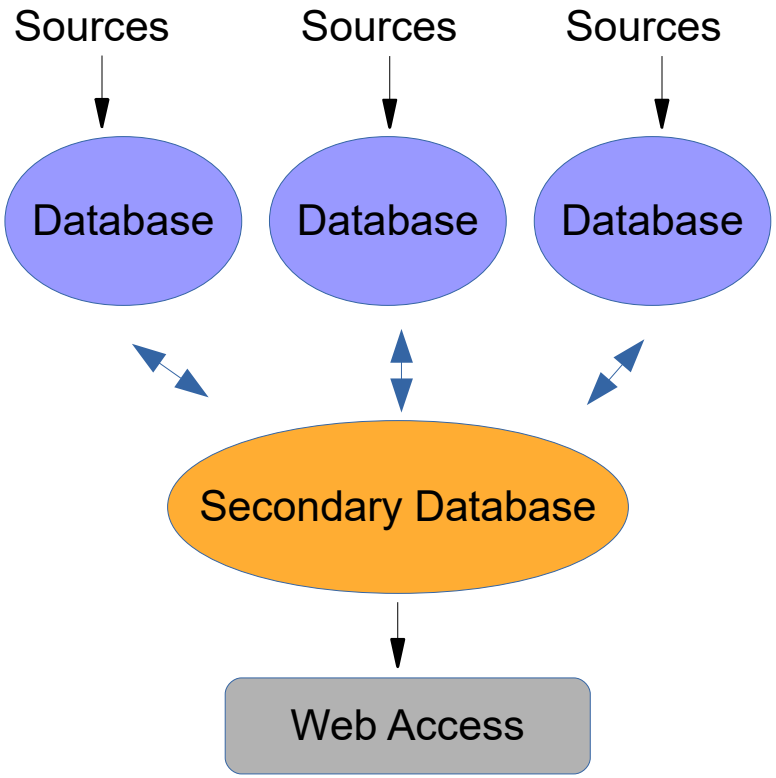
This large number of databases, though extremely useful, can lead to its own issues of redundancy and lack of integration.

- **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)

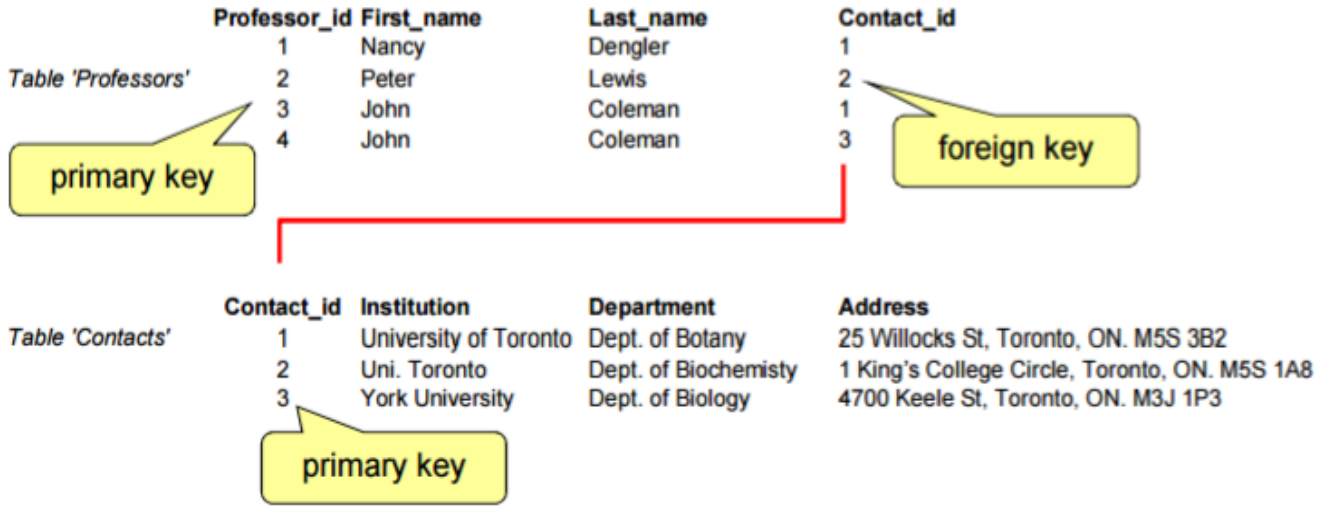
Identifier Description

sequence

```
>l36435 Mus Musculus basic domain/leucine zipper transcription factor mRNA, complete cds.
gcgccgccgcgtccccagacaaaggcttggccggcggccccggcccgctgcgccctcggctccccgcctccc
cggcttgccgctcttcgccccgcgtttggctcggcgcgtcccggccggccgcaaagttttccccgcggcag
cggcggctgagcctcgcttttagcgatggccgcggagctgagcatggggcaagagctgcccaccagcccgct
ggccatggagtacgtcaacgacttcgaccttctcaagttcgacgtgaagaaggagcccctggggcgcgcgga
gcgtccgggccggccatgcacacgcctgcagcctgctggctcggtgtcgtccaccccgctcagcactccgtg
```

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 identificator)



Integrating the information

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

Example: human gene ADH6

GenBank

```
LOCUS       AH001409                2625 bp    DNA     linear   PRI 10-JUN-2016
DEFINITION  Homo sapiens alcohol dehydrogenase 6 (ADH6) gene, complete cds.
ACCESSION   AH001409 M68895 M84402 M84403 M84404 M84405 M84406 M84407 M84408
            M84409
VERSION     AH001409.2
KEYWORDS    .
SOURCE      Homo sapiens (human)
```

UniProt

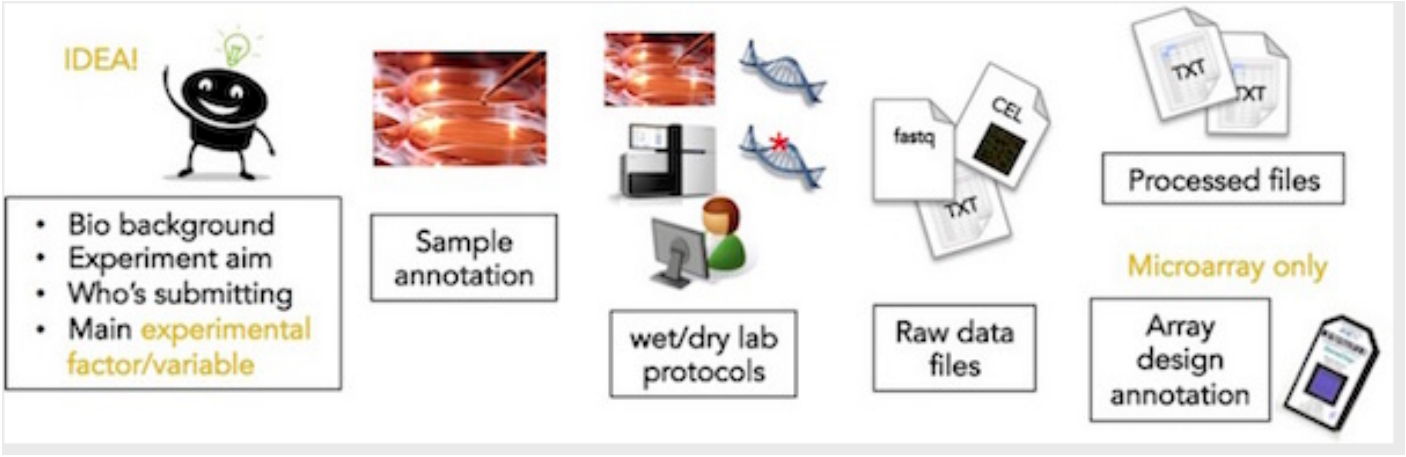
Entry	Entry name		Protein names	Gene names
P28332	ADH6_HUMAN		**Alcohol dehydrogenase 6**	**ADH6**

II. Data Annotation and Curation

Data Annotation

- 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)

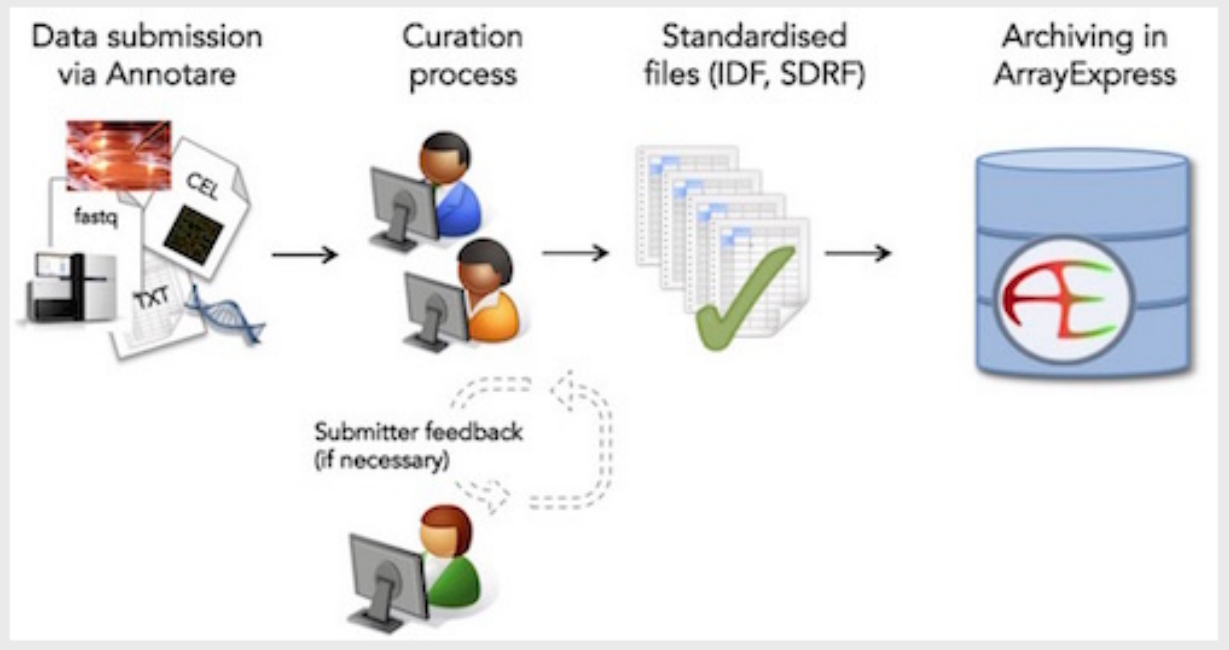


Source: https://www.ebi.ac.uk/arrayexpress/submit/overview.html

- 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.

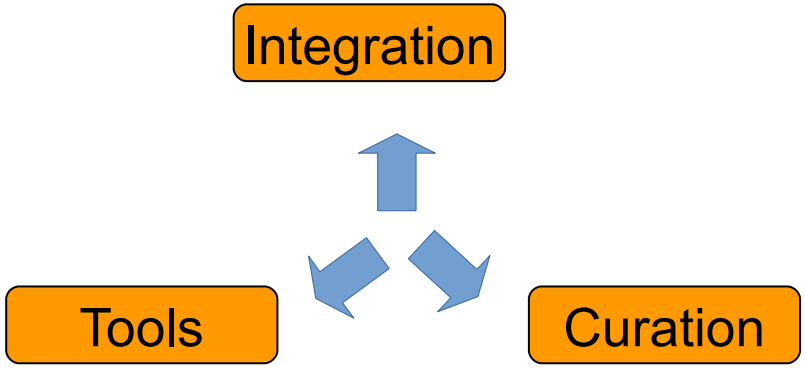


Source: https://www.ebi.ac.uk/arrayexpress/submit/overview.html

III. Centralizing data management

Centralizing data management

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



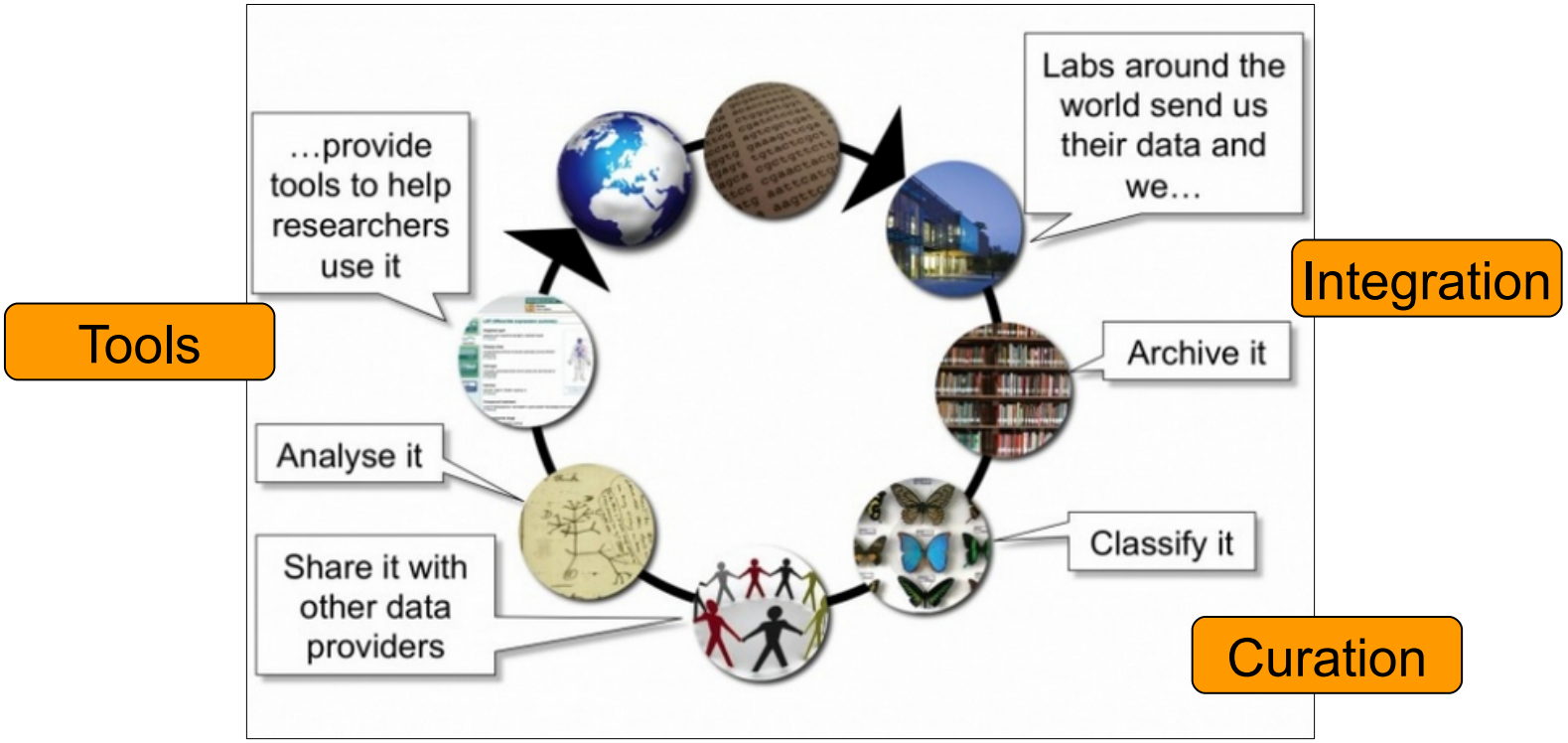
Resource providers

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

NCBI

EMBL-EBI

NATIONAL INSTITUTE OF GENETICS - JAPAN - NIG



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

Resource providers

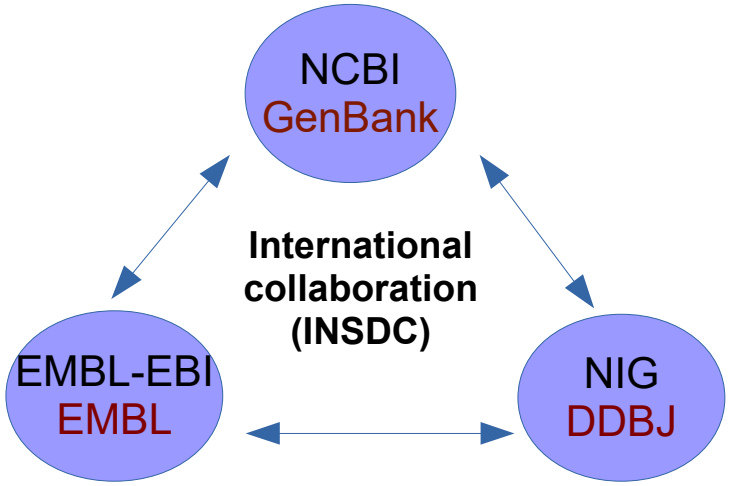
NCBI

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

EMBL-EBI

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

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



Resource providers

NCBI

- 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)

EMBL-EBI

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

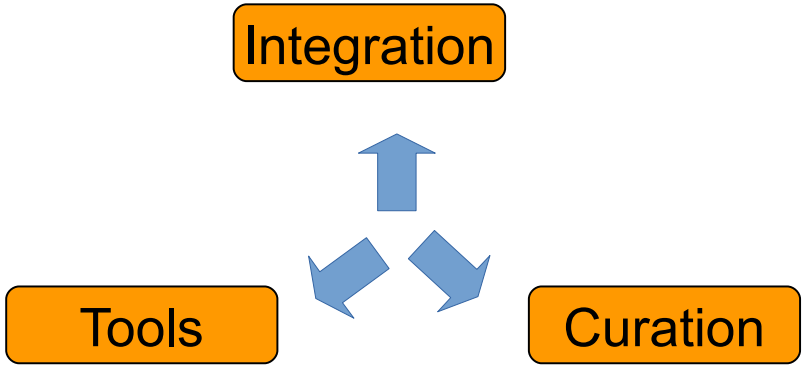
Fast sensitive protein homology searches using profile hidden Markov models (HMMs) for querying against both sequence and HMM target databases.

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



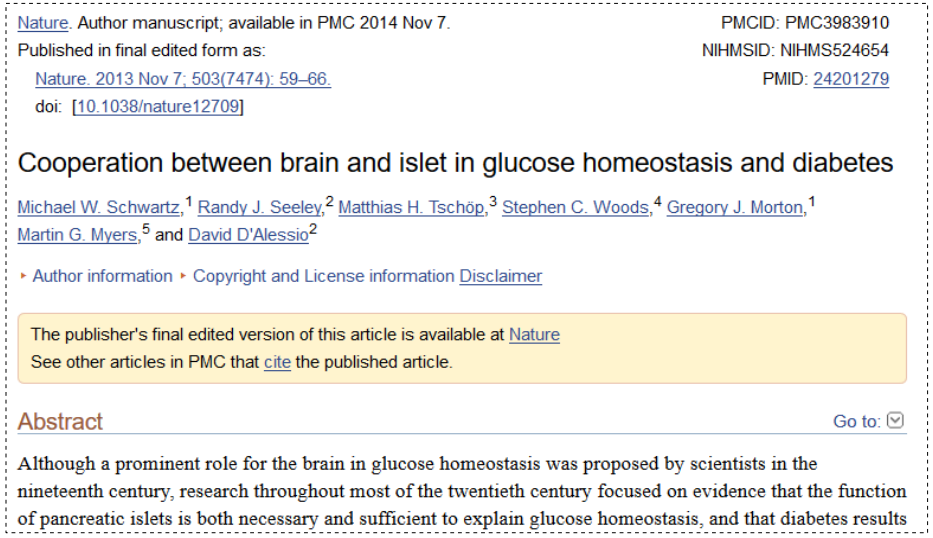
Session 3

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

Author information · Copyright and License information Disclaimer

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

Abstract Go to:

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

Examples of Databases

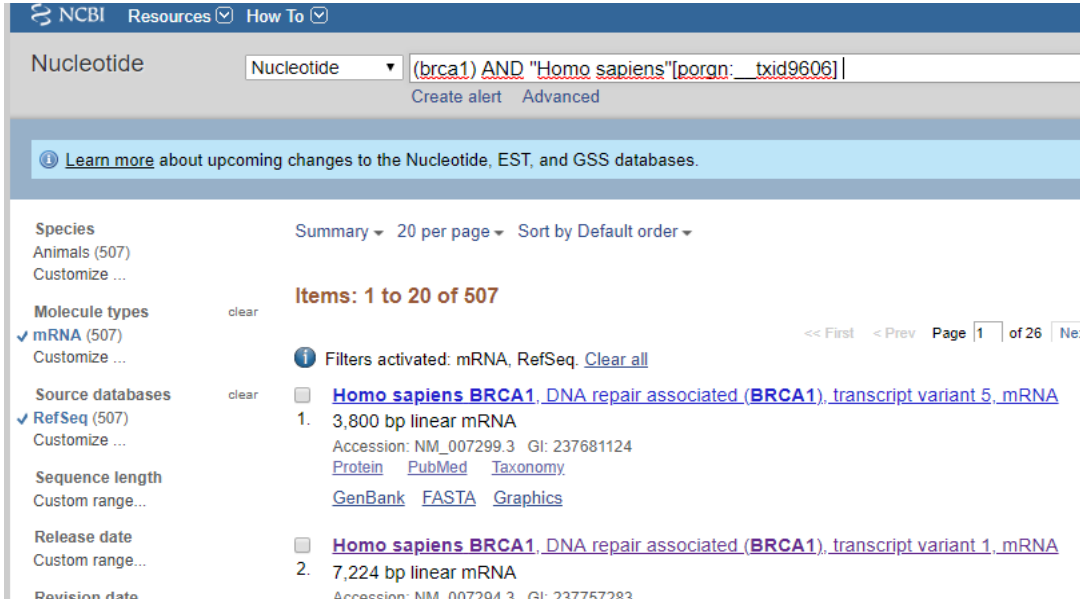
Taxonomic DB

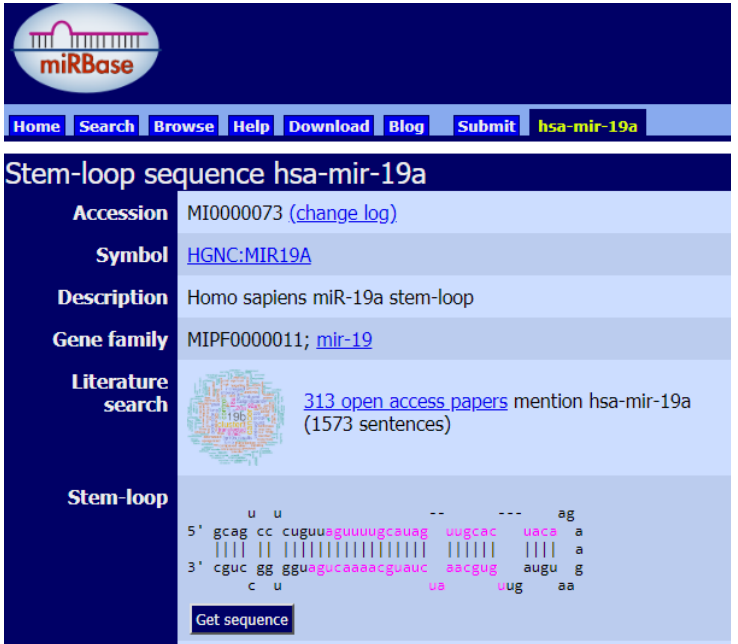
- Contain information about the classification of organisms, mainly from molecular data
- Taxonomy DB: curated classification and nomenclature for all of the organisms in the public sequence databases.
- This represents about 10% of described species

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.



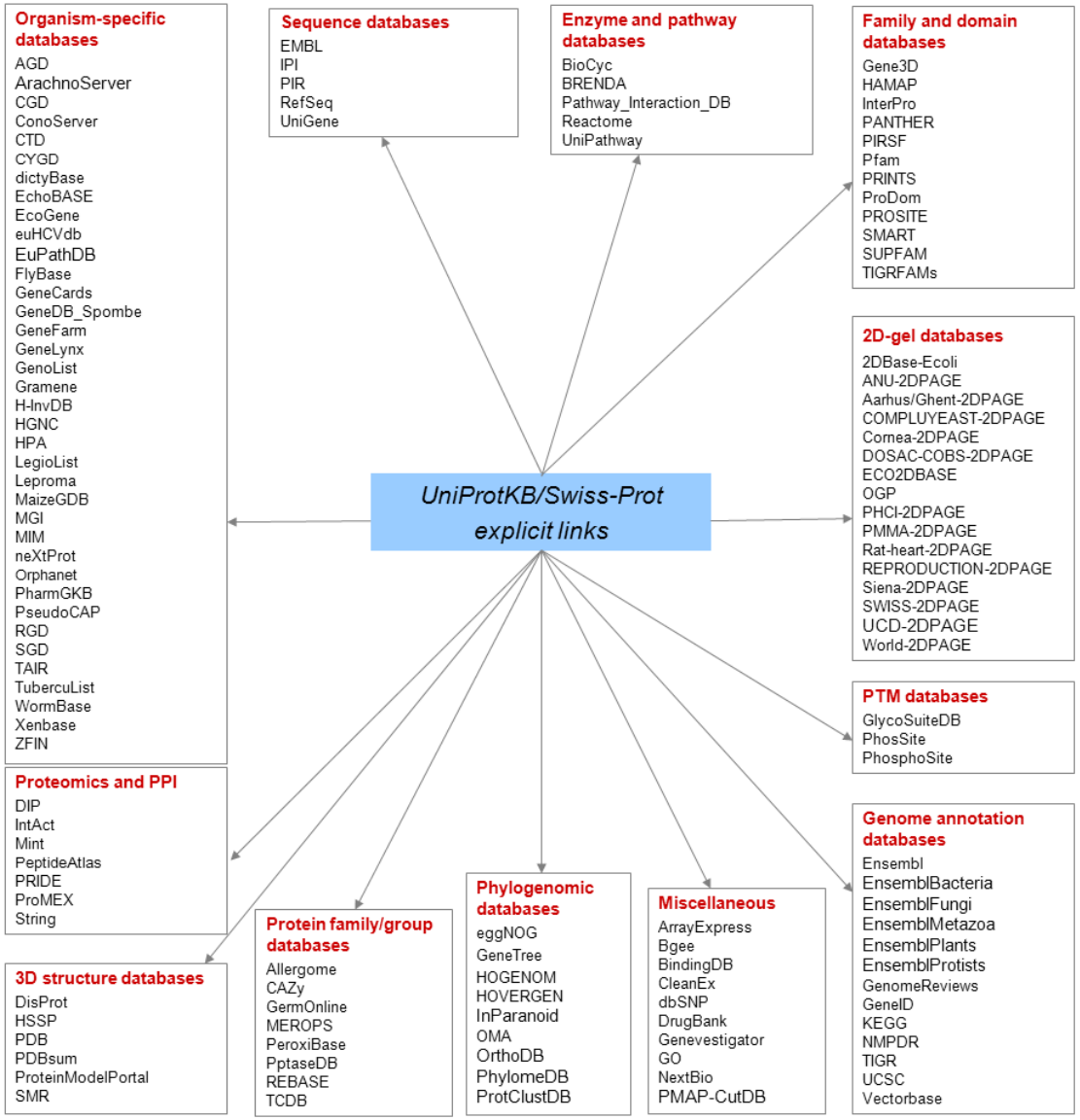


Examples of Databases

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



https://web.expasy.org/docs/userman.html

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

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*)

Vall d'Hebron Institut de Recerca VHIR

UNITAT D'ESTADÍSTICA I BIOINFORMÀTICA

Tools

NCBI

- 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)

EMBL-EBI

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

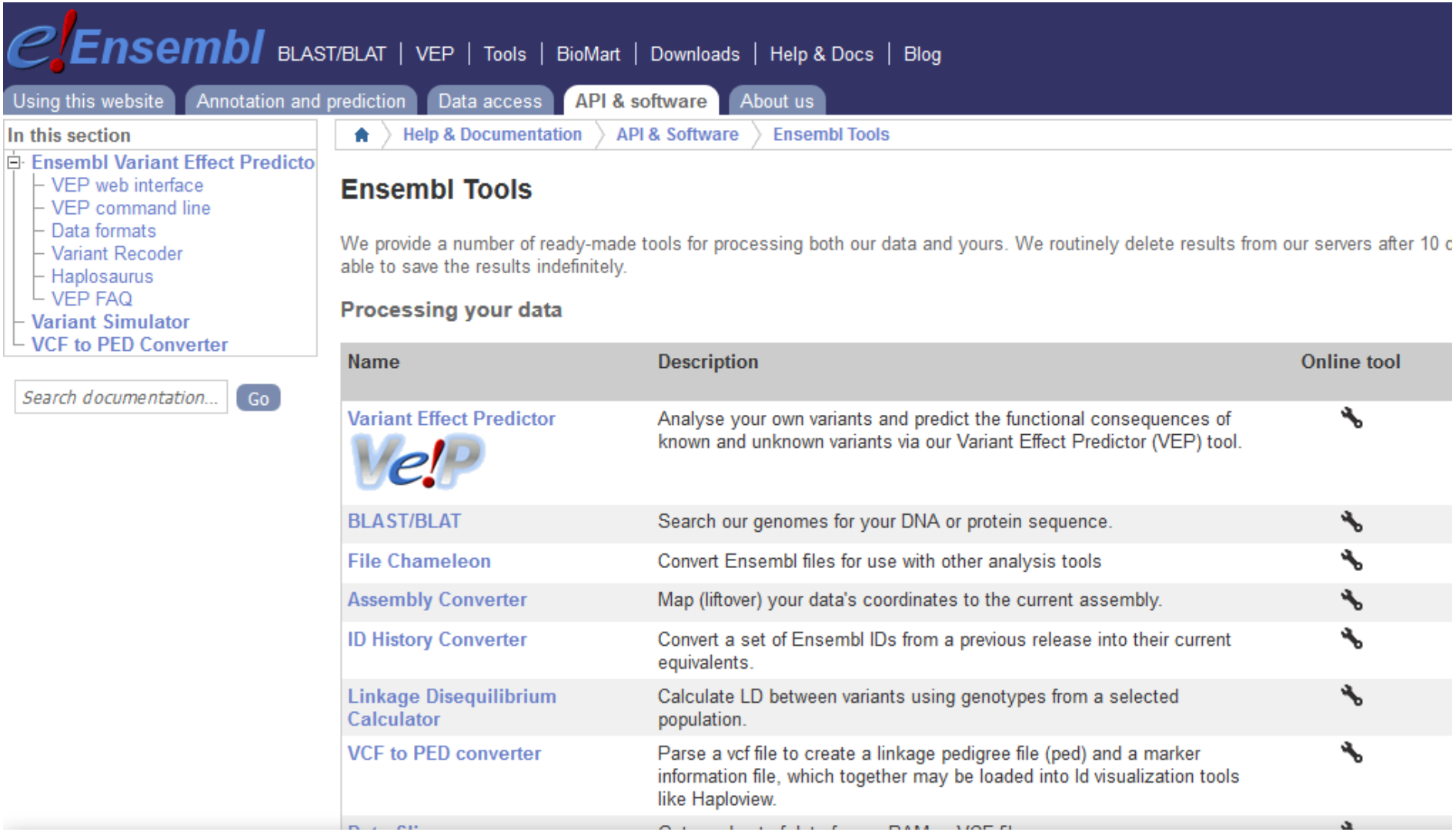
Fast sensitive protein homology searches using profile hidden Markov models (HMMs) for querying against both sequence and HMM target databases.

Sequence similarity search | Protein function analysis

Vall d'Hebron Institut de Recerca VHIR

UNITAT D'ESTADÍSTICA I BIOINFORMÀTICA

Tools



https://www.ensembl.org/info/docs/tools/index.html

Tools

Vall d'Hebron Institut de Recerca VHIR

UNITAT D'ESTADÍSTICA I BIOINFORMÀTICA

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Home | Install | Help | Developers | About

Search:

Home » BiocViews

All Packages

Bioconductor version 3.12 (Release)

Autocomplete biocViews search:

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

Packages found under FunctionalGenomics:

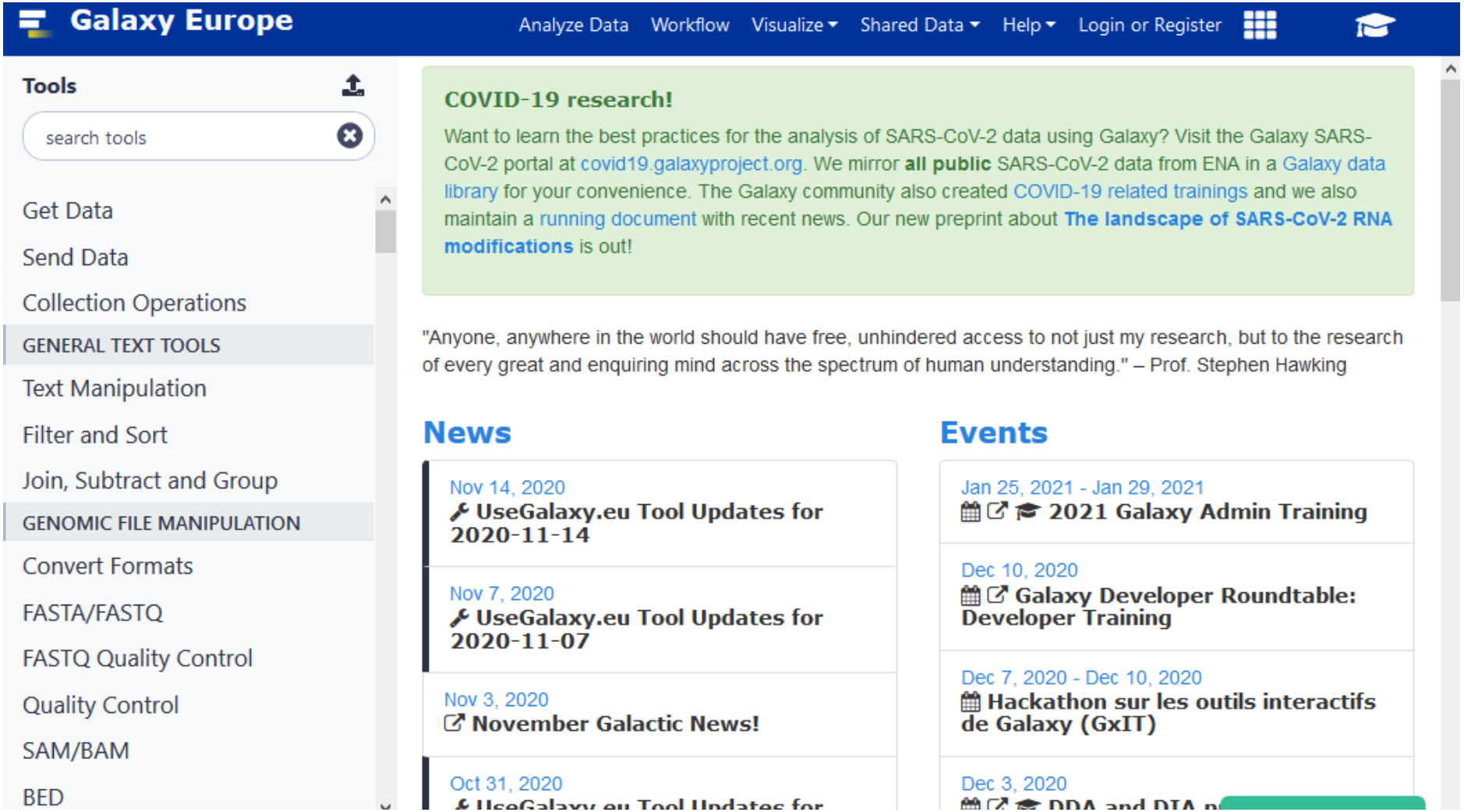
Rank based on number of downloads: lower numbers are more frequently downloaded.

Show All entries Search table:

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

https://bioconductor.org/packages/3.12/BiocViews.html#___FunctionalGenomics

Tools



https://usegalaxy.eu/

Vall d'Hebron Institut de Recerca VHIR

UNITAT D'ESTADÍSTICA I BIOINFORMÀTICA

Tools

Cytoscape App Store | Submit an App | Search the App Store | Sign In

All Apps

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

Newest Releases

Get Started with the App Store »

DKernel 3.0+

DKernel uses Diffusion Kernel algorithm to propagate sub-

PathLinker 3.0+

Reconstructs signaling pathways from protein interaction networks

XlinkCyNET 3.0+

XlinkCyNET generates residue-to-residue connections provided by

IntAct App 3.0+

BETA: Build molecular interaction networks from IntAct database.

MCODE 3.0+

Clusters a given network based on topology to find densely

OmniPath 3.0+

OmniPath: literature curated human signaling pathways

more newest releases »

https://apps.cytoscape.org/apps/

Tools

https://www.omicsdi.org/

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.

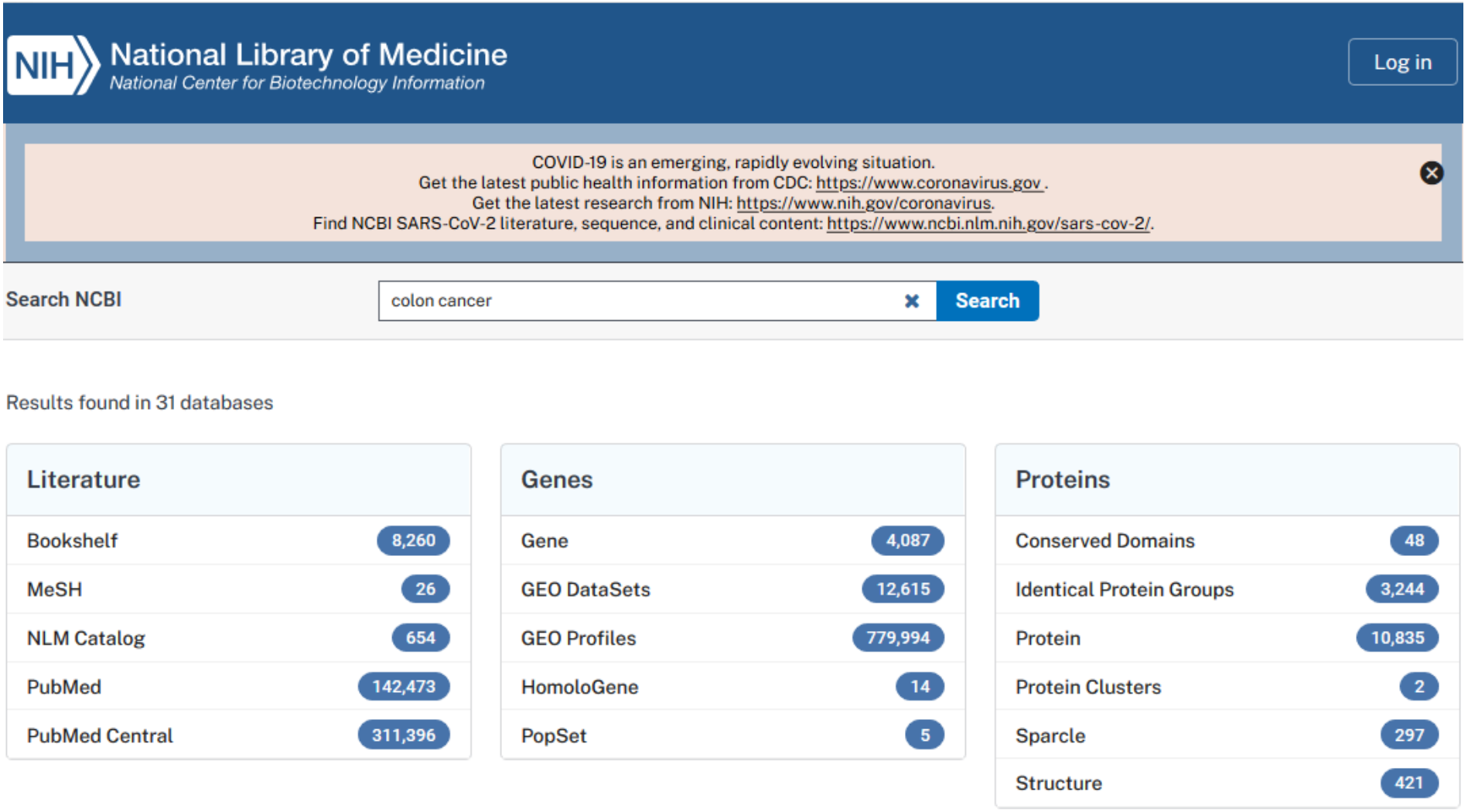
Tools

The Entrez Search and Retrieval System

		Available for Database ...				
Field	Short term	Nucleotide	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



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

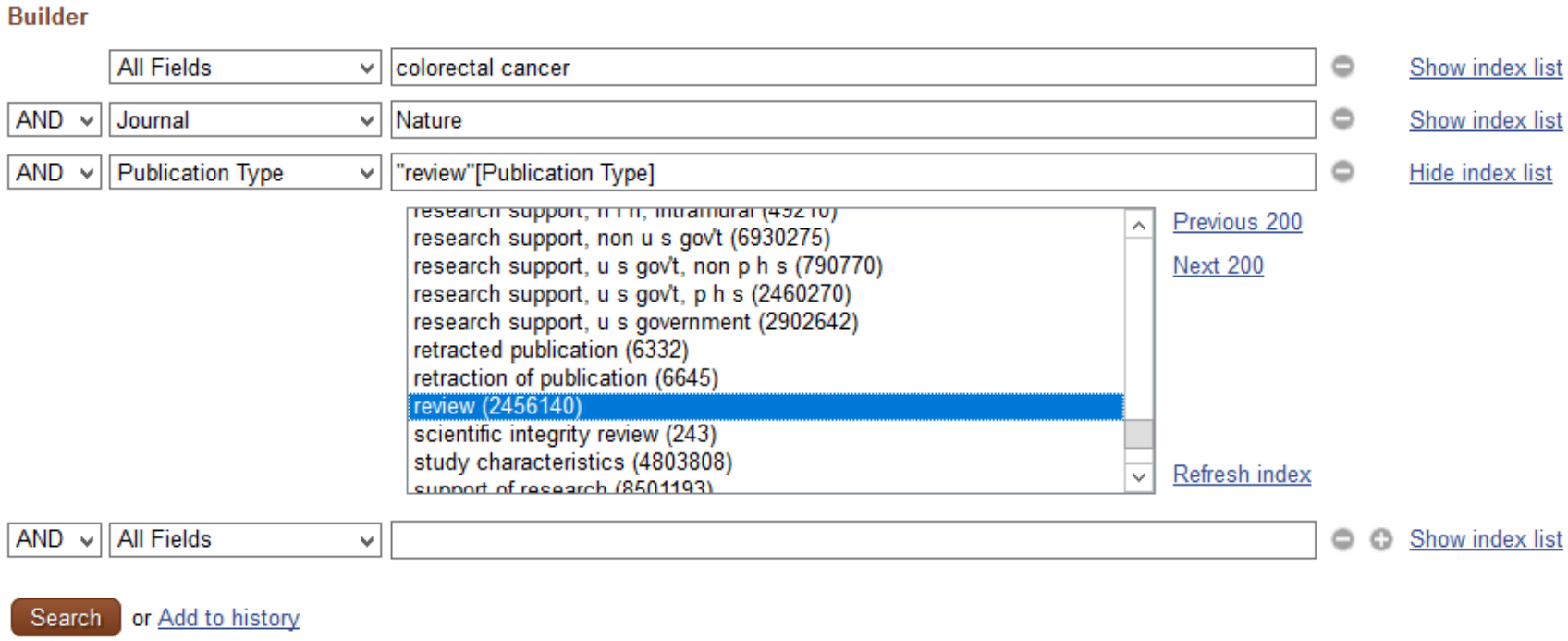
3- Look for MUTYH human protein in UniProtKB

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

Practicum

Formulating specific queries and retrieving nucleotide sequences

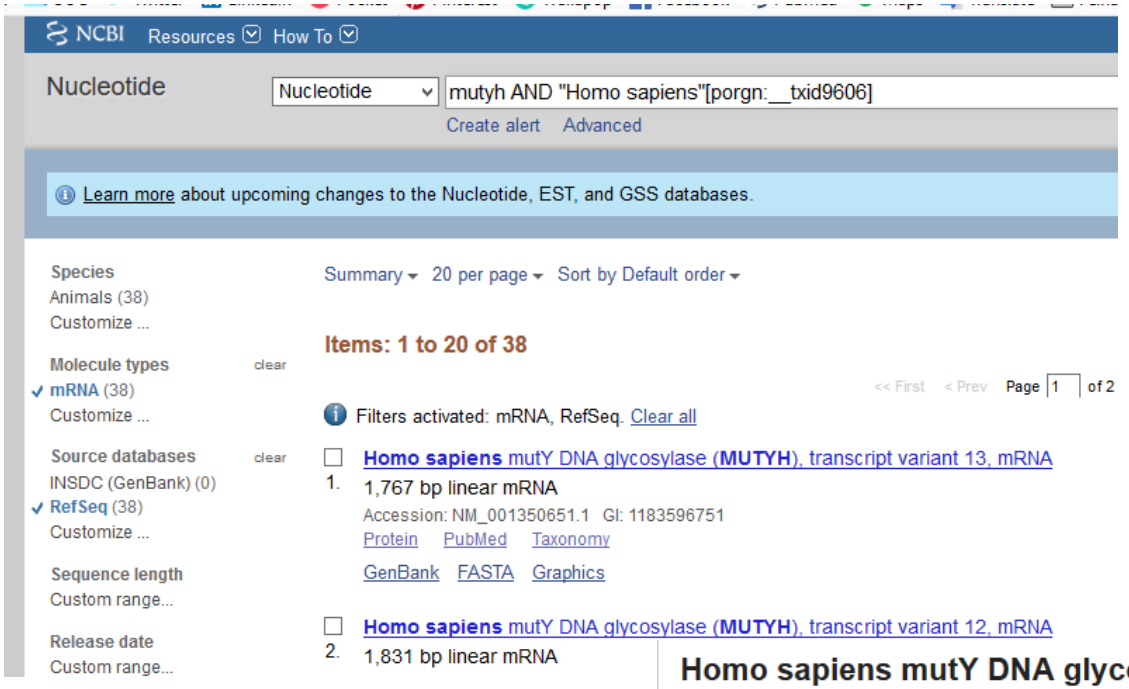
1- Using PubMed Advanced Search, look for a *review* paper published in *Nature* on *colorectal cancer* and authored by *David SS*

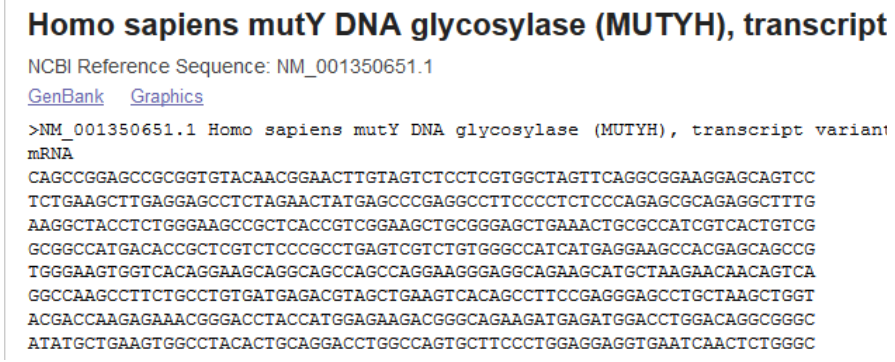


Practicum

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

Using filters



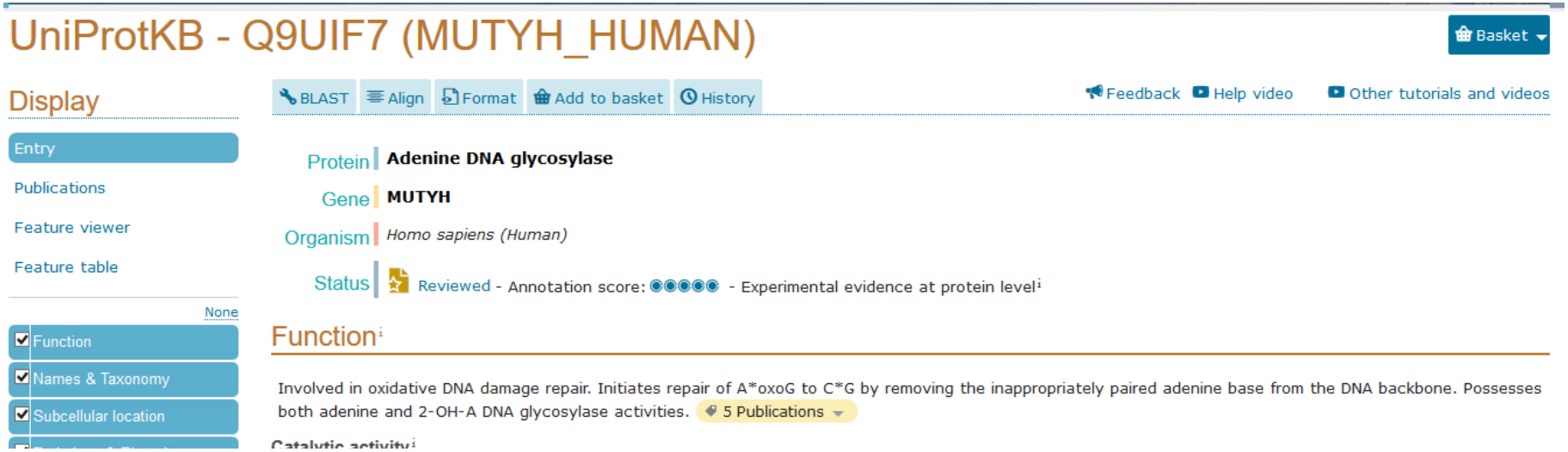


Practicum

Retrieving protein information

3- Look for MUTYH human protein in UniProtKB

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



Practicum

Interaction

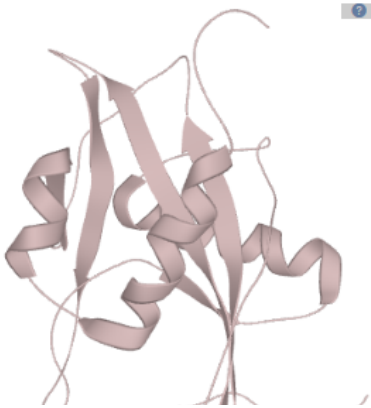
Binary interactions

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

Protein-protein interaction databases

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

Structure



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

Family and domain databases

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

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

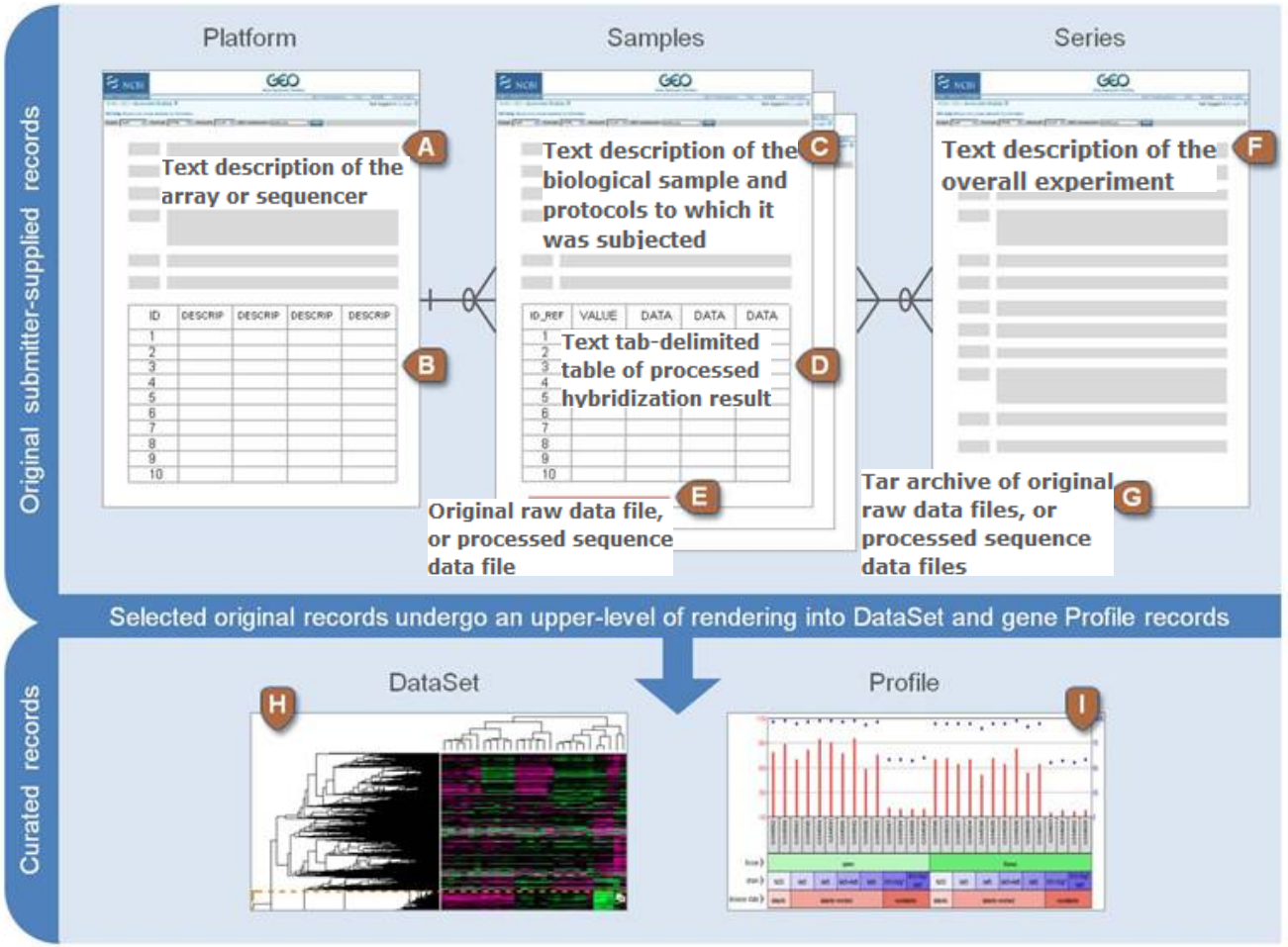
  2) Data submission: offer simple submission procedures and formats that support complete and well-annotated data deposits from the research community

  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

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

Practicum

Gene Expression Omnibus (GEO)

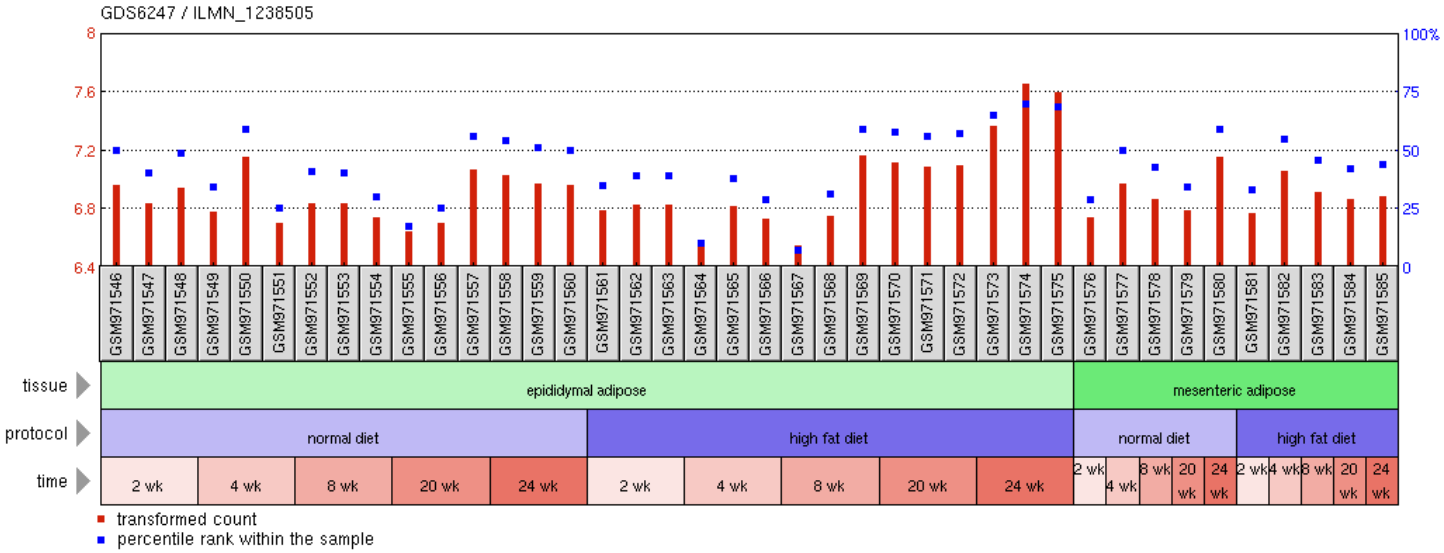
- GEO Profiles: stores individual gene expression profiles from curated DataSets.

Example: search for expression profile of IL-10

Builder

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

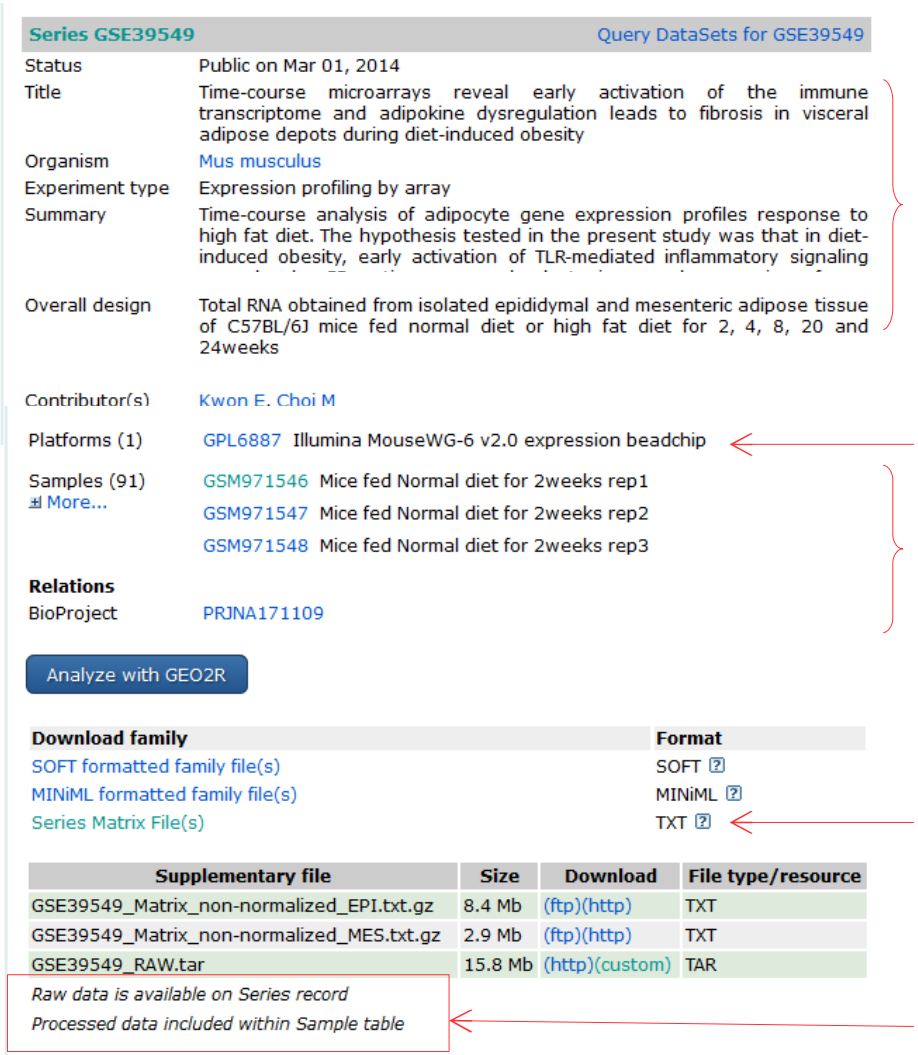
Profile GDS6247 / ILMN_1238505
Title Diet-induced obesity model: white adipose tissue
Organism Mus musculus



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.

Retrieving data from GEO



Series GSE39549 Query DataSets for GSE39549

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

Analyze with GEO2R

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

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

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

Study information

Platform used (data table with annotation of probes)

Samples

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

Info on data files

Practicum

Retrieving data from GEO

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

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

Data table header descriptions	
ID_REF	
VALUE	normalized signal

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

Data table with normalized expression values

Practicum

Retrieving data from GEO 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