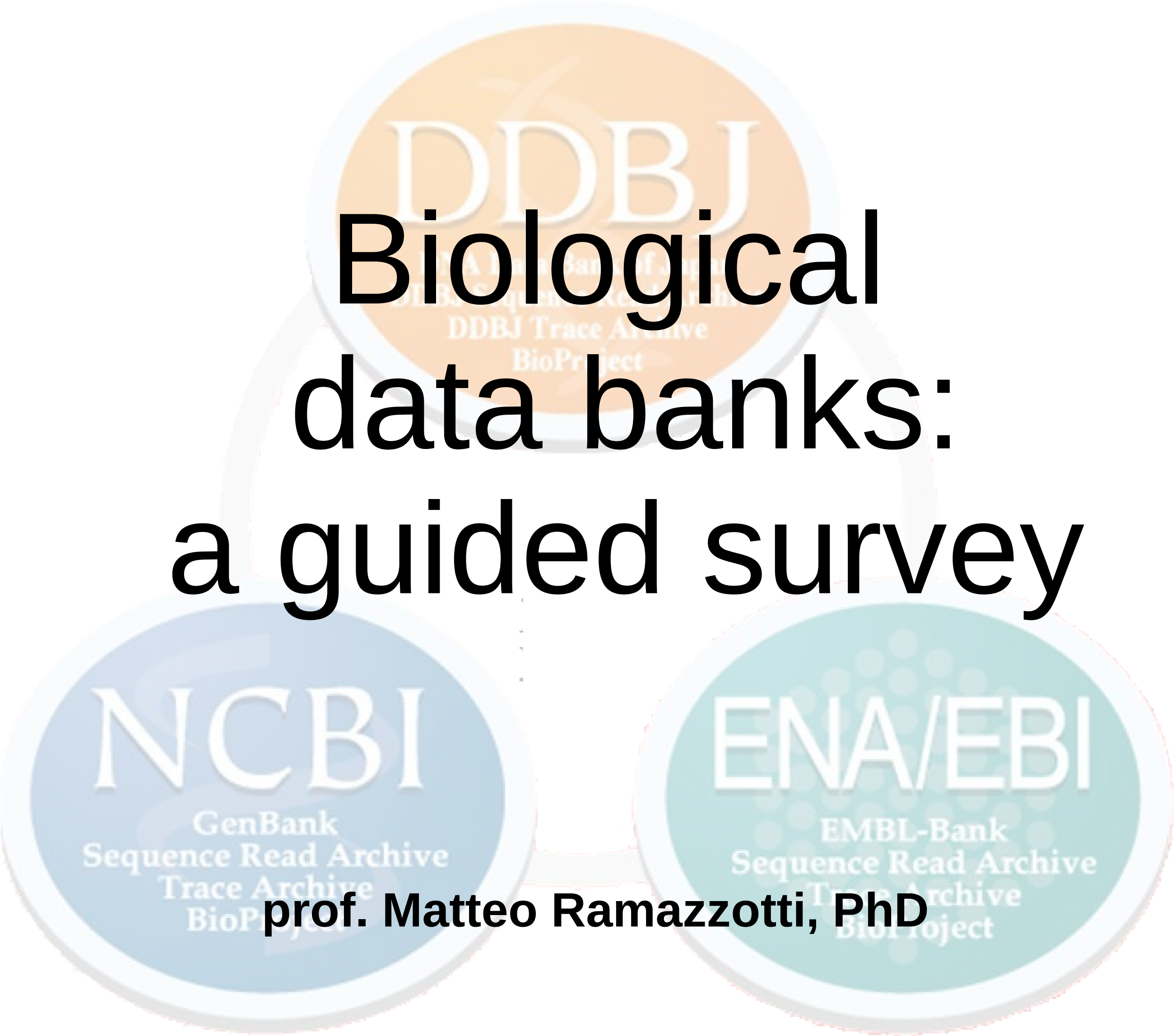


Biological data banks: a guided survey

prof. Matteo Ramazzotti, PhD

matteo.ramazzotti@unifi.it

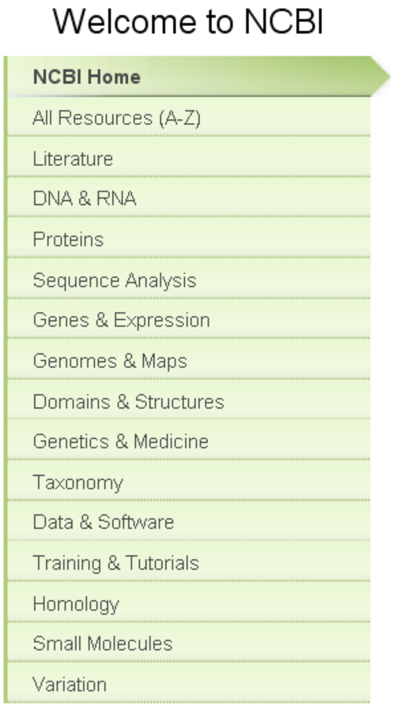
Data banks: the infrastructure

EMBnet

http://www.embnet.org

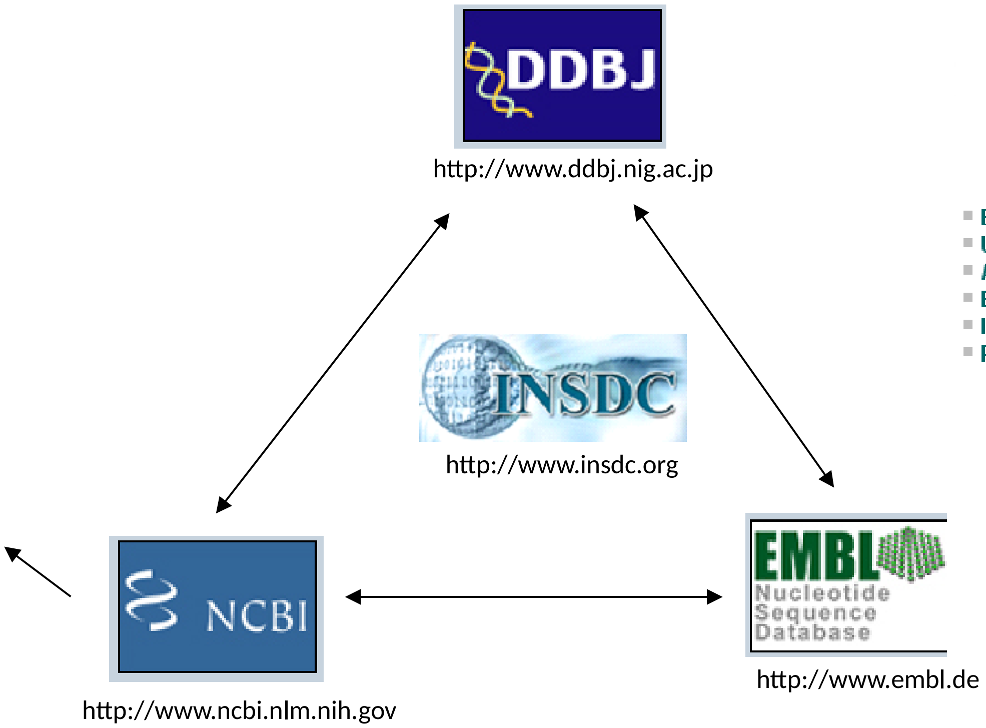
APBioNet

http://www.apbionet.org



Welcome to NCBI

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



http://www.ddbj.nig.ac.jp

http://www.insdc.org

http://www.ncbi.nlm.nih.gov

http://www.embl.de

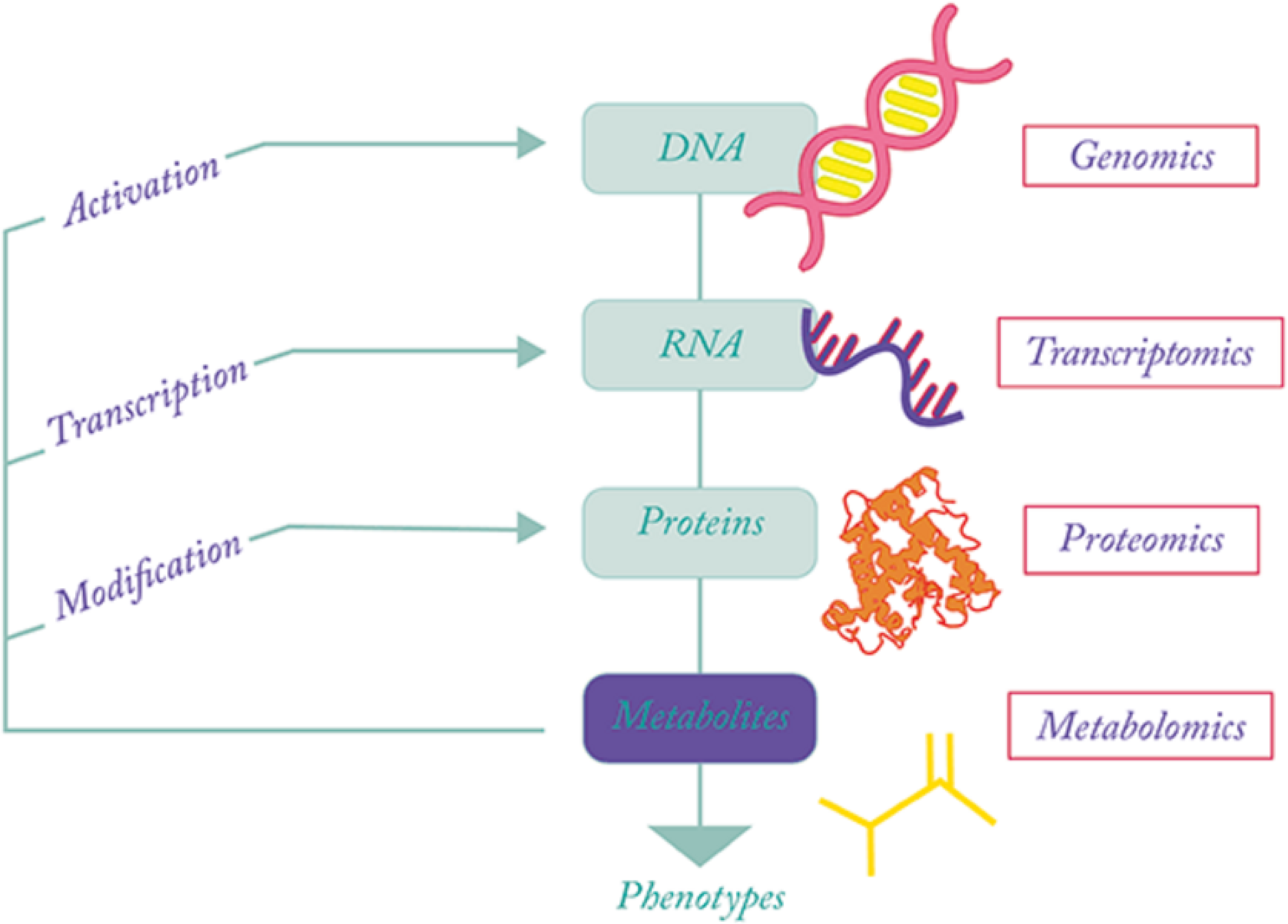
EMBL-EBI

- **ENA**
- **UniProt**
- **ArrayExpress**
- **Ensembl**
- **InterPro**
- **PDBe**

- Genomes
- Nucleotide Sequences
- Protein Sequences
- Macromolecular Structures
- Small Molecules
- Gene Expression
- Molecular Interactions
- Reactions & Pathways
- Protein Families
- Enzymes
- Literature
- Taxonomy
- Ontologies
- Patent Resources
- Sequence Similarity & Analysis
- Pattern & Motif Searches
- Structure Analysis
- Text Mining
- Downloads
- Web Services

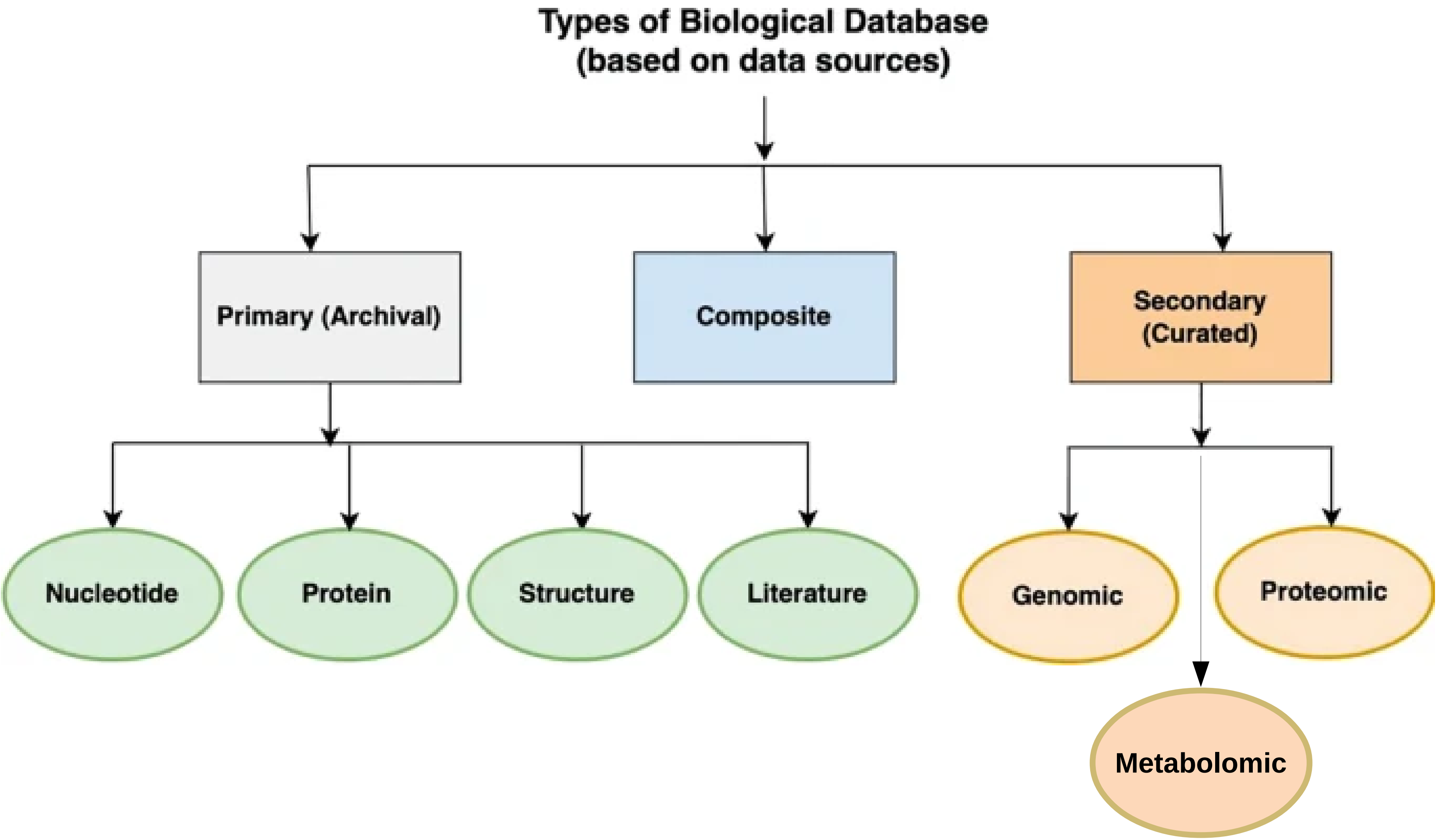
matteo.ramazzotti@unifi.it

Why we need biological databases



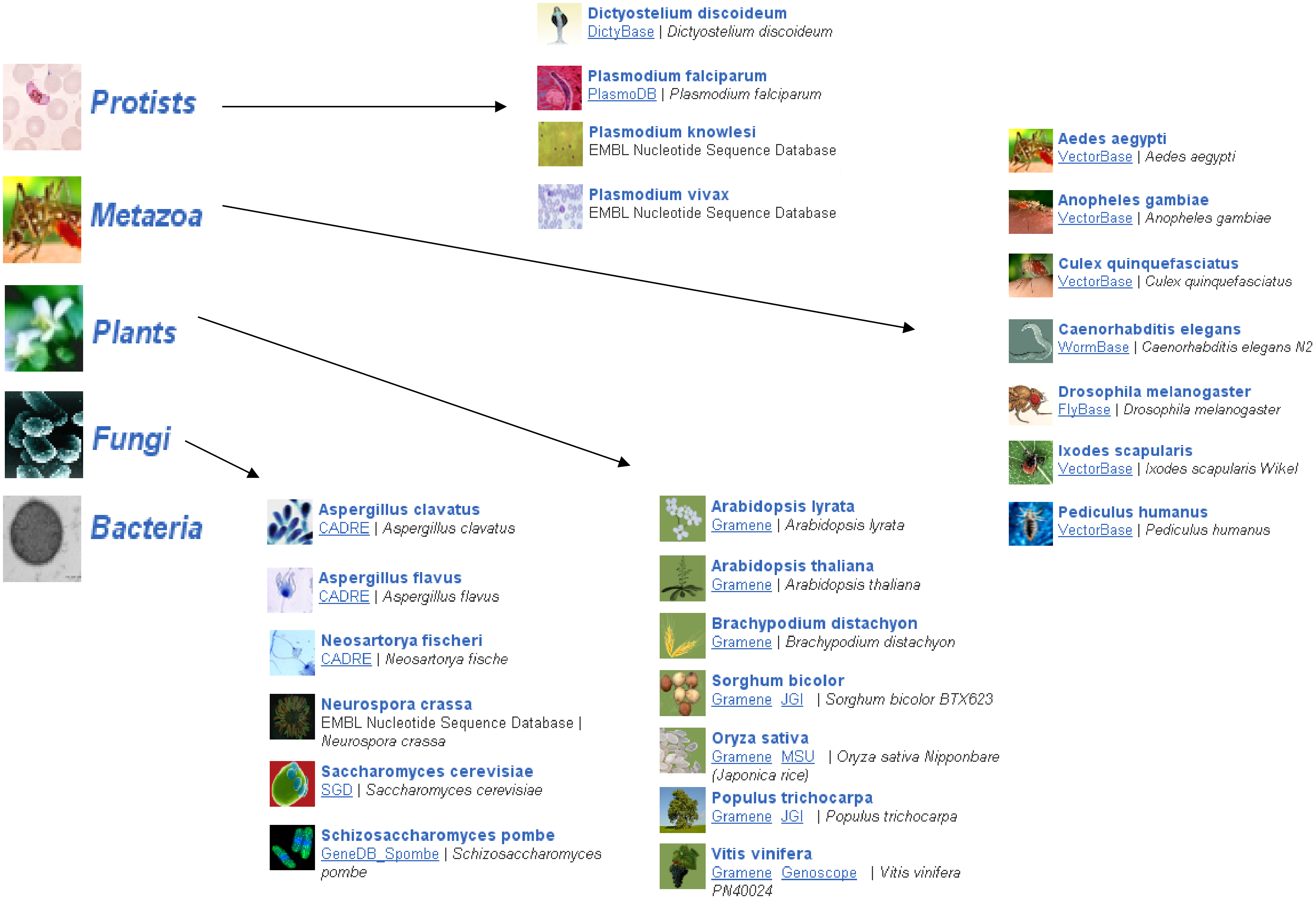
matteo.ramazzotti@unifi.it

Data bank taxonomy



matteo.ramazzotti@unifi.it

Ensembl genomes



Dictyostelium discoideum
DictyBase | *Dictyostelium discoideum*

Plasmodium falciparum
PlasmoDB | *Plasmodium falciparum*

Plasmodium knowlesi
EMBL Nucleotide Sequence Database

Plasmodium vivax
EMBL Nucleotide Sequence Database

Aedes aegypti
VectorBase | *Aedes aegypti*

Anopheles gambiae
VectorBase | *Anopheles gambiae*

Culex quinquefasciatus
VectorBase | *Culex quinquefasciatus*

Caenorhabditis elegans
WormBase | *Caenorhabditis elegans N2*

Drosophila melanogaster
FlyBase | *Drosophila melanogaster*

Ixodes scapularis
VectorBase | *Ixodes scapularis Wikel*

Pediculus humanus
VectorBase | *Pediculus humanus*

Aspergillus clavatus
CADRE | *Aspergillus clavatus*

Aspergillus flavus
CADRE | *Aspergillus flavus*

Neosartorya fischeri
CADRE | *Neosartorya fische*

Neurospora crassa
EMBL Nucleotide Sequence Database | *Neurospora crassa*

Saccharomyces cerevisiae
SGD | *Saccharomyces cerevisiae*

Schizosaccharomyces pombe
GeneDB_Spombe | *Schizosaccharomyces pombe*

Arabidopsis lyrata
Gramene | *Arabidopsis lyrata*

Arabidopsis thaliana
Gramene | *Arabidopsis thaliana*

Brachypodium distachyon
Gramene | *Brachypodium distachyon*

Sorghum bicolor
Gramene JGI | *Sorghum bicolor BTX623*

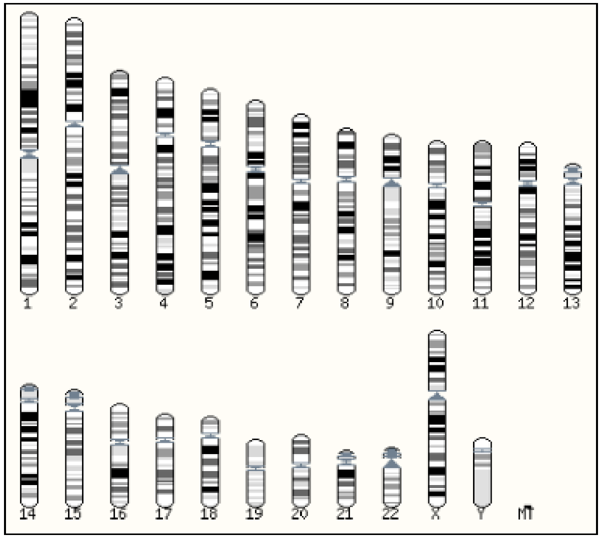
Oryza sativa
Gramene MSU | *Oryza sativa Nipponbare (Japonica rice)*

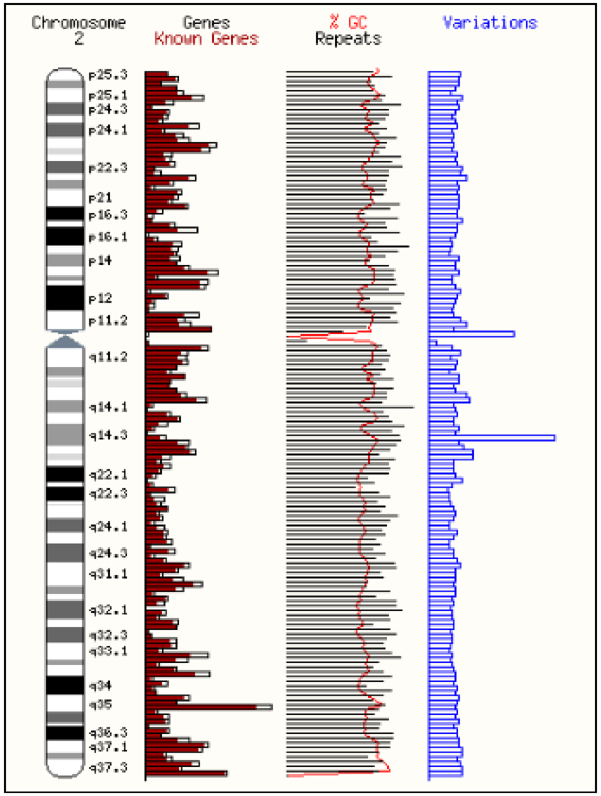
Populus trichocarpa
Gramene JGI | *Populus trichocarpa*

Vitis vinifera
Gramene Genoscope | *Vitis vinifera PN40024*

matteo.ramazzotti@unifi.it

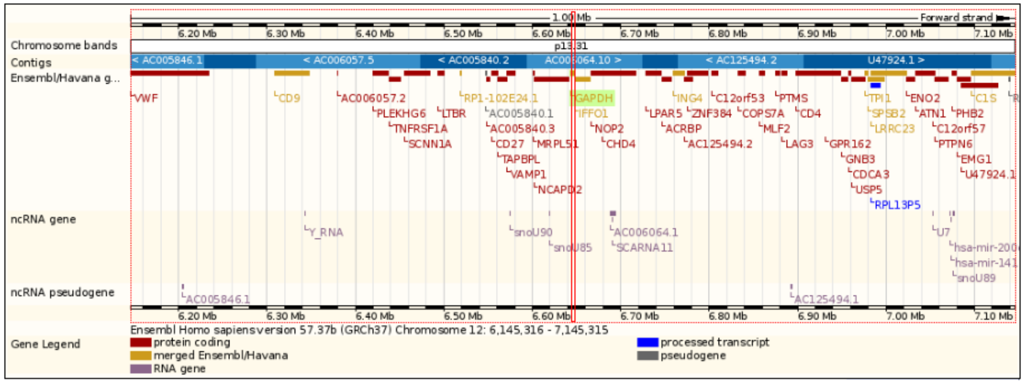
Ensembl: navigating the genomes





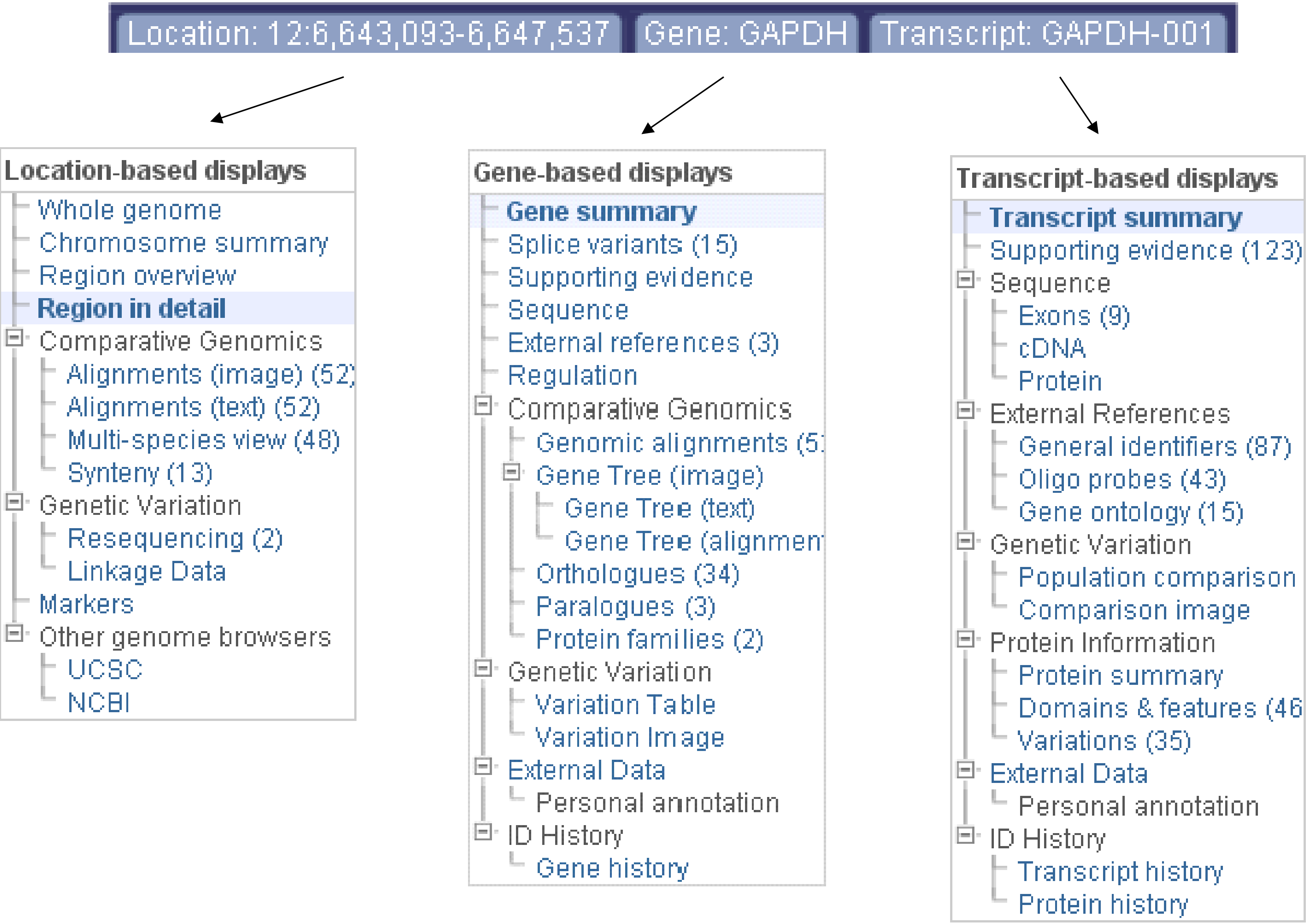
Chromosome Statistics

Length (bps):	243,199,373
Known Protein-coding Genes:	1,301
Novel Protein-coding Genes:	9
Pseudogene Genes:	902
miRNA Genes:	106
rRNA Genes:	35
snRNA Genes:	124
snoRNA Genes:	98
Misc RNA Genes:	114
SNPs:	1,459,890



Ensembl Homo sapiens version 57.37b (GRCh37) Chromosome 12: 6,145,316 - 7,145,315

ENSEMBL genomes as an access point



Location: 12:6,643,093-6,647,537 | Gene: GAPDH | Transcript: GAPDH-001

Location-based displays

- Whole genome
- Chromosome summary
- Region overview
- **Region in detail**
- Comparative Genomics
  - Alignments (image) (52
  - Alignments (text) (52)
  - Multi-species view (48)
  - Synteny (13)
- Genetic Variation
  - Resequencing (2)
  - Linkage Data
- Markers
- Other genome browsers
  - UCSC
  - NCBI

Gene-based displays

- **Gene summary**
- Splice variants (15)
- Supporting evidence
- Sequence
- External references (3)
- Regulation
- Comparative Genomics
  - Genomic alignments (5
  - Gene Tree (image)
    - Gene Tree (text)
    - Gene Tree (alignmen
  - Orthologues (34)
  - Paralogues (3)
  - Protein families (2)
- Genetic Variation
  - Variation Table
  - Variation Image
- External Data
  - Personal annotation
- ID History
  - Gene history

Transcript-based displays

- **Transcript summary**
- Supporting evidence (123)
- Sequence
  - Exons (9)
  - cDNA
  - Protein
- External References
  - General identifiers (87)
  - Oligo probes (43)
  - Gene ontology (15)
- Genetic Variation
  - Population comparison
  - Comparison image
- Protein Information
  - Protein summary
  - Domains & features (46
  - Variations (35)
- External Data
  - Personal annotation
- ID History
  - Transcript history
  - Protein history

matteo.ramazzotti@unifi.it

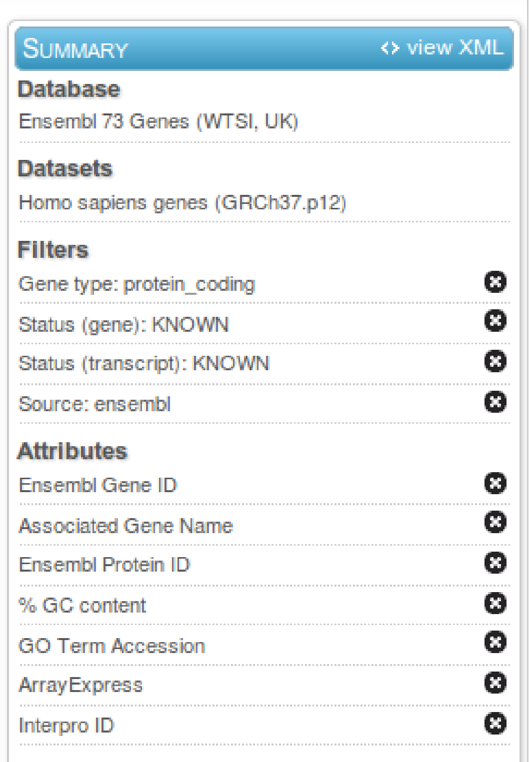
ENSEMBL genomes as an access point

Click to go to the browser

matteo.ramazzotti@unifi.it

BioMart: subsetting big databanks

BioMart is very useful for creating sub-databases of a main database. It expose all the features that can be extracted from a databank and allows to set filters to exclude part of the records.



Database
Ensembl 73 Genes (WTSI, UK)

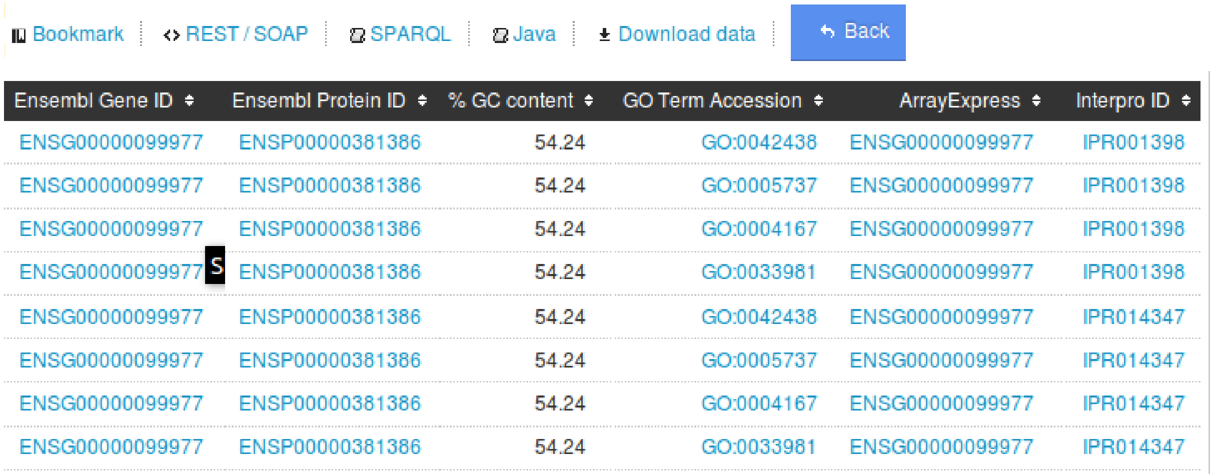
Datasets
Homo sapiens genes (GRCh37.p12)

Filters
Gene type: protein_coding
Status (gene): KNOWN
Status (transcript): KNOWN
Source: ensembl

Attributes
Ensembl Gene ID
Associated Gene Name
Ensembl Protein ID
% GC content
GO Term Accession
ArrayExpress
Interpro ID

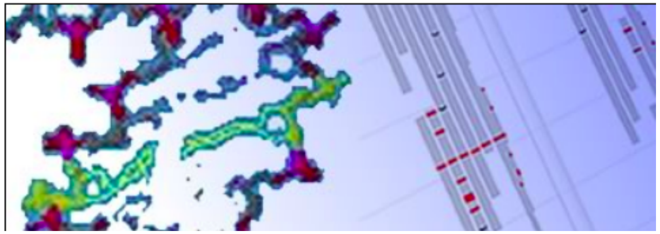
Main dataset

Selection filters



Ensembl Gene ID	Ensembl Protein ID	% GC content	GO Term Accession	ArrayExpress	Interpro ID
ENSG00000099977	ENSP00000381386	54.24	GO:0042438	ENSG00000099977	IPR001398
ENSG00000099977	ENSP00000381386	54.24	GO:0005737	ENSG00000099977	IPR001398
ENSG00000099977	ENSP00000381386	54.24	GO:0004167	ENSG00000099977	IPR001398
ENSG00000099977	ENSP00000381386	54.24	GO:0033981	ENSG00000099977	IPR001398
ENSG00000099977	ENSP00000381386	54.24	GO:0042438	ENSG00000099977	IPR014347
ENSG00000099977	ENSP00000381386	54.24	GO:0005737	ENSG00000099977	IPR014347
ENSG00000099977	ENSP00000381386	54.24	GO:0004167	ENSG00000099977	IPR014347
ENSG00000099977	ENSP00000381386	54.24	GO:0033981	ENSG00000099977	IPR014347

Genome Wide Association Studies: databases



https://www.ncbi.nlm.nih.gov/snp/

NCBI main repository of information about polymorphisms and their penetrance, possible molecular effects and functional role.

dbGaP
GENOTYPES and PHENOTYPES

www.ncbi.nlm.nih.gov/gap

“The database of Genotypes and Phenotypes (dbGaP) was developed to archive and distribute the results of studies that have investigated the interaction of genotype and phenotype”.

GWAS Catalog
The NHGRI-EBI Catalog of published genome-wide association studies

www.ebi.ac.uk/gwas/

The catalog of Published Genome-Wide Association Studies of the NIH National Human Genome Research Institute (NHGRI)

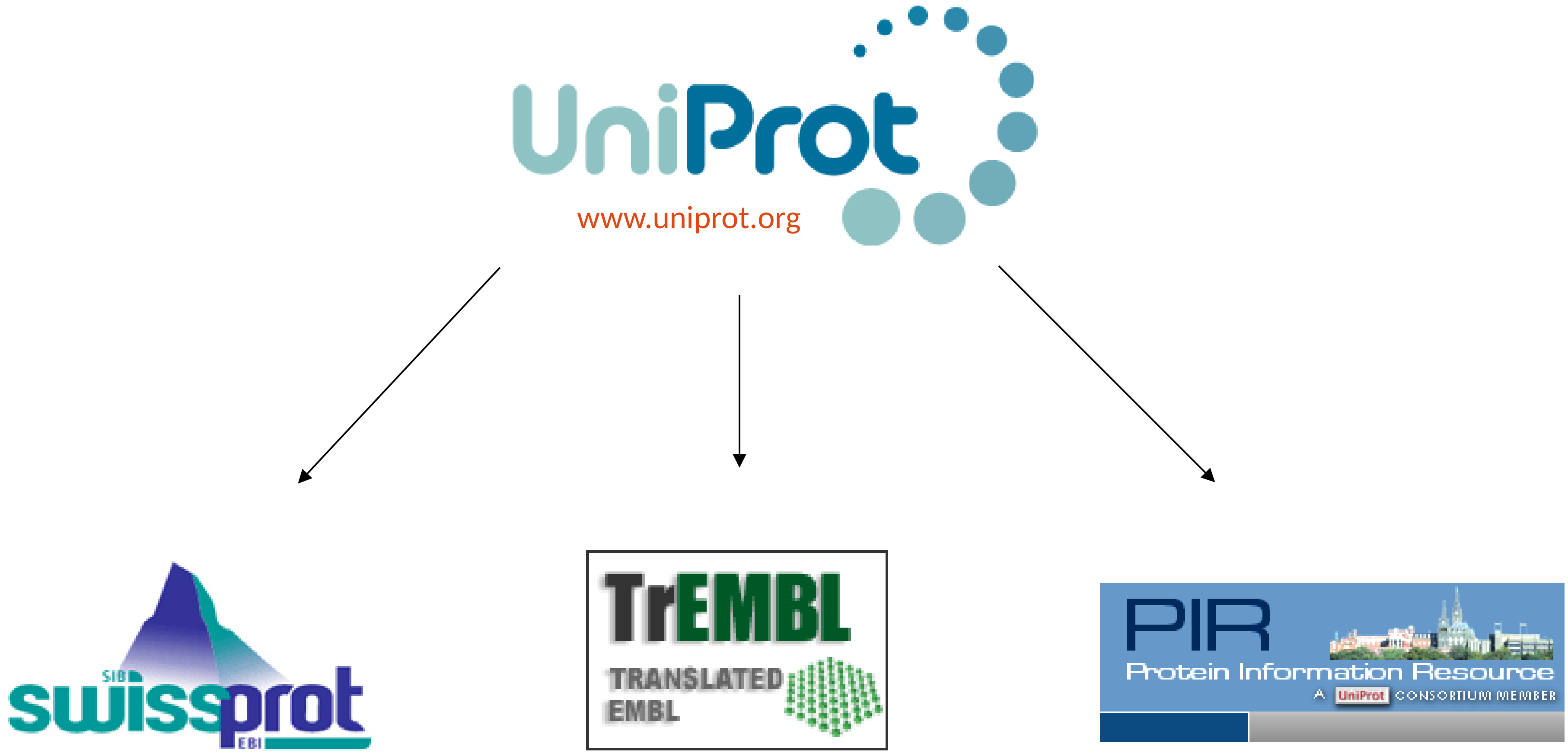
GWAS CENTRAL

www.gwascentral.org

GWAS Central provides a centralized compilation of summary level findings from genetic association studies, both large and small.

matteo.ramazzotti@unifi.it

UniProt: the reference protein database



Highly accurate and detailed, with manual annotations about function, structure, PTM and much more.

Contains *in silico* translation of each coding sequence in the EMBL primary database. Huge but inaccurate.

The direct descendent of the Dayhoff Protein Atlas, the best protein database available.

matteo.ramazzotti@unifi.it

UniProt: the reference protein database

Click to go to the browser

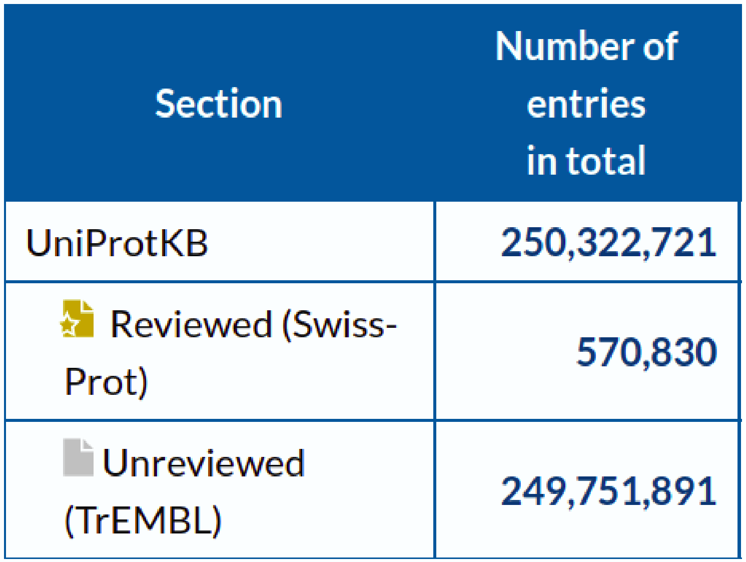
UniProt

www.uniprot.org

matteo.ramazzotti@unifi.it

UniProt: the reference protein database

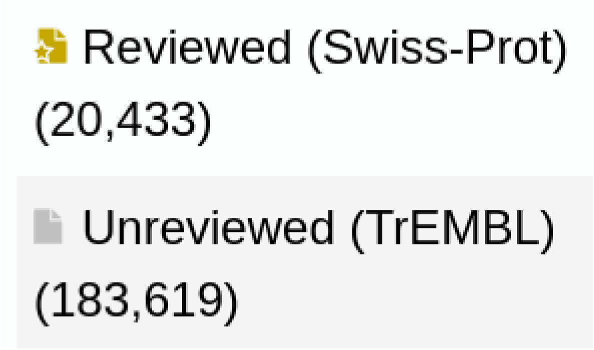
Total



Section	Number of entries in total
UniProtKB	250,322,721
Reviewed (Swiss-Prot)	570,830
Unreviewed (TrEMBL)	249,751,891

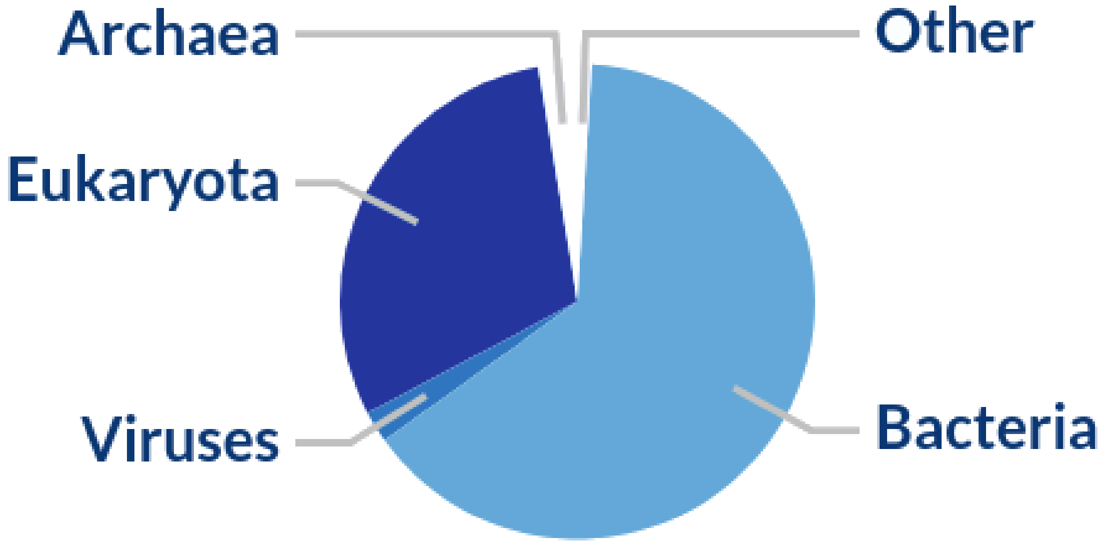
Human

UniProtKB 204,052 results



Reviewed (Swiss-Prot) (20,433)

Unreviewed (TrEMBL) (183,619)



UniProt: the reference protein database

UniProtKB collects and join Swiss-Prot and TrEMBL and is an interface and a data bank that integrate annotations coming from a huge variety of sources in an integrated framework.

The possibility of having a common interface to look at proteins makes UniProtKB the reference source for protein records.

Provide clustered sets of sequences in order to obtain complete coverage of the sequence space at several resolutions

UniRef 100: identical sequences and sub-fragments are merged.

UniRef 90: clustered if they share >90% identity.

UniRef 50: clustered if they share >50% identity.

Is a collection organized by sequence of proteins coming from databases all over the world.

1 sequence => many IDs and annotations.

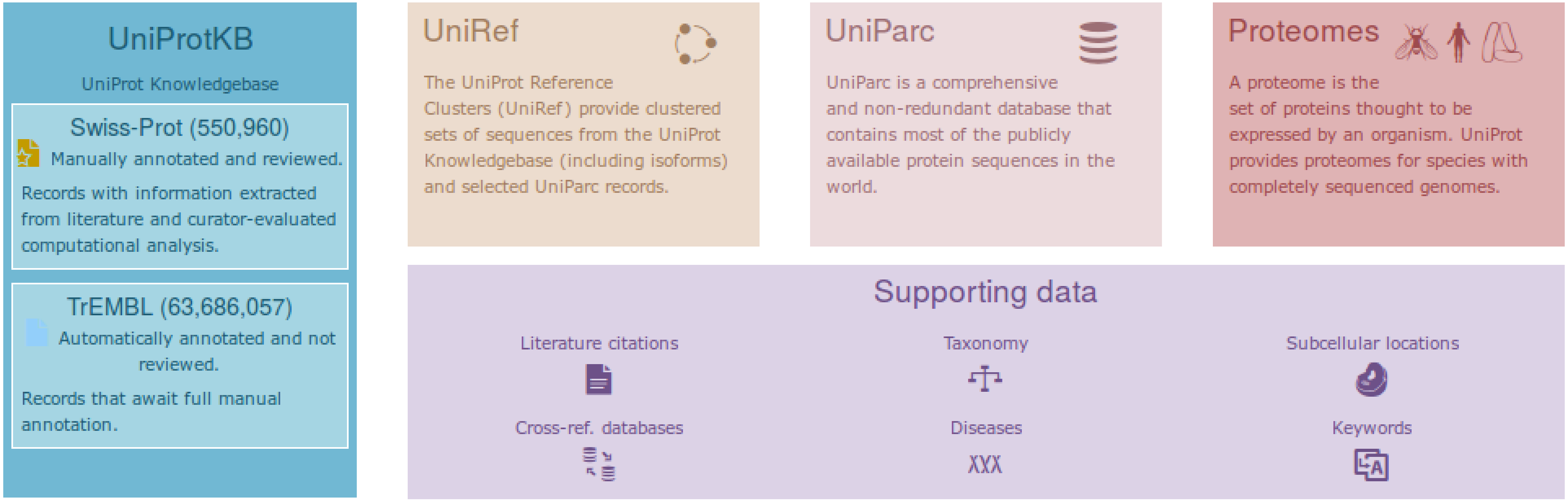
The entry is classified with a stable and unique identifier (UPI), that is never removed or changed.

Searching in uniparc is like searching in many DBs.

A proteome is the set of proteins thought to be expressed by an organism.

The set of proteins that constitute the proteome is classified using the proteome identifier (UPID), consisting in of the characters ‘UP’ followed by 9 digits.

UPIDs are stable across releases.



UniProtKB

UniProt Knowledgebase

Swiss-Prot (550,960)

Manually annotated and reviewed.

Records with information extracted from literature and curator-evaluated computational analysis.

TrEMBL (63,686,057)

Automatically annotated and not reviewed.

Records that await full manual annotation.

UniRef

The UniProt Reference Clusters (UniRef) provide clustered sets of sequences from the UniProt Knowledgebase (including isoforms) and selected UniParc records.

UniParc

UniParc is a comprehensive and non-redundant database that contains most of the publicly available protein sequences in the world.

Proteomes

A proteome is the set of proteins thought to be expressed by an organism. UniProt provides proteomes for species with completely sequenced genomes.

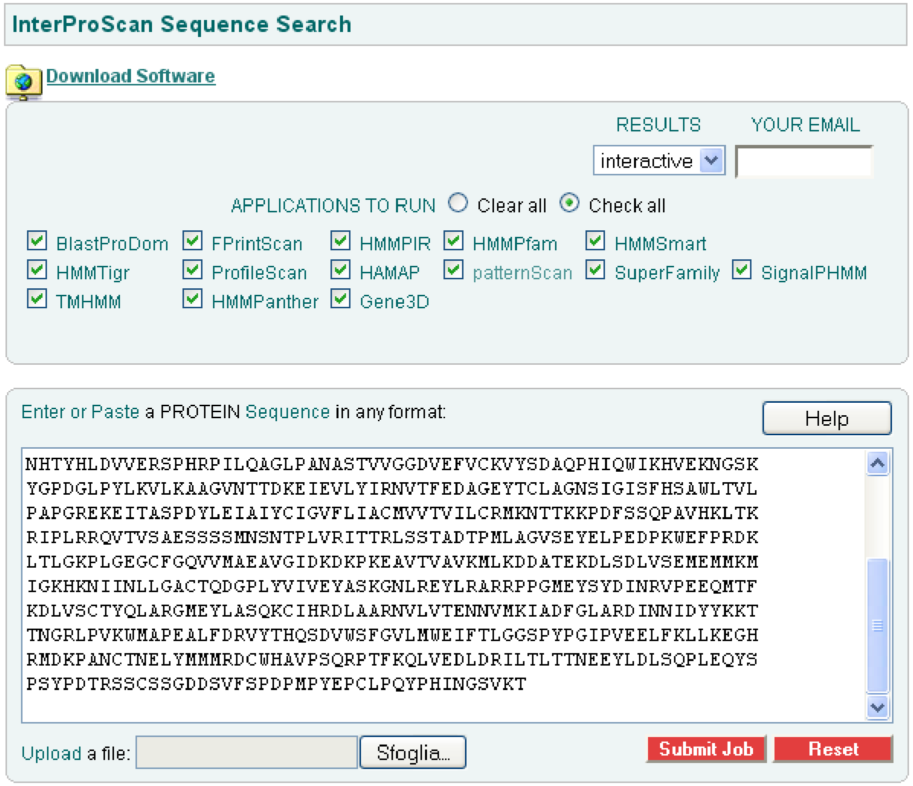
Supporting data

matteo.ramazzotti@unifi.it

InterPro: a consensus signature engine

The concept of "signature" is of paramaount importance in bioinformatics: a pattern of residues that is found in a protein can describe many of its properties and may lead to automatic functional / structural classification.

InterPro allows to automatically scan a protein against several signature data banks, searchign for signatures, domain (functional or structural) or patterns.



InterProScan Sequence Search

Download Software

RESULTS: interactive

YOUR EMAIL:

APPLICATIONS TO RUN ○ Clear all ● Check all

☑ BlastProDom ☑ FPrintScan ☑ HMMPIR ☑ HMMPfam ☑ HMMSmart
☑ HMMTigr ☑ ProfileScan ☑ HAMAP ☑ patternScan ☑ SuperFamily ☑ SignalPHMM
☑ TMHMM ☑ HMMPanther ☑ Gene3D

Enter or Paste a PROTEIN Sequence in any format:

```
NHTYHLDVVERSPHRPILQAGLPANASTVVGGDVEFVCKVYSDAQPHIQWIKHVEKNGSK
YGPDGLPYLKVLKAAGVNTTDKEIEVLYIRNVTFEDAGEYTCLAGNSIGISFHSAWLTVL
PAPGREKEITASPDYLEIAIYCIGVFLIACMVVTVILCRMKNTTKKPDFSSQPAVHKLTK
RIPLRRQVTVSAESSSSSMNSNTPLVRITTRLSSTADTPMLAGVSEYELPEDPKWEFPRDK
LTLGKPLGEGCFGQVVMAEAVGIDKDKPKEAVTVAVKMLKDDATEKDLSDLVSEMEMMKM
IGKHKNIINLLGACTQDGPLYVIVEYASKGNLREYLRARRPPGMEYSYDINRVPEEQMTF
KDLVSCTYQLARGMEYLASQKCIHRDLAARNVLVTENNVMKIADFGLARDINNIDYYKKT
TNGRLPVKWMAPEALFDRVYTHQSDVWSFGVLMWEIFTLGGSPYPGIPVEELFKLLKEGH
RMDKPANCTNELYMMMRDCWHAVPSQRPTFKQLVEDLDRILTLTTNEEYLDLSQPLEQYS
PSYPDTRSSCSSGDDSVFSPDPMPYEPCLPQYPHINGSVKT
```

Upload a file:

InterPro: many hits make a result...

SEQUENCE: FGFR2_HUMAN CRC64: 6CD5001C960ED82F LENGTH: 821 aa

Entry	Description	Signature	Name
InterPro IPR000719 Domain	Protein kinase, catalytic domain	PS50011	PROTEIN_KINASE_DOM
InterPro IPR001245 Domain	Serine-threonine/tyrosine-protein kinase	PF07714	Pkinase_Tyr
InterPro IPR003598 Domain	Immunoglobulin subtype 2	SM00408	IGc2
InterPro IPR007110 Domain	Immunoglobulin-like	PS50835	IG_LIKE
InterPro IPR011009 Domain	Protein kinase-like domain	SSF56112	Protein kinase-like (PK-like)
InterPro IPR013783 Domain	Immunoglobulin-like fold	G3DSA:2.60.40.10	no description
InterPro IPR020635 Region	Tyrosine-protein kinase	PTHR23256	TYROSINE PROTEIN KINASE
noIPR unintegrated	unintegrated	SignalP	signal-peptide
		tmhmm	transmembrane_regions

PS: ProSite

PF: Pfam

SM: SMART

PS: ProSite

SSF: SuperFamily

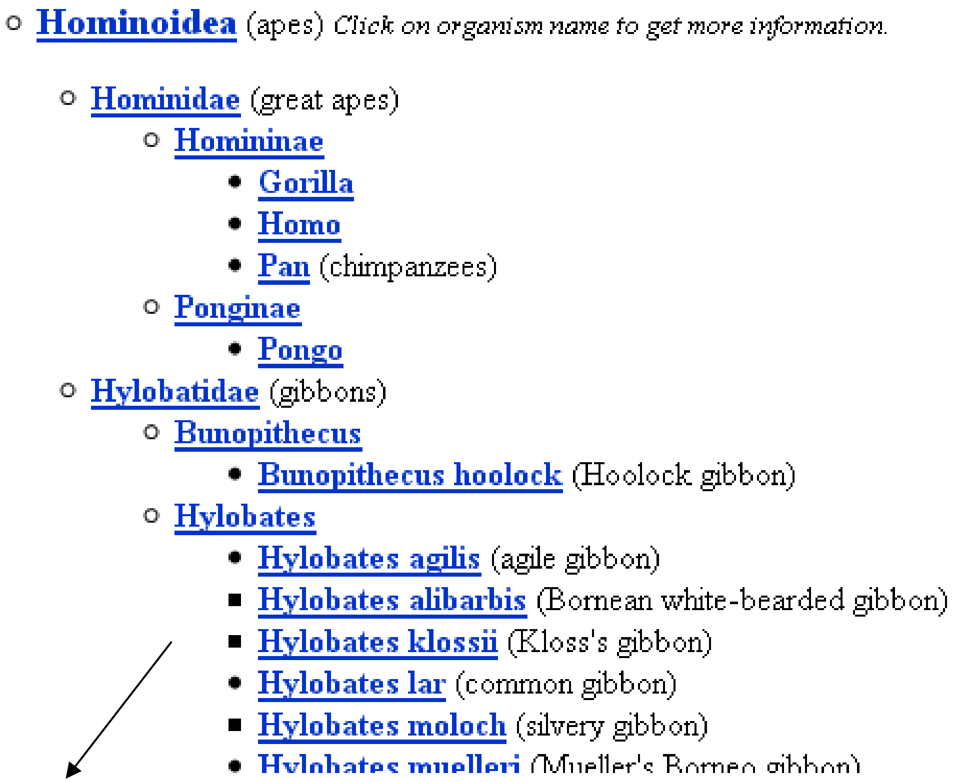
G3DSA: Gene 3D

PTHR: Panther

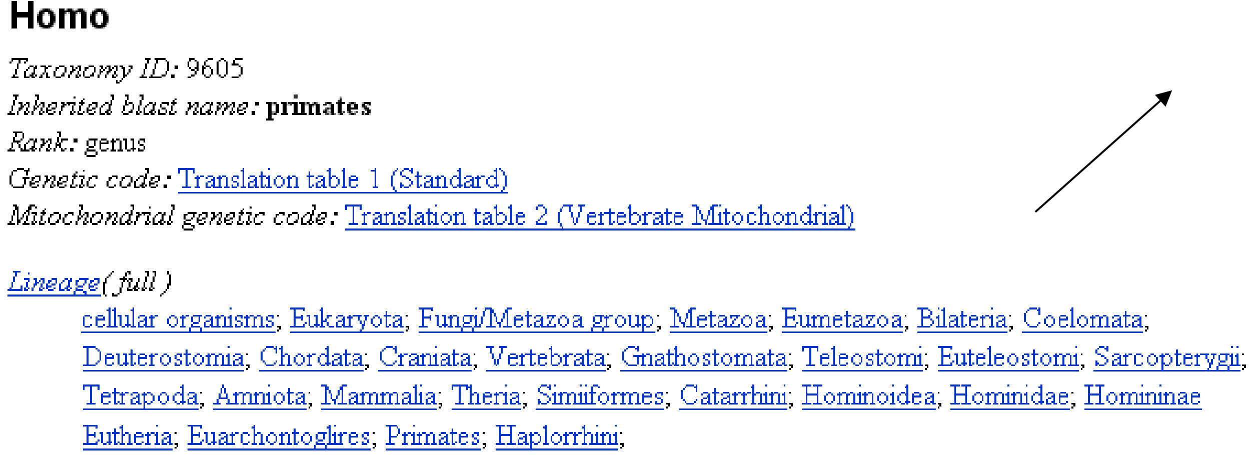
Topology predictors

It is a 820 AA residue long protein with an Ig domain (so, extracellular), with a transmembrane region and a protein kinase domain, with several phoshporylation sites and a signal peptide typic of ER produced proteins: definitely a receptor.

NCBI Taxonomy – collector, indexer and accessor



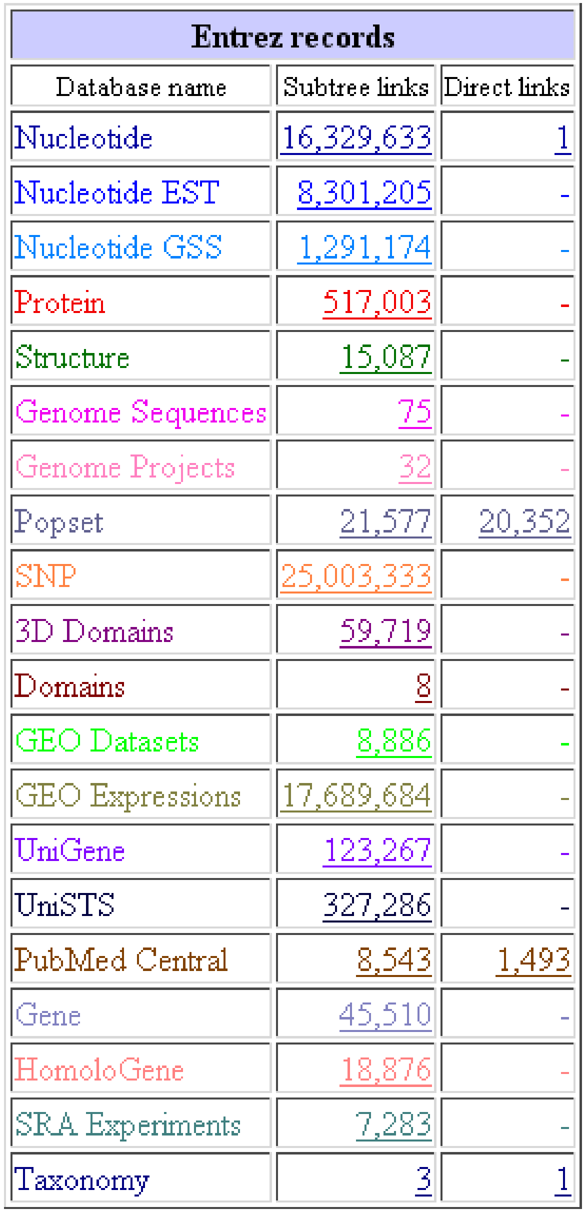
- **Hominoidea** (apes) *Click on organism name to get more information.*
  - **Hominidae** (great apes)
    - **Homininae**
      - **Gorilla**
      - **Homo**
      - **Pan** (chimpanzees)
    - **Ponginae**
      - **Pongo**
  - **Hylobatidae** (gibbons)
    - **Bunopithecus**
      - **Bunopithecus hoolock** (Hoolock gibbon)
    - **Hylobates**
      - **Hylobates agilis** (agile gibbon)
      - **Hylobates alibarbis** (Bornean white-bearded gibbon)
      - **Hylobates klossii** (Kloss's gibbon)
      - **Hylobates lar** (common gibbon)
      - **Hylobates moloch** (silvery gibbon)
      - **Hylobates muelleri** (Mueller's Borneo gibbon)



Homo

Taxonomy ID: 9605
Inherited blast name: **primates**
Rank: genus
Genetic code: Translation table 1 (Standard)
Mitochondrial genetic code: Translation table 2 (Vertebrate Mitochondrial)

Lineage (*full*)
cellular organisms; Eukaryota; Fungi/Metazoa group; Metazoa; Eumetazoa; Bilateria; Coelomata; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Tetrapoda; Amniota; Mammalia; Theria; Simiiformes; Catarrhini; Hominoidea; Hominidae; Homininae Eutheria; Euarchontoglires; Primates; Haplorrhini;

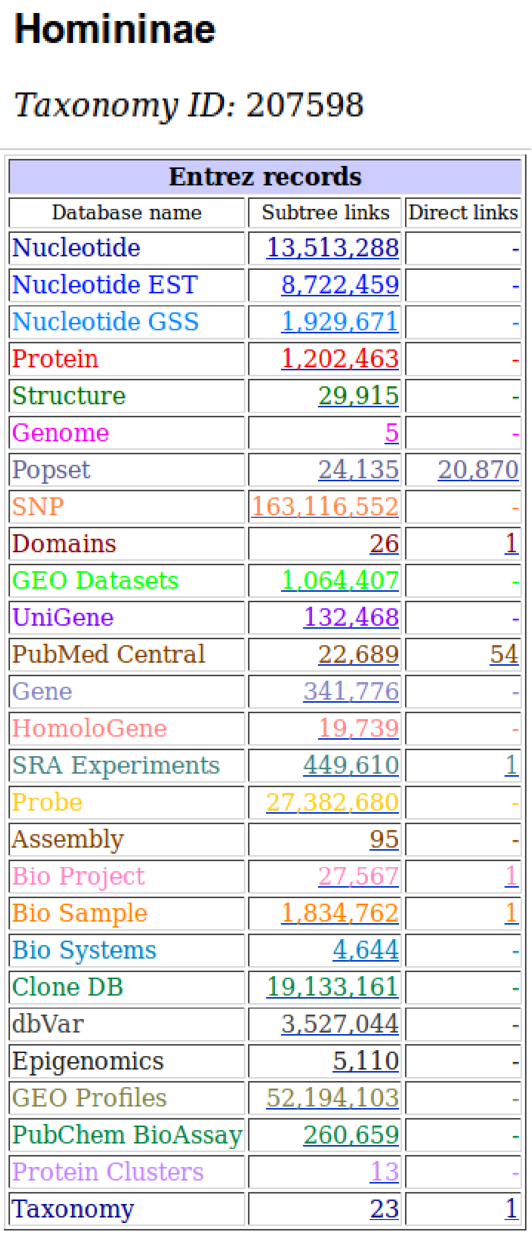


Entrez records

Database name	Subtree links	Direct links
Nucleotide	16,329,633	1
Nucleotide EST	8,301,205	-
Nucleotide GSS	1,291,174	-
Protein	517,003	-
Structure	15,087	-
Genome Sequences	75	-
Genome Projects	32	-
Popset	21,577	20,352
SNP	25,003,333	-
3D Domains	59,719	-
Domains	8	-
GEO Datasets	8,886	-
GEO Expressions	17,689,684	-
UniGene	123,267	-
UniSTS	327,286	-
PubMed Central	8,543	1,493
Gene	45,510	-
HomoloGene	18,876	-
SRA Experiments	7,283	-
Taxonomy	3	1

matteo.ramazzotti@unifi.it

NCBI Taxonomy – collector, indexer and accessor

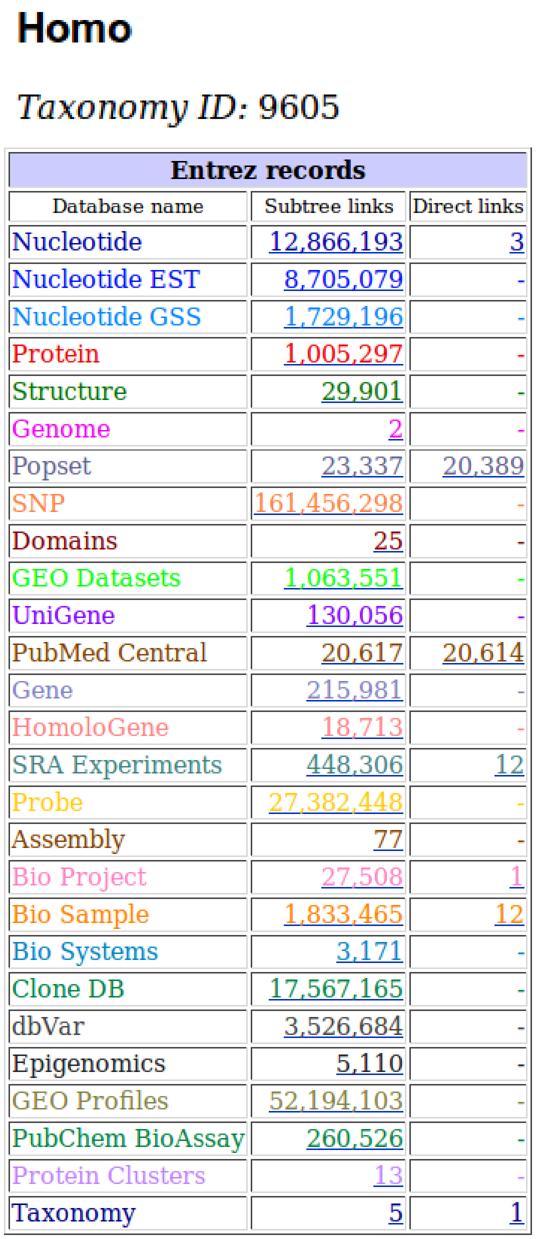


Homininae

Taxonomy ID: 207598

Entrez records

Database name	Subtree links	Direct links
Nucleotide	13,513,288	-
Nucleotide EST	8,722,459	-
Nucleotide GSS	1,929,671	-
Protein	1,202,463	-
Structure	29,915	-
Genome	5	-
Popset	24,135	20,870
SNP	163,116,552	-
Domains	26	1
GEO Datasets	1,064,407	-
UniGene	132,468	-
PubMed Central	22,689	54
Gene	341,776	-
HomoloGene	19,739	-
SRA Experiments	449,610	1
Probe	27,382,680	-
Assembly	95	-
Bio Project	27,567	1
Bio Sample	1,834,762	1
Bio Systems	4,644	-
Clone DB	19,133,161	-
dbVar	3,527,044	-
Epigenomics	5,110	-
GEO Profiles	52,194,103	-
PubChem BioAssay	260,659	-
Protein Clusters	13	-
Taxonomy	23	1

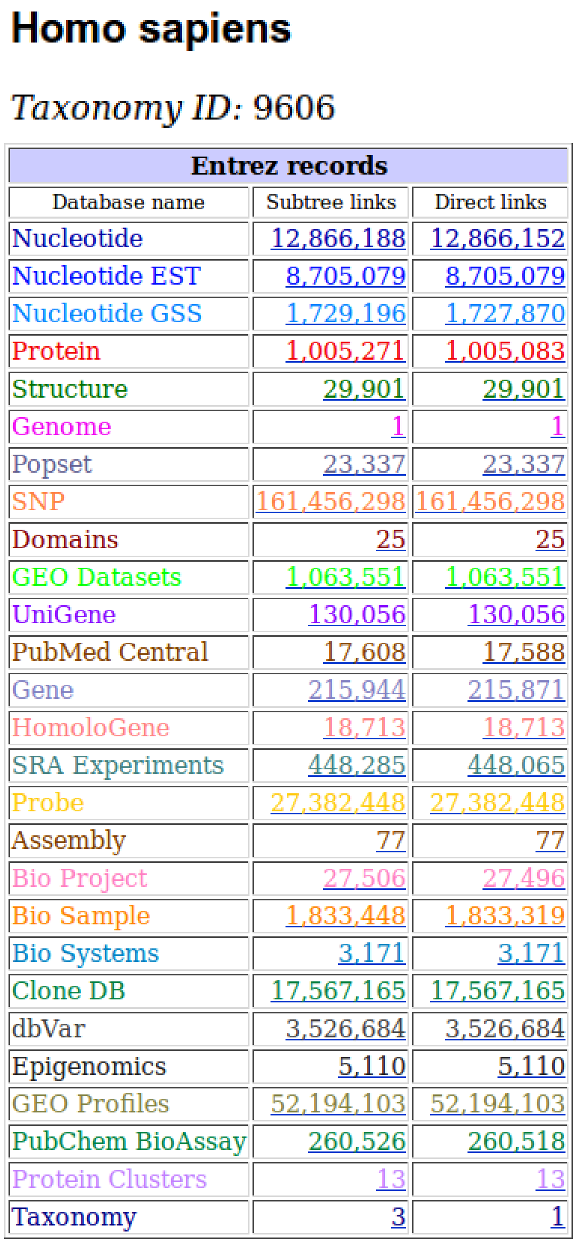


Homo

Taxonomy ID: 9605

Entrez records

Database name	Subtree links	Direct links
Nucleotide	12,866,193	3
Nucleotide EST	8,705,079	-
Nucleotide GSS	1,729,196	-
Protein	1,005,297	-
Structure	29,901	-
Genome	2	-
Popset	23,337	20,389
SNP	161,456,298	-
Domains	25	-
GEO Datasets	1,063,551	-
UniGene	130,056	-
PubMed Central	20,617	20,614
Gene	215,981	-
HomoloGene	18,713	-
SRA Experiments	448,306	12
Probe	27,382,448	-
Assembly	77	-
Bio Project	27,508	1
Bio Sample	1,833,465	12
Bio Systems	3,171	-
Clone DB	17,567,165	-
dbVar	3,526,684	-
Epigenomics	5,110	-
GEO Profiles	52,194,103	-
PubChem BioAssay	260,526	-
Protein Clusters	13	-
Taxonomy	5	1



Homo sapiens

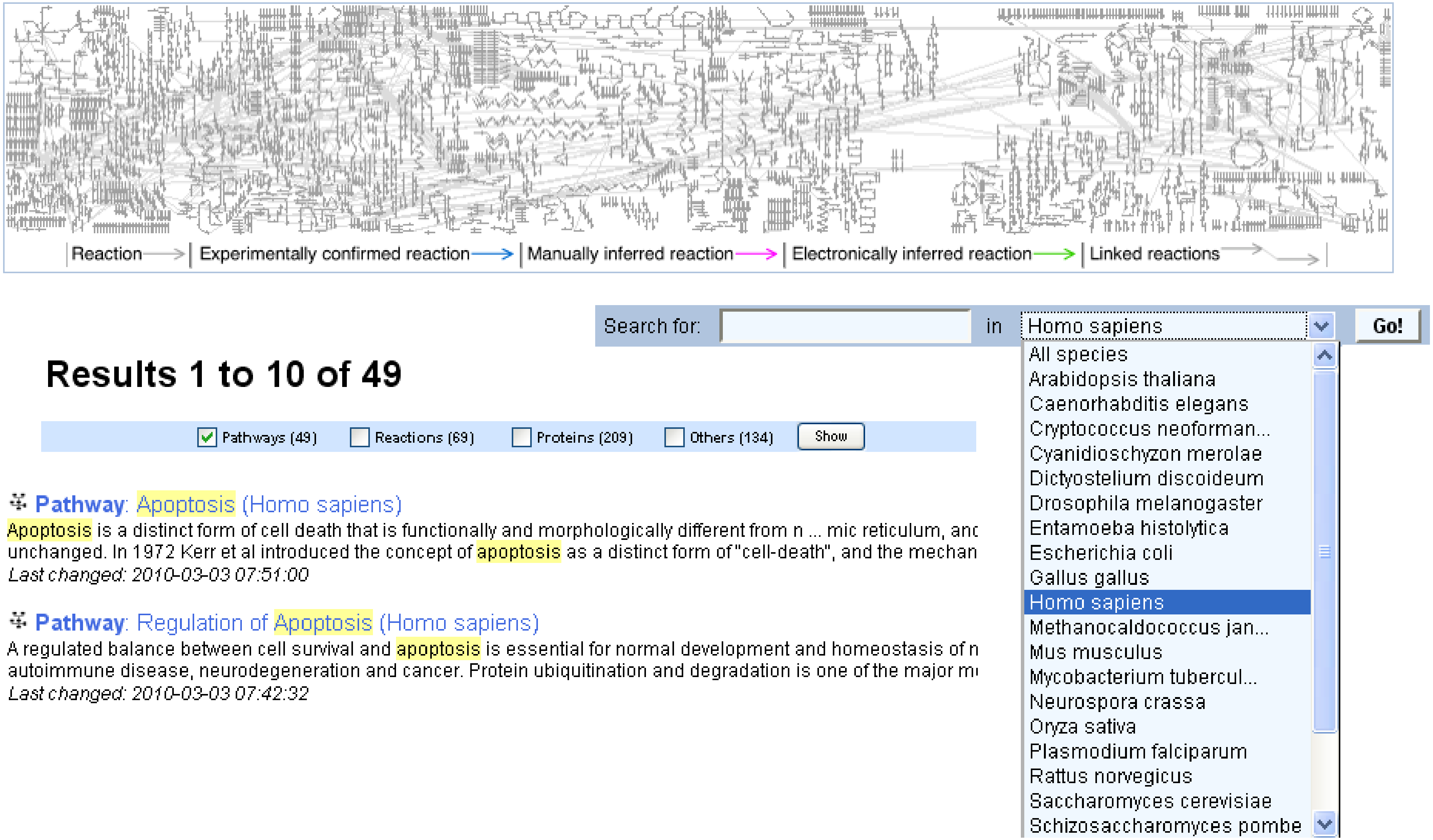
Taxonomy ID: 9606

Entrez records

Database name	Subtree links	Direct links
Nucleotide	12,866,188	12,866,152
Nucleotide EST	8,705,079	8,705,079
Nucleotide GSS	1,729,196	1,727,870
Protein	1,005,271	1,005,083
Structure	29,901	29,901
Genome	1	1
Popset	23,337	23,337
SNP	161,456,298	161,456,298
Domains	25	25
GEO Datasets	1,063,551	1,063,551
UniGene	130,056	130,056
PubMed Central	17,608	17,588
Gene	215,944	215,871
HomoloGene	18,713	18,713
SRA Experiments	448,285	448,065
Probe	27,382,448	27,382,448
Assembly	77	77
Bio Project	27,506	27,496
Bio Sample	1,833,448	1,833,319
Bio Systems	3,171	3,171
Clone DB	17,567,165	17,567,165
dbVar	3,526,684	3,526,684
Epigenomics	5,110	5,110
GEO Profiles	52,194,103	52,194,103
PubChem BioAssay	260,526	260,518
Protein Clusters	13	13
Taxonomy	3	1

matteo.ramazzotti@unifi.it

The Reactome pathway database



Results 1 to 10 of 49

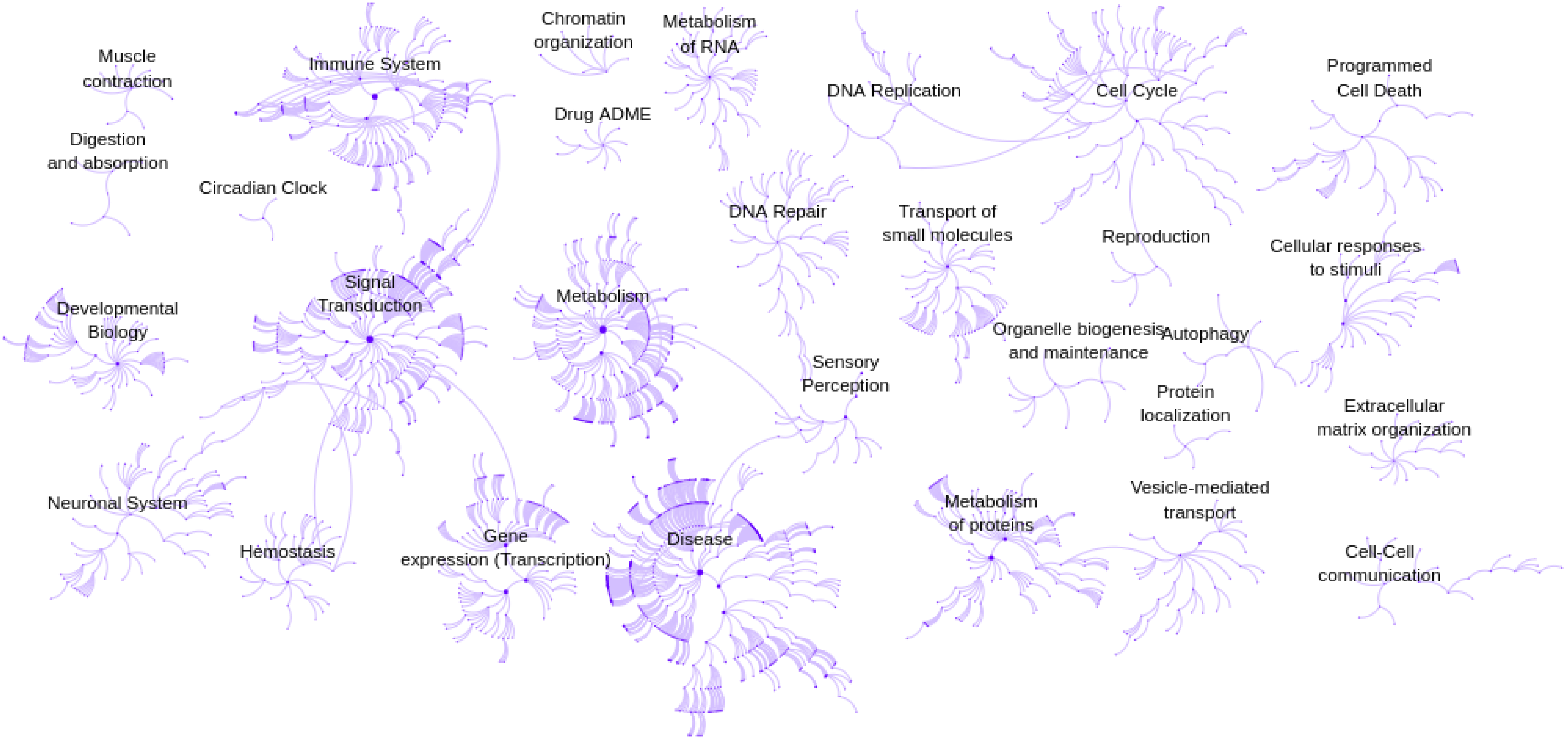
☑ Pathways (49) ☐ Reactions (69) ☐ Proteins (209) ☐ Others (134) Show

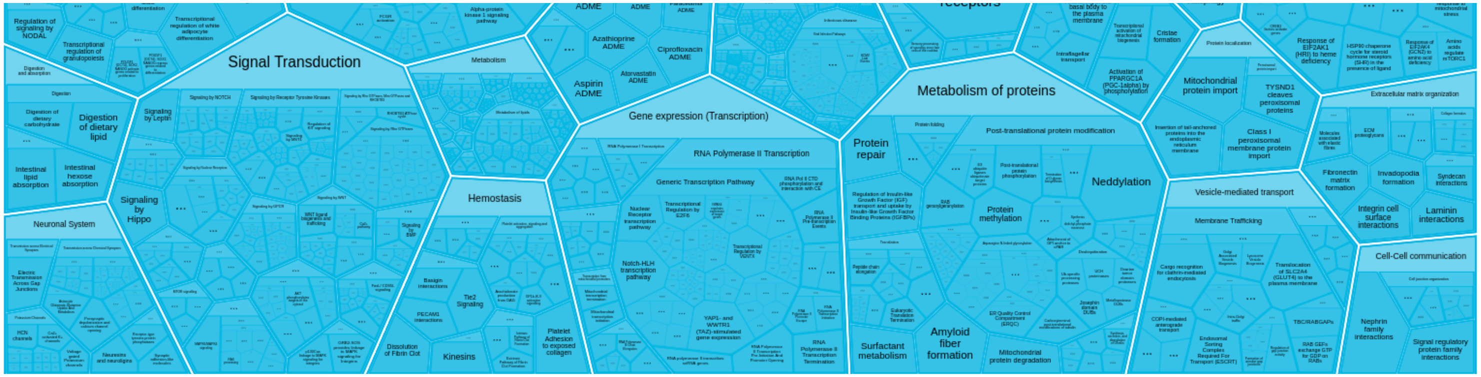
Pathway: Apoptosis (Homo sapiens)
Apoptosis is a distinct form of cell death that is functionally and morphologically different from n ... mic reticulum, anc unchanged. In 1972 Kerr et al introduced the concept of apoptosis as a distinct form of "cell-death", and the mechan
Last changed: 2010-03-03 07:51:00

Pathway: Regulation of Apoptosis (Homo sapiens)
A regulated balance between cell survival and apoptosis is essential for normal development and homeostasis of n autoimmune disease, neurodegeneration and cancer. Protein ubiquitination and degradation is one of the major mı
Last changed: 2010-03-03 07:42:32

2,673	15,046	11,392	2,120	1,046	37,933
Human Pathways	Reactions	Proteins	Small Molecules	Drugs	Literature References

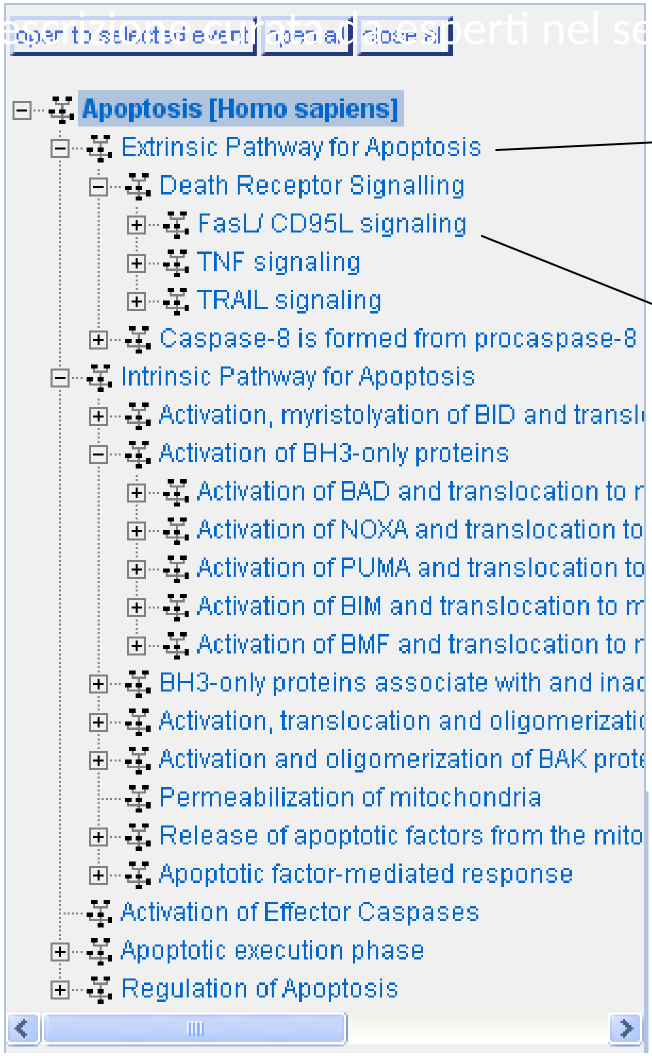
The Reactome pathway database





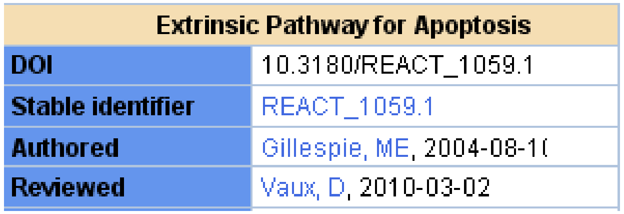
matteo.ramazzotti@unifi.it

The Reactome pathway database



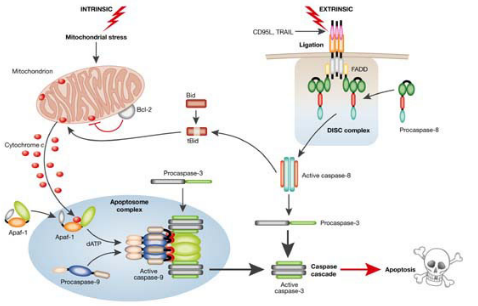
escrizione curata da esperti nel se

- Apoptosis [Homo sapiens]
  - Extrinsic Pathway for Apoptosis
    - Death Receptor Signalling
      - FasL/ CD95L signaling
      - TNF signaling
      - TRAIL signaling
    - Caspase-8 is formed from procaspase-8
  - Intrinsic Pathway for Apoptosis
    - Activation, myristolyation of BID and transl
    - Activation of BH3-only proteins
      - Activation of BAD and translocation to r
      - Activation of NOXA and translocation to
      - Activation of PUMA and translocation to
      - Activation of BIM and translocation to m
      - Activation of BMF and translocation to r
    - BH3-only proteins associate with and inac
    - Activation, translocation and oligomerizati
    - Activation and oligomerization of BAK prote
    - Permeabilization of mitochondria
    - Release of apoptotic factors from the mito
    - Apoptotic factor-mediated response
  - Activation of Effector Caspases
  - Apoptotic execution phase
  - Regulation of Apoptosis

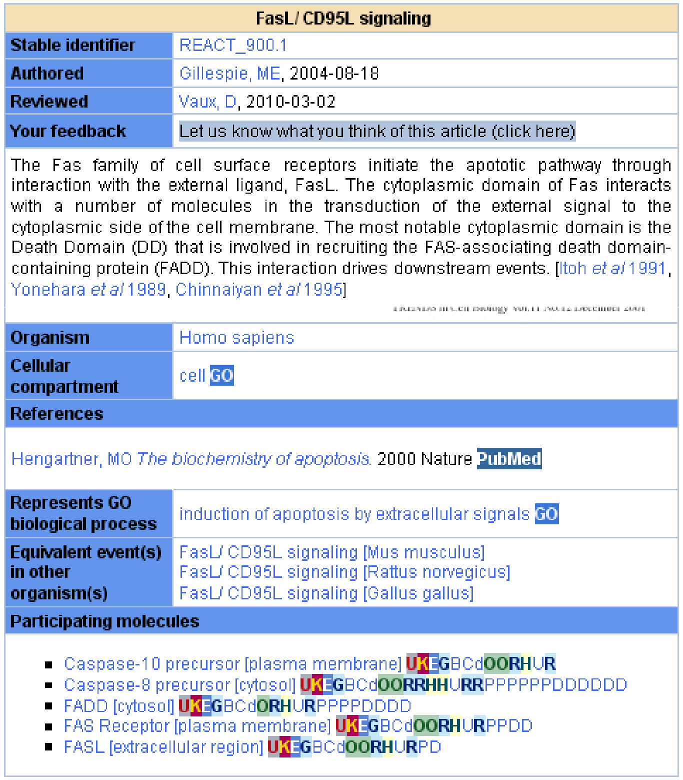


Extrinsic Pathway for Apoptosis

DOI	10.3180/REACT_1059.1
Stable identifier	REACT_1059.1
Authored	Gillespie, ME, 2004-08-1(
Reviewed	Vaux, D, 2010-03-02



Apoptosis and disease: a life or death decision.
EMBO Rep. 2004 Jul;5(7):674-8. Epub 2004 Jun 25.



FasL/ CD95L signaling

Stable identifier	REACT_900.1
Authored	Gillespie, ME, 2004-08-18
Reviewed	Vaux, D, 2010-03-02
Your feedback	Let us know what you think of this article (click here)

The Fas family of cell surface receptors initiate the apototic pathway through interaction with the external ligand, FasL. The cytoplasmic domain of Fas interacts with a number of molecules in the transduction of the external signal to the cytoplasmic side of the cell membrane. The most notable cytoplasmic domain is the Death Domain (DD) that is involved in recruiting the FAS-associating death domain-containing protein (FADD). This interaction drives downstream events. [Itoh *et al* 1991, Yonehara *et al* 1989, Chinnaiyan *et al* 1995]

Organism	Homo sapiens
Cellular compartment	cell GO

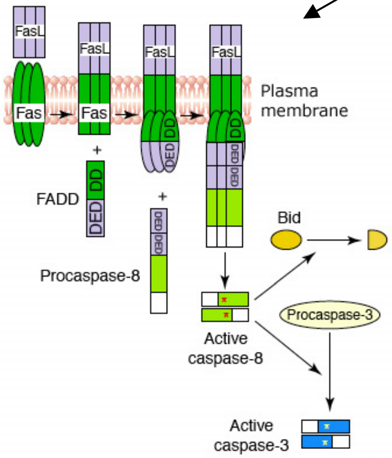
References

Hengartner, MO *The biochemistry of apoptosis.* 2000 Nature PubMed

Represents GO biological process	induction of apoptosis by extracellular signals GO
Equivalent event(s) in other organism(s)	FasL/ CD95L signaling [Mus musculus] FasL/ CD95L signaling [Rattus norvegicus] FasL/ CD95L signaling [Gallus gallus]

Participating molecules

- Caspase-10 precursor [plasma membrane] UKEGBCdOORHUR
- Caspase-8 precursor [cytosol] UKEGBCdOORRHHURRPPPPPPPDDDDDD
- FADD [cytosol] UKEGBCdORHURPPPPDDDD
- FAS Receptor [plasma membrane] UKEGBCdOORHURPPDD
- FASL [extracellular region] UKEGBCdOORHURPD



TRENDS in Cell Biology Vol.11 No.12 December

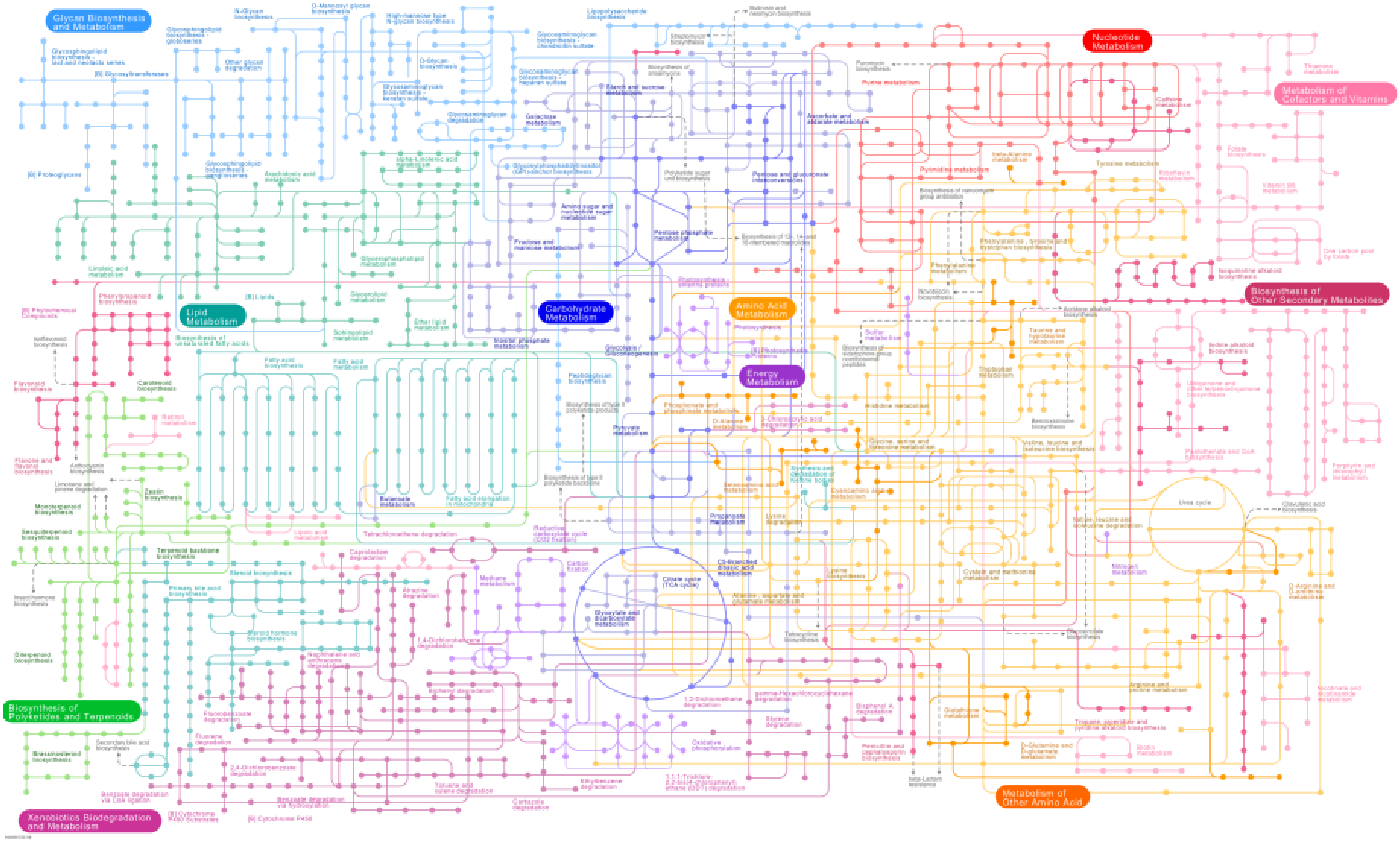
The Reactome pathway database

Click to go to the browser

matteo.ramazzotti@unifi.it

The KEGG pathway database

KEGG Pathways offers the possibility of looking at the complexity of the network in cell signaling or metabolism using predefined sets of images (maps) that are cross-connected through proteins or metabolites. The interactive maps are shared across organisms, so that organism specificity can be fully understood.



The KEGG pathway collections

1. Metabolism
- 1.1 Carbohydrate Metabolism
- 1.2 Energy Metabolism
- 1.3 Lipid Metabolism
- 1.4 Nucleotide Metabolism
- 1.5 Amino Acid Metabolism
- 1.6 Metabolism of Other Amino Acids
- 1.7 Glycan Biosynthesis and Metabolism
- 1.8 Metabolism of Cofactors and Vitamins
- 1.9 Biosynthesis of Polyketides and Terpenoids
- 1.10 Biosynthesis of Other Secondary Metabolites
- 1.11 Xenobiotics Biodegradation and Metabolism
- 1.12 Overview

2. Genetic Information Processing
- 2.1 Transcription
- 2.2 Translation
- 2.3 Folding, Sorting and Degradation
- 2.4 Replication and Repair

3. Environmental Information Processing
- 3.1 Membrane Transport
- 3.2 Signal Transduction
- 3.3 Signaling Molecules and Interaction

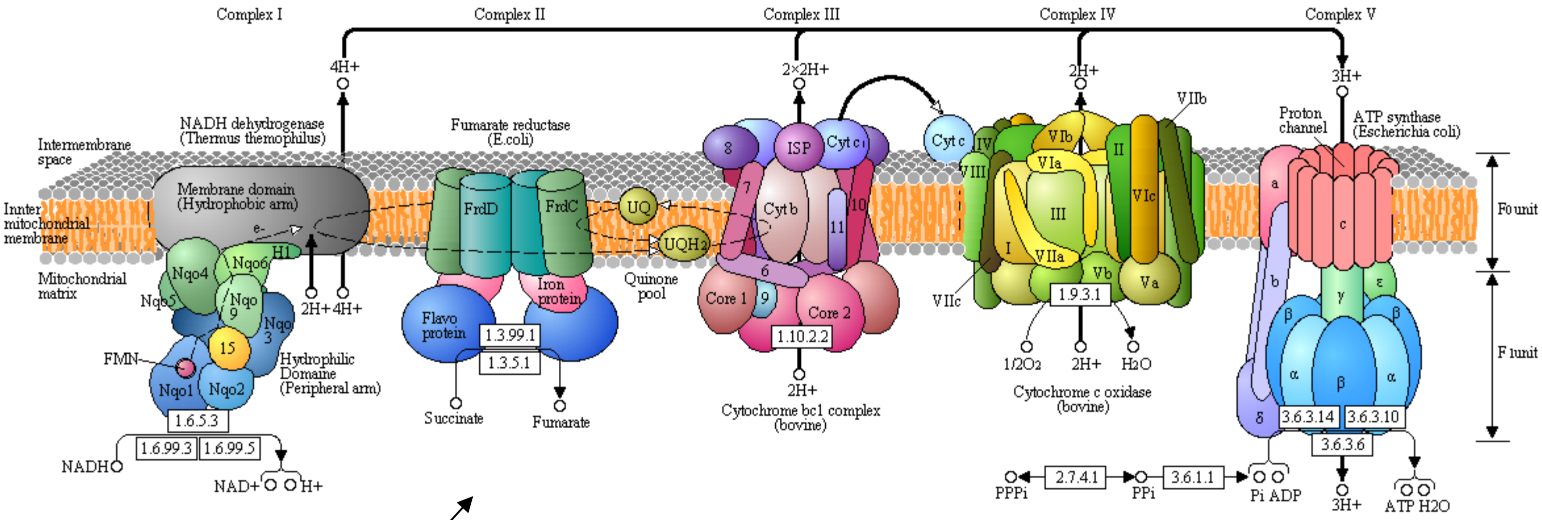
4. Cellular Processes
- 4.1 Transport and Catabolism
- 4.2 Cell Motility
- 4.3 Cell Growth and Death
- 4.4 Cell Communication

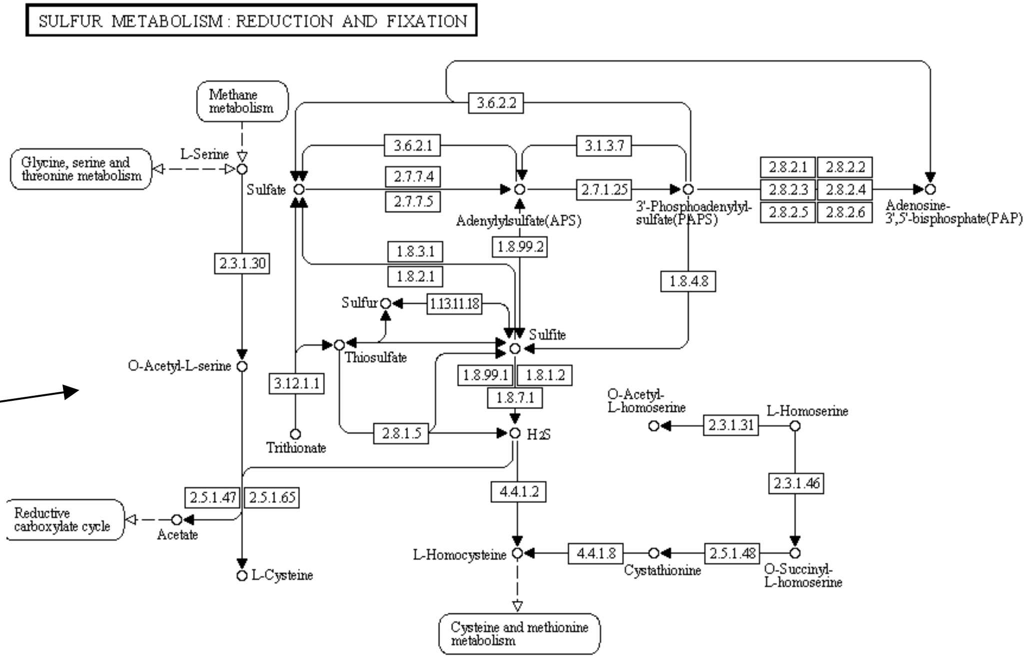
5. Organismal Systems
- 5.1 Immune System
- 5.2 Endocrine System
- 5.3 Circulatory System
- 5.4 Excretory System
- 5.5 Nervous System
- 5.6 Sensory System
- 5.7 Development
- 5.8 Environmental Adaptation

6. Human Diseases
- 6.1 Cancers
- 6.2 Immune System Diseases
- 6.3 Neurodegenerative Diseases
- 6.4 Cardiovascular Diseases
- 6.5 Metabolic Diseases
- 6.6 Infectious Diseases

7. Drug Development
- 7.1 Chronology: Antibiotics
- 7.2 Chronology: Antineoplastics
- 7.3 Chronology: Nervous System Agents
- 7.4 Chronology: Other Drugs
- 7.5 Target Based Classification
- 7.6 Structure Based Classification
- 7.7 Skeleton Based Classification

Schemes and interactive maps and diagrams





1.2 Energy Metabolism

Oxidative phosphorylation

Photosynthesis

Photosynthesis - antenna proteins

Carbon fixation in photosynthetic organisms

Reductive carboxylate cycle in photosynthetic bacteria

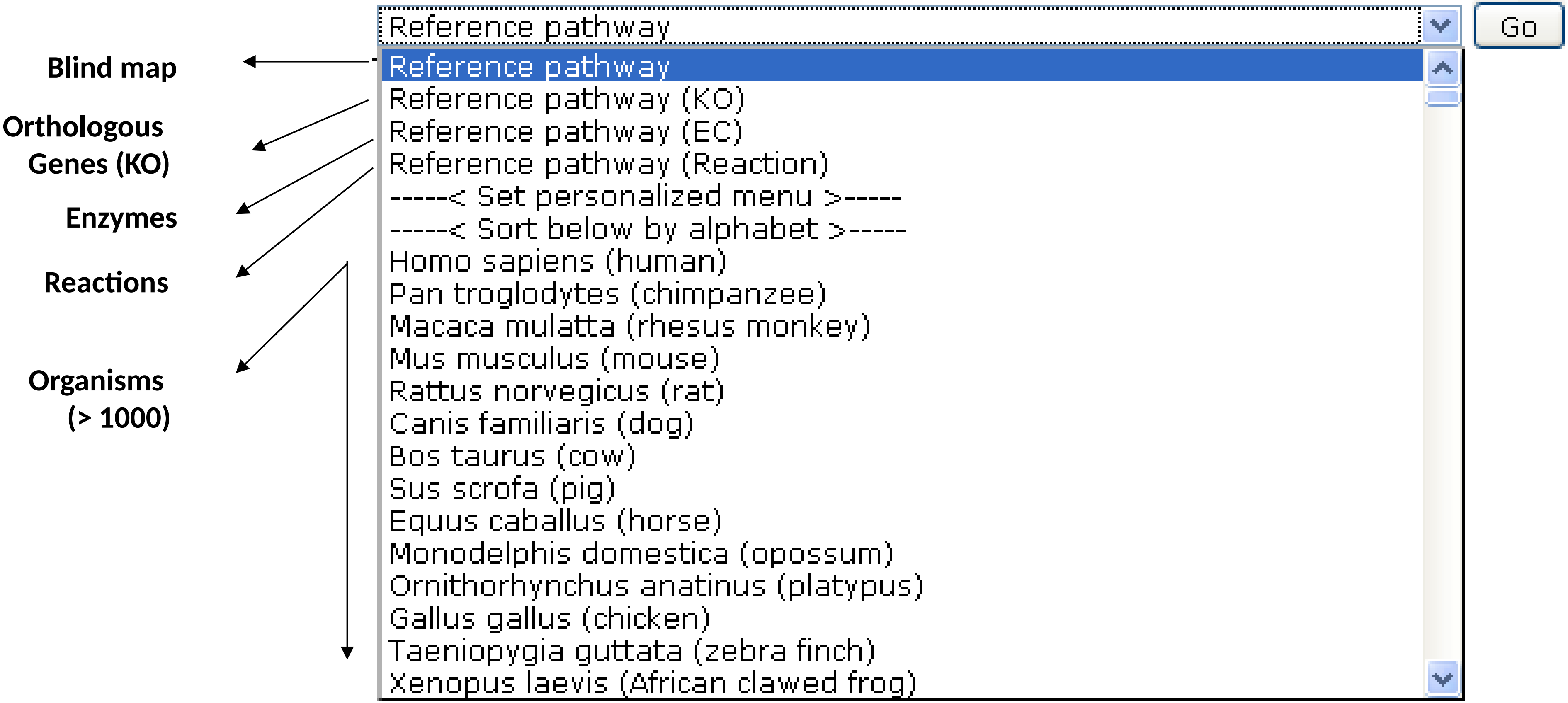
Methane metabolism

Nitrogen metabolism

Sulfur metabolism

matteo.ramazzotti@unifi.it

Reference maps and selections



Blind map

Orthologous Genes (KO)

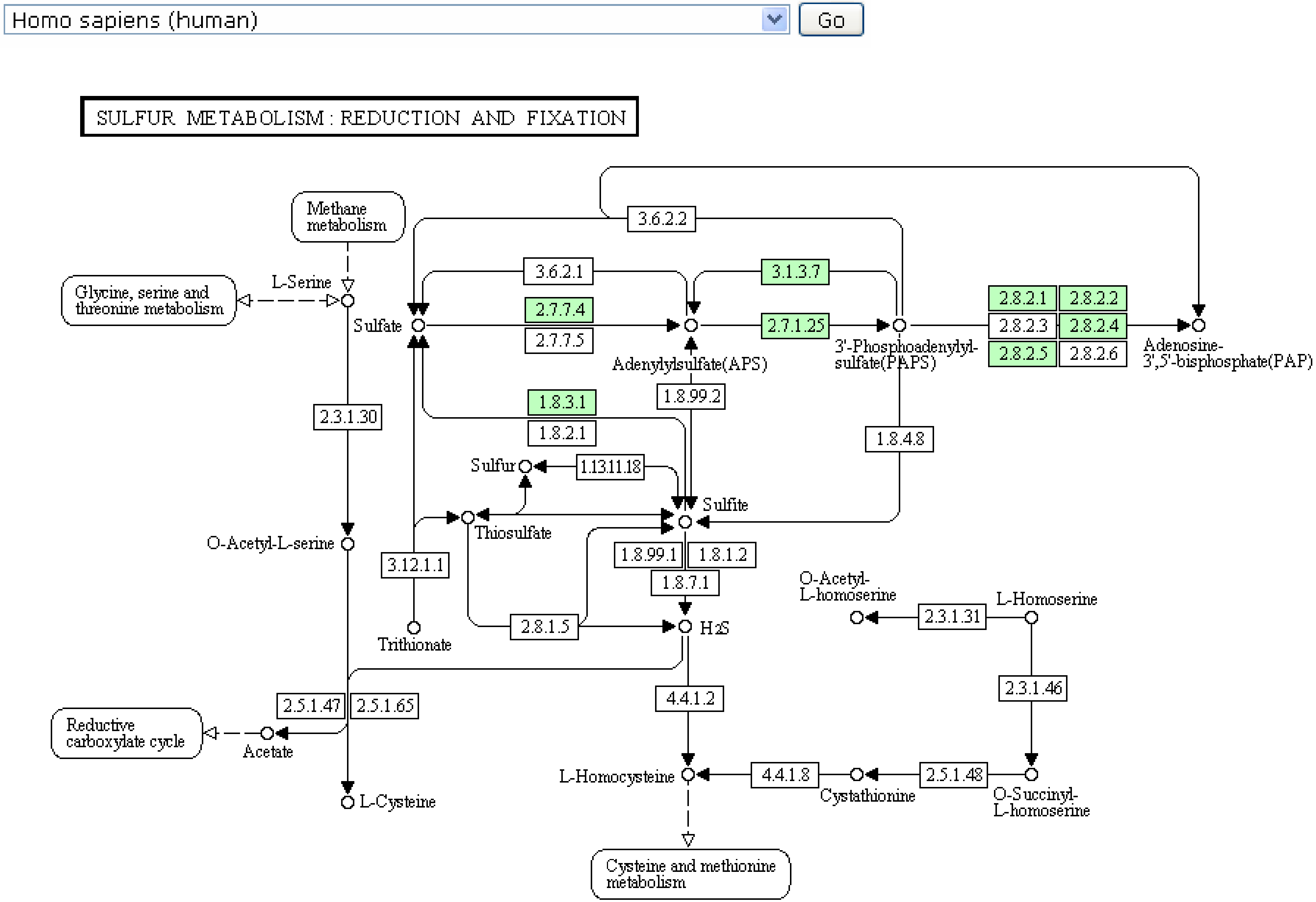
Enzymes

Reactions

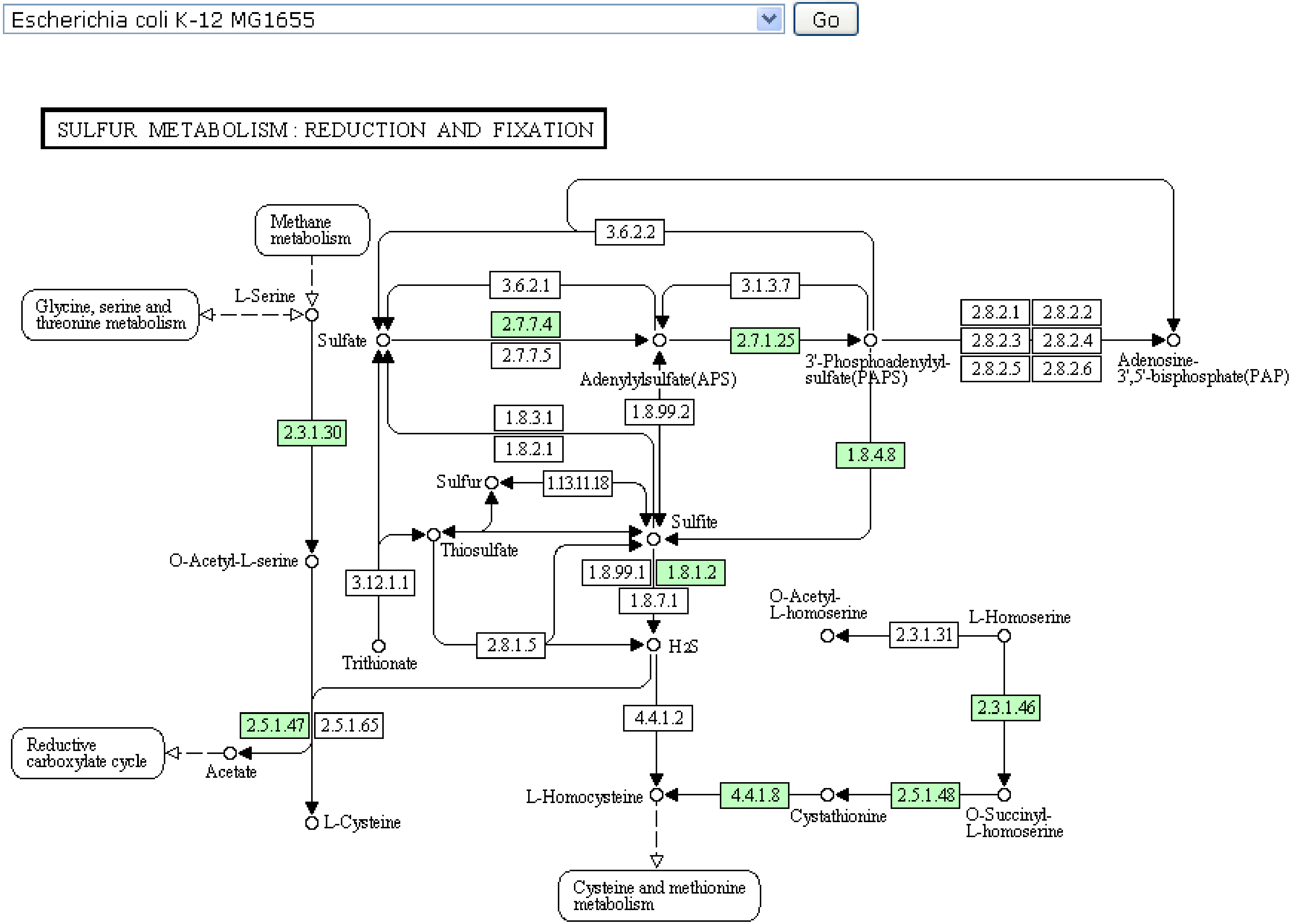
Organisms (> 1000)

matteo.ramazzotti@unifi.it

Reference maps get colors according to selections

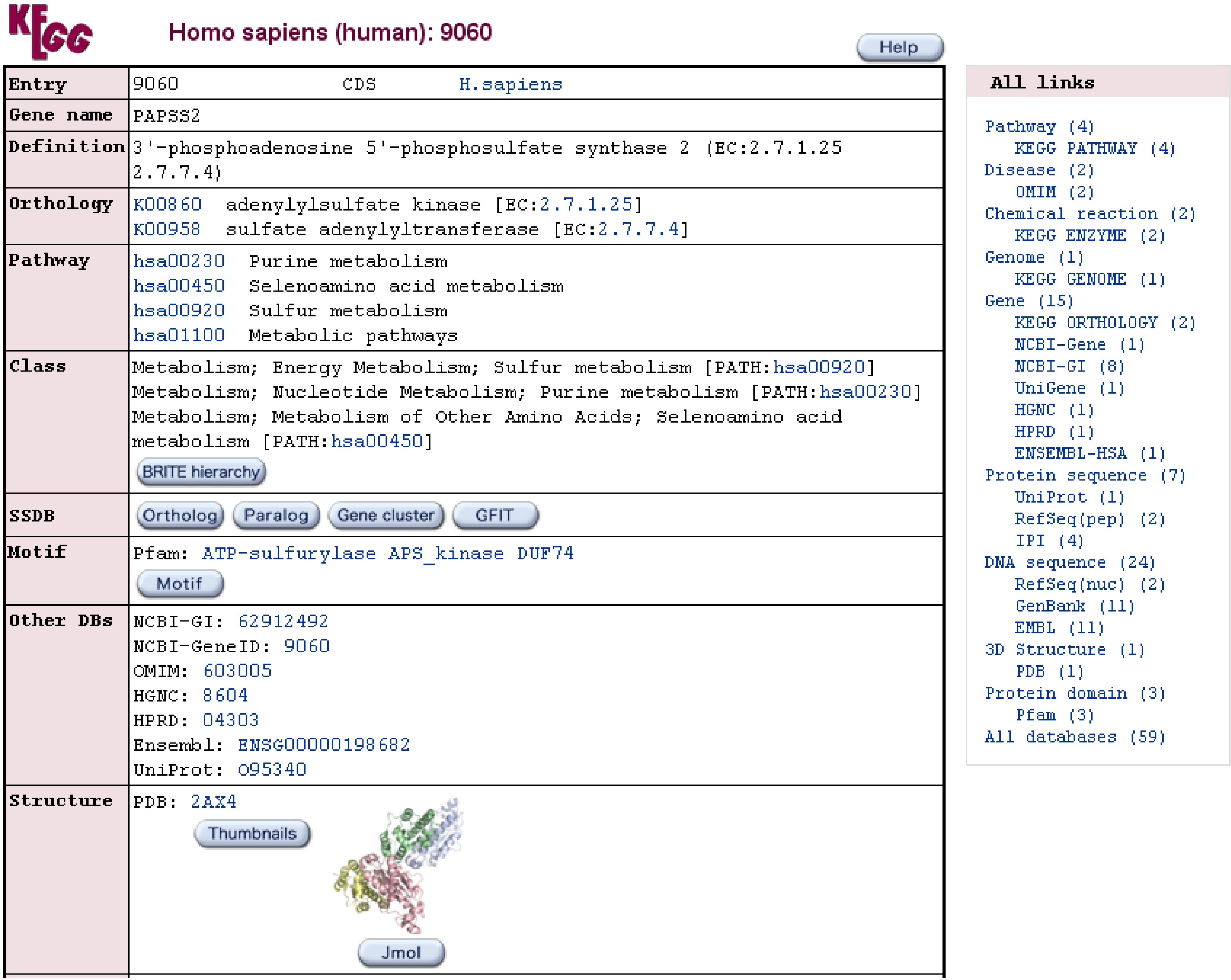


Reference maps get colors according to selections



Escherichia coli K-12 MG1655

Reference maps allow access to cross-annotations



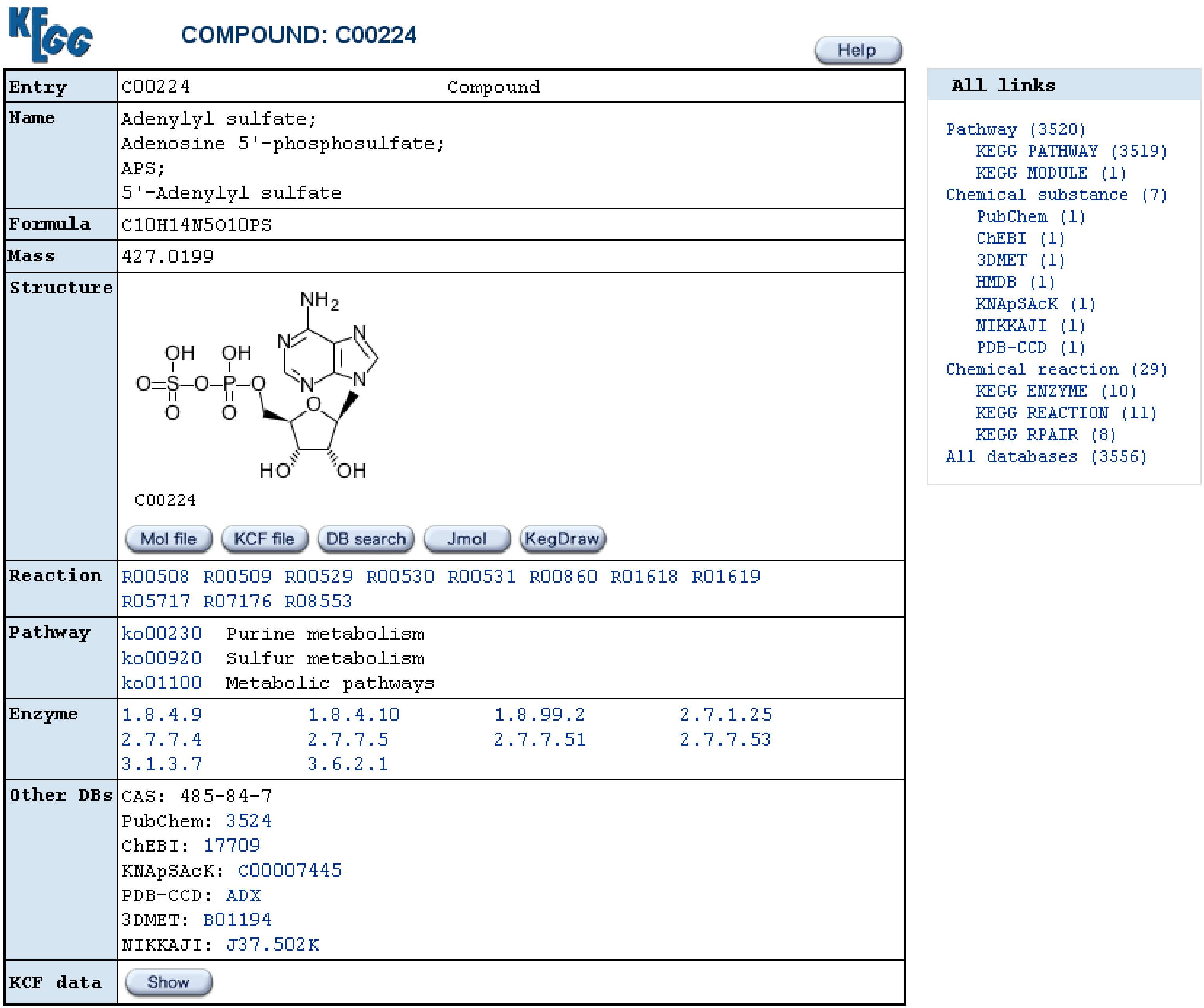
Homo sapiens (human): 9060

Entry	9060 CDS H.sapiens
Gene name	PAPSS2
Definition	3'-phosphoadenosine 5'-phosphosulfate synthase 2 (EC:2.7.1.25 2.7.7.4)
Orthology	K00860 adenylylsulfate kinase [EC:2.7.1.25] K00958 sulfate adenylyltransferase [EC:2.7.7.4]
Pathway	hsa00230 Purine metabolism hsa00450 Selenoamino acid metabolism hsa00920 Sulfur metabolism hsa01100 Metabolic pathways
Class	Metabolism; Energy Metabolism; Sulfur metabolism [PATH:hsa00920] Metabolism; Nucleotide Metabolism; Purine metabolism [PATH:hsa00230] Metabolism; Metabolism of Other Amino Acids; Selenoamino acid metabolism [PATH:hsa00450]
SSDB	
Motif	Pfam: ATP-sulfurylase APS_kinase DUF74
Other DBs	NCBI-GI: 62912492 NCBI-GeneID: 9060 OMIM: 603005 HGNC: 8604 HPRD: 04303 Ensembl: ENSG00000198682 UniProt: O95340
Structure	PDB: 2AX4

All links

- Pathway (4)
  - KEGG PATHWAY (4)
- Disease (2)
  - OMIM (2)
- Chemical reaction (2)
  - KEGG ENZYME (2)
- Genome (1)
  - KEGG GENOME (1)
- Gene (15)
  - KEGG ORTHOLOGY (2)
  - NCBI-Gene (1)
  - NCBI-GI (8)
  - UniGene (1)
  - HGNC (1)
  - HPRD (1)
  - ENSEMBL-HSA (1)
- Protein sequence (7)
  - UniProt (1)
  - RefSeq(pep) (2)
  - IPI (4)
- DNA sequence (24)
  - RefSeq(nuc) (2)
  - GenBank (11)
  - EMBL (11)
- 3D Structure (1)
  - PDB (1)
- Protein domain (3)
  - Pfam (3)
- All databases (59)

Reference maps allow access to cross-annotations



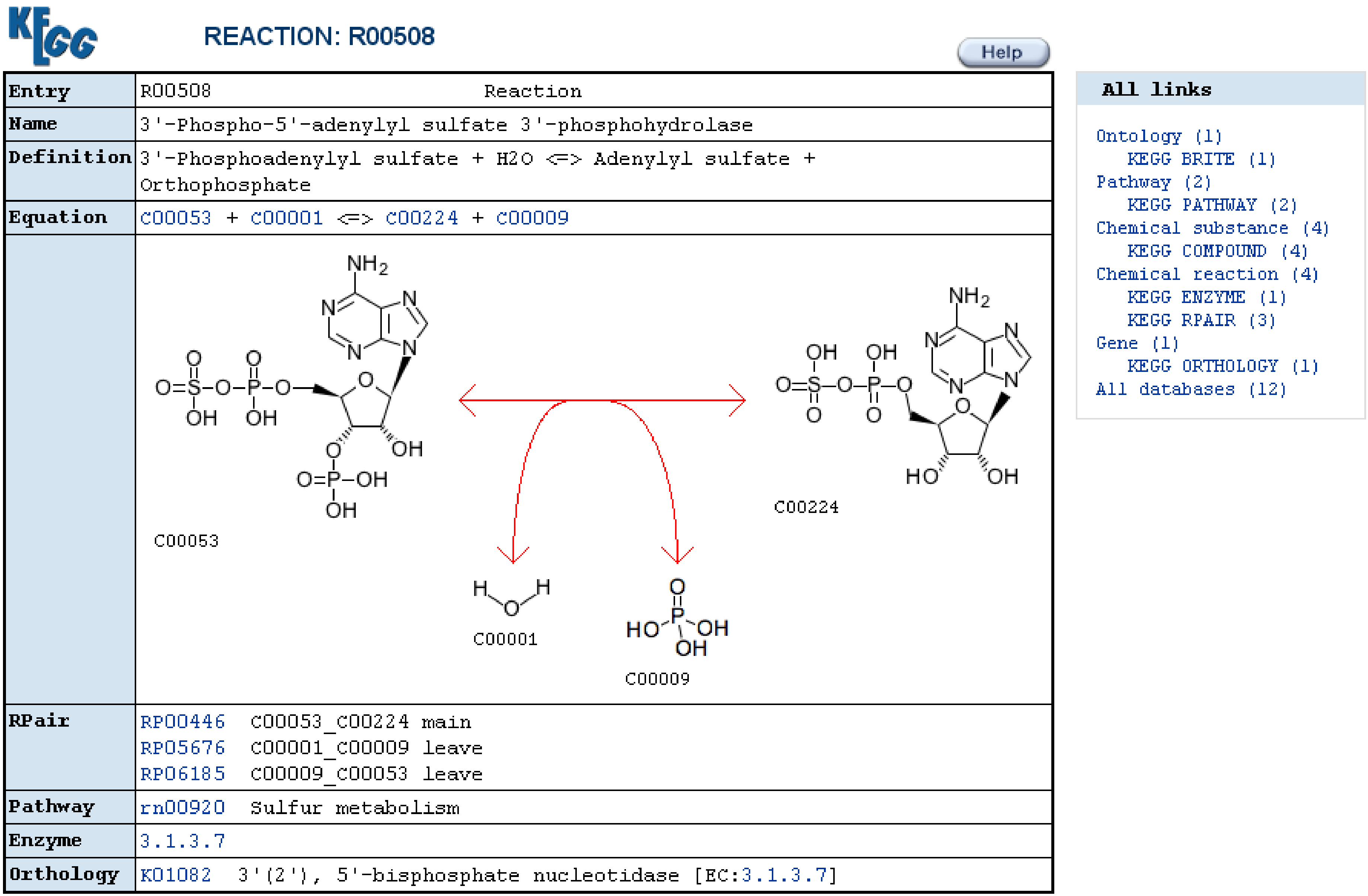
COMPOUND: C00224

Entry	C00224 Compound
Name	Adenylyl sulfate; Adenosine 5'-phosphosulfate; APS; 5'-Adenylyl sulfate
Formula	C10H14N5O10PS
Mass	427.0199
Structure	C00224
Reaction	R00508 R00509 R00529 R00530 R00531 R00860 R01618 R01619 R05717 R07176 R08553
Pathway	ko00230 Purine metabolism ko00920 Sulfur metabolism ko01100 Metabolic pathways
Enzyme	1.8.4.9 1.8.4.10 1.8.99.2 2.7.1.25 2.7.7.4 2.7.7.5 2.7.7.51 2.7.7.53 3.1.3.7 3.6.2.1
Other DBs	CAS: 485-84-7 PubChem: 3524 ChEBI: 17709 KNApSAcK: C00007445 PDB-CCD: ADX 3DMET: B01194 NIKKAJI: J37.502K
KCF data	

All links

- Pathway (3520)
  - KEGG PATHWAY (3519)
  - KEGG MODULE (1)
- Chemical substance (7)
  - PubChem (1)
  - ChEBI (1)
  - 3DMET (1)
  - HMDB (1)
  - KNApSAcK (1)
  - NIKKAJI (1)
  - PDB-CCD (1)
- Chemical reaction (29)
  - KEGG ENZYME (10)
  - KEGG REACTION (11)
  - KEGG RPAIR (8)
- All databases (3556)

Reference maps allow access to cross-annotations



REACTION: R00508

Entry	R00508 Reaction
Name	3'-Phospho-5'-adenylyl sulfate 3'-phosphohydrolase
Definition	3'-Phosphoadenylyl sulfate + H2O <=> Adenylyl sulfate + Orthophosphate
Equation	C00053 + C00001 <=> C00224 + C00009
RPair	RP00446 C00053_C00224 main RP05676 C00001_C00009 leave RP06185 C00009_C00053 leave
Pathway	rn00920 Sulfur metabolism
Enzyme	3.1.3.7
Orthology	K01082 3'(2'), 5'-bisphosphate nucleotidase [EC:3.1.3.7]

All links

- Ontology (1)
  - KEGG BRITE (1)
- Pathway (2)
  - KEGG PATHWAY (2)
- Chemical substance (4)
  - KEGG COMPOUND (4)
- Chemical reaction (4)
  - KEGG ENZYME (1)
  - KEGG RPAIR (3)
- Gene (1)
  - KEGG ORTHOLOGY (1)
- All databases (12)

matteo.ramazzotti@unifi.it

The Gene Ontology Database (GO)

Annotation@EBI

GeneDB

J. Craig Venter INSTITUTE

FlyBase

BDGP

EMBL-EBI

A Comparative mapping resource GRAMENE

the Gene Ontology

dictyBase

AstraZeneca

SGD

CGD

WormBase

PAMGO PLANT-ASSOCIATED MICROBE GENE ONTOLOGY

IGS

TGD

RGD

MGI

tair

The Gene Ontology Database (GO)

In informatics the term “ontology” refers to a controlled vocabulary of terms and definitions that can be used to describe objects and relationships among them. GO is an accurate and complete collection of definitions used to describe genes.

Gene Ontology

Cellular component

A cellular component is just that, a component of a cell, but with the proviso that it is part of some larger object; this may be an anatomical structure (e.g. rough endoplasmic reticulum or nucleus) or a gene product group (e.g. ribosome, proteasome or a protein dimer). See the Documentation on the cellular component ontology for more details.

Biological process

A biological process is series of events accomplished by one or more ordered assemblies of molecular functions. Examples of broad biological process terms are cellular physiological process or signal transduction. […] It can be difficult to distinguish between a biological process and a molecular function, but the general rule is that a process must have more than one distinct steps.

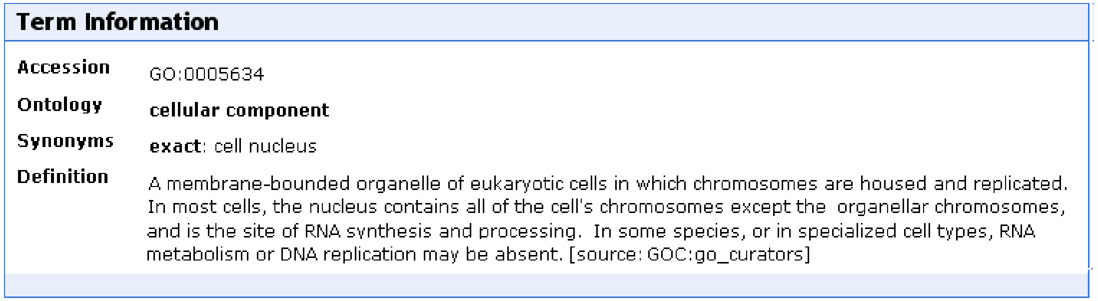
A biological process is not equivalent to a pathway; at present, GO does not try to represent the dynamics or dependencies that would be required to fully describe a pathway.

Molecular function

Molecular function describes activities, such as catalytic or binding activities, that occur at the molecular level. GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where or when, or in what context, the action takes place. Molecular functions generally correspond to activities that can be performed by individual gene products, but some activities are performed by assembled complexes of gene products. Examples of broad functional terms are catalytic activity, transporter activity, or binding; examples of narrower functional terms are adenylate cyclase activity or Toll receptor binding.

matteo.ramazzotti@unifi.it

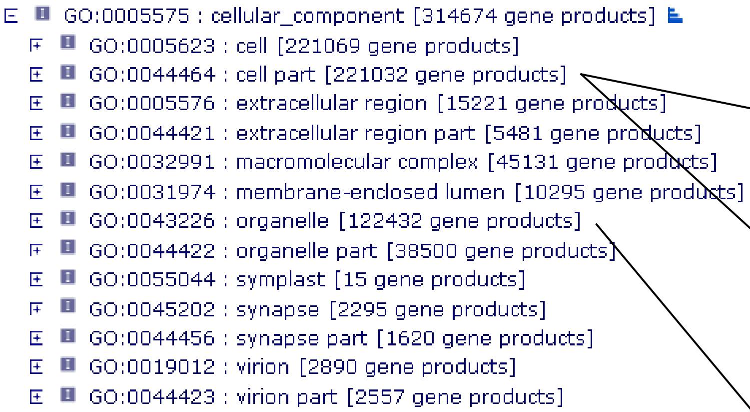
GO Cellular Component (CC)



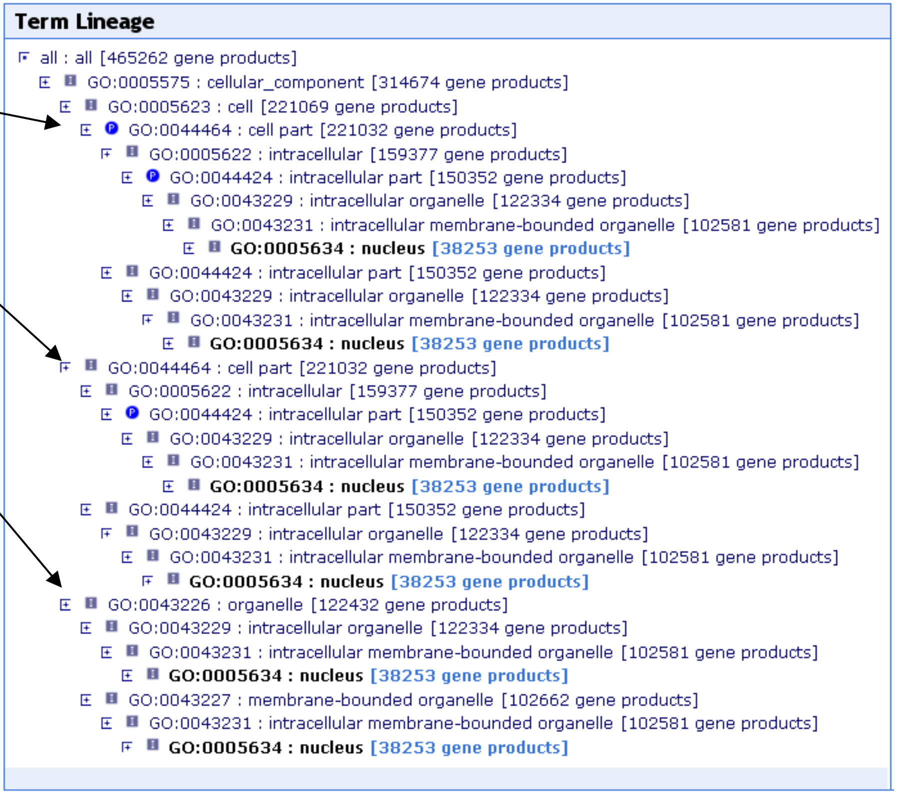
Term Information

Accession	GO:0005634
Ontology	**cellular component**
Synonyms	**exact**: cell nucleus
Definition	A membrane-bounded organelle of eukaryotic cells in which chromosomes are housed and replicated. In most cells, the nucleus contains all of the cell's chromosomes except the organellar chromosomes, and is the site of RNA synthesis and processing. In some species, or in specialized cell types, RNA metabolism or DNA replication may be absent. [source: GOC:go_curators]

Here is how GO define the term “nucleus”. If I refer to the nucleus as “GO:0005634”, al ambiguities are removed.



```
GO:0005575 : cellular_component [314674 gene products]
   GO:0005623 : cell [221069 gene products]
   GO:0044464 : cell part [221032 gene products]
   GO:0005576 : extracellular region [15221 gene products]
   GO:0044421 : extracellular region part [5481 gene products]
   GO:0032991 : macromolecular complex [45131 gene products]
   GO:0031974 : membrane-enclosed lumen [10295 gene products]
   GO:0043226 : organelle [122432 gene products]
   GO:0044422 : organelle part [38500 gene products]
   GO:0055044 : symplast [15 gene products]
   GO:0045202 : synapse [2295 gene products]
   GO:0044456 : synapse part [1620 gene products]
   GO:0019012 : virion [2890 gene products]
   GO:0044423 : virion part [2557 gene products]
```



Term Lineage

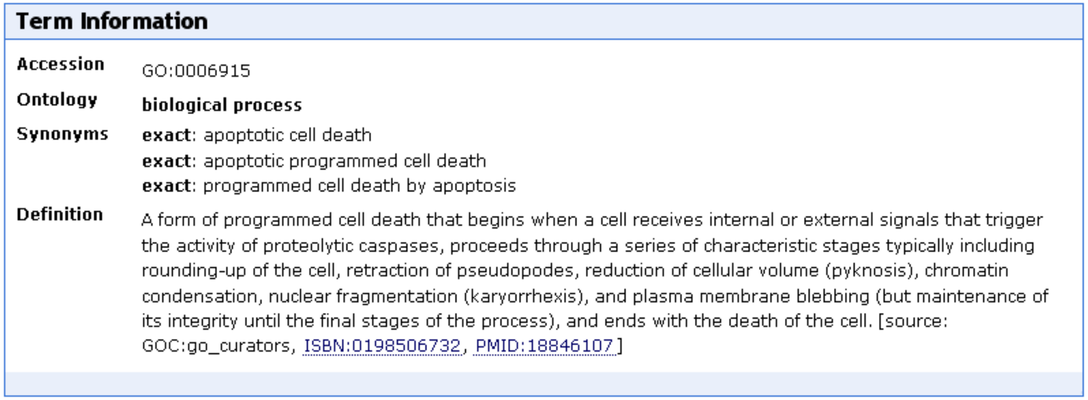
```
all : all [465262 gene products]
  GO:0005575 : cellular_component [314674 gene products]
    GO:0005623 : cell [221069 gene products]
      GO:0044464 : cell part [221032 gene products]
        GO:0005622 : intracellular [159377 gene products]
          GO:0044424 : intracellular part [150352 gene products]
            GO:0043229 : intracellular organelle [122334 gene products]
              GO:0043231 : intracellular membrane-bounded organelle [102581 gene products]
                GO:0005634 : nucleus [38253 gene products]
        GO:0044424 : intracellular part [150352 gene products]
          GO:0043229 : intracellular organelle [122334 gene products]
            GO:0043231 : intracellular membrane-bounded organelle [102581 gene products]
              GO:0005634 : nucleus [38253 gene products]
    GO:0044464 : cell part [221032 gene products]
      GO:0005622 : intracellular [159377 gene products]
        GO:0044424 : intracellular part [150352 gene products]
          GO:0043229 : intracellular organelle [122334 gene products]
            GO:0043231 : intracellular membrane-bounded organelle [102581 gene products]
              GO:0005634 : nucleus [38253 gene products]
      GO:0044424 : intracellular part [150352 gene products]
        GO:0043229 : intracellular organelle [122334 gene products]
          GO:0043231 : intracellular membrane-bounded organelle [102581 gene products]
            GO:0005634 : nucleus [38253 gene products]
    GO:0043226 : organelle [122432 gene products]
      GO:0043229 : intracellular organelle [122334 gene products]
        GO:0043231 : intracellular membrane-bounded organelle [102581 gene products]
          GO:0005634 : nucleus [38253 gene products]
      GO:0043227 : membrane-bounded organelle [102662 gene products]
        GO:0043231 : intracellular membrane-bounded organelle [102581 gene products]
          GO:0005634 : nucleus [38253 gene products]
```

Clearly, the nucleus is contained in other cell structures (components) that will have an appropriate GO term and code.

A hierarchical structure forms, that exemplify the reasoning behind GO.

matteo.ramazzotti@unifi.it

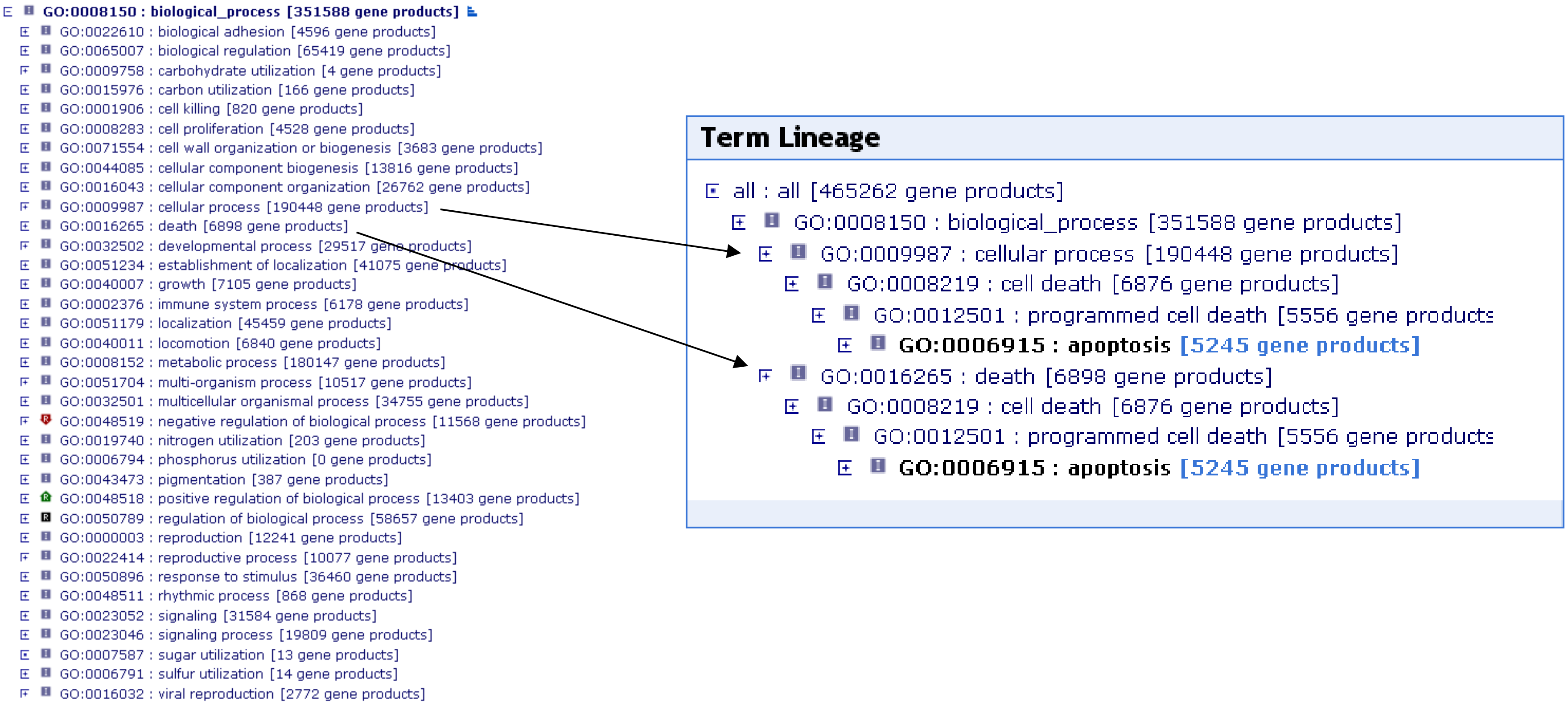
GO Biological Process (BP)



Term Information

Accession	GO:0006915
Ontology	**biological process**
Synonyms	**exact**: apoptotic cell death **exact**: apoptotic programmed cell death **exact**: programmed cell death by apoptosis
Definition	A form of programmed cell death that begins when a cell receives internal or external signals that trigger the activity of proteolytic caspases, proceeds through a series of characteristic stages typically including rounding-up of the cell, retraction of pseudopodes, reduction of cellular volume (pyknosis), chromatin condensation, nuclear fragmentation (karyorrhexis), and plasma membrane blebbing (but maintenance of its integrity until the final stages of the process), and ends with the death of the cell. [source: GOC:go_curators, ISBN:0198506732, PMID:18846107]

Here the definition of apoptosis (GO:000915): different criteria have to be used for a biological process, but the same hierarchical structure is presented.



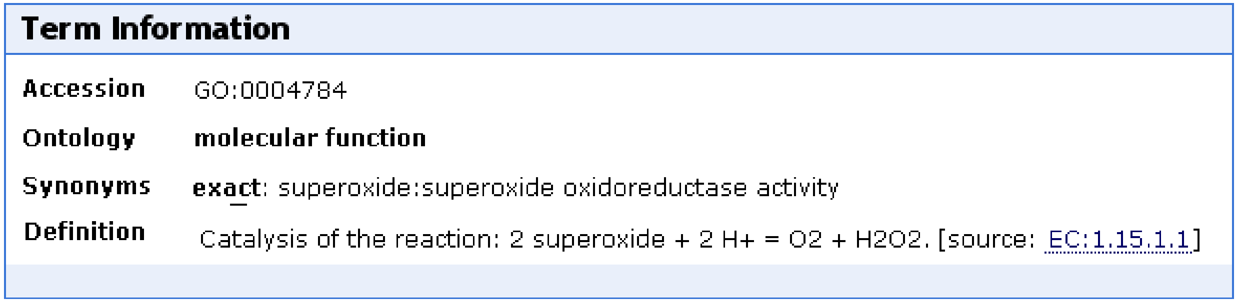
GO:0008150 : biological_process [351588 gene products]
- GO:0022610 : biological adhesion [4596 gene products]
- GO:0065007 : biological regulation [65419 gene products]
- GO:0009758 : carbohydrate utilization [4 gene products]
- GO:0015976 : carbon utilization [166 gene products]
- GO:0001906 : cell killing [820 gene products]
- GO:0008283 : cell proliferation [4528 gene products]
- GO:0071554 : cell wall organization or biogenesis [3683 gene products]
- GO:0044085 : cellular component biogenesis [13816 gene products]
- GO:0016043 : cellular component organization [26762 gene products]
- GO:0009987 : cellular process [190448 gene products]
- GO:0016265 : death [6898 gene products]
- GO:0032502 : developmental process [29517 gene products]
- GO:0051234 : establishment of localization [41075 gene products]
- GO:0040007 : growth [7105 gene products]
- GO:0002376 : immune system process [6178 gene products]
- GO:0051179 : localization [45459 gene products]
- GO:0040011 : locomotion [6840 gene products]
- GO:0008152 : metabolic process [180147 gene products]
- GO:0051704 : multi-organism process [10517 gene products]
- GO:0032501 : multicellular organismal process [34755 gene products]
- GO:0048519 : negative regulation of biological process [11568 gene products]
- GO:0019740 : nitrogen utilization [203 gene products]
- GO:0006794 : phosphorus utilization [0 gene products]
- GO:0043473 : pigmentation [387 gene products]
- GO:0048518 : positive regulation of biological process [13403 gene products]
- GO:0050789 : regulation of biological process [58657 gene products]
- GO:0000003 : reproduction [12241 gene products]
- GO:0022414 : reproductive process [10077 gene products]
- GO:0050896 : response to stimulus [36460 gene products]
- GO:0048511 : rhythmic process [868 gene products]
- GO:0023052 : signaling [31584 gene products]
- GO:0023046 : signaling process [19809 gene products]
- GO:0007587 : sugar utilization [13 gene products]
- GO:0006791 : sulfur utilization [14 gene products]
- GO:0016032 : viral reproduction [2772 gene products]

Term Lineage

- all : all [465262 gene products]
  - GO:0008150 : biological_process [351588 gene products]
    - GO:0009987 : cellular process [190448 gene products]
      - GO:0008219 : cell death [6876 gene products]
        - GO:0012501 : programmed cell death [5556 gene products
          - **GO:0006915 : apoptosis [5245 gene products]**
    - GO:0016265 : death [6898 gene products]
      - GO:0008219 : cell death [6876 gene products]
        - GO:0012501 : programmed cell death [5556 gene products
          - **GO:0006915 : apoptosis [5245 gene products]**

GO Molecular Function (MF)

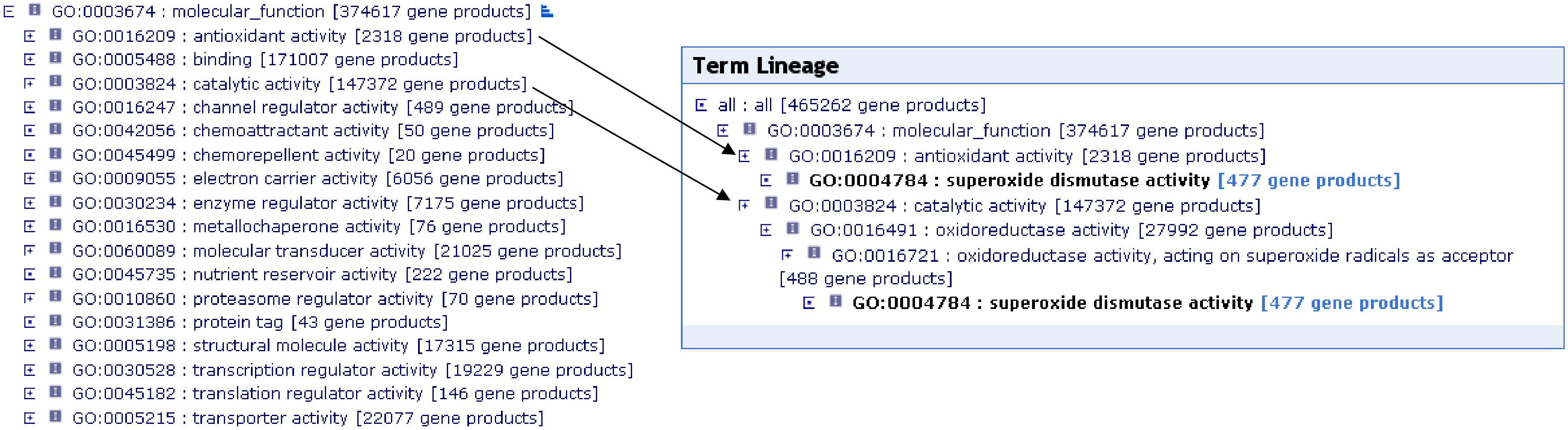
Note that each lemma in the “molecular function” class end with the term “activity”. We are not talking about specific gene or proteins, they are abstract cathegories.



Term Information

Accession	GO:0004784
Ontology	**molecular function**
Synonyms	**exact**: superoxide:superoxide oxidoreductase activity
Definition	Catalysis of the reaction: 2 superoxide + 2 H+ = O2 + H2O2. [source: EC:1.15.1.1]

Molecular function: now go defines how molecules or complexes acts. This is the deinition of superoxide dismutase activity (GO:0004784)



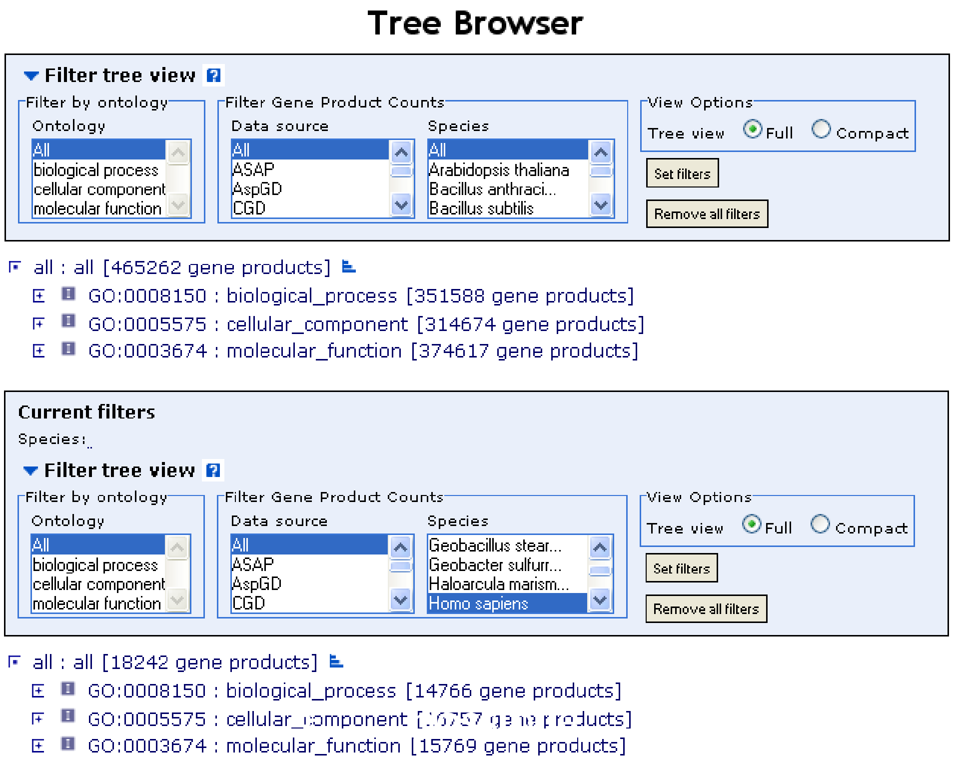
- GO:0003674 : molecular_function [374617 gene products]
  - GO:0016209 : antioxidant activity [2318 gene products]
  - GO:0005488 : binding [171007 gene products]
  - GO:0003824 : catalytic activity [147372 gene products]
  - GO:0016247 : channel regulator activity [489 gene products]
  - GO:0042056 : chemoattractant activity [50 gene products]
  - GO:0045499 : chemorepellent activity [20 gene products]
  - GO:0009055 : electron carrier activity [6056 gene products]
  - GO:0030234 : enzyme regulator activity [7175 gene products]
  - GO:0016530 : metallochaperone activity [76 gene products]
  - GO:0060089 : molecular transducer activity [21025 gene products]
  - GO:0045735 : nutrient reservoir activity [222 gene products]
  - GO:0010860 : proteasome regulator activity [70 gene products]
  - GO:0031386 : protein tag [43 gene products]
  - GO:0005198 : structural molecule activity [17315 gene products]
  - GO:0030528 : transcription regulator activity [19229 gene products]
  - GO:0045182 : translation regulator activity [146 gene products]
  - GO:0005215 : transporter activity [22077 gene products]

Term Lineage

- all : all [465262 gene products]
  - GO:0003674 : molecular_function [374617 gene products]
    - GO:0016209 : antioxidant activity [2318 gene products]
      - **GO:0004784 : superoxide dismutase activity** [477 gene products]
    - GO:0003824 : catalytic activity [147372 gene products]
      - GO:0016491 : oxidoreductase activity [27992 gene products]
        - GO:0016721 : oxidoreductase activity, acting on superoxide radicals as acceptor [488 gene products]
          - **GO:0004784 : superoxide dismutase activity** [477 gene products]

matteo.ramazzotti@unifi.it

AmiGO – search and browse GO



Tree Browser

Tree view ● Full ○ Compact

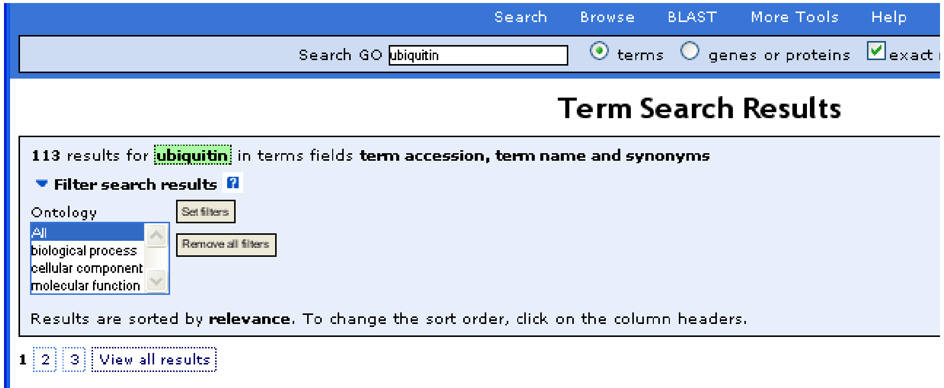
all : all [465262 gene products]
- GO:0008150 : biological_process [351588 gene products]
- GO:0005575 : cellular_component [314674 gene products]
- GO:0003674 : molecular_function [374617 gene products]

Current filters

Species:

Tree view ● Full ○ Compact

all : all [18242 gene products]
- GO:0008150 : biological_process [14766 gene products]
- GO:0005575 : cellular_component [16757 gene products]
- GO:0003674 : molecular_function [15769 gene products]



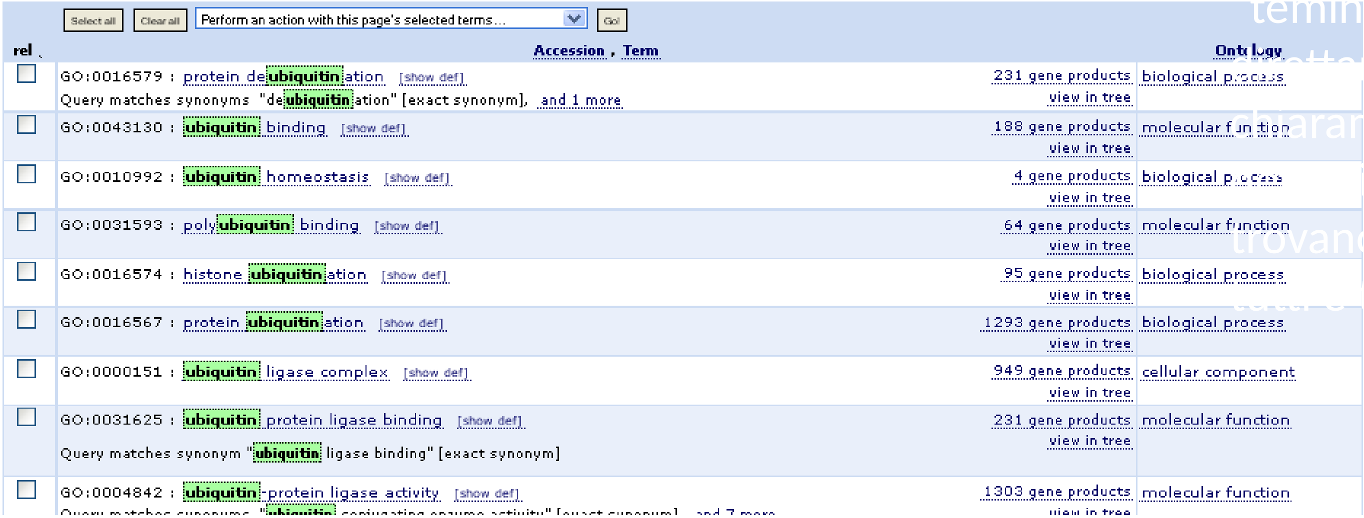
Search GO ubiquitin ● terms ○ genes or proteins ☑ exact

Term Search Results

113 results for **ubiquitin** in terms fields **term accession, term name and synonyms**

Results are sorted by **relevance**. To change the sort order, click on the column headers.

1 | 2 | 3 | View all results

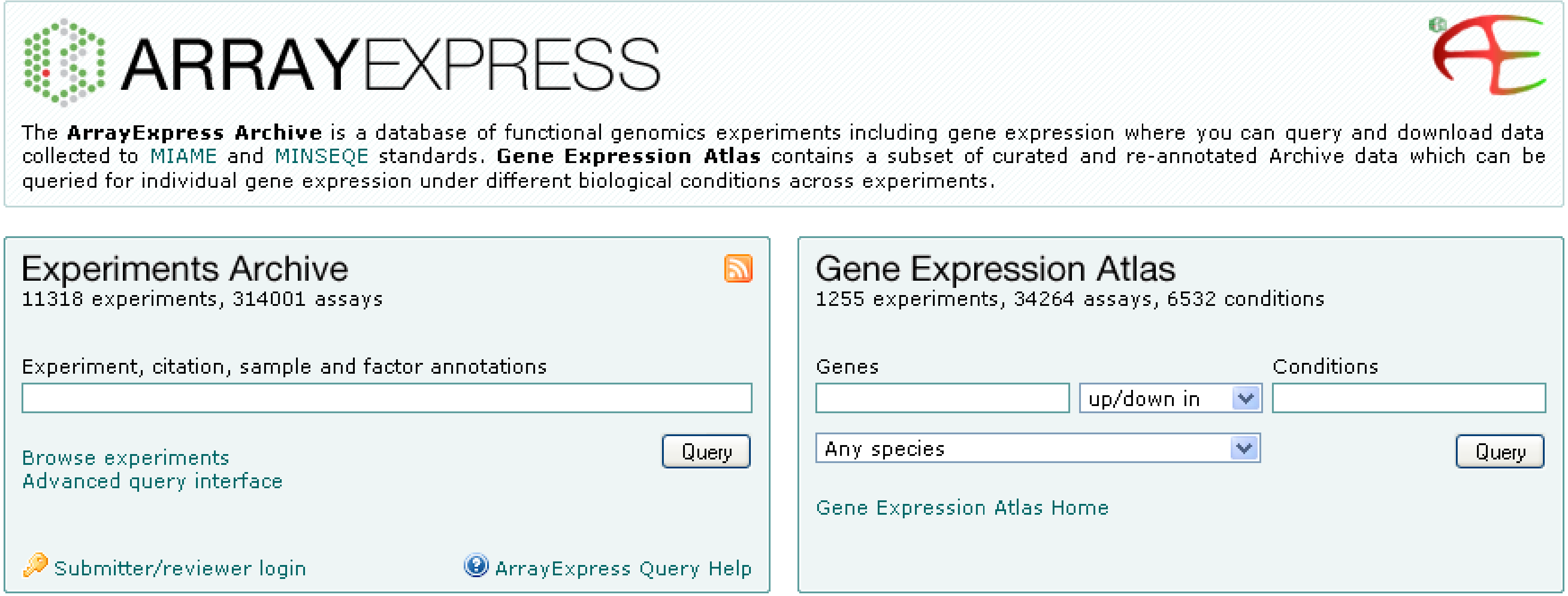


rel	Accession, Term		Ontology
☐	GO:0016579 : protein de**ubiquitin**ation [show def] Query matches synonyms "de**ubiquitin**ation" [exact synonym], and 1 more	231 gene products view in tree	biological process
☐	GO:0043130 : **ubiquitin** binding [show def]	188 gene products view in tree	molecular function
☐	GO:0010992 : **ubiquitin** homeostasis [show def]	4 gene products view in tree	biological process
☐	GO:0031593 : poly**ubiquitin** binding [show def]	64 gene products view in tree	molecular function
☐	GO:0016574 : histone **ubiquitin**ation [show def]	95 gene products view in tree	biological process
☐	GO:0016567 : protein **ubiquitin**ation [show def]	1293 gene products view in tree	biological process
☐	GO:0000151 : **ubiquitin** ligase complex [show def]	949 gene products view in tree	cellular component
☐	GO:0031625 : **ubiquitin** protein ligase binding [show def] Query matches synonym "**ubiquitin** ligase binding" [exact synonym]	231 gene products view in tree	molecular function
☐	GO:0004842 : **ubiquitin**-protein ligase activity [show def] Query matches synonyms "**ubiquitin** conjugating enzyme activity" [exact synonym], and 7 more	1303 gene products view in tree	molecular function

matteo.ramazzotti@unifi.it

EBI ArrayExpress

It is one of the most important transcriptomics repository. The International journals nowadays force authors of high throughput data to expose them into specialized data banks.



ARRAYEXPRESS

The **ArrayExpress Archive** is a database of functional genomics experiments including gene expression where you can query and download data collected to MIAME and MINSEQE standards. **Gene Expression Atlas** contains a subset of curated and re-annotated Archive data which can be queried for individual gene expression under different biological conditions across experiments.

Experiments Archive

11318 experiments, 314001 assays

Gene Expression Atlas

1255 experiments, 34264 assays, 6532 conditions

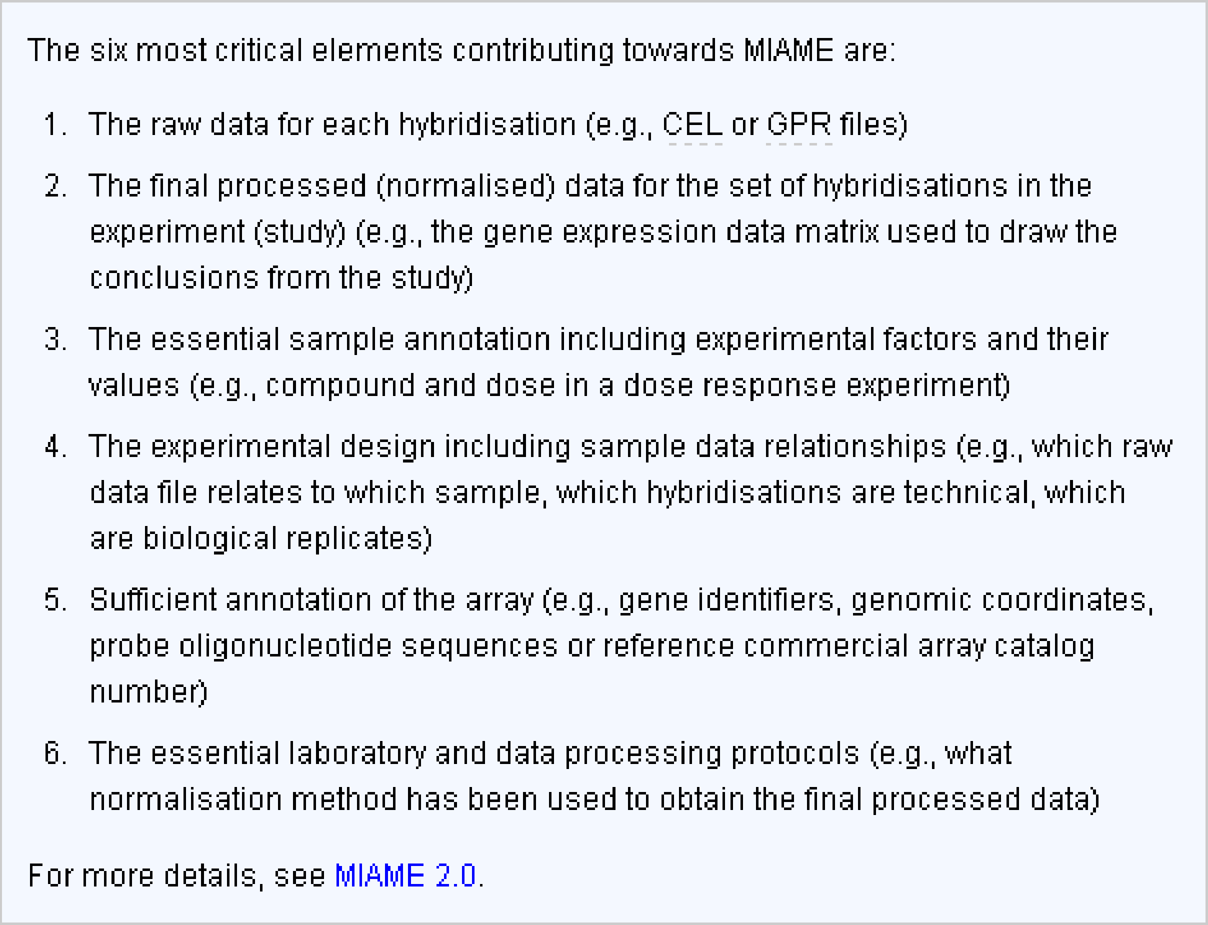
This is the the true data bank, that stores, collects and organizes experiments according to rigid submission protocols. All the experimental details, sample descriptions, and statistical procedures are also stored.

This is a gene centric version of the data bank where the measurement for a given gene is normalized across experiments to allow comparison.

matteo.ramazzotti@unifi.it

MIAME - Minimum Information About a Microarray Experiment

The submission of microarray data (not only) follows a strict check: the MIAME standard ensure that all deposited experiment is correct and clearly understandable (and reproducible, in principle) by every user of the data bank...



The six most critical elements contributing towards MIAME are:

1. The raw data for each hybridisation (e.g., CEL or GPR files)
2. The final processed (normalised) data for the set of hybridisations in the experiment (study) (e.g., the gene expression data matrix used to draw the conclusions from the study)
3. The essential sample annotation including experimental factors and their values (e.g., compound and dose in a dose response experiment)
4. The experimental design including sample data relationships (e.g., which raw data file relates to which sample, which hybridisations are technical, which are biological replicates)
5. Sufficient annotation of the array (e.g., gene identifiers, genomic coordinates, probe oligonucleotide sequences or reference commercial array catalog number)
6. The essential laboratory and data processing protocols (e.g., what normalisation method has been used to obtain the final processed data)

For more details, see MIAME 2.0.

Raw data

Preprocessed data

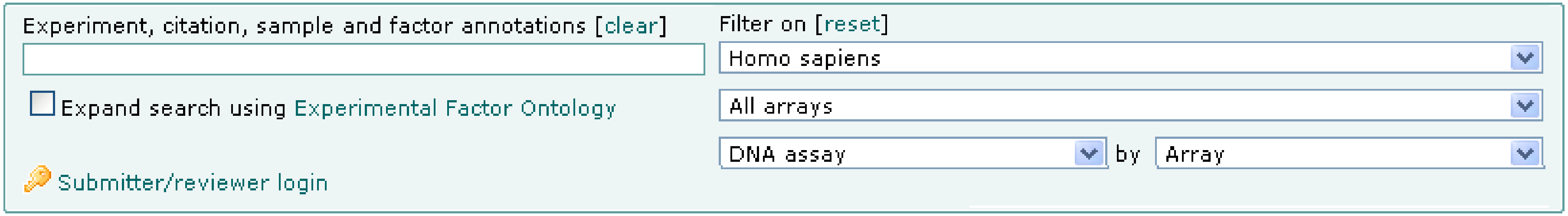
Ontology driven annotations

Experiment design

Array specifications

Experimental and informatic protocols

ArrayExpress: experiment archive



Experiment, citation, sample and factor annotations [clear]

☐ Expand search using Experimental Factor Ontology

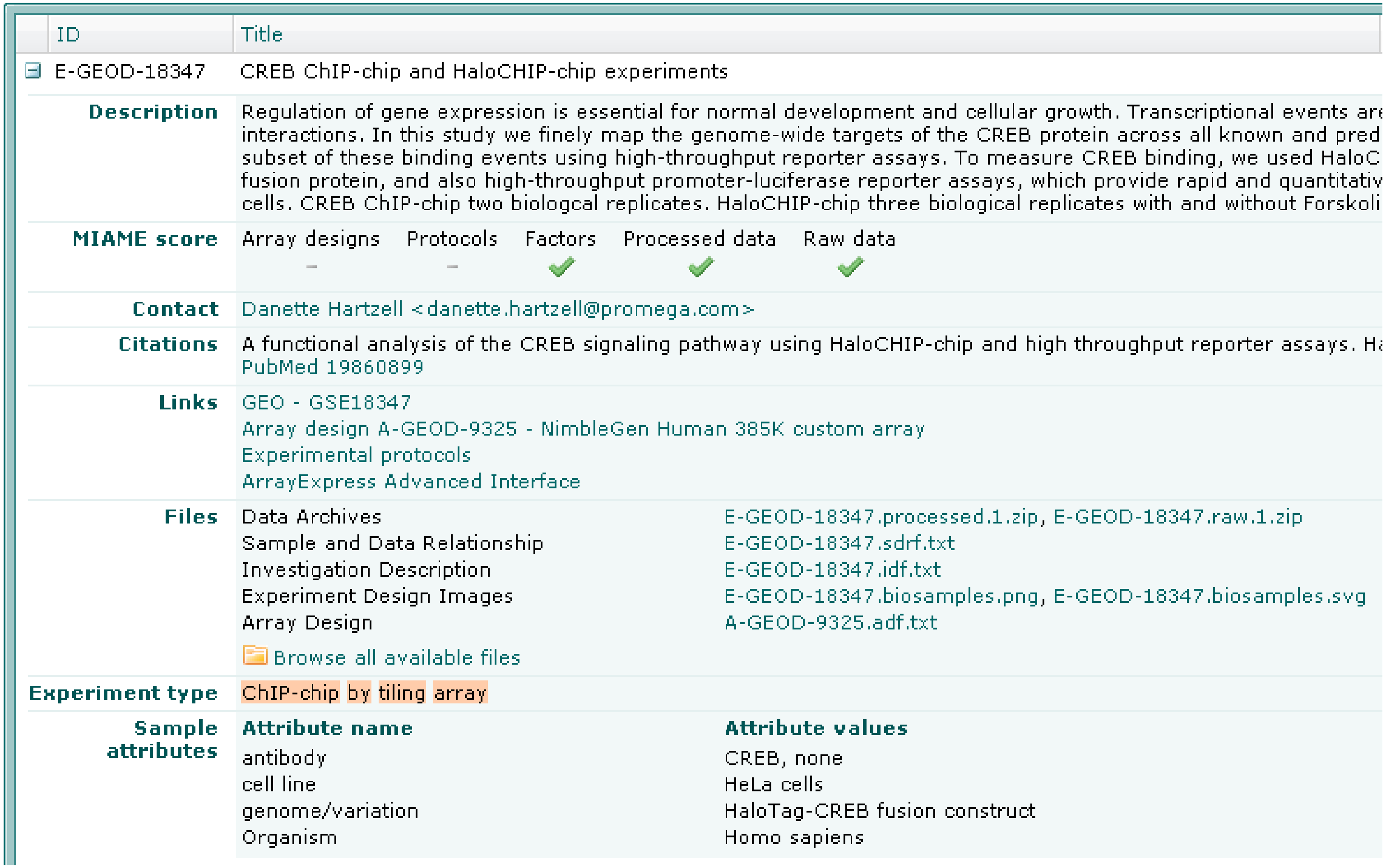
Submitter/reviewer login

Filter on [reset]

Homo sapiens

All arrays

DNA assay by Array



ID	Title
E-GEOD-18347	CREB ChIP-chip and HaloCHIP-chip experiments

Description Regulation of gene expression is essential for normal development and cellular growth. Transcriptional events are interactions. In this study we finely map the genome-wide targets of the CREB protein across all known and pred subset of these binding events using high-throughput reporter assays. To measure CREB binding, we used HaloC fusion protein, and also high-throughput promoter-luciferase reporter assays, which provide rapid and quantitativ cells. CREB ChIP-chip two biologcal replicates. HaloCHIP-chip three biological replicates with and without Forskoli

MIAME score

Array designs	Protocols	Factors	Processed data	Raw data
–	–	✓	✓	✓

Contact Danette Hartzell <danette.hartzell@promega.com>

Citations A functional analysis of the CREB signaling pathway using HaloCHIP-chip and high throughput reporter assays. Ha
PubMed 19860899

Links
GEO - GSE18347
Array design A-GEOD-9325 - NimbleGen Human 385K custom array
Experimental protocols
ArrayExpress Advanced Interface

Files

Data Archives	E-GEOD-18347.processed.1.zip, E-GEOD-18347.raw.1.zip
Sample and Data Relationship	E-GEOD-18347.sdrf.txt
Investigation Description	E-GEOD-18347.idf.txt
Experiment Design Images	E-GEOD-18347.biosamples.png, E-GEOD-18347.biosamples.svg
Array Design	A-GEOD-9325.adf.txt

Browse all available files

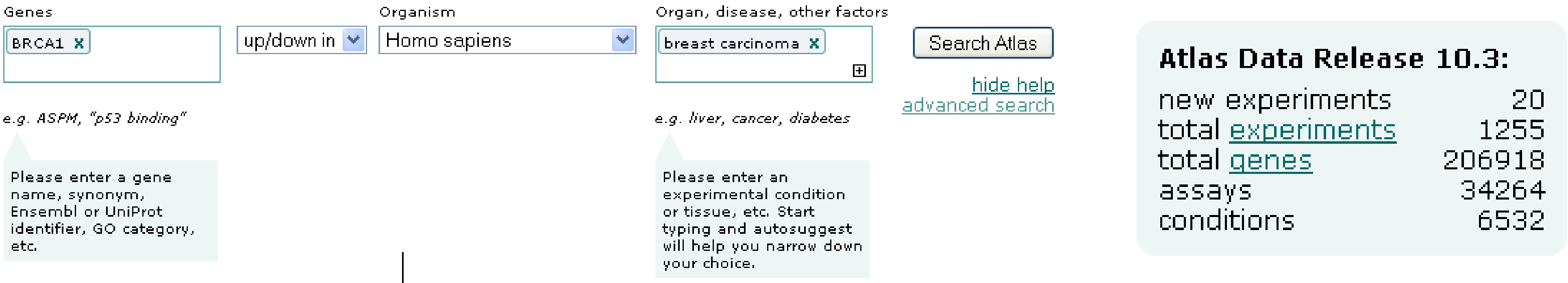
Experiment type ChIP-chip by tiling array

Sample attributes

Attribute name	Attribute values
antibody	CREB, none
cell line	HeLa cells
genome/variation	HaloTag-CREB fusion construct
Organism	Homo sapiens

matteo.ramazzotti@unifi.it

ArrayExpress: the gene expression atlas



e.g. ASPM, "p53 binding"

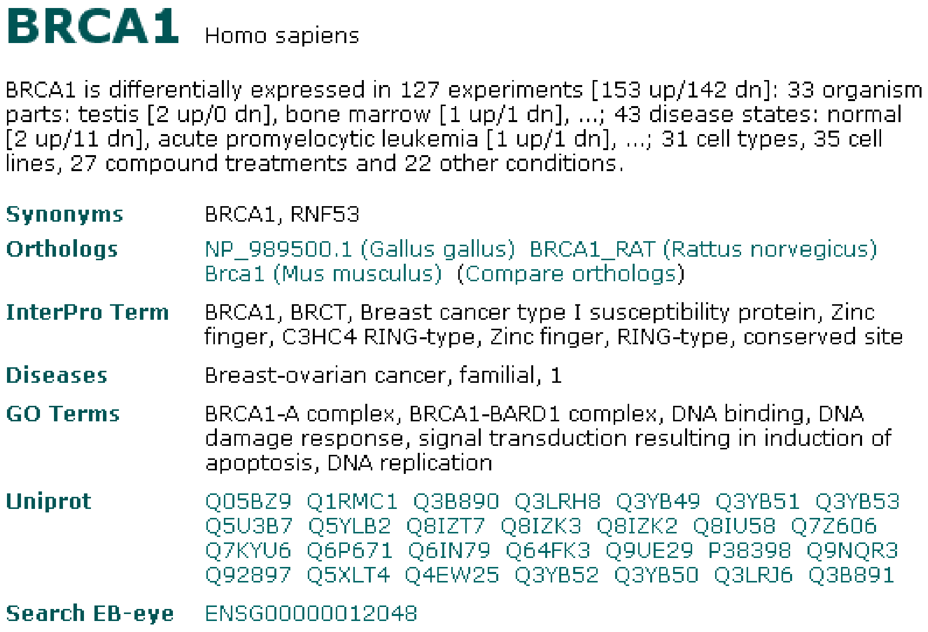
Please enter a gene name, synonym, Ensembl or UniProt identifier, GO category, etc.

e.g. liver, cancer, diabetes

Please enter an experimental condition or tissue, etc. Start typing and autosuggest will help you narrow down your choice.

Atlas Data Release 10.3:

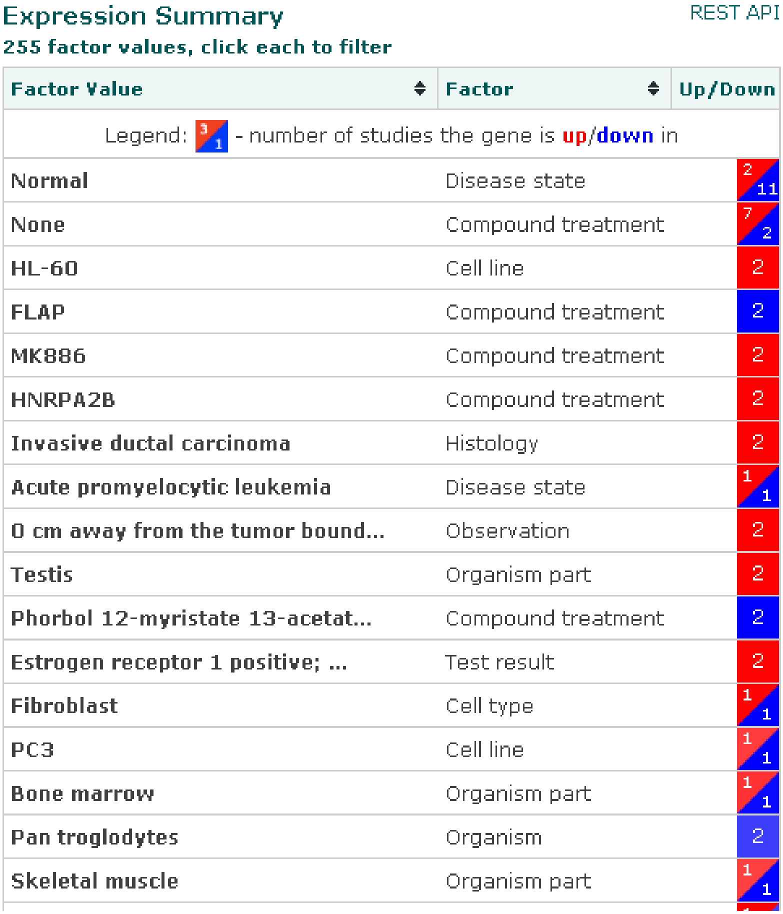
new experiments	20
total experiments	1255
total genes	206918
assays	34264
conditions	6532



BRCA1 Homo sapiens

BRCA1 is differentially expressed in 127 experiments [153 up/142 dn]: 33 organism parts: testis [2 up/0 dn], bone marrow [1 up/1 dn], ...; 43 disease states: normal [2 up/11 dn], acute promyelocytic leukemia [1 up/1 dn], ...; 31 cell types, 35 cell lines, 27 compound treatments and 22 other conditions.

Synonyms	BRCA1, RNF53
Orthologs	NP_989500.1 (Gallus gallus) BRCA1_RAT (Rattus norvegicus) Brca1 (Mus musculus) (Compare orthologs)
InterPro Term	BRCA1, BRCT, Breast cancer type I susceptibility protein, Zinc finger, C3HC4 RING-type, Zinc finger, RING-type, conserved site
Diseases	Breast-ovarian cancer, familial, 1
GO Terms	BRCA1-A complex, BRCA1-BARD1 complex, DNA binding, DNA damage response, signal transduction resulting in induction of apoptosis, DNA replication
Uniprot	Q05BZ9 Q1RMC1 Q3B890 Q3LRH8 Q3YB49 Q3YB51 Q3YB53 Q5U3B7 Q5YLB2 Q8IZT7 Q8IZK3 Q8IZK2 Q8IU58 Q7Z606 Q7KYU6 Q6P671 Q6IN79 Q64FK3 Q9UE29 P38398 Q9NQR3 Q92897 Q5XLT4 Q4EW25 Q3YB52 Q3YB50 Q3LRJ6 Q3B891
Search EB-eye	ENSG00000012048



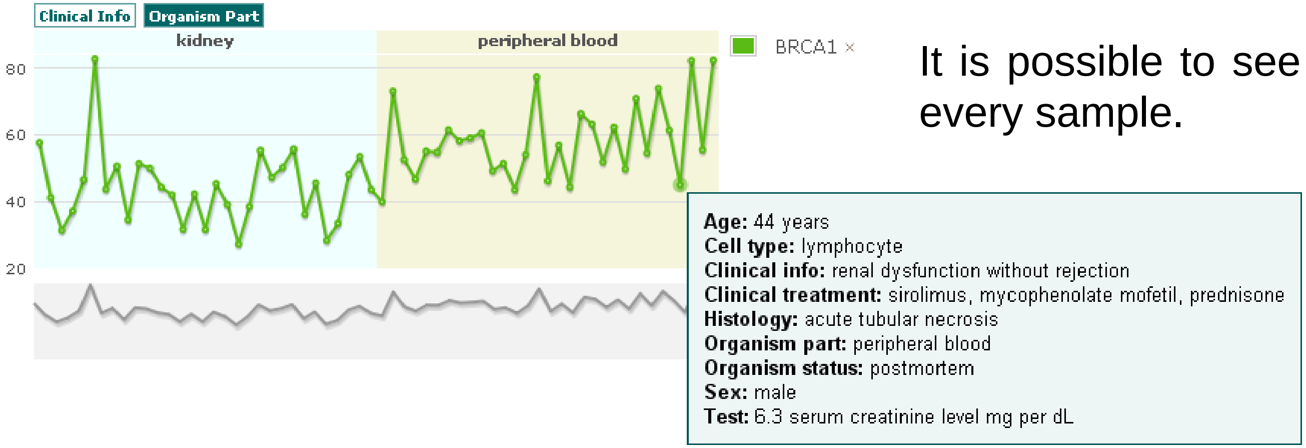
Expression Summary REST API

255 factor values, click each to filter

Factor Value	**Factor**	**Up/Down**
Legend: 🟥3 🟦1 - number of studies the gene is **up**/**down** in		
Normal	Disease state	🟥 2 🟦 11
None	Compound treatment	🟥 7 🟦 2
HL-60	Cell line	🟥 2
FLAP	Compound treatment	🟦 2
MK886	Compound treatment	🟥 2
HNRPA2B	Compound treatment	🟥 2
Invasive ductal carcinoma	Histology	🟥 2
Acute promyelocytic leukemia	Disease state	🟥 1 🟦 1
0 cm away from the tumor bound...	Observation	🟥 2
Testis	Organism part	🟥 2
Phorbol 12-myristate 13-acetat...	Compound treatment	🟦 2
Estrogen receptor 1 positive; ...	Test result	🟥 2
Fibroblast	Cell type	🟥 1 🟦 1
PC3	Cell line	🟥 1 🟦 1
Bone marrow	Organism part	🟥 1 🟦 1
Pan troglodytes	Organism	🟦 2
Skeletal muscle	Organism part	🟥 1 🟦 1

matteo.ramazzotti@unifi.it

ArrayExpress: gene expression atlas



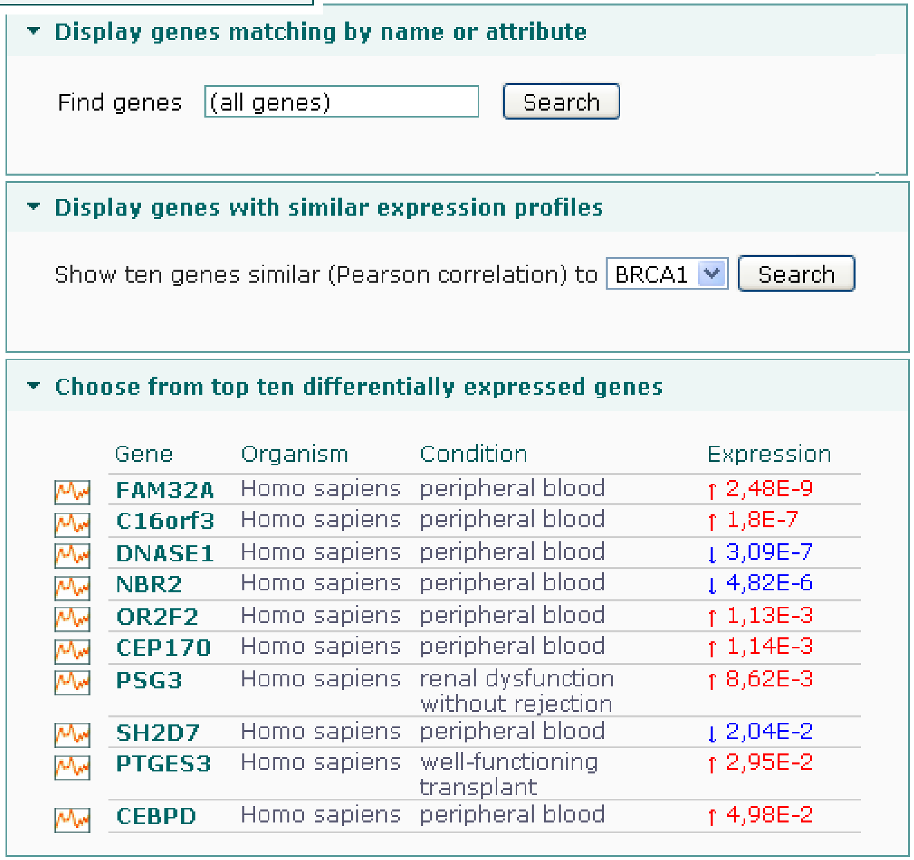
Age: 44 years
Cell type: lymphocyte
Clinical info: renal dysfunction without rejection
Clinical treatment: sirolimus, mycophenolate mofetil, prednisone
Histology: acute tubular necrosis
Organism part: peripheral blood
Organism status: postmortem
Sex: male
Test: 6.3 serum creatinine level mg per dL

It is possible to see the gene expresion profiles of every sample.

Experimental data can be easily accessed.

Using the menus, it is possible to add the profiles of other genes selected by:

- Name
- Similar expression profile
- Top variable genea in the sample



Display genes matching by name or attribute

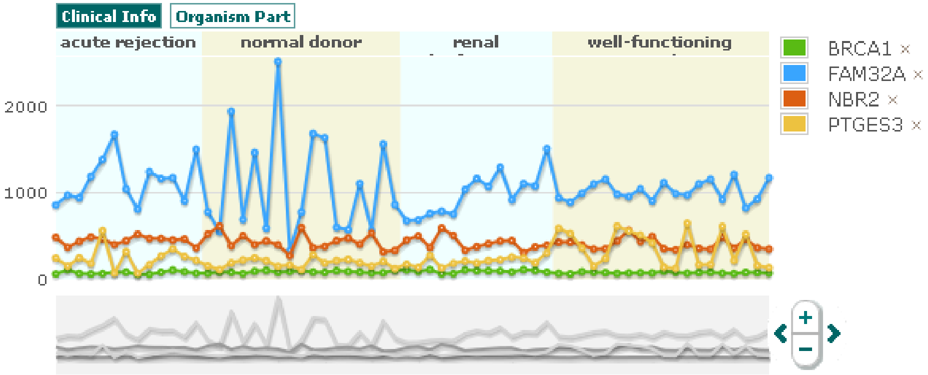
Find genes (all genes) Search

Display genes with similar expression profiles

Show ten genes similar (Pearson correlation) to BRCA1 Search

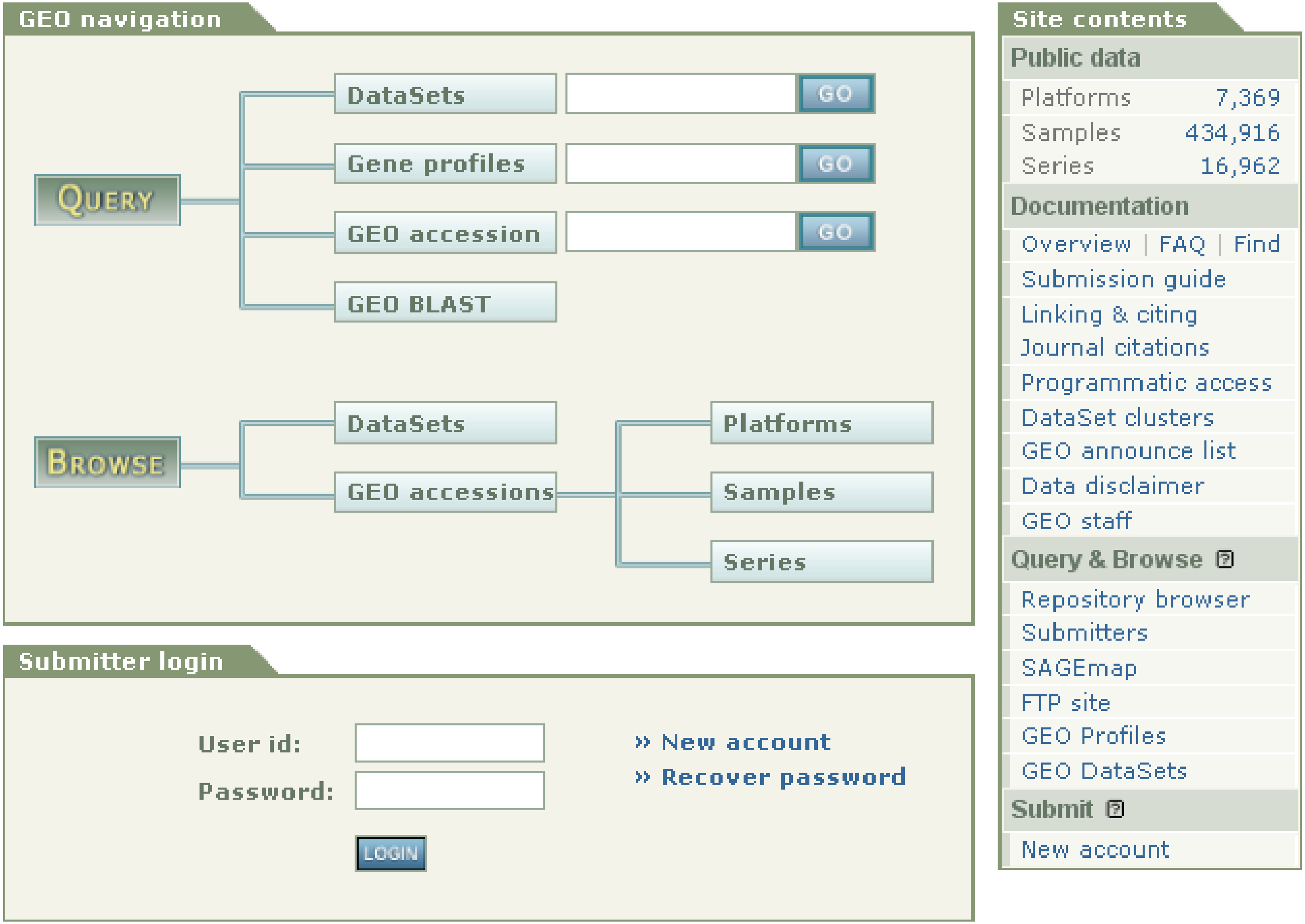
Choose from top ten differentially expressed genes

Gene	Organism	Condition	Expression
FAM32A	Homo sapiens	peripheral blood	↑ 2,48E-9
C16orf3	Homo sapiens	peripheral blood	↑ 1,8E-7
DNASE1	Homo sapiens	peripheral blood	↓ 3,09E-7
NBR2	Homo sapiens	peripheral blood	↓ 4,82E-6
OR2F2	Homo sapiens	peripheral blood	↑ 1,13E-3
CEP170	Homo sapiens	peripheral blood	↑ 1,14E-3
PSG3	Homo sapiens	renal dysfunction without rejection	↑ 8,62E-3
SH2D7	Homo sapiens	peripheral blood	↓ 2,04E-2
PTGES3	Homo sapiens	well-functioning transplant	↑ 2,95E-2
CEBPD	Homo sapiens	peripheral blood	↑ 4,98E-2



matteo.ramazzotti@unifi.it

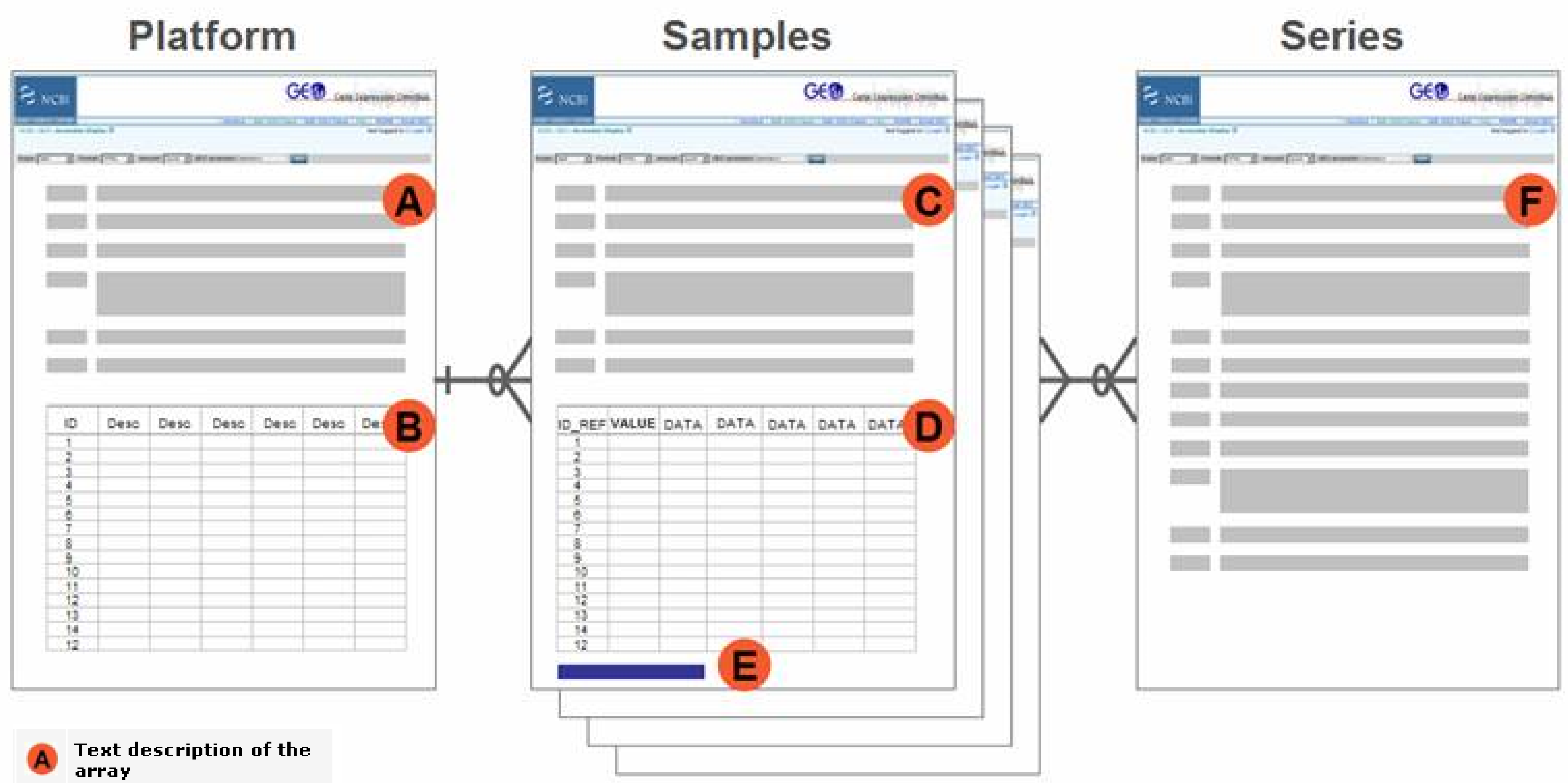
Gene Expression Omnibus (GEO)



Public data	
Platforms	7,369
Samples	434,916
Series	16,962

matteo.ramazzotti@unifi.it

Gene Expression Omnibus (GEO)



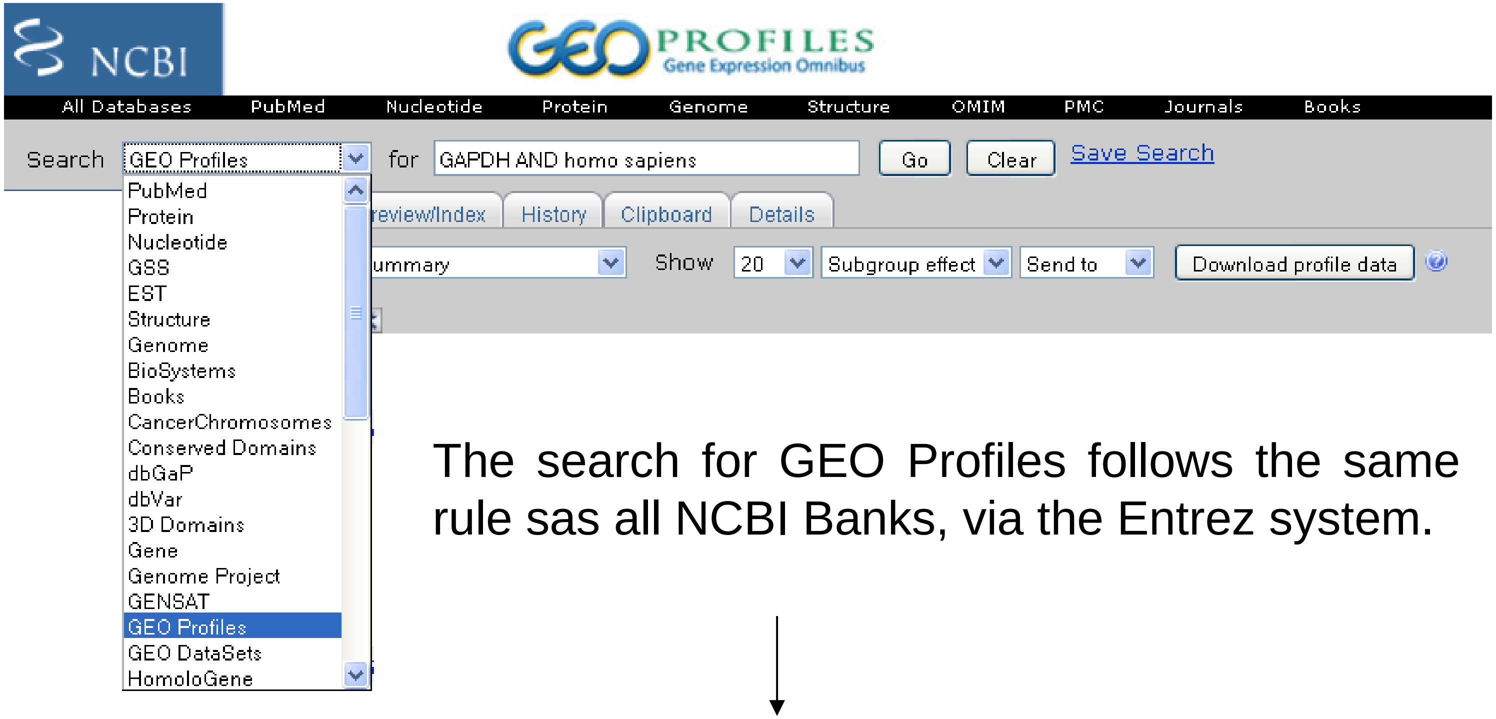
A	Text description of the array
B	Text tab-delimited table of the array template
C	Text description of a biological sample
D	Text tab-delimited table of processed hybridization result (may optionally include raw data columns)
E	Original raw data file
F	Text description of the overall experiment

GEO data are very clearly organized: platforms (GPL, types of chip) where samples (GPS) have been hybridized in series (GSE, the experiments).

matteo.ramazzotti@unifi.it

GEO: datasets and profiles

GEO curators continuously check the GSE series in order to verify wheter they can be combined (due to experimetal and technical reasons) into **dataset** (GDS), the heart of the GEO gene **profiles**, i.e. the expression level of genes across different conditions.

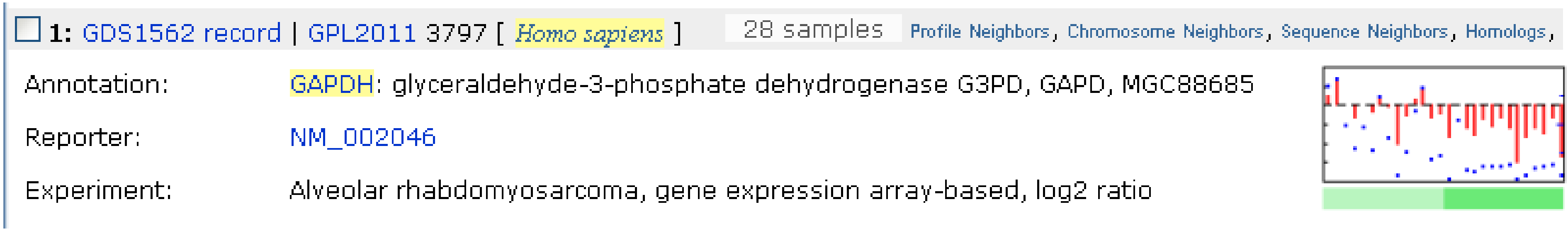


Search GEO Profiles for GAPDH AND homo sapiens

ummary

Show 20 Subgroup effect Send to

The search for GEO Profiles follows the same rule sas all NCBI Banks, via the Entrez system.



☐ **1:** GDS1562 record | GPL2011 3797 [*Homo sapiens*] 28 samples

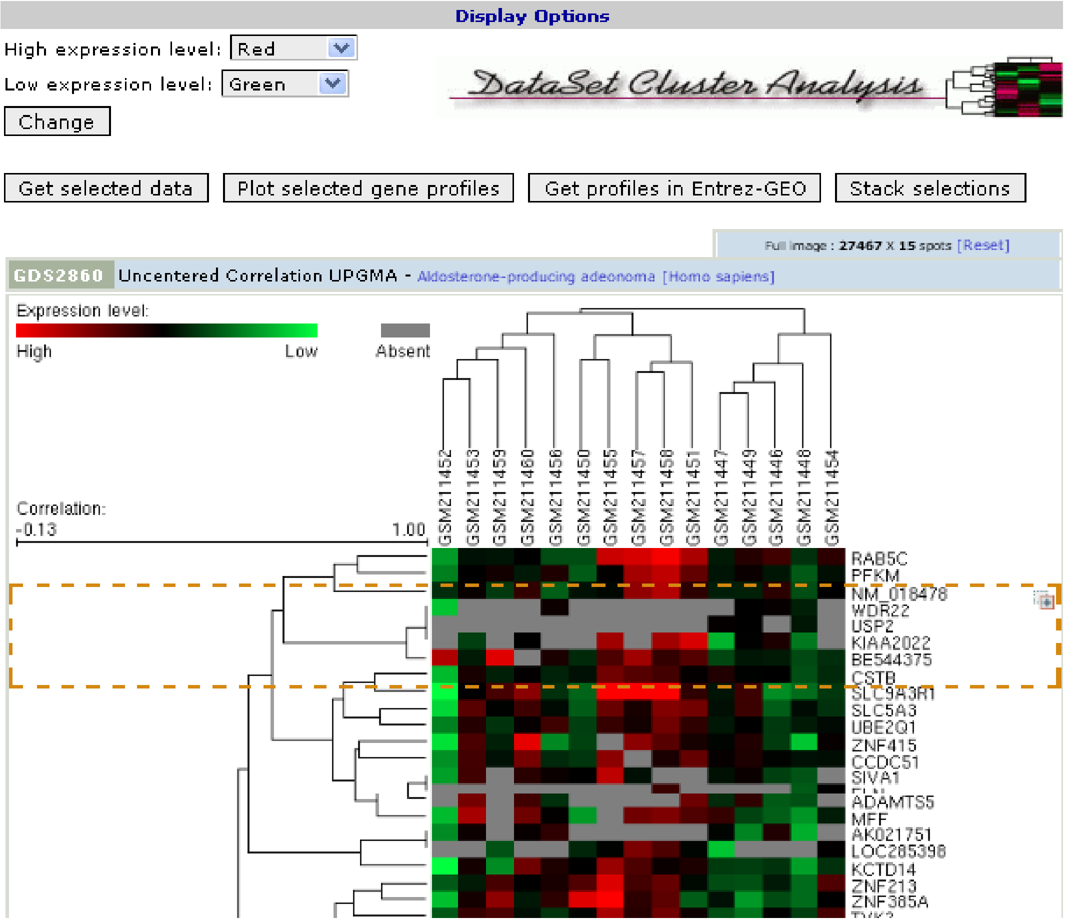
Annotation: GAPDH: glyceraldehyde-3-phosphate dehydrogenase G3PD, GAPD, MGC88685

Reporter: NM_002046

Experiment: Alveolar rhabdomyosarcoma, gene expression array-based, log2 ratio

GEO: cluster analysis

Microarray data, after background correction and normalization, can be analyzed in several multivariate ways, among which cluster analysis. Genes with similar profiles are precomputed in GEO.

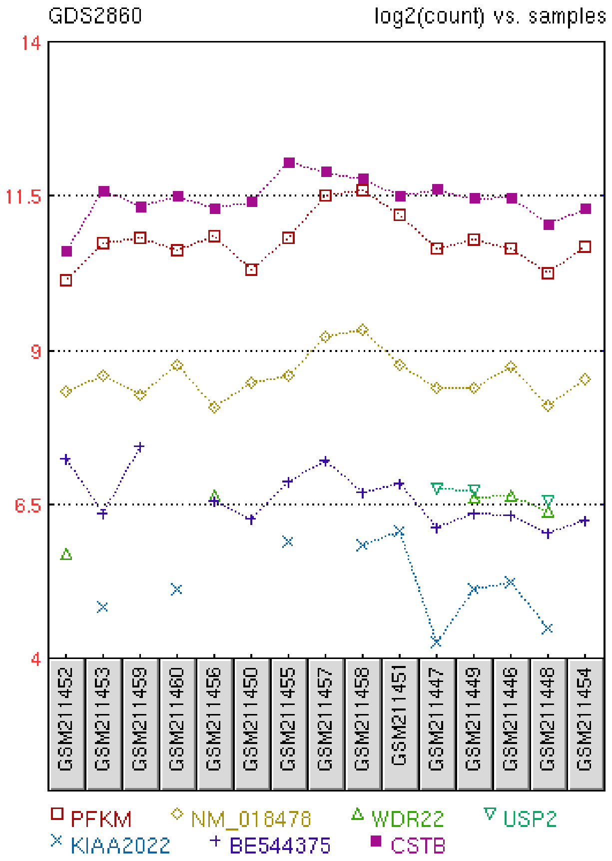


High expression level: Red

Low expression level: Green

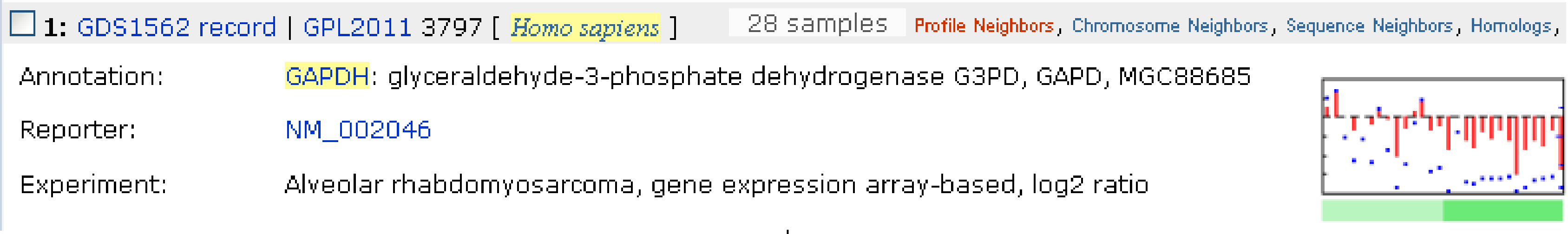
Full image : 27467 X 15 spots [Reset]

GDS2860 Uncentered Correlation UPGMA - Aldosterone-producing adeonoma [Homo sapiens]



GEO: neighbour analysis

The GEO systesm also extends the gene-based analysis by searching for **neighbours** in term of profile, sequence or position in the chromosome.



☐ **1:** GDS1562 record | GPL2011 3797 [*Homo sapiens*] 28 samples Profile Neighbors, Chromosome Neighbors, Sequence Neighbors, Homologs,

Annotation: GAPDH: glyceraldehyde-3-phosphate dehydrogenase G3PD, GAPD, MGC88685

Reporter: NM_002046

Experiment: Alveolar rhabdomyosarcoma, gene expression array-based, log2 ratio

Profile Neighbors, Chromosome Neighbors, Sequence Neighbors, Homologs,

Genes with similar profiles in the same dataset

Profiles of the neighboring genes in the chromosome (same organism)

Profiles of genes with high sequence similarity (same organism)

Profiles of ortholog genes (high sequence similarity in (other organisms)

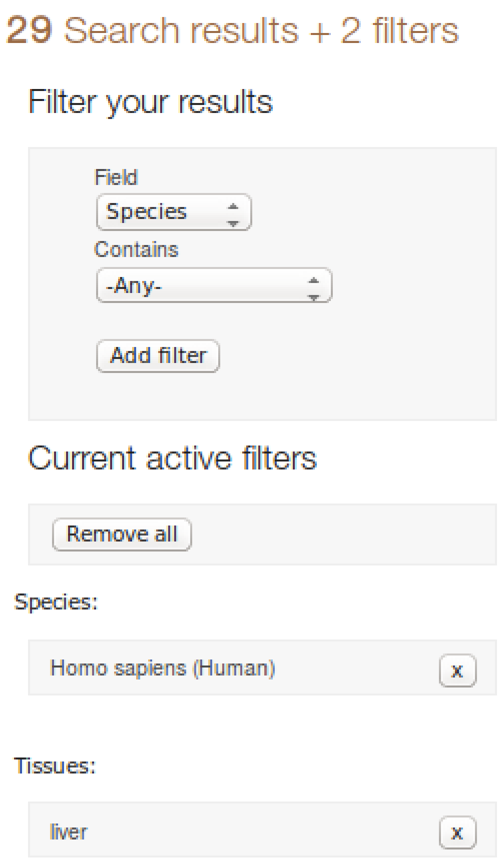
matteo.ramazzotti@unifi.it

PRIDE: PRoteomics IDEntifications database

This is the central repository for shotgun proteomics. Experiments are archived as in ArrayExpress and can be searched using filters and then download as raw or processd files.

Datasets

- 4175 projects
- 58096 assays



29 Search results + 2 filters

Filter your results

Field: Species

Contains: -Any-

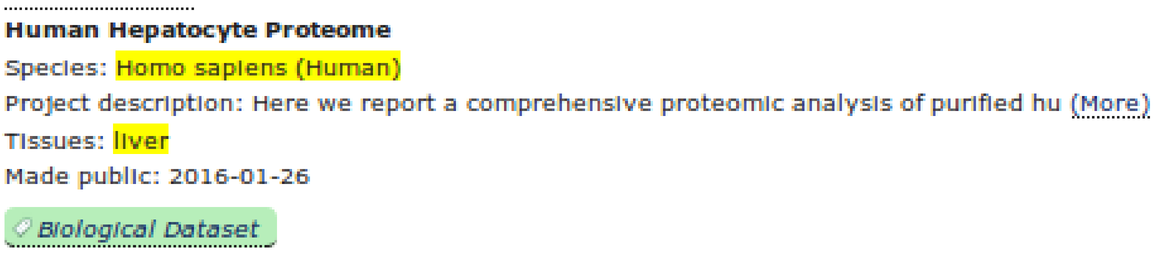
Current active filters

Species:

Homo sapiens (Human) x

Tissues:

liver x



PXD001874

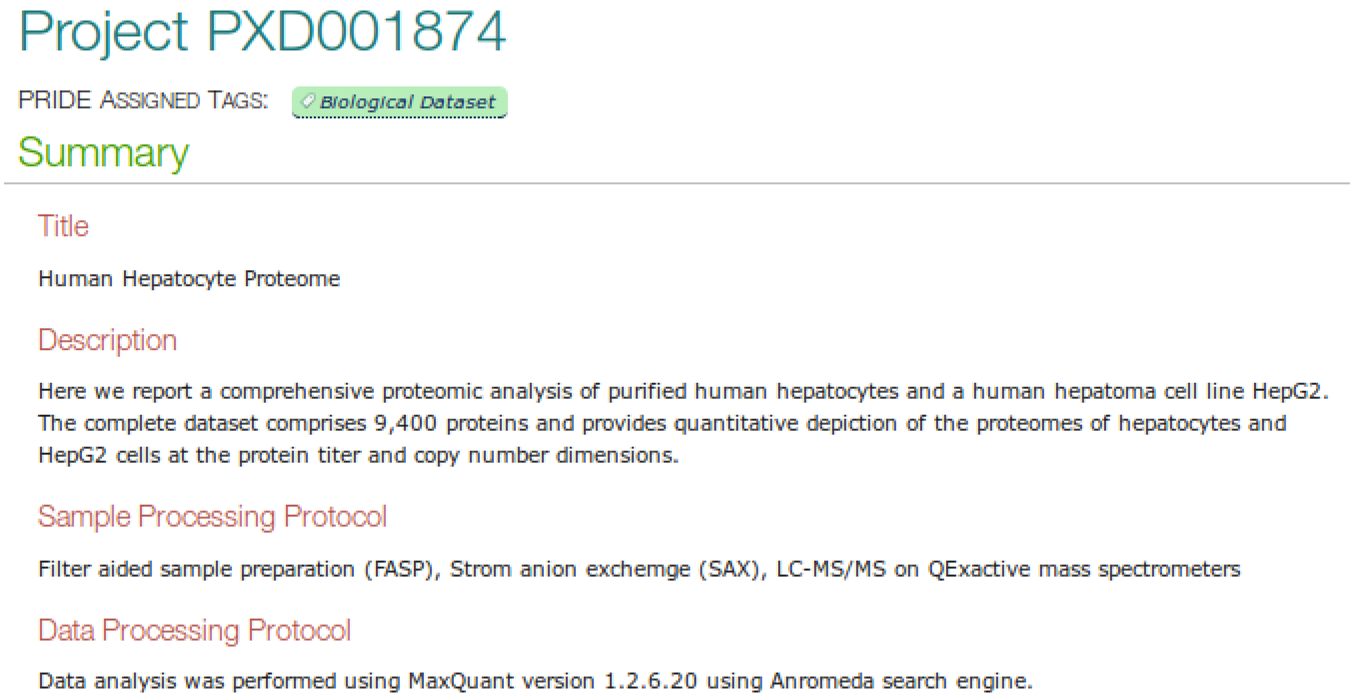
Human Hepatocyte Proteome

Species: Homo sapiens (Human)

Project description: Here we report a comprehensive proteomic analysis of purified hu (More)

Tissues: liver

Made public: 2016-01-26



Project PXD001874

PRIDE ASSIGNED TAGS: Biological Dataset

Summary

Title

Human Hepatocyte Proteome

Description

Here we report a comprehensive proteomic analysis of purified human hepatocytes and a human hepatoma cell line HepG2. The complete dataset comprises 9,400 proteins and provides quantitative depiction of the proteomes of hepatocytes and HepG2 cells at the protein titer and copy number dimensions.

Sample Processing Protocol

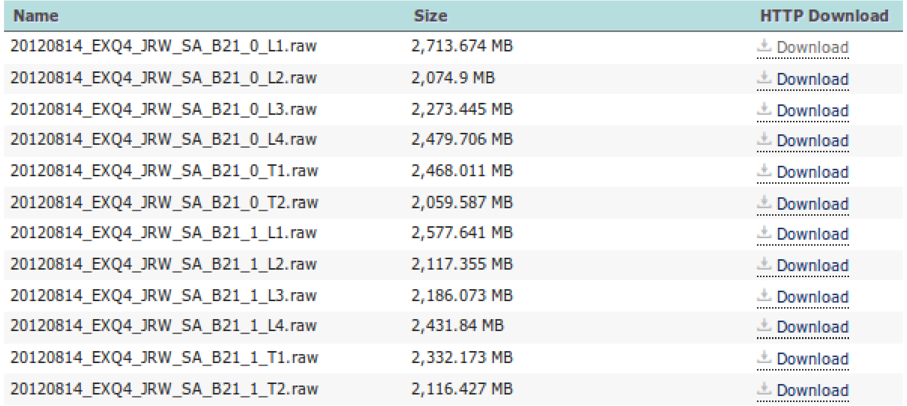
Filter aided sample preparation (FASP), Strom anion exchemge (SAX), LC-MS/MS on QExactive mass spectrometers

Data Processing Protocol

Data analysis was performed using MaxQuant version 1.2.6.20 using Anromeda search engine.

matteo.ramazzotti@unifi.it

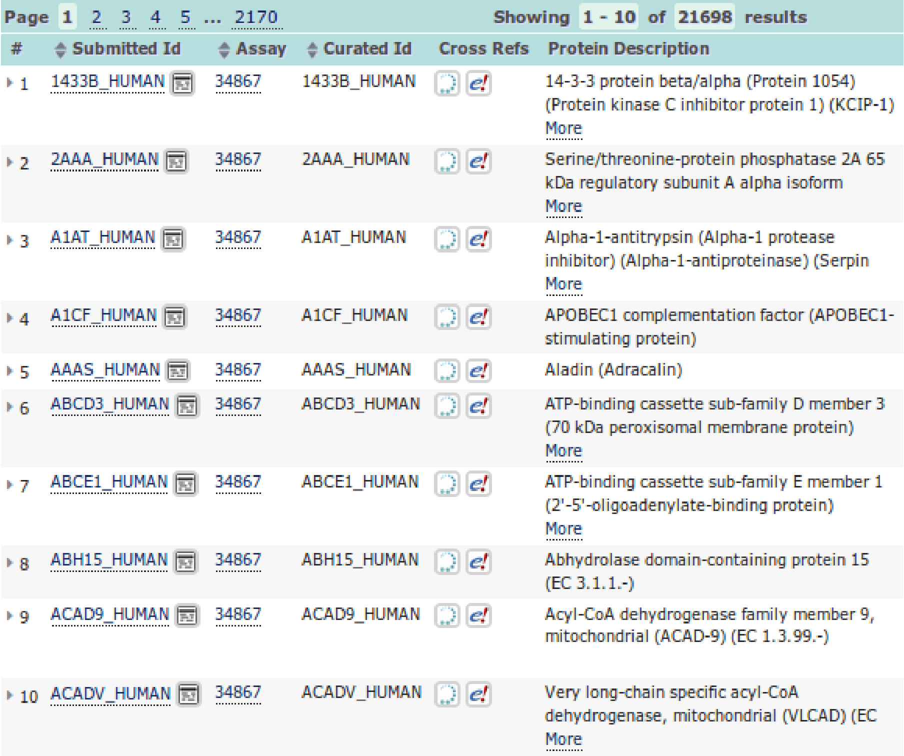
PRIDE: PRoteomics IDEntifications database



Name	Size	HTTP Download
20120814_EXQ4_JRW_SA_B21_0_L1.raw	2,713.674 MB	Download
20120814_EXQ4_JRW_SA_B21_0_L2.raw	2,074.9 MB	Download
20120814_EXQ4_JRW_SA_B21_0_L3.raw	2,273.445 MB	Download
20120814_EXQ4_JRW_SA_B21_0_L4.raw	2,479.706 MB	Download
20120814_EXQ4_JRW_SA_B21_0_T1.raw	2,468.011 MB	Download
20120814_EXQ4_JRW_SA_B21_0_T2.raw	2,059.587 MB	Download
20120814_EXQ4_JRW_SA_B21_1_L1.raw	2,577.641 MB	Download
20120814_EXQ4_JRW_SA_B21_1_L2.raw	2,117.355 MB	Download
20120814_EXQ4_JRW_SA_B21_1_L3.raw	2,186.073 MB	Download
20120814_EXQ4_JRW_SA_B21_1_L4.raw	2,431.84 MB	Download
20120814_EXQ4_JRW_SA_B21_1_T1.raw	2,332.173 MB	Download
20120814_EXQ4_JRW_SA_B21_1_T2.raw	2,116.427 MB	Download

Experiments can be accessed as raw machine files (gigabytes of data, actually) or in processed forms, displaying the assignment (e.g. via Mascot) of peptides and the corresponding proteins, as the result of a mass fingerprint.

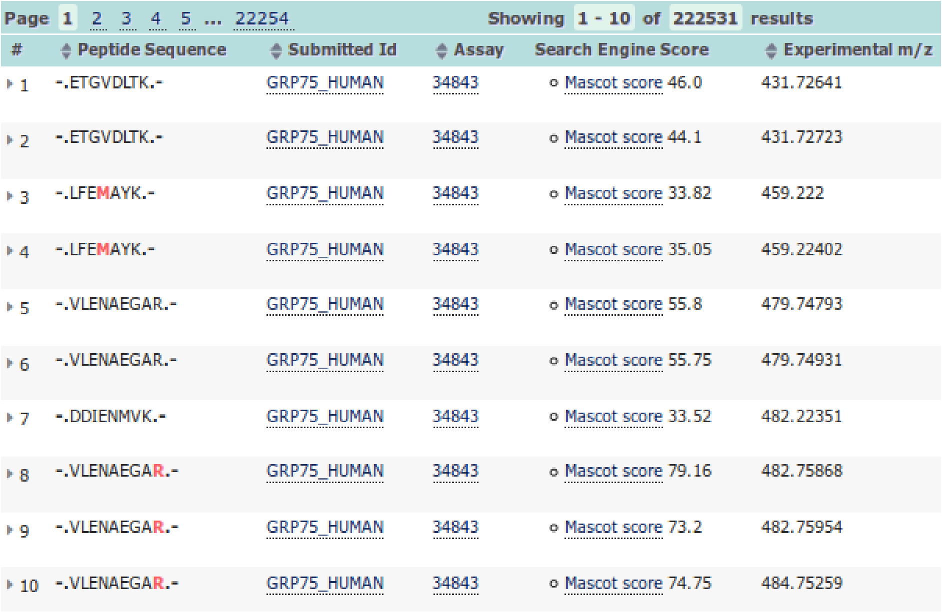
Protein view



Page 1 2 3 4 5 ... 2170 Showing 1 - 10 of 21698 results

#	Submitted Id	Assay	Curated Id	Cross Refs	Protein Description
1	1433B_HUMAN	34867	1433B_HUMAN		14-3-3 protein beta/alpha (Protein 1054) (Protein kinase C inhibitor protein 1) (KCIP-1) More
2	2AAA_HUMAN	34867	2AAA_HUMAN		Serine/threonine-protein phosphatase 2A 65 kDa regulatory subunit A alpha isoform More
3	A1AT_HUMAN	34867	A1AT_HUMAN		Alpha-1-antitrypsin (Alpha-1 protease inhibitor) (Alpha-1-antiproteinase) (Serpin More
4	A1CF_HUMAN	34867	A1CF_HUMAN		APOBEC1 complementation factor (APOBEC1-stimulating protein)
5	AAAS_HUMAN	34867	AAAS_HUMAN		Aladin (Adracalin)
6	ABCD3_HUMAN	34867	ABCD3_HUMAN		ATP-binding cassette sub-family D member 3 (70 kDa peroxisomal membrane protein) More
7	ABCE1_HUMAN	34867	ABCE1_HUMAN		ATP-binding cassette sub-family E member 1 (2'-5'-oligoadenylate-binding protein) More
8	ABH15_HUMAN	34867	ABH15_HUMAN		Abhydrolase domain-containing protein 15 (EC 3.1.1.-)
9	ACAD9_HUMAN	34867	ACAD9_HUMAN		Acyl-CoA dehydrogenase family member 9, mitochondrial (ACAD-9) (EC 1.3.99.-)
10	ACADV_HUMAN	34867	ACADV_HUMAN		Very long-chain specific acyl-CoA dehydrogenase, mitochondrial (VLCAD) (EC More

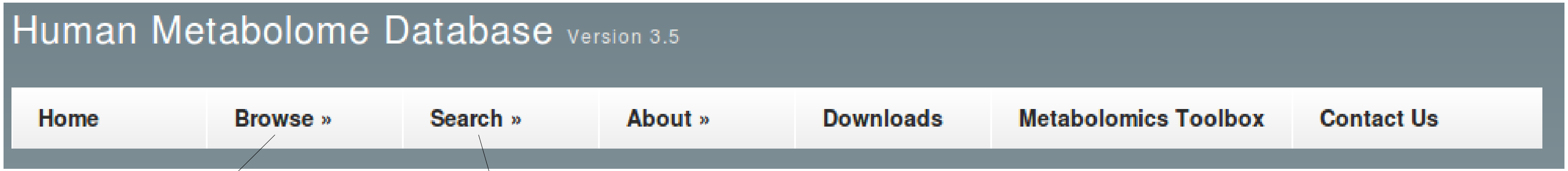
Peptide view



Page 1 2 3 4 5 ... 22254 Showing 1 - 10 of 222531 results

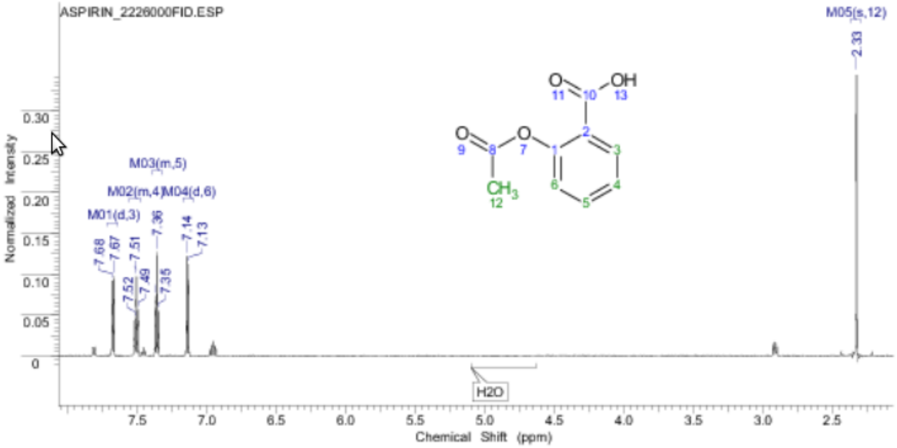
#	Peptide Sequence	Submitted Id	Assay	Search Engine Score	Experimental m/z
1	-.ETGVDLTK.-	GRP75_HUMAN	34843	Mascot score 46.0	431.72641
2	-.ETGVDLTK.-	GRP75_HUMAN	34843	Mascot score 44.1	431.72723
3	-.LFEMAYK.-	GRP75_HUMAN	34843	Mascot score 33.82	459.222
4	-.LFEMAYK.-	GRP75_HUMAN	34843	Mascot score 35.05	459.22402
5	-.VLENAEGAR.-	GRP75_HUMAN	34843	Mascot score 55.8	479.74793
6	-.VLENAEGAR.-	GRP75_HUMAN	34843	Mascot score 55.75	479.74931
7	-.DDIENMVK.-	GRP75_HUMAN	34843	Mascot score 33.52	482.22351
8	-.VLENAEGAR.-	GRP75_HUMAN	34843	Mascot score 79.16	482.75868
9	-.VLENAEGAR.-	GRP75_HUMAN	34843	Mascot score 73.2	482.75954
10	-.VLENAEGAR.-	GRP75_HUMAN	34843	Mascot score 74.75	484.75259

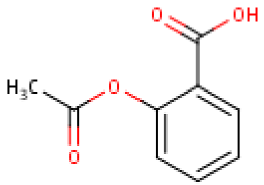
Human Metabolome Database


Human Metabolome Database Version 3.5

- Metabolite Browse
- Diseases
- Pathways
- Biofluids
- Classes
- Proteins
- Reactions
- Metabolite Library (HML)

- ChemQuery Structure Search
- Text Query
- Sequence Search
- Advanced Search
- MS Search
- MS/MS Search
- GC/MS Search
- 1D NMR Search
- 2D NMR Search

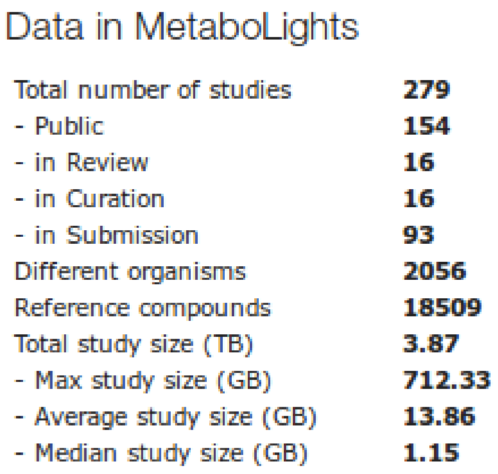




Showing metabocard for Aspirin (HMDB01879)

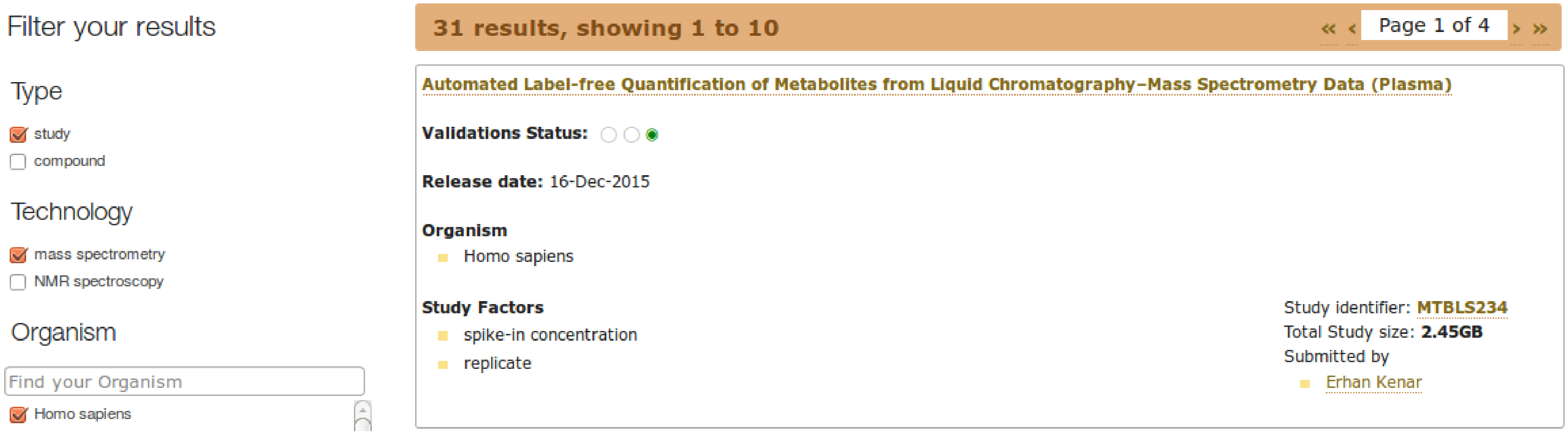
MetaboLights

MetaboLights is a general purpose, open access repository for metabolomics studies, their raw experimental data and associated metadata. As all EBI repositories, it is highly interactive and allow to browse data by



Data in MetaboLights

Total number of studies	**279**
- Public	**154**
- in Review	**16**
- in Curation	**16**
- in Submission	**93**
Different organisms	**2056**
Reference compounds	**18509**
Total study size (TB)	**3.87**
- Max study size (GB)	**712.33**
- Average study size (GB)	**13.86**
- Median study size (GB)	**1.15**



Filter your results

Type

☑ study
☐ compound

Technology

☑ mass spectrometry
☐ NMR spectroscopy

Organism

☑ Homo sapiens

31 results, showing 1 to 10

Automated Label-free Quantification of Metabolites from Liquid Chromatography–Mass Spectrometry Data (Plasma)

Validations Status: ○ ○ ●

Release date: 16-Dec-2015

Organism

- Homo sapiens

Study Factors

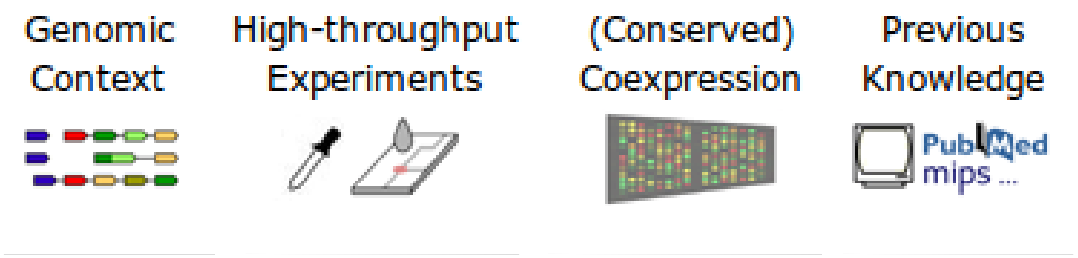
- spike-in concentration
- replicate

Study identifier: **MTBLS234**
Total Study size: **2.45GB**
Submitted by

- Erhan Kenar

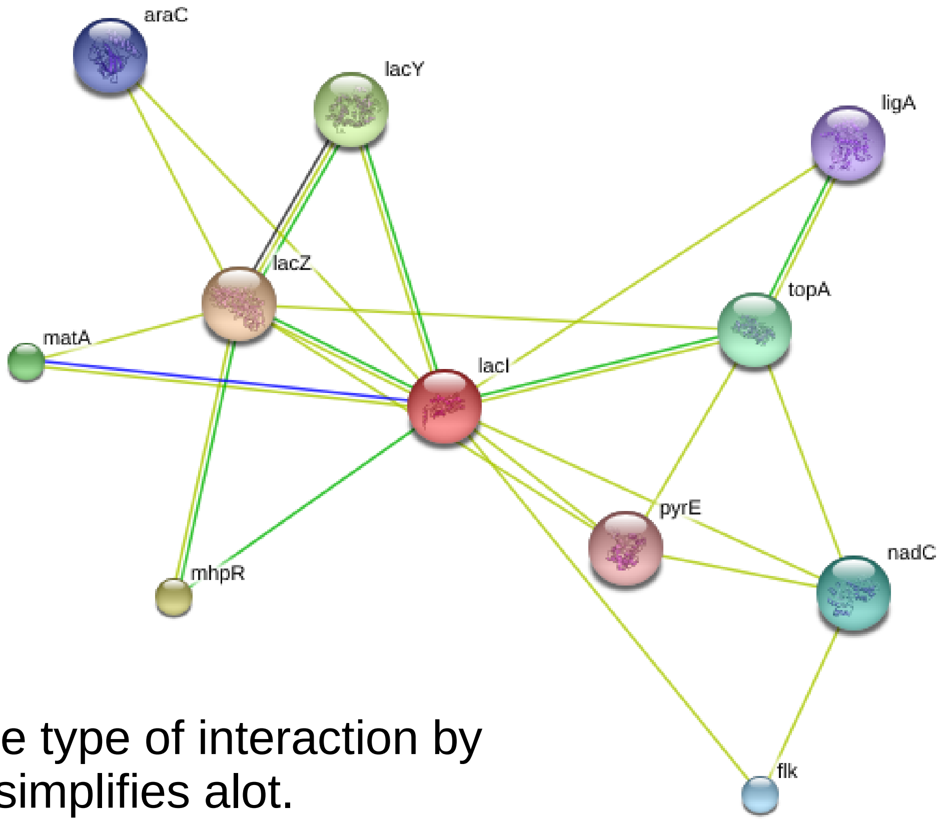
STRING: interaction network

The interactions include direct (physical) and indirect (functional) associations and are derived from four several experimental or computational sources.



Neighborhood
Gene Fusion
Cooccurrence
Coexpression
Experiments
Databases
Textmining
[Homology]

It sometimes risk to be dominated by computational interactions such as “co-citation” or “homology” that partly hinders the true biological interactions



The network can be restricted to some or one type of interaction by removing information sources. The network simplifies alot.

matteo.ramazzotti@unifi.it