

# **EBI web resources I: databases and tools**

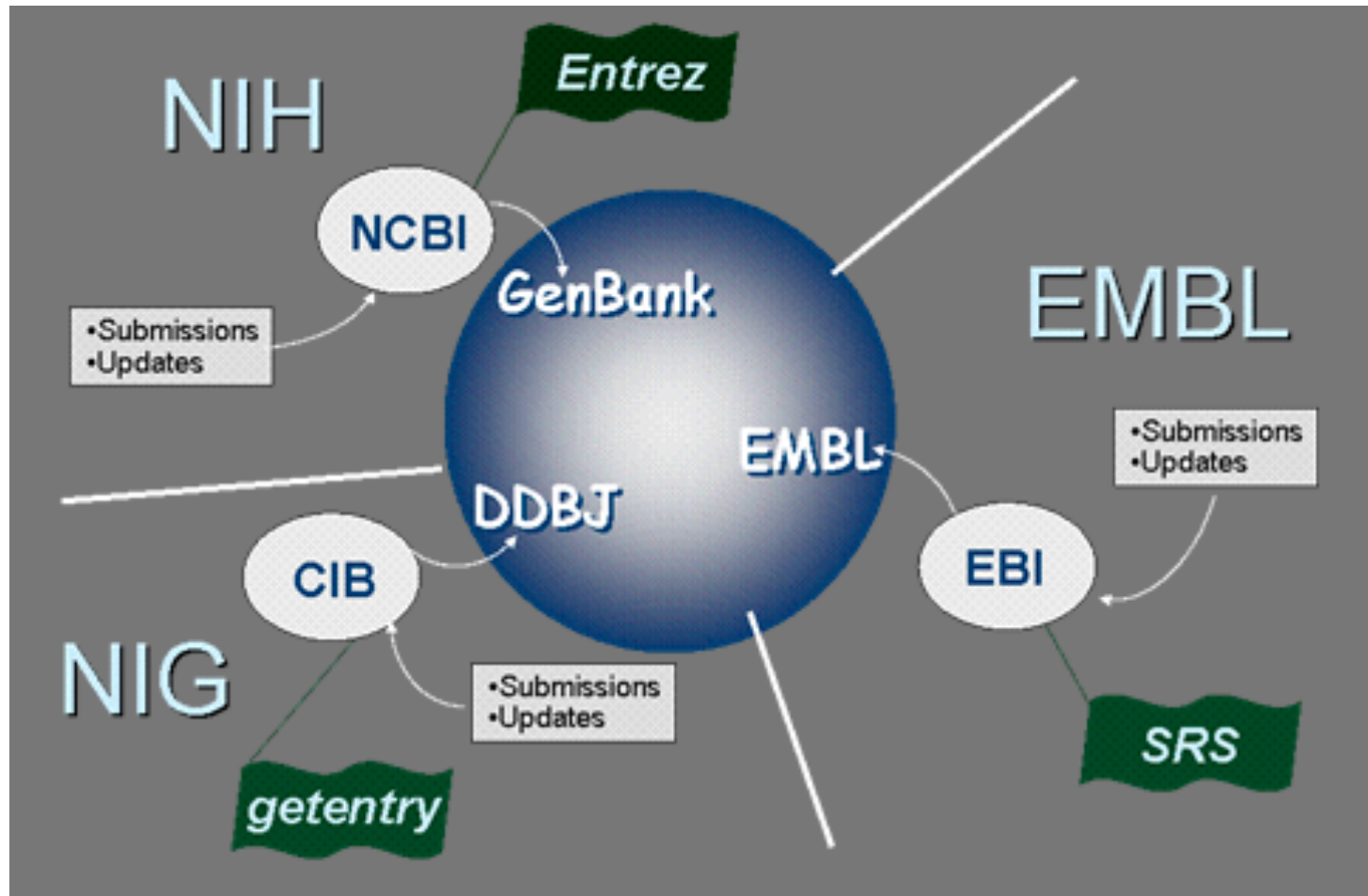
Yanbin Yin

# Outline

- Intro to EBI
- Databases and web tools
  - UniProt
  - Gene Ontology
- Hands on Practice

[MOST MATERIALS ARE FROM: http://www.ebi.ac.uk/training/online/course-list](http://www.ebi.ac.uk/training/online/course-list)

# Three international nucleotide sequence databases



# The European Bioinformatics Institute (EBI)



Created in 1992 as part of [European Molecular Biology Laboratory](#) (EMBL)

EMBL was created in 1974 and is a [molecular biology](#) research institution supported by 20 European countries and Australia

[Wellcome Trust Genome Campus, Hinxton, Cambridge, UK](#)  
Neighbor of [Wellcome Trust Sanger Institute](#)



# The European Bioinformatics Institute

EMBL-EBI

[Other EMBL locations >](#)

The home for big data in biology

At EMBL-EBI, we use bioinformatics — the science of storing, sharing and analysing biological data — to help people everywhere understand how living systems work, and what makes them change.

Find a gene, protein or chemical:



Examples: blast, keratin, bf11, Janet Thornton ...

Explore EMBL-EBI

[Services >](#)

[Research >](#)

[Training >](#)

[Industry >](#)

[ELIXIR >](#)

## Featured events

15 Sep 2016 - 15 Sep 2016

Understanding the mechanistic consequences of genetic variations in diseased and healthy individuals

Understanding the mechanistic consequences of genetic variations

15 Sep 2016 - 15 Sep 2016

Mitochondrial control of gene expression









Research Tea Talk - Stegle group

19 Sep 2016 - 19 Sep 2016

Cambridge New Therapeutics Forum

The Cambridge New Therapeutics Forum (CamNTF) welcomes all scientists in Cambridge and the local area, including those...

## Popular

-  [Services](#)
-  [Data submission](#)
-  [Research](#)
-  [Training](#)
-  [News](#)
-  [Jobs](#)
-  [Visit us](#)
-  [EMBL](#)

# Research groups in EBI

	Group/team leader	Area of research	
<b>Genomes</b>	<a href="#">Ewan Birney</a>	Algorithmic methods for genome analysis	InterPro
	<a href="#">Paul Flicek</a>	Vertebrate genomics	
	<a href="#">Nick Goldman</a>	Evolutionary tools for sequence analysis	
<b>Transcriptomes</b>	<a href="#">Alvis Brazma</a>	Functional genomics	miRBase
	<a href="#">Anton Enright</a>	Functional genomics and analysis of small RNA function	
	<a href="#">John Marioni</a>	Computational and evolutionary genomics	
	<a href="#">Oliver Stegle</a>	Statistical genomics and systems genetics	
<b>Proteins</b>	<b>Janet Thornton</b>	Computational biology of proteins: structure, function and evolution	
	<a href="#">Rolf Apweiler</a>	Protein sequence analysis and functional annotation	UniProt
	<a href="#">Gerard Kleywegt</a>	Structural validation of proteins; protein-ligand interactions	
<b>Pathways and systems</b>	<a href="#">Nicolas Le Novère</a>	Computational systems neurobiology	
	<a href="#">Nick Luscombe</a>	Genomics and regulatory systems	
	<a href="#">Paul Bertone</a>	Pluripotency, reprogramming and differentiation	
	<a href="#">Julio Saez-Rodriguez</a>	Systems biomedicine	
<b>Literature</b>	<a href="#">Dietrich Rebholz-Schuhmann</a>	Literature analysis and semantic data integration in life science research	
<b>Chemistry</b>	<a href="#">Christoph Steinbeck</a>	Cheminformatics and metabolism	
	<a href="#">John Overington</a>	Chemogenomics and drug discovery	

# Major databases in EBI

GenBank	<a href="#">EMBL-Bank</a> (DNA and RNA sequences)
Genome MapView	<a href="#">Ensembl</a> (genomes)
GEO	<a href="#">ArrayExpress</a> (microarray-based gene-expression data)
nr (GenPept)	<a href="#">UniProt</a> (protein sequences)
CDD	<a href="#">InterPro</a> (protein families, domains and motifs)
MMDB	<a href="#">PDBe</a> (macromolecular structures)

Others, such as

[IntAct](#) (protein–protein interactions)

[Reactome](#) (pathways)

[ChEBI](#) (small molecules)

[IntEnz](#) (enzyme classification)

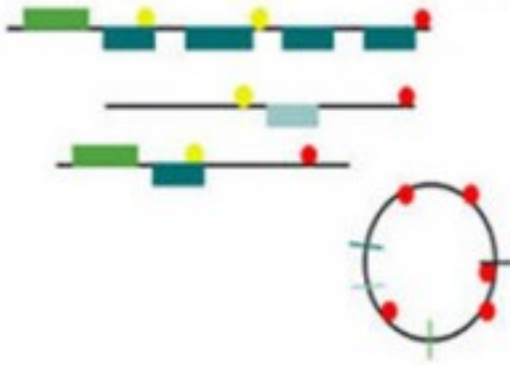
[GO](#) (gene ontology)

Swiss Institute of Bioinformatics

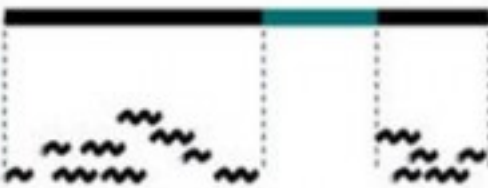
Sanger Institute



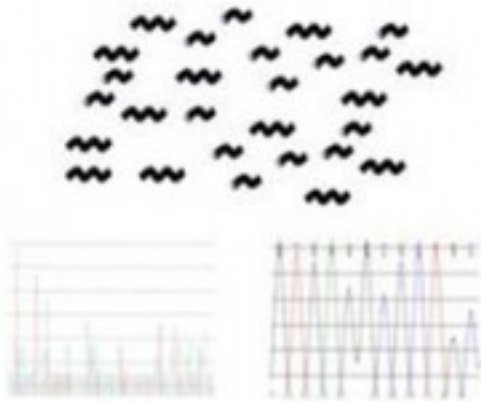
Feature  
annotation



Assembly  
information



Sequencing  
and sampling  
information



1) **EMBL-Bank**

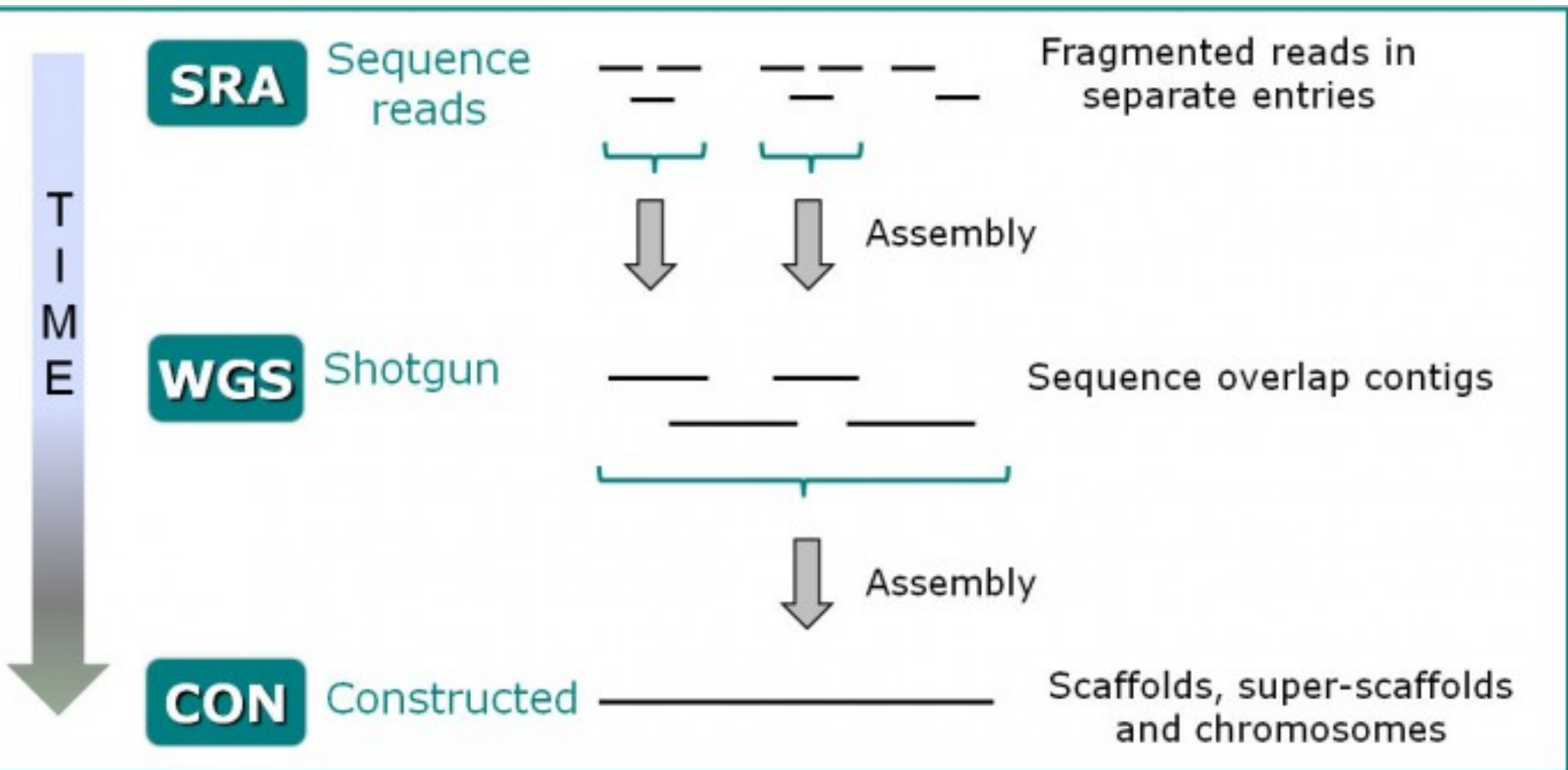
2) **Sequence Read  
Archive**

3) **Trace Archive**  
chromatograms





Sequence might first enter ENA as **SRA** (Sequence Read Archive) **fragmented** sequence reads; it might be re-submitted as **assembled WGS** (Whole Genome Shotgun) sequence overlap **contigs**; it might be re-submitted again with **further assembly** as **CON** (Constructed) sequence entries, with the older WGS entries being consigned to the Sequence Version Archive



Data is first split into **classes**, then it is split into intersecting slices by **taxonomy**

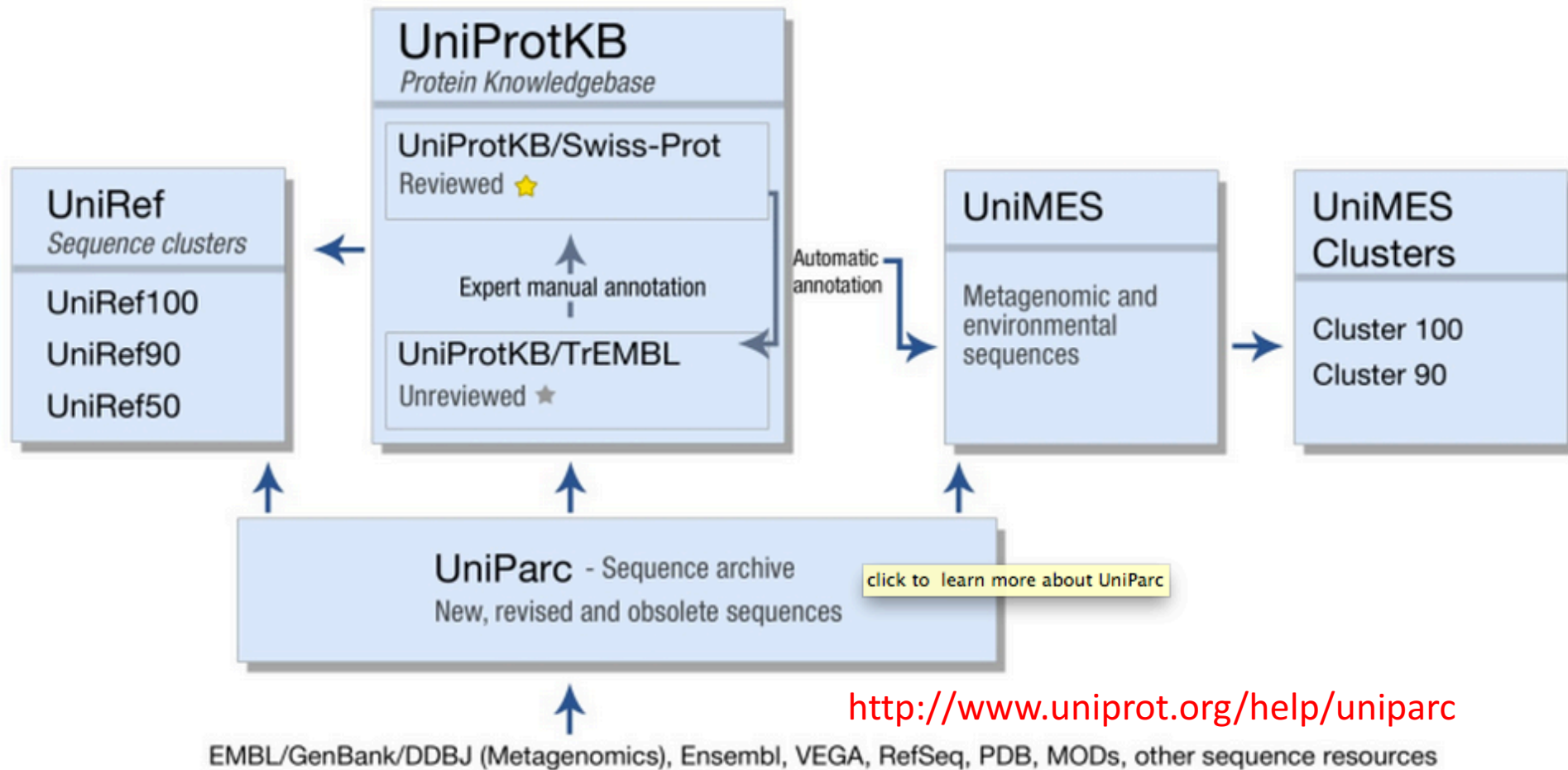
## EMBL-Bank:

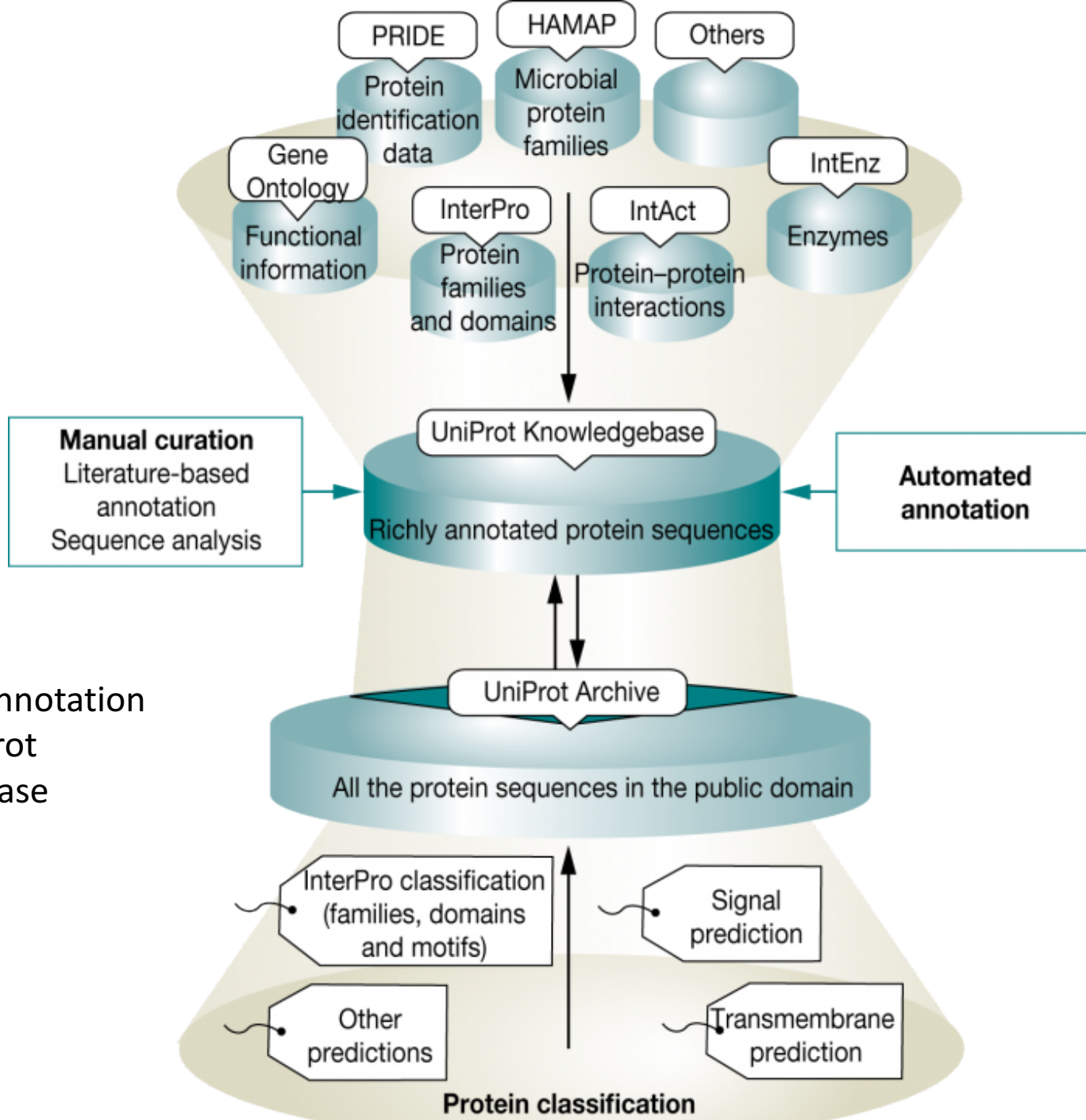
### Data classes

	CON	EST	GSS	HTC	HTG	MSA	PAT	STS	STD	TSA	WGA
A											
HUM											
MUS											
ROD											
MAM											
VRT											
FUN											
INV											
⋮											

### Taxonomic Divisions

# UniProt





Sources of annotation for the UniProt Knowledgebase

Curation generation

<http://cys.bios.niu.edu/yyin/teach/PBB/Bioinformatics%20Curation%20generation.pdf>

Life as a **Scientific Curator**

<http://www.ebi.ac.uk/about/jobs/career-profiles/scientific-curator>

Scientific Database Curator job : Cambridge, United Kingdom

<http://www.nature.com/naturejobs/science/jobs/589083-hgnc-gene-nomenclature-advisor>

# Hands on practice 1: UniProt

[www.uniprot.org](http://www.uniprot.org)

[http://www.uniprot.org/docs/uniprot\\_flyer.pdf](http://www.uniprot.org/docs/uniprot_flyer.pdf)

<http://www.uniprot.org/help/about>

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein

**UniProtKB**  
UniProt Knowledgebase


Swiss-Prot  
(549,215)  
Manually annotated and reviewed.

TrEMBL  
(50,825,784)  
Automatically annotated and not reviewed.


**UniRef**  
Sequence clusters



**UniParc**  
Sequence archive



**Proteomes**



**Supporting data**

Literature citations 	Taxonomy 	Subcellular locations 
Cross-ref. databases 	Diseases 	Keywords 

**News**

Forthcoming  
Planned changes

UniProt releases  
Life (and death) of protein  
variation files

UniProt releases  
Pseudo-allergens  
access to UniProt  
of human variation

News archive

## Getting started

 **Text search**

Our basic text search allows you to search all the resources available

 **BLAST**

Find regions of similarity between your sequences




## UniProt data

 **Download latest release**

Get the UniProt data

 **Statistics**

View Swiss-Prot and TrEMBL statistics

 **How to cite us**

## Protein structure



We are going to do ID mapping

UniProtKB  Advanced

BLAST Align **Retrieve/ID Mapping** Help Contact

Welcome to the new UniProt website! We hope you enjoy the new design. If you're not quite ready yet, you can still [go back to the old site](#).

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

<b>UniProtKB</b> Swiss-Prot (546,238) Manually annotated and reviewed. TrEMBL (82,126,897) Automatically annotated and not reviewed.	<b>UniRef</b> Sequence clusters	<b>UniParc</b> Sequence archive	<b>Proteomes</b>	<b>News</b> Ubiquitin caught at its own game   New human variant types available on the FTP site <a href="#">UniProt release 2014_08</a> Lark or owl? PER3 is the answer   Cross-references to CCDS and GeneReviews   UniParc cross-references with protein and gene names   FTP ... <a href="#">UniProt release 2014_07</a> <a href="#">News archive</a>
<b>Supporting data</b>				
Literature citations	Taxonomy	Subcellular locations		
Cross-ref. databases	Diseases	Keywords		

## Getting started



### Text search

Our basic text search allows you to search all the resources available

### BLAST

Find regions of similarity between your sequences

### Sequence alignments

Align two or more protein sequences using the Clustal Omega program

## UniProt data

### Download latest release

Get the UniProt data

### Statistics

View Swiss-Prot and TrEMBL statistics

### Forthcoming changes

Planned changes for the UniProt knowledgebase

### Submit your data

## Protein spotlight



### Two's Company

August 2014

Pairing up is sometimes paramount to life. On the molecular scale, dimerization in our bodies is at the heart of many fundamental biological processes, such as the transduction of signals from the outside of a cell to the inside for instance. Split two molecules apart and,



<http://cys.bios.niu.edu/yyin/teach/PBB/at-id.txt>

## 1. Provide your identifiers

At2g37040  
At3g53260  
At5g04230  
At3g10340  
At2g30490  
At1g51680  
At3g21240  
At3g21230  
At1g65060

OR upload your own file:  No file chosen

Run in a new window.

## 2. Select options

Choose Araport here and UniProtKB here

From

To

Araport

UniProtKB

# These are UniProt IDs

61 out of 61 TAIR identifiers were successfully mapped to 79 UniProtKB IDs.

## Results

### Filter by<sup>i</sup>

Reviewed (40)  
Swiss-Prot

Unreviewed (39)  
TrEMBL

Popular organisms  
A. thaliana (79)

### View by

- Taxonomy
- Keywords
- Gene Ontology
- Enzyme class
- Pathway

### UniRef

Your results in sequence clusters with identity of: 100%, 90% or 50%

Columns BLAST Align Download Add to basket

1 to 25 of 79 Show 25

Your list:...	Entry	Entry name	Protein names	Gene names	Organism	Length
<input type="checkbox"/> AT2G37040	P35510	PAL1_ARATH	Phenylalanine ammonia-lyase 1	PAL1, At2g37040, T1J8.22	Arabidopsis thaliana (Mouse-ear cress)	725
<input type="checkbox"/> AT3G53260	P45724	PAL2_ARATH	Phenylalanine ammonia-lyase 2	PAL2, At3g53260, T4D2.190	Arabidopsis thaliana (Mouse-ear cress)	717
<input type="checkbox"/> AT5G04230	F4JW69	F4JW69_ARATH	Phenylalanine ammonia-lyase	PAL3, At5g04230	Arabidopsis thaliana (Mouse-ear cress)	698
<input type="checkbox"/> AT5G04230	P45725	PAL3_ARATH	Phenylalanine ammonia-lyase 3	PAL3, At5g04230, F21E1_150	Arabidopsis thaliana (Mouse-ear cress)	694
<input type="checkbox"/> AT3G10340	Q9SS45	PAL4_ARATH	Phenylalanine ammonia-lyase 4	PAL4, At3g10340, F1J82.6	Arabidopsis thaliana (Mouse-ear cress)	707

## Select the PAL proteins and align them

61 out of 61 TAIR identifiers were successfully mapped to 79 UniProtKB IDs.

Basket ▾

Columns BLAST **Align** Download Add to basket

◀ 1 to 25 of 79 ▶ Show 25 ▾

<input type="checkbox"/>	Your list:...APA6C	Entry	Entry name		Protein names	Gene names	Organism	Length	
<input checked="" type="checkbox"/>	At2G37040	P35510	PAL1_ARATH		<b>Phenylalanine ammonia-lyase 1</b>	<b>PAL1</b> , At2g37040, T1J8.22	Arabidopsis thaliana (Mouse-ear cress)	725	
<input checked="" type="checkbox"/>	At3G53260	P45724	PAL2_ARATH		<b>Phenylalanine ammonia-lyase 2</b>	<b>PAL2</b> , At3g53260, T4D2.190	Arabidopsis thaliana (Mouse-ear cress)	717	
<input checked="" type="checkbox"/>	At5G04230	F4JW69	F4JW69_ARATH		<b>Phenylalanine ammonia-lyase</b>	<b>PAL3</b> , At5g04230	Arabidopsis thaliana (Mouse-ear cress)	698	
<input checked="" type="checkbox"/>	At5G04230	P45725	PAL3_ARATH		<b>Phenylalanine ammonia-lyase 3</b>	<b>PAL3</b> , At5g04230, F21E1_150	Arabidopsis thaliana (Mouse-ear cress)	694	
<input checked="" type="checkbox"/>	At3G10340	Q9SS45	PAL4_ARATH		<b>Phenylalanine ammonia-lyase 4</b>	<b>PAL4</b> , At3g10340, F14P13.6	Arabidopsis thaliana (Mouse-ear cress)	707	

Clustal omega program will be called to align the selected protein seqs  
May take 1 min to finish

Toggle these options on will add colors in the alignment

- ALIGNMENT
- TREE
- RESULT INFO

## Alignment

 How to print an alignment in color

## Highlight

### Annotation

- Sequence conflict
- Binding site
- Chain
- Active site
- Modified residue
- Cross-link

### Amino acid properties

- Similarity
- Hydrophobic
- Negative
- Positive
- Aliphatic
- Tiny
- Aromatic
- Charged
- Small
- Polar
- Big
- Serine Threonine

P35510	PAL1_ARATH	1	MEINGAHKSNNGGVDAMLCGGDIKTKNMVI--NAEDPLNWGAAAEQMKGSHLDEVKRMVA	58
P45724	PAL2_ARATH	1	-----MDQIEAMLCGGGEKTKVAVTTKTLADPLNWGLAADQMKGSHLDEVKKMVE	50
F4JW69	F4JW69_ARATH	1	-----MEFR---QPNATALSDPLNWNVAAEALKGSHLEEVKKMVK	37
P45725	PAL3_ARATH	1	-----MEFR---QPNATALSDPLNWNVAAEALKGSHLEEVKKMVK	37
Q9SS45	PAL4_ARATH	1	-----MELCNQNNHITAVSGDPLNWNATAEALKGSHLDEVKRMVK	40
			***** :* :*****:***:**	
P35510	PAL1_ARATH	59	EFRKPVVNLGGETLTIGQVAAISTIGNSVKVELSETARAGVNASDQVWVMSMNKGTDSYG	118
P45724	PAL2_ARATH	51	EYRRPVVNLGGETLTIGQVAAISTVGGSVKVELAETSRAQVKASSDQVWVMSMNKGTDSYG	110
F4JW69	F4JW69_ARATH	38	DYRKGTVQLGGETLTIGQVAVAS--GGPTVELSEEARGGVKASSDQVWVMSMNRDQDQY	95
P45725	PAL3_ARATH	38	DYRKGTVQLGGETLTIGQVAVAS--GGPTVELSEEARGGVKASSDQVWVMSMNRDQDQY	95
Q9SS45	PAL4_ARATH	41	EYRKEAVKLGGETLTIGQVAAVARGGGSTVELAEARAGVKASSEWVMSMNRGTDSYG	100
			::*: .*:*****:***:** . .***:* :*.***:***:*****:***:**	
P35510	PAL1_ARATH	119	VTTGFGATSHRRTKNGVALQKELIRFLNAGIFGSTK---ETSHTLPHSATRAAMLVRINT	175
P45724	PAL2_ARATH	111	VTTGFGATSHRRTKNGTALQTELIRFLNAGIFGNTK---ETCHTLQSATRAAMLVRVNT	167
F4JW69	F4JW69_ARATH	96	ITTFGSSSSRRRTDQGAALQKELIRYLNAGIFATGNEDDDRSNTLPRPATRAAMLIRVNT	155
P45725	PAL3_ARATH	96	ITTFGSSSSRRRTDQGAALQKELIRYLNAGIFATGNEDDDRSNTLPRPATRAAMLIRVNT	155
Q9SS45	PAL4_ARATH	101	VTTGFGATSHRRTKQGGALQNELIRFLNAGIFGPGAG--DTSHTLPKPTTRAAMLVRVNT	158
			:*****:***:**.* ***.***:*****. : ..***: *****:***:**	
P35510	PAL1_ARATH	176	LLQGFSGIRFEILEAITSFLNNNITPSLPLRGTITASGDLVPLSYIAGLLTGRPNSKATG	235
P45724	PAL2_ARATH	168	LLQYSGIRFEILEAITSLLNHNISPSLPLRGTITASGDLVPLSYIAGLLTGRPNSKATG	227
F4JW69	F4JW69_ARATH	156	LLQYSGIRFEILEAITLLNCKITPLPLRGTITASGDLVPLSYIAGFLIGRPNSRSVG	215
P45725	PAL3_ARATH	156	LLQYSGIRFEILEAITLLNCKITPLPLRGTITASGDLVPLSYIAGFLIGRPNSRSVG	215
Q9SS45	PAL4_ARATH	159	LLQYSGIRFEILEAITKLLNHEITPCLPLRGTITASGDLVPLSYIAGLLTGRPNSKAVG	218
			****:*****.***:**.* *****:*****:***:**	
P35510	PAL1_ARATH	236	PNGEALTAEEAFKLAGISSGFFDLQPKEGLALVNGTAVGSGMAMVLFETNVLSVLAEIL	295
P45724	PAL2_ARATH	228	PDGESLTAKEAFEKAGISTGFFDLQPKEGLALVNGTAVGSGMAMVLFANVQAVLAEVL	287
F4JW69	F4JW69_ARATH	216	PSGEILTAEAFKLAGVS--SFFELRPKEGLALVNGTAVGSALASTVLYDANILVVFSEVA	274
P45725	PAL3_ARATH	216	PSGEILTAEAFKLAGVS--SFFELRPKEGLALVNGTAVGSALASTVLYDANILVVFSEVA	274
Q9SS45	PAL4_ARATH	219	PSGETLTASEAFKLAGVS--SFFELQPKEGLALVNGTAVGSGLASTVLYDANILAVLSEVM	277
			*.*** ** ***:***:* .***:*****:***:** **::** :***:**	


Go back to the protein list page  
Selecting one protein will enable the BLAST button


## Results

61 out of 61 TAIR identifiers were successfully mapped to 79 UniProtKB IDs.

 Basket ▾

### Filter by<sup>i</sup>

 Reviewed (40)  
Swiss-Prot





 Unreviewed (39)  
TrEMBL

Popular organisms  
*A. thaliana* (79)

### View by

Taxonomy  
Keywords  
Gene Ontology  
Enzyme class  
Pathway

Columns BLAST Align Download Add to basket 1 to 25 of 79 Show 25

<input type="checkbox"/>	Your list	Entry name	Protein names	Gene names	Organism	Length
<input checked="" type="checkbox"/>	AT2G3704	P45724 PAL1_ARATH	 Phenylalanine ammonia-lyase 1	<b>PAL1</b> , At2g37040, T1J8.22	Arabidopsis thaliana (Mouse-ear cress)	725
<input type="checkbox"/>	AT3G53260	P45724 PAL2_ARATH	 Phenylalanine ammonia-lyase 2	<b>PAL2</b> , At3g53260, T4D2.190	Arabidopsis thaliana (Mouse-ear cress)	717
<input type="checkbox"/>	AT5G04230	F4JW69 F4JW69_ARATH	 Phenylalanine ammonia-lyase	<b>PAL3</b> , At5g04230	Arabidopsis thaliana (Mouse-ear cress)	698
<input type="checkbox"/>	AT5G04230	P45725 PAL3_ARATH	 Phenylalanine ammonia-lyase 3	<b>PAL3</b> , At5g04230,	Arabidopsis thaliana	694

Choose advanced will allow to change BLAST parameters

Here you can make changes

### How to use this tool

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences, which can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

1. Enter either a protein or nucleotide sequence or a UniProt identifier (e.g.P00750 or A4\_HUMAN or UPI0000000001) into the form field.
2. Optionally, change the program parameters with the dropdown menus under the form.
3. Click the *Run BLAST* button.

# BLAST

P35510

Target database

UniProtKB

E-Threshold

10

Matrix

Auto

Filtering

None

Gapped

yes

Hits

250

Run Blast in a separate window.

[Run BLAST](#)

We are going to search UniProt proteomes for human protein set

Click on Advanced you will see a pop-out window

The screenshot displays the UniProt website interface. At the top, the UniProt logo is on the left, and a search bar is on the right. The search bar has 'Proteomes' selected in a dropdown menu. To the right of the search bar is an 'Advanced' button, which is highlighted with a red box. A red arrow points from the text 'Click on Advanced you will see a pop-out window' to this button. Below the search bar is a navigation bar with links for 'BLAST', 'Align', 'Upload Lists', 'Help', and 'Contact'. A yellow banner below the navigation bar reads: 'Welcome to the new UniProt website! We hope you enjoy the new design. If you're not quite ready yet, you can still go back to the old site.' Below the banner is a paragraph: 'The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.' The main content area features several sections: 'UniProtKB' (Swiss-Prot: 546,238; TrEMBL: 82,1...), 'UniRef' (Sequence clusters), 'UniParc' (Sequence archive), and 'Proteomes'. A 'Supporting data' section includes 'Literature citations', 'Taxonomy', and 'Subcellular locations'. A 'News' section on the right contains two articles. A pop-out window titled 'Searching in Proteomes' is open, showing search options. The 'Organism [OS]' dropdown is set to 'Human [9606]'. The search type is set to 'All'. A red arrow points from the text 'Here you can specify search terms' to the search input fields in this window.

Here you can specify search terms

### What are proteome sets?

A proteome consists of the set of proteins thought to be expressed by an organism whose genome has been completely sequenced. [Help](#)

### What are reference proteome sets?

Some proteomes have been (manually and algorithmically) selected as **reference proteomes**. They cover well-studied model organisms and other organisms of interest for biomedical research and phylogeny. [Help](#)

[Tutorials and Videos](#) [Downloads](#)

Click here to get help

BLAST Align Upload Lists

Help Contact

Show help for Proteomes

## Results

### Filter by

1 Reference proteomes

### Map To

UniProtKB

### Demo

[Help video](#)

Repeat search in UniProtKB (140,991)

[Basket](#)

Download

1 to 1 of 1 Show 25

Proteome ID	Organism	Last modified	Protein Count
UP000005640	Homo sapiens (Human)	2014-07-09	68049

1 to 1 of 1 Show 25

## Results

### Filter by

Reviewed (20,187)

Swiss-Prot

Unreviewed (47,862)

TrEMBL

### Popular organisms

Human (68,049)

### Proteomes

UP000005640 (68,049)

### View by

Taxonomy

Keywords

Gene Ontology

Enzyme class

Pathway

### UniRef

Your results in sequence clusters with identity of: 100%, 90% or 50%

### Demo

[Help video](#)

Click here to open a new page

Columns BLAST Align Download Add to basket

1 to 25 of 68,049 Show 25

Entry	Entry name	Protein names	Gene names	Organism	Length
P31946	1433B_HUMAN	14-3-3 protein beta/alpha	YWHAB	Homo sapiens (Human)	246
P62258	1433E_HUMAN	14-3-3 protein epsilon	YWHAE	Homo sapiens (Human)	255
Q04917	1433F_HUMAN	14-3-3 protein eta	YWHAH, YWHA1	Homo sapiens (Human)	246
P61981	1433G_HUMAN	14-3-3 protein gamma	YWHAG	Homo sapiens (Human)	247
P31947	1433S_HUMAN	14-3-3 protein sigma	SFN, HME1	Homo sapiens (Human)	248
P27348	1433T_HUMAN	14-3-3 protein theta	YWHAQ	Homo sapiens (Human)	245
P63104	1433Z_HUMAN	14-3-3 protein zeta/delta	YWHAZ	Homo sapiens (Human)	245
P30443	1A01_HUMAN	HLA class I histocompatibility anti...	HLA-A, HLAA	Homo sapiens (Human)	365
P01892	1A02_HUMAN	HLA class I histocompatibility anti...	HLA-A, HLAA	Homo sapiens (Human)	365
P04439	1A03_HUMAN	HLA class I histocompatibility anti...	HLA-A, HLAA	Homo sapiens (Human)	365
P13746	1A11_HUMAN	HLA class I histocompatibility anti...	HLA-A, HLAA	Homo sapiens (Human)	365
Q960U6	1A11.1_HUMAN	1-aminocyclopropane-1-carboxylate S...	ACC5, PHACS	Homo sapiens (Human)	501



# Gene Ontology

<http://geneontology.org/page/documentation>

The Gene Ontology (GO) project is a collaborative effort to address the need for **consistent descriptions of gene products** in different databases

The project began as a collaboration between three model organism databases, [FlyBase](#) (*Drosophila*), the [Saccharomyces Genome Database](#) (SGD) and the [Mouse Genome Database](#) (MGD), in 1998

Three structured **controlled vocabularies** (ontologies) that describe gene products in terms of their associated **biological processes**, **cellular components** and **molecular functions** in a **species-independent** manner.

There are three separate aspects to this effort:

- 1, the development and maintenance of the **ontologies** themselves;
- 2, the **annotation** of gene products, which entails **making associations between the ontologies and the genes and gene products** in the collaborating databases; and
- 3, development of **tools** that facilitate the creation, maintenance and use of ontologies.

# The scope of GO

Gene Ontology covers three domains:

**cellular component**, the parts of a cell or its extracellular environment;

**molecular function**, the elemental activities of a gene product at the molecular level, such as binding or catalysis;

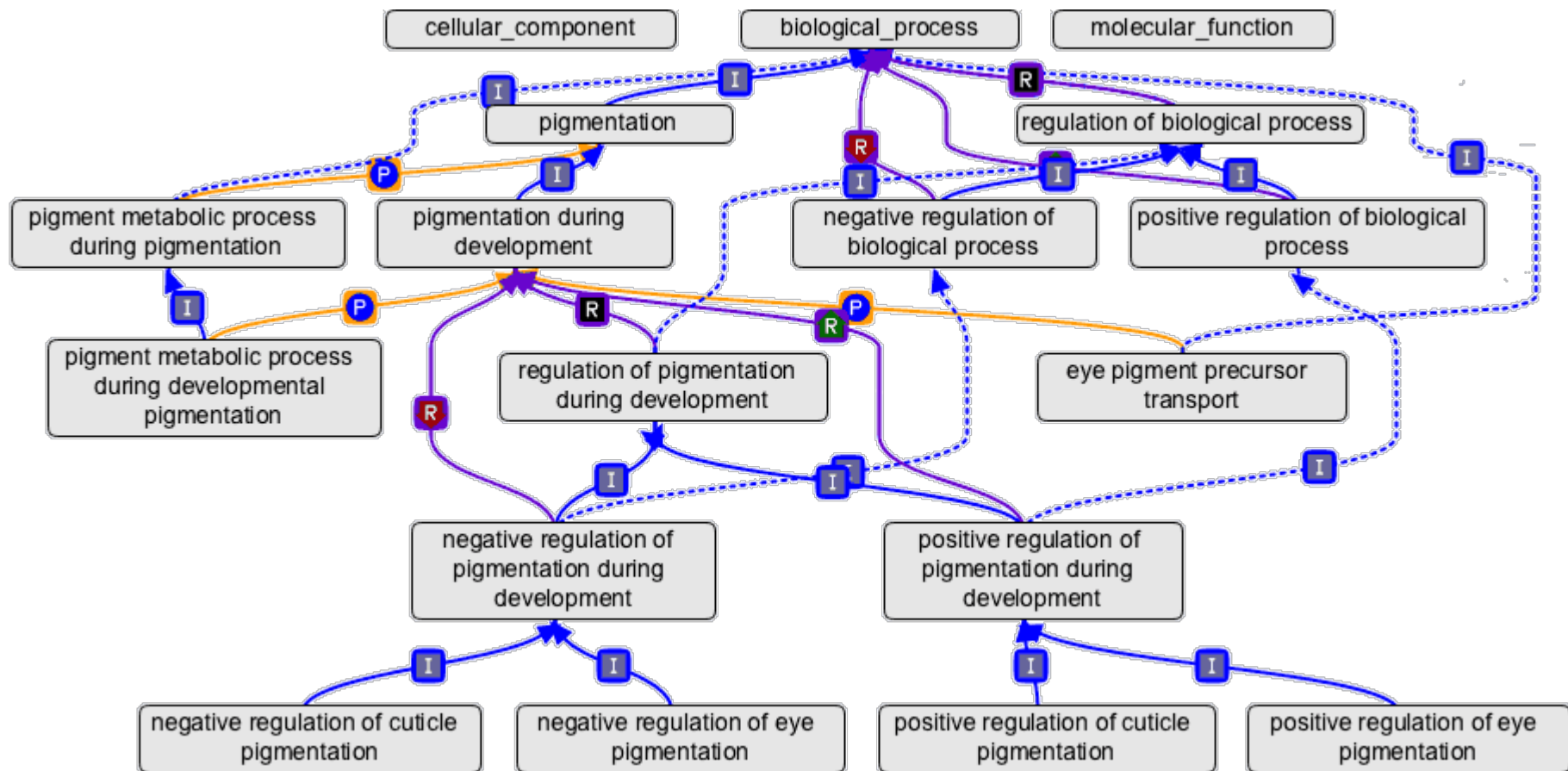
**biological process**, operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms

GO is **not a database of gene sequences**, nor a catalog of gene products. Rather, GO **describes how gene products behave** in a cellular context.

GO is not a dictated standard, mandating nomenclature across databases. Groups participate because of self-interest, and cooperate to arrive at a **consensus**.

GO is not a way to unify biological databases (i.e. GO is not a 'federated solution'). Sharing vocabulary is a step towards unification, but is not, in itself, sufficient.

The structure of GO can be described in terms of a graph, where each GO term is a node, and the relationships between the terms are edges between the nodes. GO is loosely hierarchical, with 'child' terms being more specialized than their 'parent' terms, but unlike a strict hierarchy, a term may have more than one parent term



```
id: GO:0000016
name: lactase activity namespace: molecular_function
def: "Catalysis of the reaction: lactose + H2O = D-glucose + D-galactose."
[EC:3.2.1.108]
synonym: "lactase-phlorizin hydrolase activity" BROAD [EC:3.2.1.108]
synonym: "lactose galactohydrolase activity" EXACT [EC:3.2.1.108]
xref: EC:3.2.1.108
xref: MetaCyc:LACTASE-RXN
xref: Reactome:20536
is_a: GO:0004553 ! hydrolase activity, hydrolyzing O-glycosyl compounds
```

## What can I do with GO?

### What can I do with GO?

One of the most popular uses of **GO** is to find significant shared GO terms (or parents of those GO terms) that are annotated to **genes** in a particular query set (e.g. a set of genes that are overexpressed in a microarray experiment). This process helps you to find out what those genes may have in common and is known as a **GO enrichment analysis**.

GO is also used for purposes as diverse as:

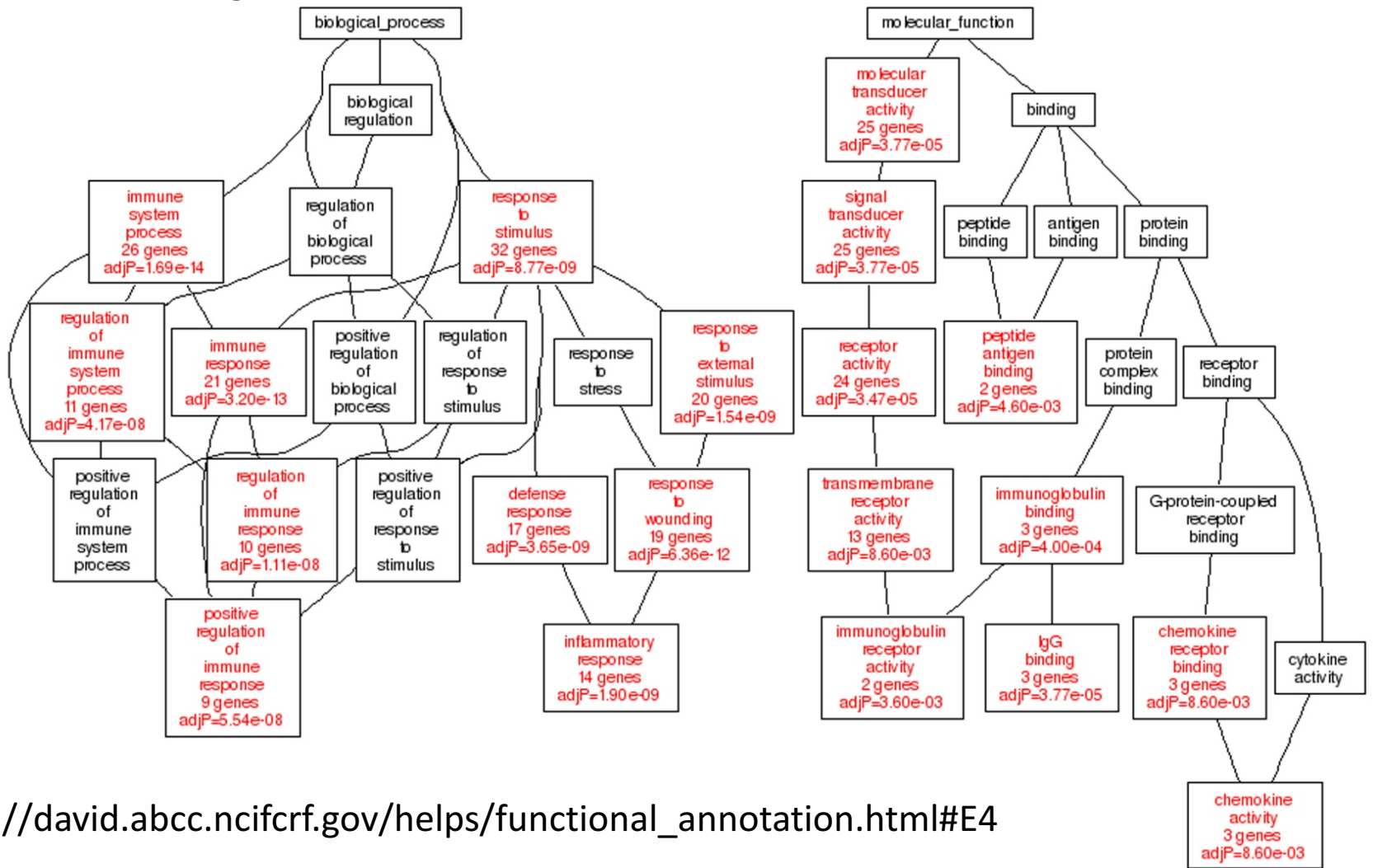
- integrating proteomic information from different organisms;
- assigning functions to protein domains;
- finding functional similarities in genes that are overexpressed or underexpressed in diseases and as we age;
- analysing groups of genes that are co-expressed during development;
- developing automated ways of deriving information about gene function from the literature;
- verifying models of genetic, metabolic and product interaction networks.

The [GO tools web page](#) lists the tools that you can use to analyse the data from GO.

<http://www.ebi.ac.uk/training/online/course/go-quick-tour/what-can-i-do-go>

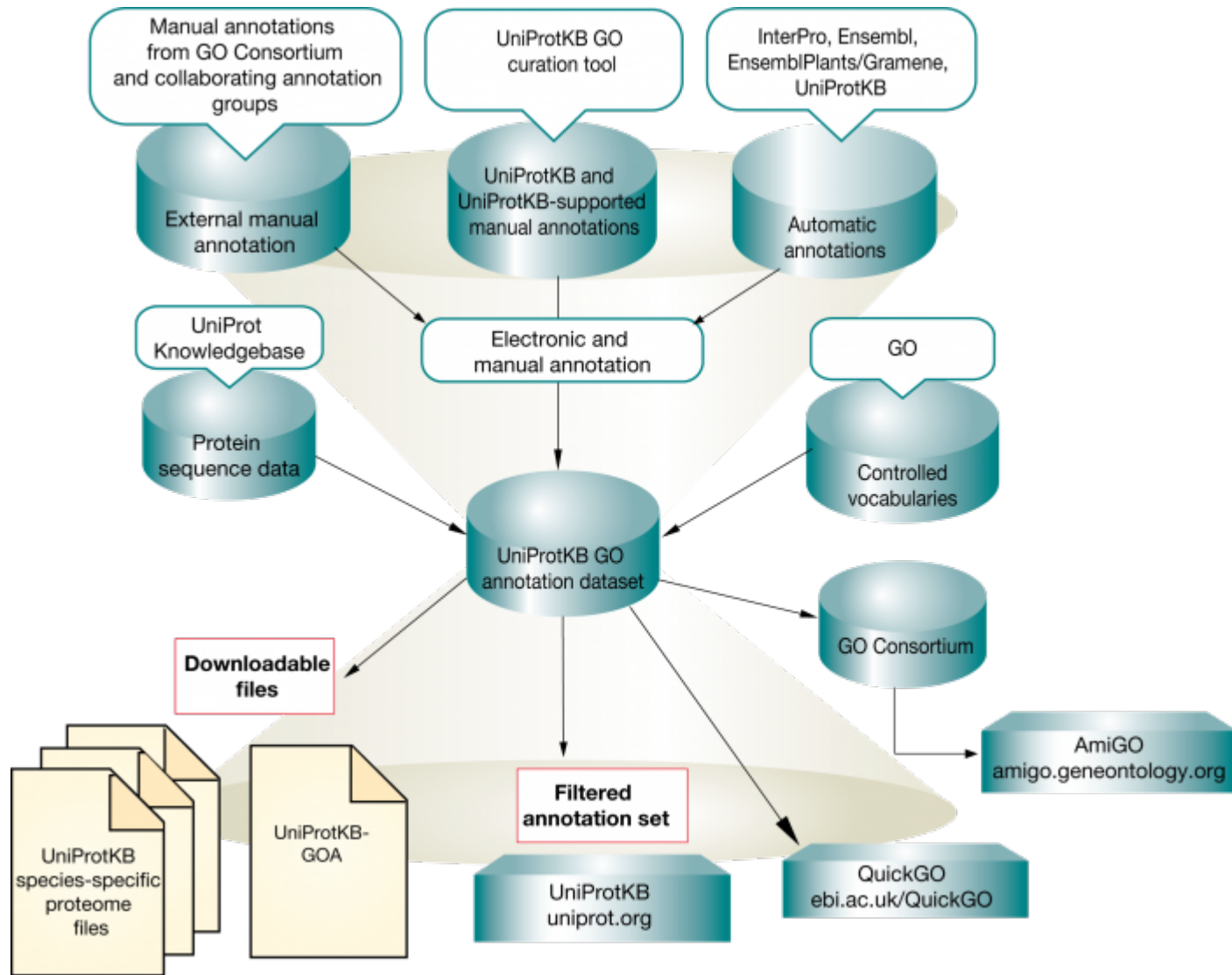
Enrichment analysis: use statistical test e.g. Fisher exact test

Example: in human genome background (20,000 gene total), 40 genes are involved in p53 signaling pathway. A given gene list has found that 3 out of 300 belong to p53 signaling pathway. Then we ask the question if 3/300 is more than random chance comparing to the human background of 40/20000



[http://david.abcc.ncifcrf.gov/helps/functional\\_annotation.html#E4](http://david.abcc.ncifcrf.gov/helps/functional_annotation.html#E4)

# UniProt-GO annotation (GOA)



# UniProt-GOA format

The *reference* used to make the annotation (e.g. a journal article)

An *evidence code* denoting the type of evidence upon which the annotation is based

The date and the creator of the annotation

```
Gene product: Actin, alpha cardiac muscle 1, UniProtKB:P68032  
GO term: heart contraction ; GO:0060047 (biological process)  
Evidence code: Inferred from Mutant Phenotype (IMP) Reference: PMID 17611253  
Assigned by: UniProtKB, June 6, 2008
```

# The idea of GO annotation for new sequences

If you have a new genome/transcriptome sequenced, how do you perform a GO annotation for it?

1. Find a closet model organism which has been annotated by GO
2. BLAST your data against this closest organism
3. Transfer the GO annotation of the best match to your query sequences

For instance, if we want to annotate fern transcriptome with GO function descriptions ....

1. Find Arabidopsis UniProt protein dataset
2. Find the Arabidopsis GOA association file
3. BLASTx fern reads (or assembled UniGenes) against the UniProt set
4. Analyze BLAST result to link fern reads GO terms



# Hands on practice 2: GO annotation



## Search GO data

terms and gene products

Search

## Enrichment analysis (beta)

Your genes here...

biological process

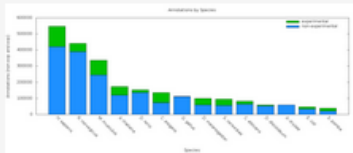
H. sapiens

Submit

Advanced options

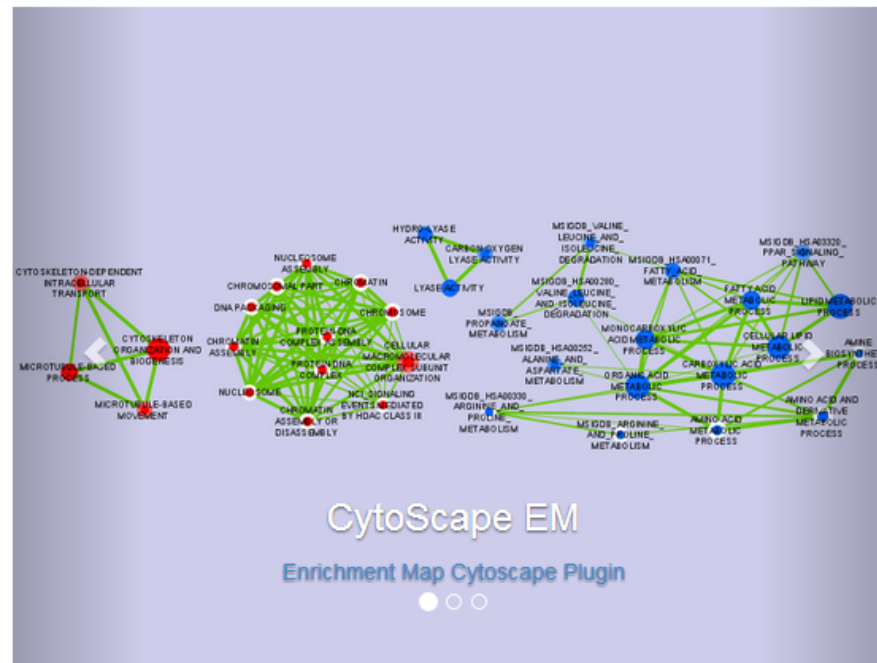
Powered by PANTHER

## Statistics



# Gene Ontology Consortium

Search



## Highlighted GO term

Representing "phases" in GO biological process

The GOC has recently introduced a new term [biological phase \(GO:0044848\)](#), as a direct subclass of biological process.

This class represents a distinct period or stage during which biological processes can occur.

[more](#)

## On the web

[Analysis of Tumor Suppressor Genes Based on <b>Gene Ontology</b> and the KEGG Pathway](#)

[Analysis of Tumor Suppressor Genes Based on <b>Gene Ontology</b> and the KEGG Pathway](#)

[An association analysis between psychophysical characteristics and genome-wide <b>gene</b> <b>...</b>](#)

[An association analysis between psychophysical characteristics and genome-wide <b>gene</b> <b>...</b>](#)

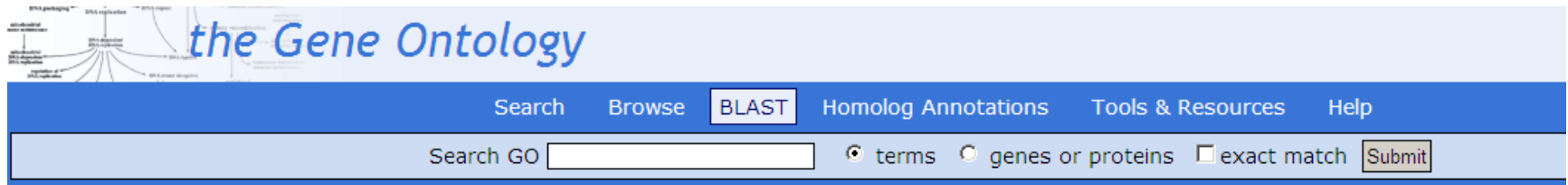
[Differentiation of the two rice subspecies indica and japonica](#)

## What is the Gene Ontology?

- [An introduction to the Gene Ontology](#)
- [What are annotations?](#)
- [Ten quick tips for using the Gene Ontology](#) Important
- [Gene Ontology tools](#)
- [Enrichment analysis](#)
- [Downloads](#)

## Recent news


<http://amigo1.geneontology.org/cgi-bin/amigo/blast.cgi>



## BLAST Search

The sequence search is performed using either BLASTP or BLASTX (from the [WU-BLAST](#) package), depending on the type of the input sequence.

### BLAST Query

**Enter your query** 

Enter a UniProtKB accession **or** upload a text file of queries **or** paste in FASTA sequence(s)

UniProtKB accession:

Text file (maximum file size 500K):  No file chosen

FASTA sequence(s):  
Sequences should be separated with an empty line.

```
>AT5G22740.1|AT5G22740.1|cs1A
MDGVSPKFEVLPETFDGVRMEITGQLGMIWELVKAPVIVPLLQLAVYICLL
MSVMLLCERVYMGIVIVLVKLFWKKPKDKRYKFPIHDDEELGSSNFPVVL
VQIPMFNEREVYKLSIGAACGLSWPSDRLVIQVLDDSTDPTVKQMEVEEC
QRWASKGINIRYQIRENRVGYKAGALKEGLKRSYVKHCEYVVI FDADFQP
EPDFLRRSIPFLMHNPNIALVQARWRVNSDECLLTRMQEMSLDYHFTVE
QEVGSSSTHAFFGFNGTAGIWRIAAINEAGGWKDRITVEDMDLAVRASLRG
MKRTYI GDT QVKSRT PSTPBAEPDROCHDMSGGDANT EDKXUMFTVDNKKY
```

Get an example protein sequence file from <http://cys.bios.niu.edu/yyin/teach/PBB/csl-pr.fa>

# BLAST Query Submission

## Success!

Your job has been successfully submitted to the BLAST queue.

Please be patient as your job may take several minutes to complete. This page will automatically refresh with the BLAST results when the job is done.

[Try retrieving your job now](#)

## Query Summary

Your job contains 2 sequences.

Parameters

Threshold: 0.1

Maximum number of alignments shown: 50

BLAST filter: on

AmiGO version: [1.8](#)

[Try AmiGO Labs](#)

## High Scoring Gene Products

Perform an action with this page's selected gene products...

	Symbol, full name	Information	P value
<input type="checkbox"/>	<a href="#">CSLA02</a> cellulose synthase-like A02	<a href="#">BLAST match</a> ↓ <a href="#">view associations</a> → <a href="#">BLAST with CSLA02</a> ↗	3.6e-295
<input type="checkbox"/>	<a href="#">ATCSLA09</a>	<a href="#">BLAST match</a> ↓ <a href="#">view associations</a> → <a href="#">BLAST with ATCSLA09</a> ↗	4.5e-217
<input type="checkbox"/>	<a href="#">CSLA03</a> cellulose synthase-like A3	<a href="#">BLAST match</a> ↓ <a href="#">view associations</a> → <a href="#">BLAST with CSLA03</a> ↗	4.1e-191
<input type="checkbox"/>	<a href="#">ATCSLA15</a>	<a href="#">BLAST match</a> ↓ <a href="#">view associations</a> → <a href="#">BLAST with ATCSLA15</a> ↗	2.9e-183
<input type="checkbox"/>	<a href="#">CSLA07</a> cellulose synthase like	<a href="#">BLAST match</a> ↓ <a href="#">view associations</a> → <a href="#">BLAST with CSLA07</a> ↗	1.2e-182
<input type="checkbox"/>	<a href="#">CSLA10</a> cellulose synthase-like A10	<a href="#">BLAST match</a> ↓ <a href="#">view associations</a> → <a href="#">BLAST with CSLA10</a> ↗	3.3e-173
<input type="checkbox"/>	<a href="#">CSLA01</a> cellulose synthase-like A01	<a href="#">BLAST match</a> ↓ <a href="#">view associations</a> → <a href="#">BLAST with CSLA01</a> ↗	4.0e-170
<input type="checkbox"/>	<a href="#">CSLA14</a> cellulose synthase like A14	<a href="#">BLAST match</a> ↓ <a href="#">view associations</a> → <a href="#">BLAST with CSLA14</a> ↗	7.6e-167

This is easy. Now let's try to get a list of differentially expressed genes and then find what's common in this list of genes in terms of functions.

We're gonna use NCBI GEO website to get the gene list and then feed the gene list to GO enrichment analysis tools

Go to NCBI home page, search GEO DataSets with keyword "GDS4831", and hit search

Secure | <https://www.ncbi.nlm.nih.gov>

NCBI Resources ▾ How To ▾

NCBI  
National Center for Biotechnology Information

GEO DataSets ▾ GDS4831

**NCBI Home**

**Resource List (A-Z)**

- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps

**Welcome to NCBI**

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

**Submit**  
Deposit data or manuscripts into NCBI databases

**Download**  
Transfer NCBI data to your computer

**Learn**  
Find help documents, attend a class or watch a tutorial

Choose "Compare 2 sets of samples"

Choose "Value means difference"

Choose "8+ fold"

Choose "higher"

Then go to Step 2

Select to choose group A: three samples for COP 1 depletion and Huh7 cell line

Group B: three samples for negative control and Huh7 cell line

Hit ok, and go to Step 3

**DataSet Record GDS4831:** Expression Profiles | Data Analysis Tools | Sample Subsets

**Title:** COP1 depletion effect on hepatocellular carcinoma cell lines

**Summary:** Analysis of Huh7, HepG2, and Hep3B hepatocellular carcinoma (HCC) cells depleted for the ubiquitin modulator COP1. COP1 regulates p53 activity by ubiquitination. p53 is wild type in HepG2, mutated in Huh7, and lacking in Hep3B. Results provide insight into the role of COP1 in HCC pathogenesis.

**Organism:** *Homo sapiens*

**Platform:** GPL6883: Illumina HumanRef-8 v3.0 expression beadchip

**Citation:** Lee YH, Andersen JB, Song HT, Judge AD et al. Definition of ubiquitination modulator COP1 as a novel therapeutic target in hepatocellular carcinoma (HCC) cell lines. *PLoS One* 11(7):e216264. doi:10.1371/journal.pone.0170(21):8264-9. PMID: 20959491

**Reference Series:** GSE21955

**Value type:** count

**Sample count:** 22

**Series published:** 2013

**Data Analysis Tools**

Find genes

**Compare 2 sets of samples** ?

Cluster heatmaps

Experiment design and value distribution

**Step 1:** Select test and significance level

Value means difference | A vs B: | 8+ fold | higher

**Step 2:** Select which samples to put in Group A and Group B

**Step 3:** Query Group A vs. B

Click on accessions to select samples individually, click on colored blocks and then on blinking arrows to select groups of samples.

Samples, Group A	Factors		Samples, Group B
	protocol	cell line	
GSM545954	COP1 depletion	Huh7	GSM545954
GSM545955			GSM545955
GSM545956			GSM545956
GSM545960		HepG2	GSM545960
GSM545961			GSM545961
GSM545962			GSM545962
GSM545963	negative control	Hep3B	GSM545963
GSM545968			GSM545968
GSM545969			GSM545969
GSM545970		Huh7	GSM545970
GSM545971			GSM545971
GSM545975			GSM545975
GSM545958	Huh7	GSM545958	
GSM545959		GSM545959	
GSM545964		GSM545964	
GSM545965	HepG2	GSM545965	
GSM545966		GSM545966	
GSM545967		GSM545967	
GSM545972	Hep3B	GSM545972	
GSM545973		GSM545973	
GSM545974		GSM545974	
GSM545975			GSM545975

Ok

Reset

Cancel

**DataSet Record**

cell lines

carcinoma (HCC)

sults provide insig

idchip

nition of ubiquitin

elect test and sign

ans difference

elect which Samp

Total 256 gene profiles are found with 8+ fold higher expression in COP 1 depletion than in negative control in Huh7 cell line

To get the list of genes, choose Gene database and hit Find items

Display Settings:  Summary, 20 per page, Sorted by Default order

Results: 1 to 20 of 256

[UBE2G2 - COP1 depletion effect on hepatocellular carcinoma cell lines](#)

1. Annotation: UBE2G2, ubiquitin-conjugating enzyme E2G 2  
Organism: Homo sapiens  
Reporter: GPL6883, ILMN\_2297824 (ID\_REF), GDS4831, NM\_182688  
DataSet type: Expression profiling by array, count, 22 samples  
ID: 104862213  
[GEO DataSets](#) [Gene](#) [UniGene](#) [Profile neighbors](#) [Chromosome neighbors](#) [Homologene neighbors](#)

[MS4A6A - COP1 depletion effect on hepatocellular carcinoma cell lines](#)

2. Annotation: MS4A6A, membrane-spanning 4-domains, subfamily A, member 6A  
Organism: Homo sapiens  
Reporter: GPL6883, ILMN\_2359800 (ID\_REF), GDS4831, NM\_152851  
DataSet type: Expression profiling by array, count, 22 samples  
ID: 104862285  
[GEO DataSets](#) [Gene](#) [UniGene](#) [Profile neighbors](#) [Chromosome neighbors](#) [Homologene neighbors](#)

[VIP - COP1 depletion effect on hepatocellular carcinoma cell lines](#)

3. Annotation: VIP, vasoactive intestinal peptide  
Organism: Homo sapiens  
Reporter: GPL6883, ILMN\_1794638 (ID\_REF), GDS4831, NM\_194435  
DataSet type: Expression profiling by array, count, 22 samples  
ID: 104862502  
[GEO DataSets](#) [Gene](#) [UniGene](#) [Chromosome neighbors](#) [Homologene neighbors](#)

[HIST1H2BO - COP1 depletion effect on hepatocellular carcinoma cell lines](#)

4. Annotation: HIST1H2BO, histone cluster 1, H2B-

Send to:  Filters: [Manage Filters](#)  
<< First < Prev Page 1 of 13 Next > Last >>

Profile data

[Download profile data](#)

Profile pathways

[Find pathways](#)

Find related data

Database:

based on gene annotation from platform profile accessio

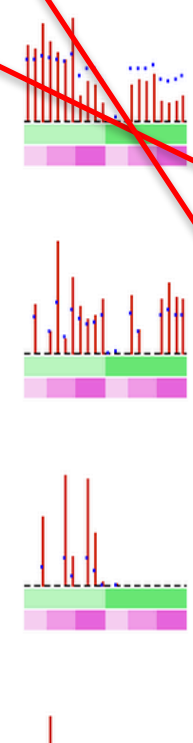
[Find items](#)

Recent activity

[Turn Off](#)

COP1 depletion effect on hepatocellular carcinoma cell lines GDS

(GDS4831[ACCN]) AND GDS[filter] (1) GDS






Total 225 genes correspond to 256 gene profiles

To download the list of Gene IDs, hit Send to, choose UI list as format and hit Create file

Display Settings:  Tabular, 20 per page, Sort by Relevance

Results: 1 to 20 of 225

 Showing Current items.

Name/Gene ID	Description	Location		
<input type="checkbox"/> <a href="#">ADAM19</a> ID: 8728	ADAM metallopeptidase domain 19 [ <i>Homo sapiens</i> (human)]	Chromosome 5, NC_000005.10 (157477304..157575823, complement)		
<input type="checkbox"/> <a href="#">C3orf14</a> ID: 57415	chromosome 3 open reading frame 14 [ <i>Homo sapiens</i> (human)]	Chromosome 3, NC_000003.12 (62318973..62336213)		
<input type="checkbox"/> <a href="#">NT5DC2</a> ID: 64943	5'-nucleotidase domain containing 2 [ <i>Homo sapiens</i> (human)]	Chromosome 3, NC_000003.12 (52524369..52535077, complement)		
<input type="checkbox"/> <a href="#">SP140</a> ID: 11262	SP140 nuclear body protein [ <i>Homo sapiens</i> (human)]	Chromosome 2, NC_000002.12 (230202742..230316616)	LYSP100, LYSP100-A, LYSP100-B	608602
<input type="checkbox"/> <a href="#">ADAMTS6</a> ID: 11174	ADAM metallopeptidase with thrombospondin type 1 motif, 6 [ <i>Homo sapiens</i> (human)]	Chromosome 5, NC_000005.10 (65148736..65482027, complement)	ADAM-TS 6, ADAM-TS6, ADAMTS-6	605008
<input type="checkbox"/> <a href="#">GAS6</a> ID: 2621	growth arrest-specific 6 [ <i>Homo sapiens</i> (human)]	Chromosome 13, NC_000013.11 (113820549..113864103, complement)	AXLLG, AXSF	600441
<input type="checkbox"/> <a href="#">BST2</a> ID: 684	bone marrow stromal cell antigen 2 [ <i>Homo sapiens</i> (human)]	Chromosome 19, NC_000019.10 (17402939..17405648, complement)	CD317, TETHERIN	600534

Send to:

**Choose Destination**

File  Clipboard

Collections

Download 225 items.

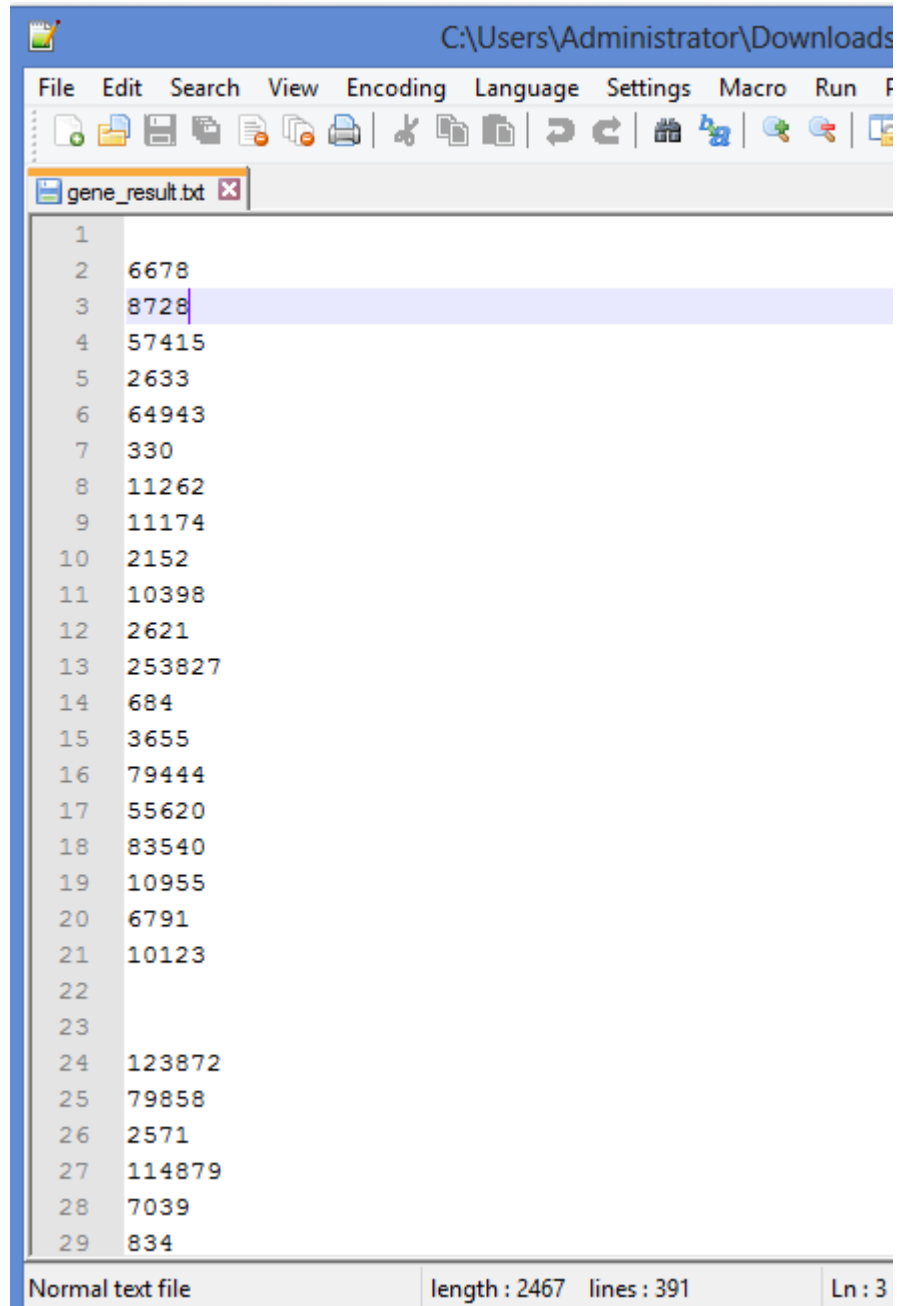
Format  
**UI List**

Sort by  
Relevance

Create File

A file named "gene\_result.txt" will be automatically downloaded to your local computer  
Find out where it is downloaded to, open it using notepad++

View the file using notepad++



Next we will use DAVID to perform function enrichment analysis

# The Database for Annotation, Visualization and Integrated Discovery (DAVID)

← → ↻ 🏠 [david.abcc.ncifcrf.gov/home.jsp](http://david.abcc.ncifcrf.gov/home.jsp)

CSR Internet - Study... Sample Applications... Bioinformatics 1 Co... Bioinformatics Cour... I519: Introduction to... BMIF 310: Foundatio... NIU Libraries Advisor

## DAVID Bioinformatics Resources 6.7

National Institute of Allergy and Infectious Diseases (NIAID), NIH

Home Start Analysis Shortcut to DAVID Tools Technical Center Downloads & APIs Term of Service Why DAVID? About Us

### Shortcut to DAVID Tools

- Functional Annotation**  
Gene-annotation enrichment analysis, functional annotation clustering, BioCarta & KEGG pathway mapping, gene-disease association, homologue match, ID translation, literature match and [more](#)
- Gene Functional Classification**  
Provide a rapid means to reduce large lists of genes into functionally related groups of genes to help unravel the biological content captured by high throughput technologies. [More](#)
- Gene ID Conversion**  
Convert list of gene ID/accessions to others of your choice with the most comprehensive gene ID mapping repository. The ambiguous accessions in the list can also be determined semi-automatically. [More](#)
- Gene Name Batch Viewer**  
Display gene names for a given gene list; Search functionally related genes within your list or not in your list; Deep links to enriched detailed information. [More](#)

Recommending: A [paper](#) published in *Nature Protocols* describes step-by-step procedure to use DAVID!

## Welcome to DAVID 6.7

2003 - 2014

The Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 is an update to the sixth version of our original web-accessible programs. DAVID now provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes. For any given gene list, DAVID tools are able to:

- Identify enriched biological themes, particularly GO terms
- Discover enriched functional-related gene groups
- Cluster redundant annotation terms
- Visualize genes on BioCarta & KEGG pathway maps
- Display related many-genes-to-many-terms on 2-D view.
- Search for other functionally related genes not in the list
- List interacting proteins
- Explore gene names in batch
- Link gene-disease associations
- Highlight protein functional domains and motifs
- Redirect to related literatures
- Convert gene identifiers from one type to another.
- And more

### What's Important in DAVID?

- [Current \(v 6.7\) release note](#)
- [New requirement to cite DAVID](#)
- [IDs of Affy Exon and Gene arrays supported](#)
- [Novel Classification Algorithms](#)
- [Pre-built Affymetrix and Illumina backgrounds](#)
- [User's customized gene background](#)
- [Enhanced calculating speed](#)

### Statistics of DAVID

DAVID Bioinformatic Resources Citations

Year	Citations
2004	~100
2005	~200
2006	~300
2007	~400
2008	~500
2009	~700
2010	~1000
2011	~1500
2012	~2200
2013	~3182

Hit start analysis

Upload | **List** | Background

**Upload Gene List**

[Demolist 1](#) [Demolist 2](#)

[Upload Help](#)

**Step 1: Enter Gene List**

A: Paste a list

Clear

Or

B: Choose From a File

Choose File [gene\\_result.txt](#)

Multi-List File ?

**Step 2: Select Identifier**

ENTREZ\_GENE\_ID ▼

**Step 3: List Type**

Gene List

Background

**Step 4: Submit List**

Submit List

**Analysis Wizard**

[Tell us how you like the tool](#)  
[Contact us for questions](#)

← Step 1. Submit your gene list through left panel.

An example:

Copy/paste IDs to "box A" -> Select Identifier as "Affy\_ID" -> List Type as "Gene List" -> Click "Submit" button

1007\_s\_at  
1053\_at  
117\_at  
121\_at  
1255\_g\_at  
1294\_at  
1316\_at  
1320\_at  
1405\_i\_at  
1431\_at  
1438\_at  
1487\_at  
1494\_f\_at  
1598\_g\_at

Upload the list of Gene IDs

Select ENTREZ\_GENE\_ID

Click on Gene list

**Upload** **List** Background

### Gene List Manager

Select to limit annotations by one or more species [Help](#)

- Use All Species -  
 Homo sapiens(354)  
 Unknown(1)

Select Species

**List Manager** [Help](#)

gene\_result

Select List to:

Use Rename  
 Remove Combine

Show Gene List

[View Unmapped Ids](#)

## Analysis Wizard

[Tell us how you like the tool](#)  
[Contact us for questions](#)

**Step 1. Successfully submitted gene list**

Current Gene List: gene\_result

Current Background: Homo sapiens

**Step 2. Analyze above gene list with one of DAVID tools**



[Which DAVID tools to use?](#)

↷ [Functional Annotation Tool](#)

- [Functional Annotation Clustering](#)
- [Functional Annotation Chart](#)
- [Functional Annotation Table](#)

↷ [Gene Functional Classification Tool](#)

↷ [Gene ID Conversion Tool](#)

↷ [Gene Name Batch Viewer](#)

This allows you to view functional annotation from various resources including GO

Check the submitted gene list

If you have clicked on Functional Annotation tool, you are at this page

**Upload** **List** **Background**

### Gene List Manager

Select to limit annotations by one or more species [Help](#)

- Use All Species -
- Homo sapiens(225)
- Unknown(1)

Select Species

### List Manager [Help](#)

- gene\_result (6)

Select List to:

Use Rename

Remove Combine

Show Gene List

[View Unmapped Ids](#)

## Annotation Summary Results

[Help and Tool Manual](#)

**Current Gene List: gene\_result (6)** **225 DAVID IDs**

**Current Background: Homo sapiens** **Check Defaults**  [Clear All](#)

- ⊕ **Disease** (0 selected)
- ⊕ **Functional\_Categories** (0 selected)
- ⊕ **Gene\_Ontology** (0 selected)
- ⊕ **General Annotations** (0 selected)
- ⊕ **Literature** (0 selected)
- ⊕ **Main\_Accessions** (0 selected)
- ⊕ **Pathways** (0 selected)
- ⊕ **Protein\_Domains** (0 selected)
- ⊕ **Protein\_Interactions** (0 selected)
- ⊕ **Tissue\_Expression** (0 selected)

\*\*\*Red annotation categories denote DAVID defined defaults\*\*\*

### Combined View for Selected Annotation

Functional Annotation Clustering

Functional Annotation Chart

Functional Annotation Table

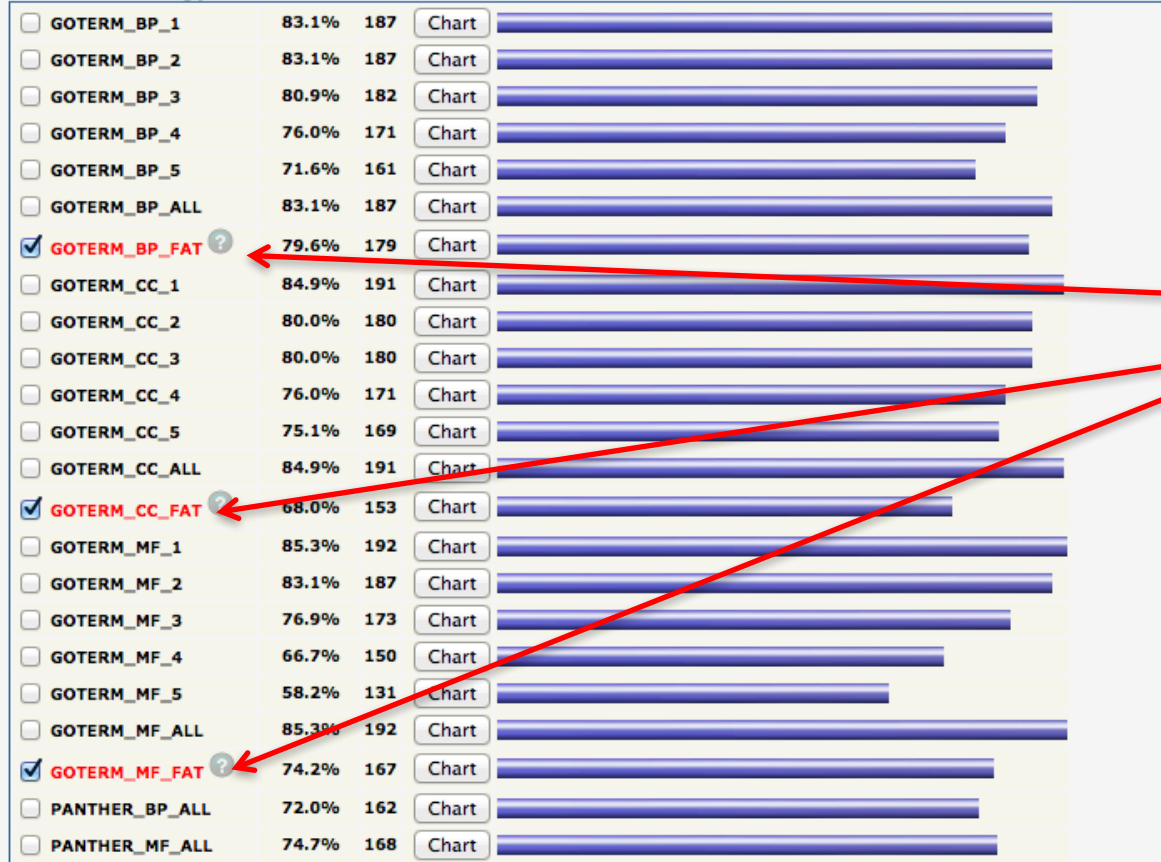
Uncheck this

All these can be changed by users (to show or not to show and show what)

Disease (0 selected)

Functional\_Categories (0 selected)

Gene\_Ontology (3 selected)



Select just GO

General Annotations (0 selected)

Literature (0 selected)

Main\_Accessions (0 selected)

Pathways (0 selected)

Protein\_Domains (0 selected)

Protein\_Interactions (0 selected)

Tissue\_Expression (0 selected)

\*\*\*Red annotation categories denote DAVID defined defaults\*\*\*

Combined View for Selected Annotation

Functional Annotation Clustering

Functional Annotation Chart

Functional Annotation Table

Click here will open a new window to show the 225 differentially expressed genes are enriched in what GO

# Functional Annotation Chart

[Help and Manual](#)

Current Gene List: gene\_result (6)

Current Background: Homo sapiens

225 DAVID IDs

Options



















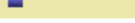





Genes are enriched in what GO categories  
(compared to the genome background)?

Rerun Using Options

Create Sublist

50 chart records

 [Download File](#)

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
<input type="checkbox"/>	GOTERM_BP_FAT	<a href="#">integrin-mediated signaling pathway</a>	RT		7	3.1	3.1E-4	3.6E-1
<input type="checkbox"/>	GOTERM_CC_FAT	<a href="#">plasma membrane</a>	RT		65	28.9	6.2E-4	1.4E-1
<input type="checkbox"/>	GOTERM_CC_FAT	<a href="#">integral to plasma membrane</a>	RT		28	12.4	7.5E-4	8.8E-2
<input type="checkbox"/>	GOTERM_CC_FAT	<a href="#">intrinsic to plasma membrane</a>	RT		28	12.4	1.0E-3	8.3E-2
<input type="checkbox"/>	GOTERM_BP_FAT	<a href="#">cell surface receptor linked signal transduction</a>	RT		38	16.9	5.8E-3	9.8E-1
<input type="checkbox"/>	GOTERM_BP_FAT	<a href="#">G-protein coupled receptor protein signaling pathway</a>	RT		26	11.6	6.5E-3	9.6E-1
<input type="checkbox"/>	GOTERM_CC_FAT	<a href="#">nucleosome</a>	RT		5	2.2	6.6E-3	3.4E-1
<input type="checkbox"/>	GOTERM_BP_FAT	<a href="#">positive regulation of protein kinase activity</a>	RT		9	4.0	9.4E-3	9.7E-1
<input type="checkbox"/>	GOTERM_BP_FAT	<a href="#">positive regulation of kinase activity</a>	RT		9	4.0	1.1E-2	9.6E-1
<input type="checkbox"/>	GOTERM_CC_FAT	<a href="#">integral to membrane</a>	RT		78	34.7	1.3E-2	4.8E-1
<input type="checkbox"/>	GOTERM_BP_FAT	<a href="#">cell activation</a>	RT		10	4.4	1.4E-2	9.6E-1
<input type="checkbox"/>	GOTERM_BP_FAT	<a href="#">positive regulation of transferase activity</a>	RT		9	4.0	1.4E-2	9.5E-1
<input type="checkbox"/>	GOTERM_BP_FAT	<a href="#">leukocyte activation</a>	RT		9	4.0	1.5E-2	9.3E-1
<input type="checkbox"/>	GOTERM_CC_FAT	<a href="#">plasma membrane part</a>	RT		38	16.9	1.6E-2	4.8E-1
<input type="checkbox"/>	GOTERM_BP_FAT	<a href="#">positive regulation of epithelial cell proliferation</a>	RT		4	1.8	1.7E-2	9.3E-1
<input type="checkbox"/>	GOTERM_BP_FAT	<a href="#">activation of protein kinase activity</a>	RT		6	2.7	1.7E-2	9.2E-1
<input type="checkbox"/>	GOTERM_BP_FAT	<a href="#">DNA packaging</a>	RT		6	2.7	1.9E-2	9.2E-1
<input type="checkbox"/>	GOTERM_CC_FAT	<a href="#">protein-DNA complex</a>	RT		5	2.2	1.9E-2	5.0E-1
<input type="checkbox"/>	GOTERM_CC_FAT	<a href="#">intrinsic to membrane</a>	RT		79	35.1	2.2E-2	5.0E-1
<input type="checkbox"/>	GOTERM_BP_FAT	<a href="#">heart development</a>	RT		8	3.6	2.4E-2	9.4E-1
<input type="checkbox"/>	GOTERM_BP_FAT	<a href="#">nucleosome assembly</a>	RT		5	2.2	2.5E-2	9.4E-1
<input type="checkbox"/>	GOTERM_BP_FAT	<a href="#">chromatin assembly</a>	RT		5	2.2	2.8E-2	9.4E-1
<input type="checkbox"/>	GOTERM_BP_FAT	<a href="#">locomotory behavior</a>	RT		9	4.0	2.9E-2	9.4E-1
<input type="checkbox"/>	GOTERM_BP_FAT	<a href="#">leukocyte differentiation</a>	RT		6	2.7	2.9E-2	9.3E-1



**Next lecture:** EBI web  
resources II (ENSEMBL  
and InterPro)