EBI web resources I: databases and tools

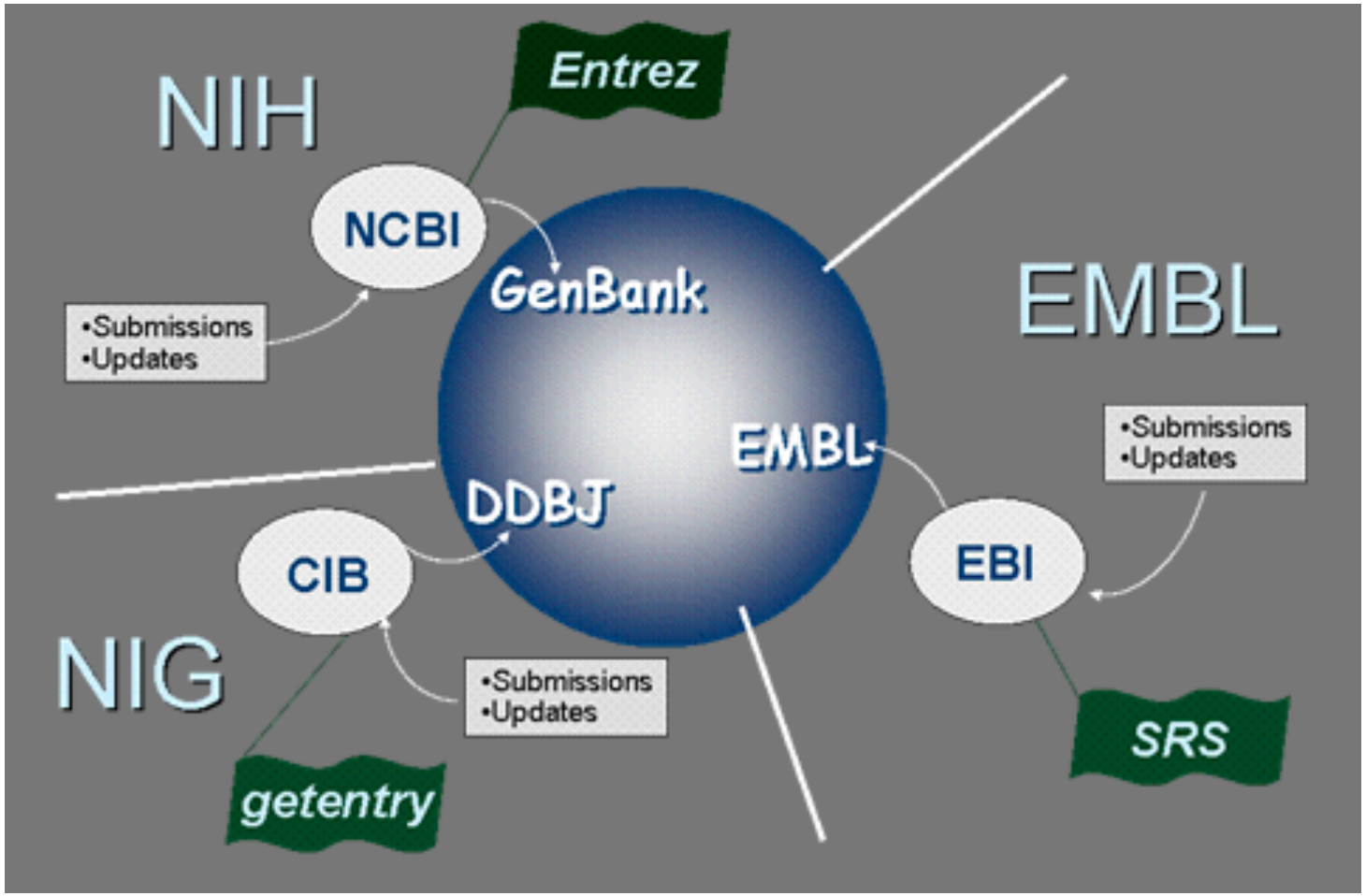
Yanbin Yin

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

Three international nucleotide sequence databases



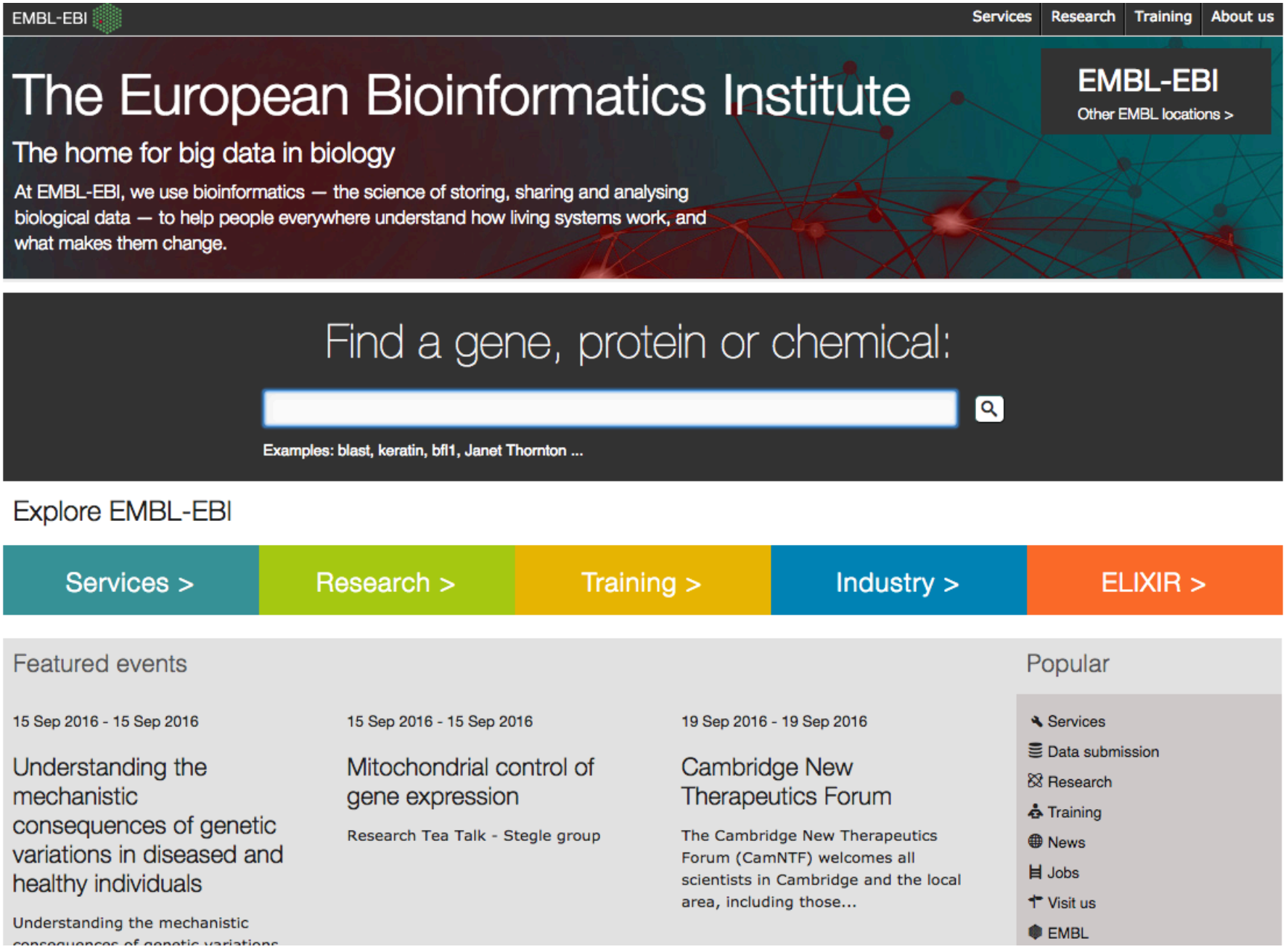
The European Bioinformatics Institute (EBI)

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](#)

http://www.ebi.ac.uk/



Research groups in EBI

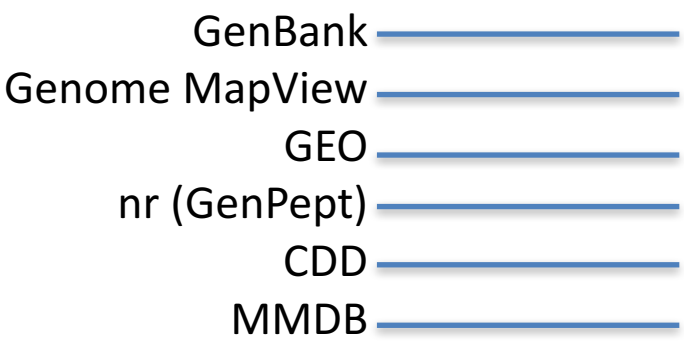
	Group/team leader	Area of research
Genomes	Ewan Birney	Algorithmic methods for genome analysis
	Paul Flicek	Vertebrate genomics
	Nick Goldman	Evolutionary tools for sequence analysis
Transcriptomes	Alvis Brazma	Functional genomics
	Anton Enright	Functional genomics and analysis of small RNA function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

InterPro

miRBase

UniProt

Major databases in EBI

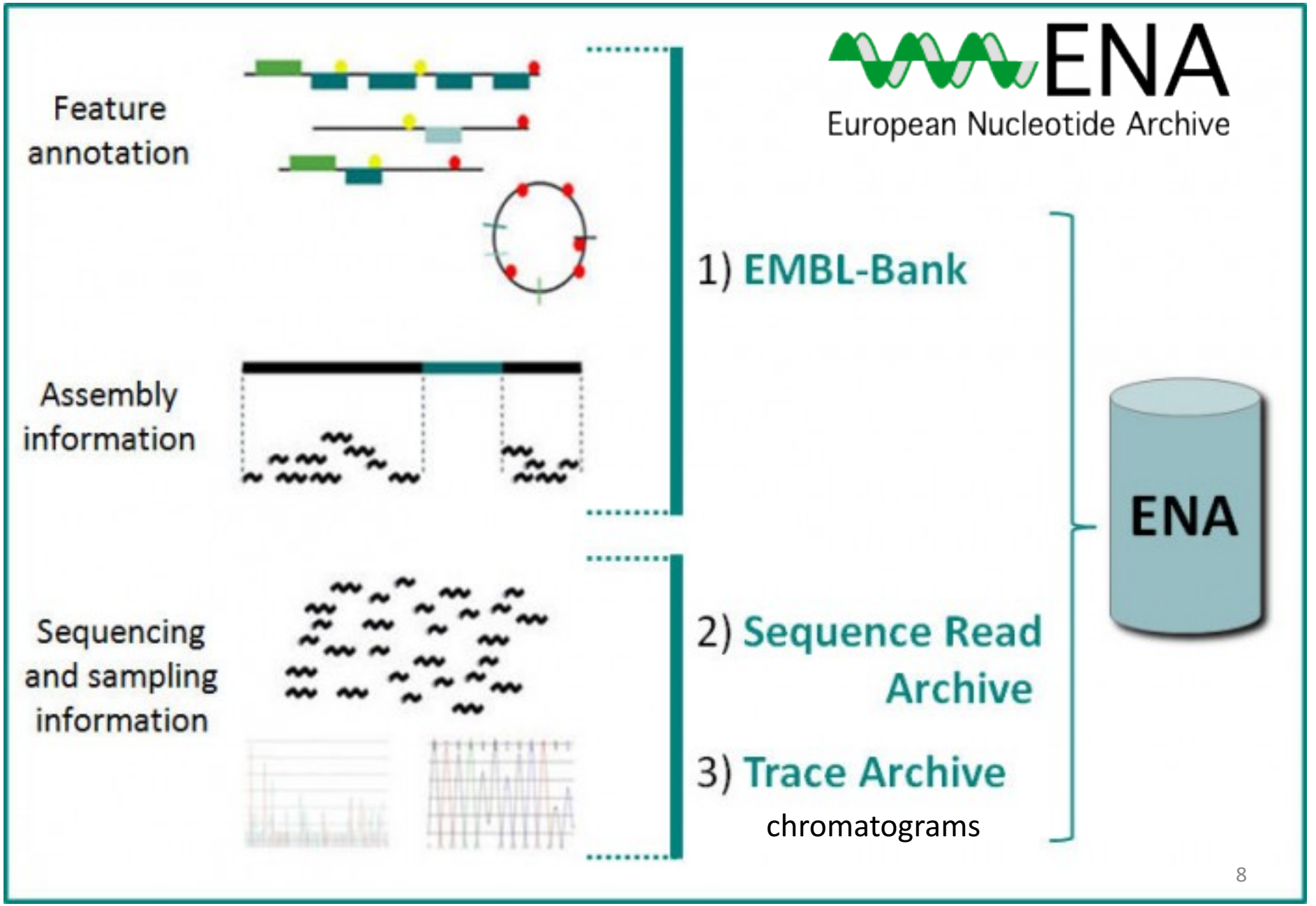


EMBL-Bank (DNA and RNA sequences)
Ensembl (genomes)
ArrayExpress(microarray-based gene-expression data)
UniProt (protein sequences)
InterPro(protein families, domains and motifs)
PDBe (macromolecular structures)

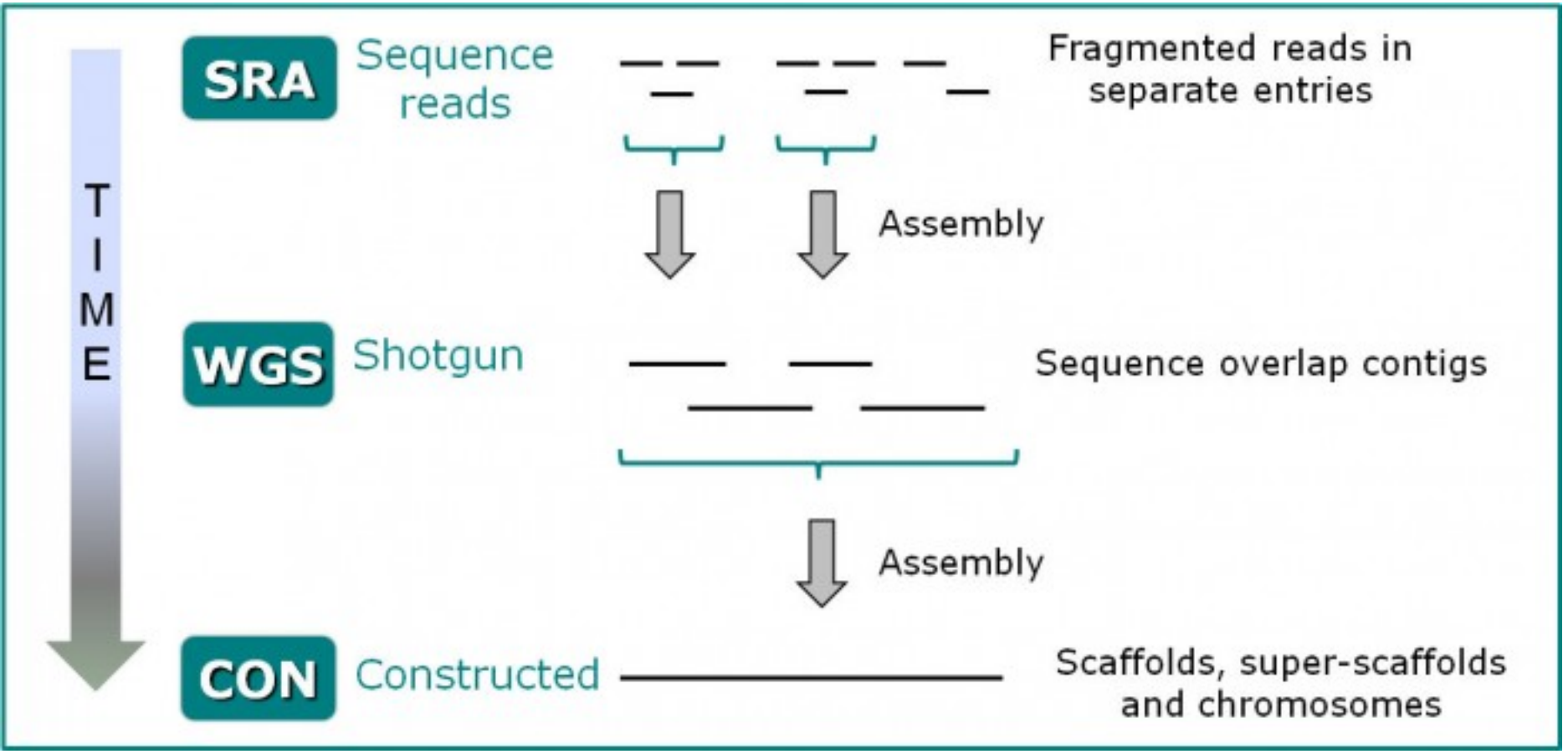
Others, such as
IntAct (protein–protein interactions)
Reactome (pathways)
ChEBI (small molecules)
IntEnz (enzyme classification)
GO (gene ontology)

Swiss Institute of Bioinformatics
Sanger Institute

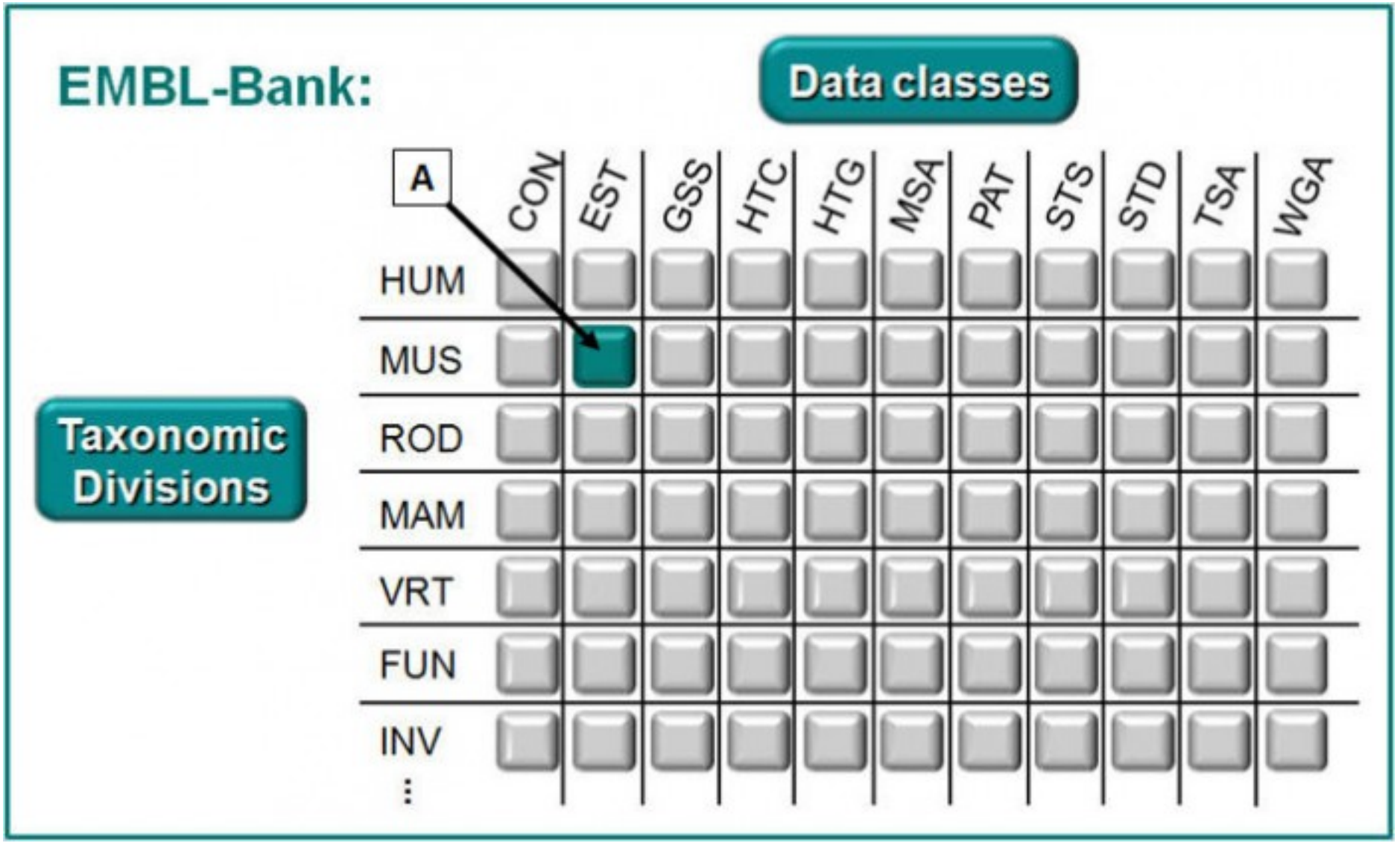
http://www.ebi.ac.uk/training/online/course/nucleotide-sequence-data-resources-ebi



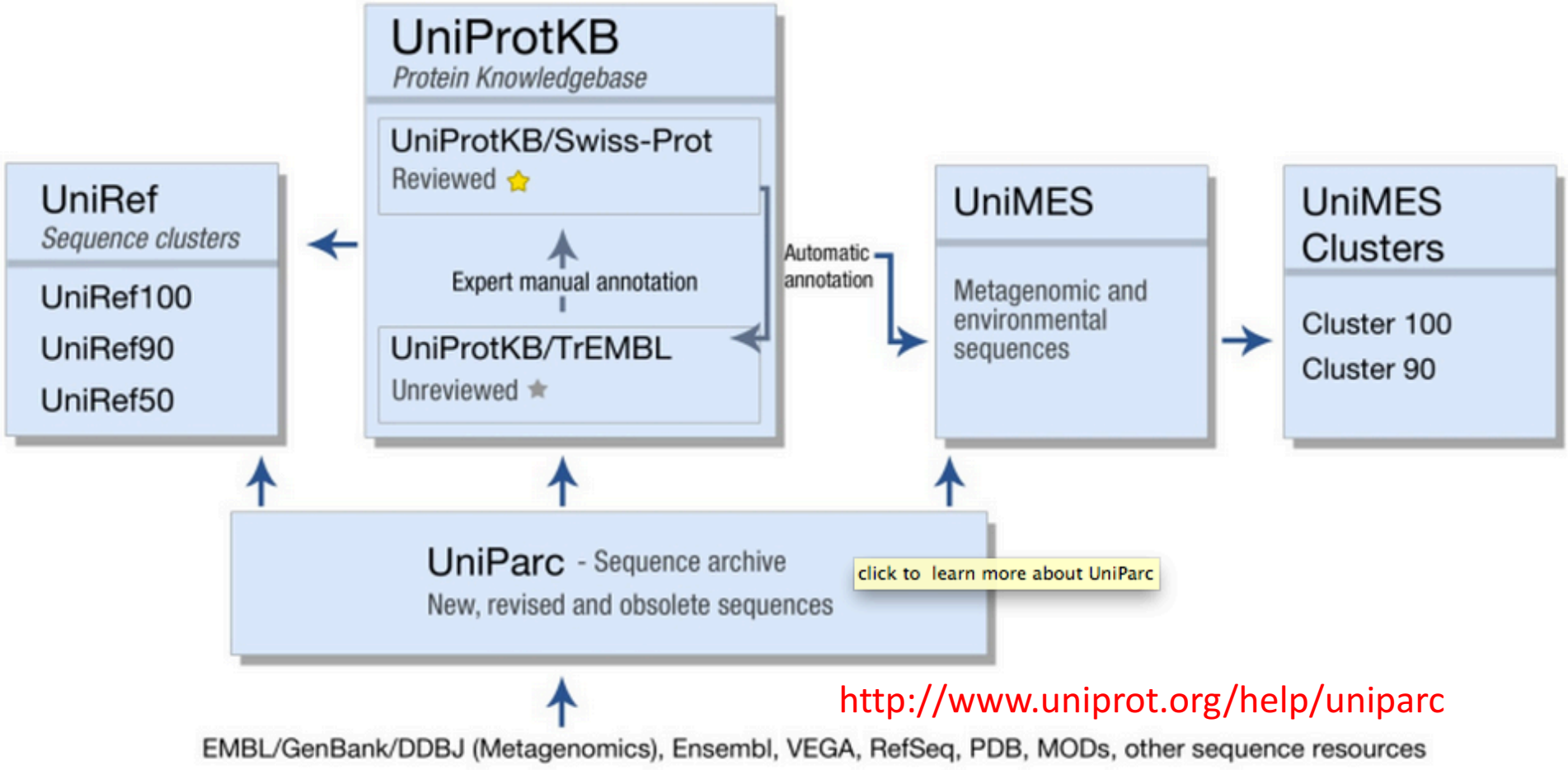
Sequence might first enter ENA as SRA (Sequence Read Archive) fragmented sequence reads; it might be re-submitted as assembled WGS (Whole Genome Shotgun) sequence overlap contigs; it might be re-submitted again with further assembly as CON (Constructed) sequence entries, with the older WGS entries being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



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



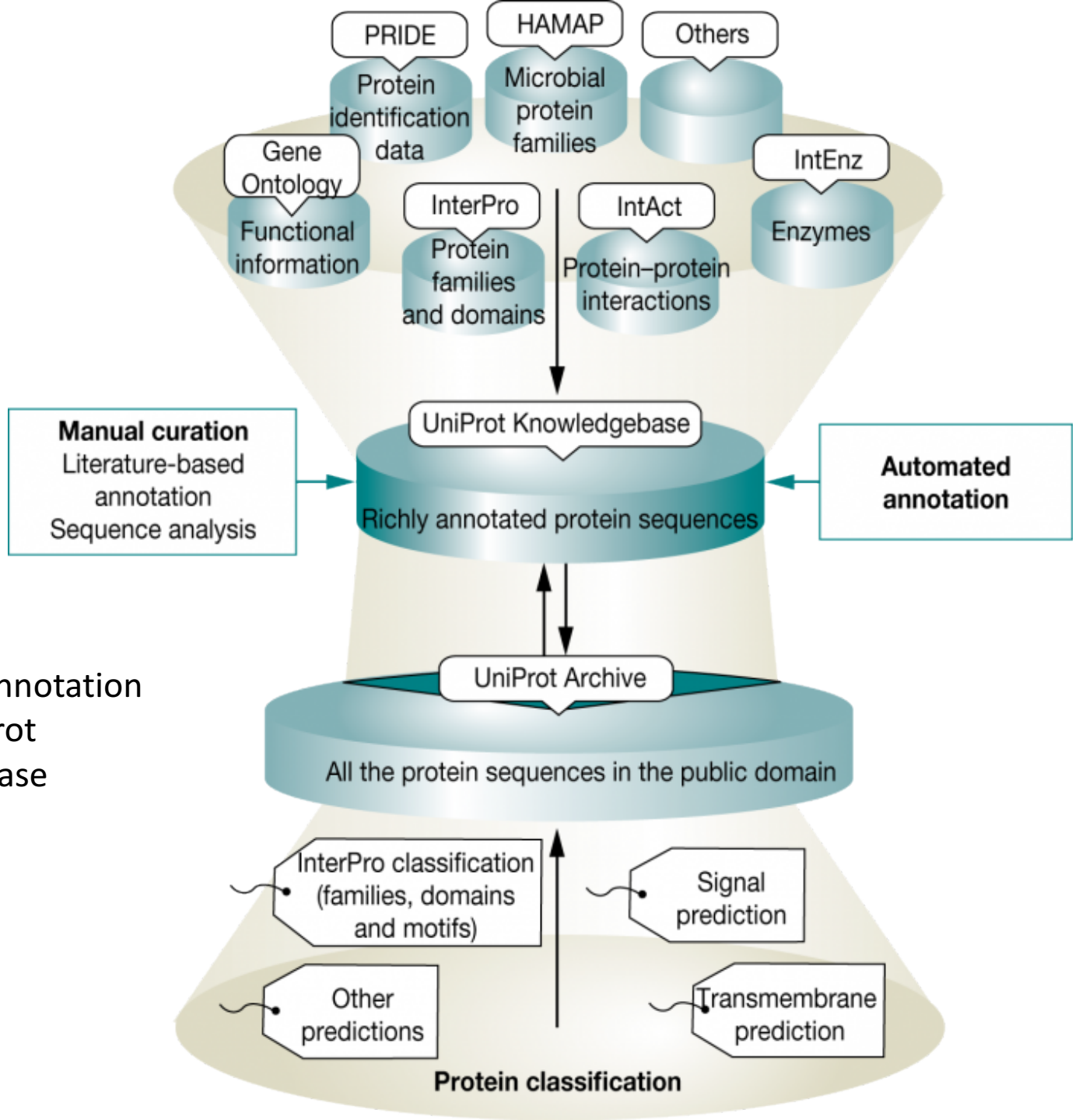
UniProt



http://www.uniprot.org/help/uniparc

EMBL-EBI SIB PIR

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/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 chan

UniProt relea

Life (and dea

variation files

UniProt relea

Pseudo-allerg

access to Uni

of human var

News arch

Getting started

YouTube

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

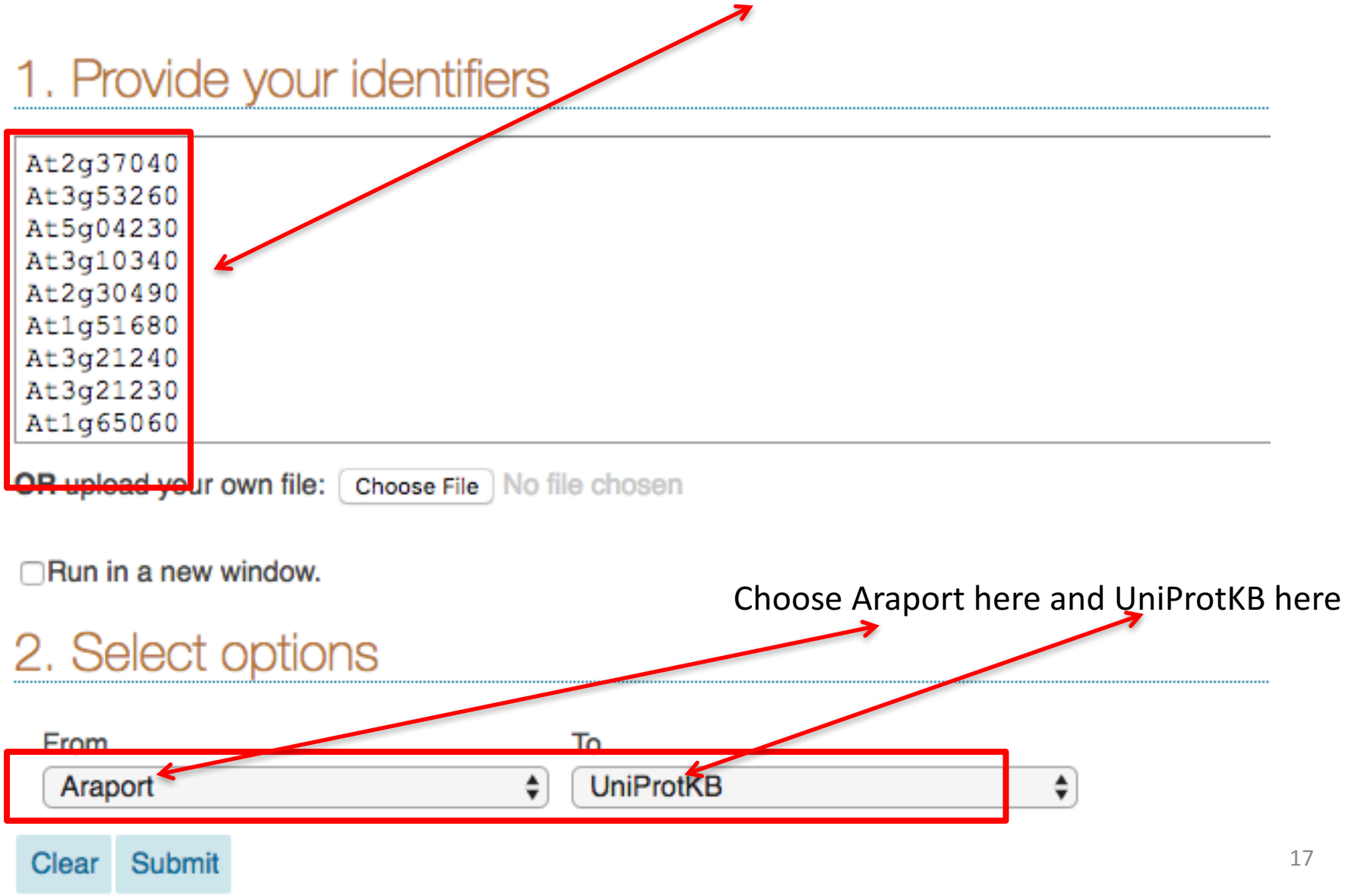
View Swiss-Prot and TrEMBL statistics

How to cite us

Protein sp

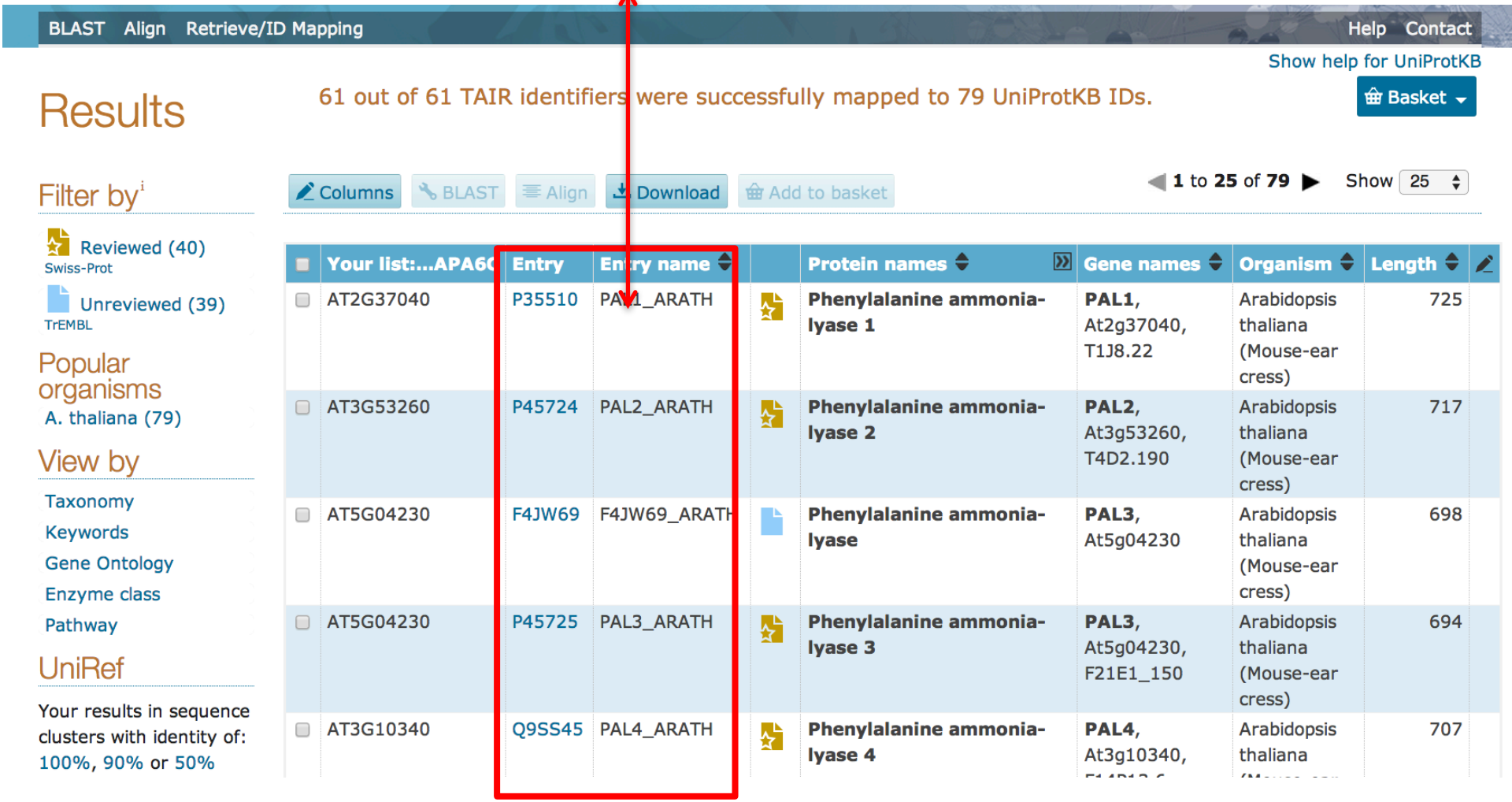
We are going to do ID mapping

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

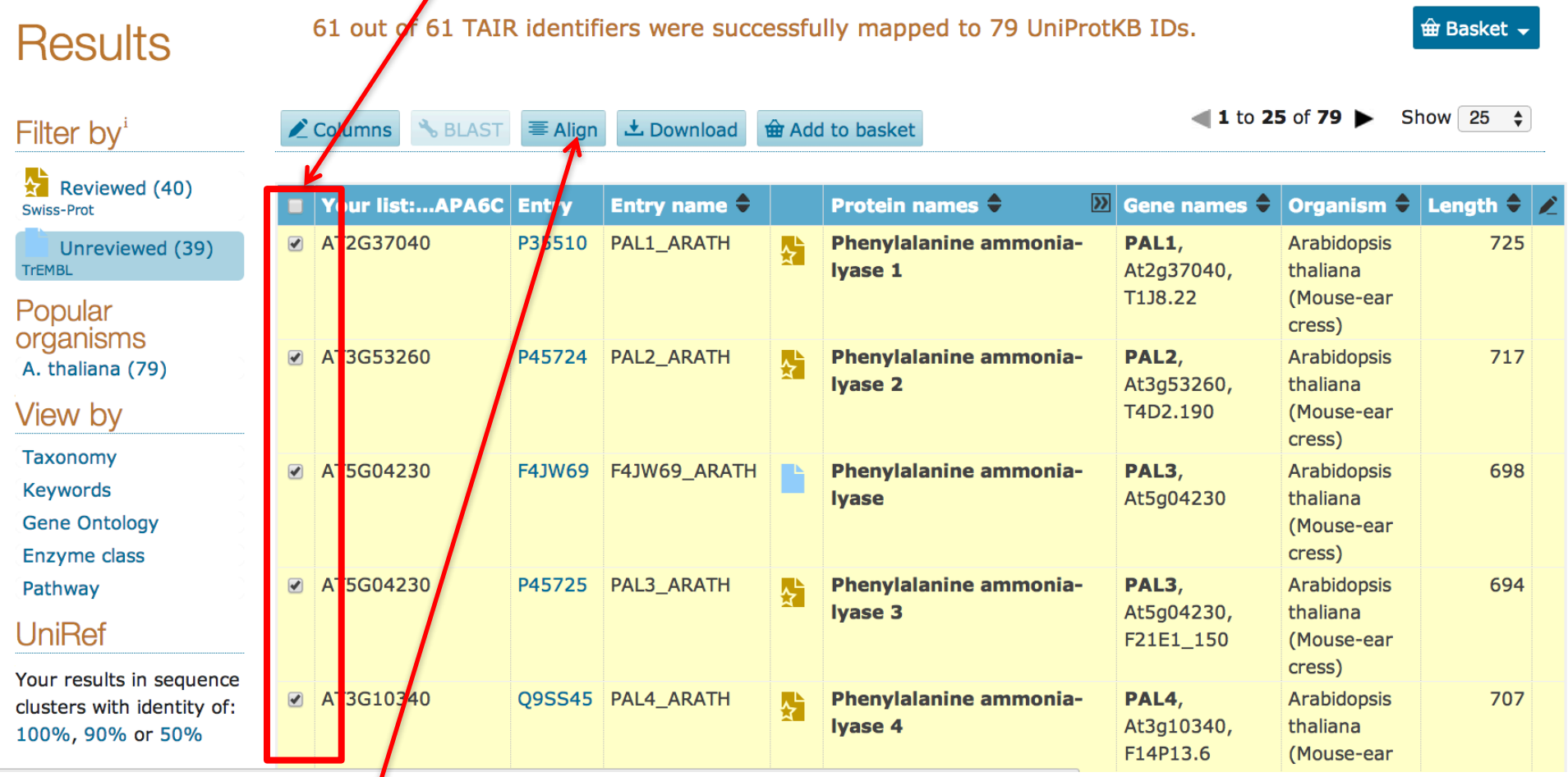


Choose Araport here and UniProtKB here

These are UniProt IDs



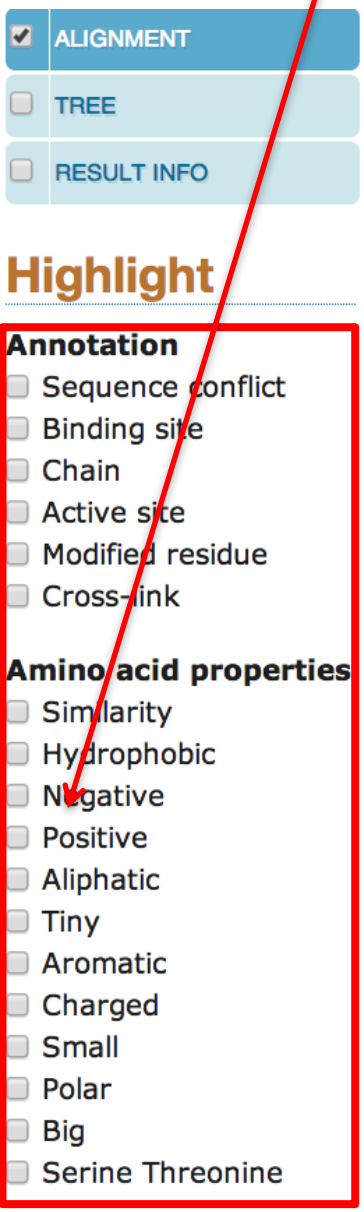
Select the PAL proteins and align them



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

This is the MSA result page

Toggle these options on will add colors in the alignment

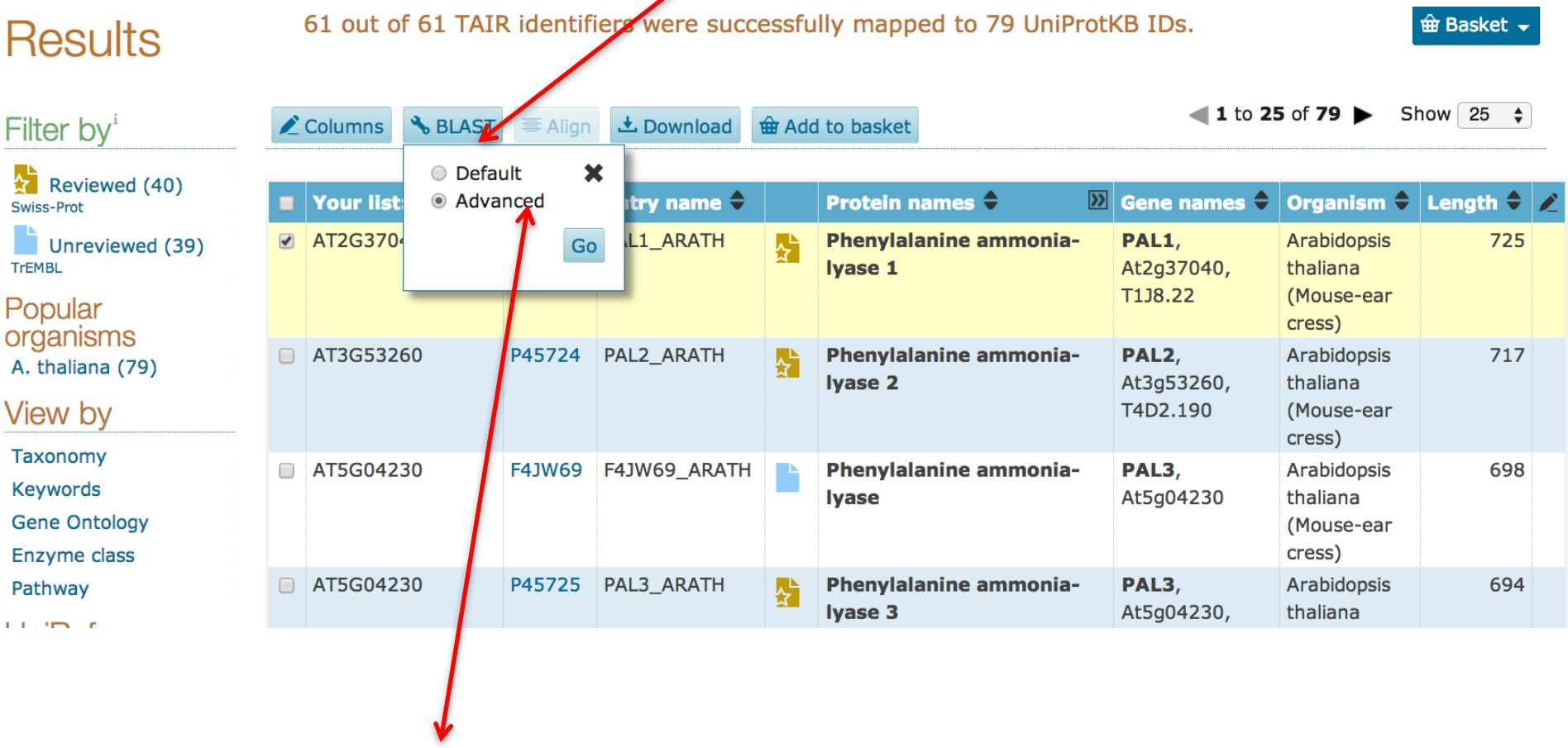


Alignment

How to print an alignment in color

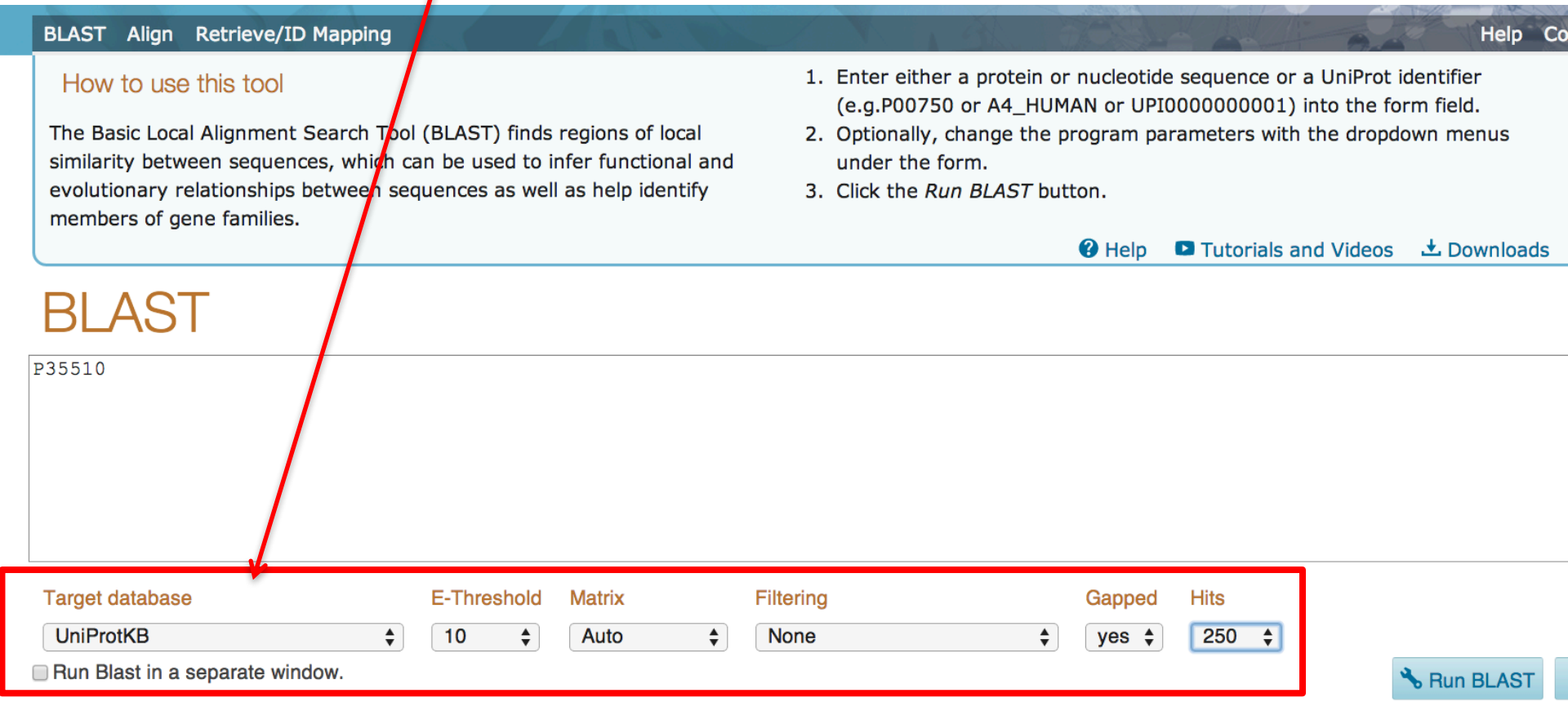
```
P35510 PAL1_ARATH    1   MEINGAHKSNGGGVDAMLCGGDIKTKNMVI--NAEDPLNWGAAAEQMKGSHLDEVKRMVA   58
P45724 PAL2_ARATH    1   ----------MDQIEAMLCGGGEKTKVAVTTKTLADPLNWGLAADQMKGSHLDEVKKMVE   50
F4JW69 F4JW69_ARATH  1   --------------------MEFR---QPNATALSDPLNWNVAAEALKGSHLEEVKKMVK   37
P45725 PAL3_ARATH    1   --------------------MEFR---QPNATALSDPLNWNVAAEALKGSHLEEVKKMVK   37
Q9SS45 PAL4_ARATH    1   --------------------MELCNQNNHITAVSGDPLNWNATAEALKGSHLDEVKRMVK   40
                                                     *****  :*: :*****:***:**

P35510 PAL1_ARATH   59   EFRKPVVNLGGETLTIGQVAAISTIGNSVKVELSETARAGVNASSDWVMESMNKGTDSYG   118
P45724 PAL2_ARATH   51   EYRRPVVNLGGETLTIGQVAAISTVGGSVKVELAETSRAGVKASSDWVMESMNKGTDSYG   110
F4JW69 F4JW69_ARATH 38   DYRKGTVQLGGETLTIGQVAAVAS--GGPTVELSEEARGGVKASSDWVMESMNRDTDTYG   95
P45725 PAL3_ARATH   38   DYRKGTVQLGGETLTIGQVAAVAS--GGPTVELSEEARGGVKASSDWVMESMNRDTDTYG   95
Q9SS45 PAL4_ARATH   41   EYRKEAVKLGGETLTIGQVAAVARGGGGSTVELAEEARAGVKASSEWVMESMNRGTDSYG   100
                         ::*: .*:**************::    . .***:* :*.**:***:*******: **:**

P35510 PAL1_ARATH   119  VTTGFGATSHRRTKNGVALQKELIRFLNAGIFGSTK---ETSHTLPHSATRAAMLVRINT   175
P45724 PAL2_ARATH   111  VTTGFGATSHRRTKNGTALQTELIRFLNAGIFGNTK---ETCHTLPQSATRAAMLVRVNT   167
F4JW69 F4JW69_ARATH 96   ITTGFGSSSRRRTDQGAALQKELIRYLNAGIFATGNEDDDRSNTLPRPATRAAMLIRVNT   155
P45725 PAL3_ARATH   96   ITTGFGSSSRRRTDQGAALQKELIRYLNAGIFATGNEDDDRSNTLPRPATRAAMLIRVNT   155
Q9SS45 PAL4_ARATH   101  VTTGFGATSHRRTKQGGALQNELIRFLNAGIFGPGAG--DTSHTLPKPTTRAAMLVRVNT   158
                         :******::*:***.:* ***.****:******.      : ..***: :*******:*:**

P35510 PAL1_ARATH   176  LLQGFSGIRFEILEAITSFLNNNITPSLPLRGTITASGDLVPLSYIAGLLTGRPNSKATG   235
P45724 PAL2_ARATH   168  LLQGYSGIRFEILEAITSLLNHNISPSLPLRGTITASGDLVPLSYIAGLLTGRPNSKATG   227
F4JW69 F4JW69_ARATH 156  LLQGYSGIRFEILEAITTLLNCKITPLLPLRGTITASGDLVPLSYIAGFLIGRPNSRSVG   215
P45725 PAL3_ARATH   156  LLQGYSGIRFEILEAITTLLNCKITPLLPLRGTITASGDLVPLSYIAGFLIGRPNSRSVG   215
Q9SS45 PAL4_ARATH   159  LLQGYSGIRFEILEAITKLLNHEITPCLPLRGTITASGDLVPLSYIAGLLTGRPNSKAVG   218
                         ****:************.:** :*:* **********************:* *****::.*

P35510 PAL1_ARATH   236  PNGEALTAEEAFKLAGISSGFFDLQPKEGLALVNGTAVGSGMASMVLFETNVLSVLAEIL   295
P45724 PAL2_ARATH   228  PDGESLTAKEAFEKAGISTGFFDLQPKEGLALVNGTAVGSGMASMVLFEANVQAVLAEVL   287
F4JW69 F4JW69_ARATH 216  PSGEILTALEAFKLAGVS-SFFELRPKEGLALVNGTAVGSALASTVLYDANILVVFSEVA   274
P45725 PAL3_ARATH   216  PSGEILTALEAFKLAGVS-SFFELRPKEGLALVNGTAVGSALASTVLYDANILVVFSEVA   274
Q9SS45 PAL4_ARATH   219  PSGETLTASEAFKLAGVS-SFFELQPKEGLALVNGTAVGSGLASTVLFDANILAVLSEVM   277
                         *.** *** ***: **:* .**:*:***************.:** **:::*:  *::*:
```

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



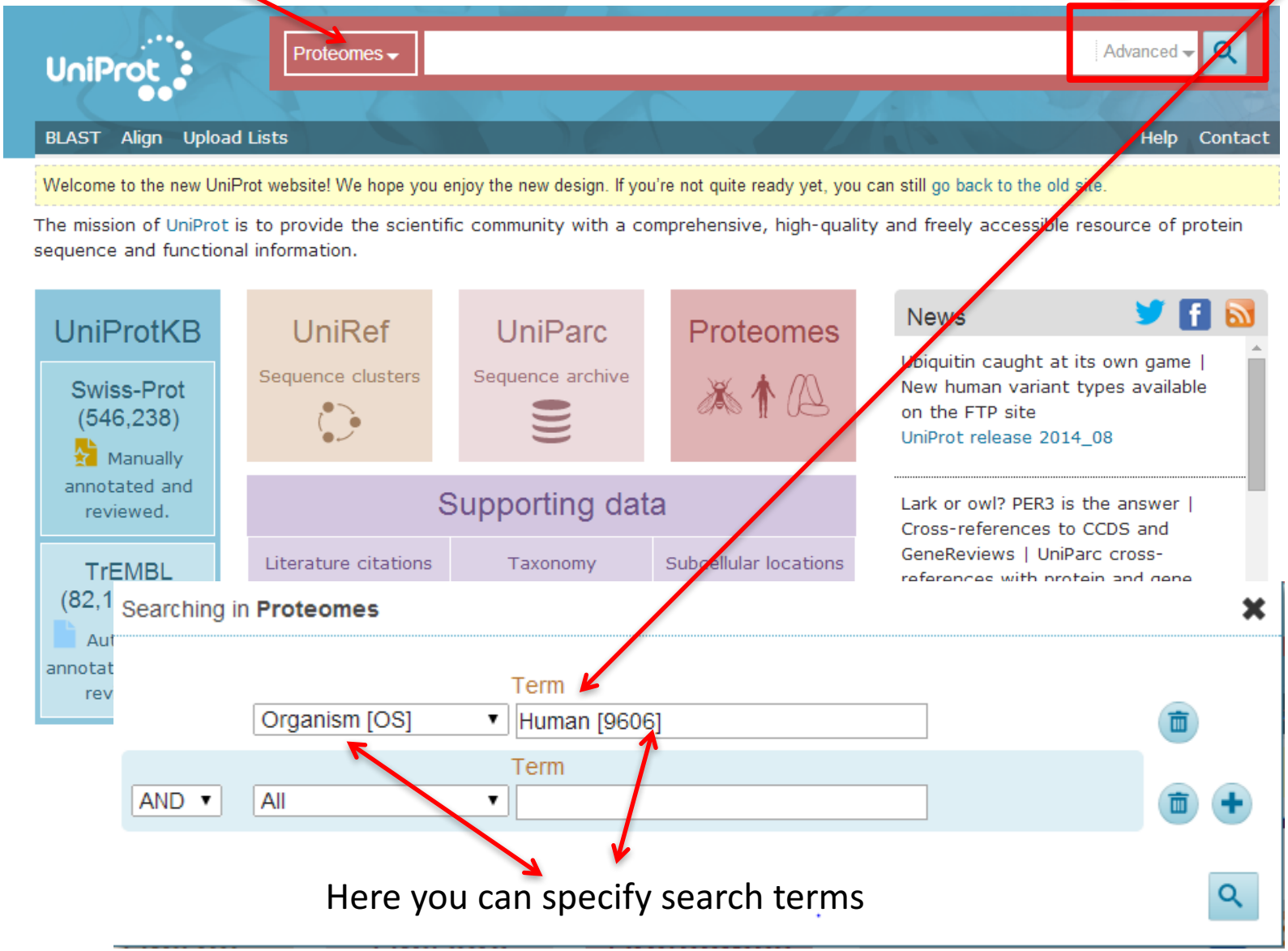
Choose advanced will allow to change BLAST parameters

Here you can make changes

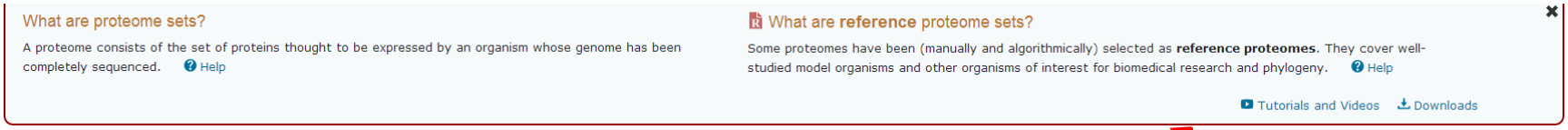


We are going to search UniProt proteomes for human protein set

Click on Advanced you will see a pop-out window

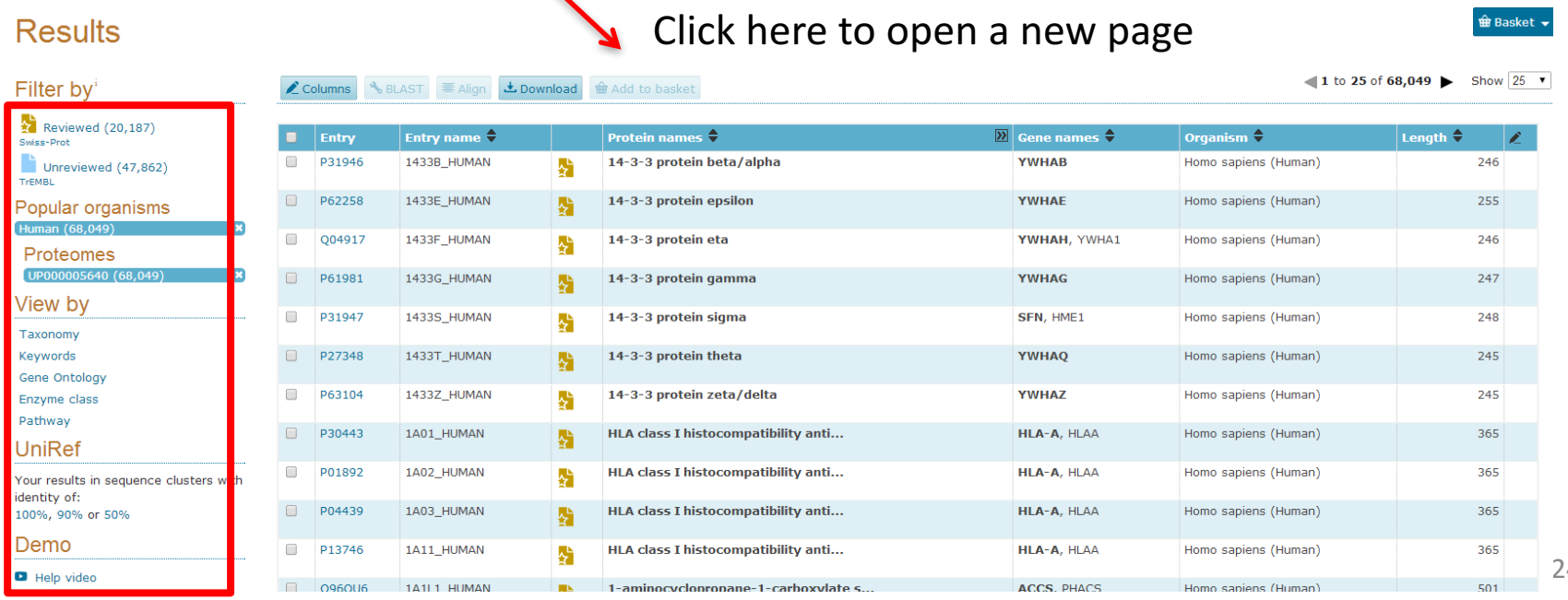


Here you can specify search terms



Click here to get help

Click here to open a new page



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;

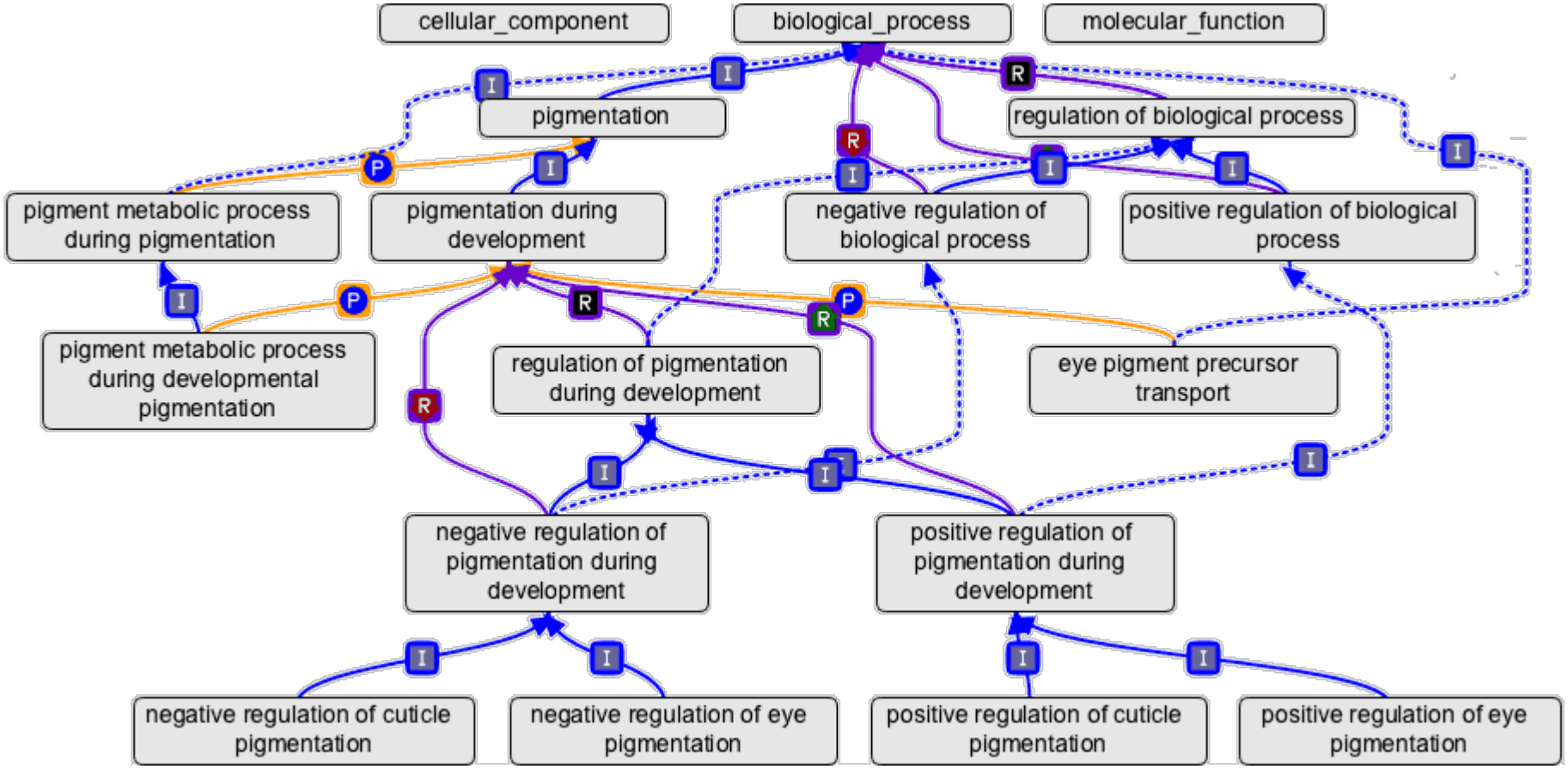
biological process, operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms

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



http://geneontology.org/page/ontology-structure

```
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:

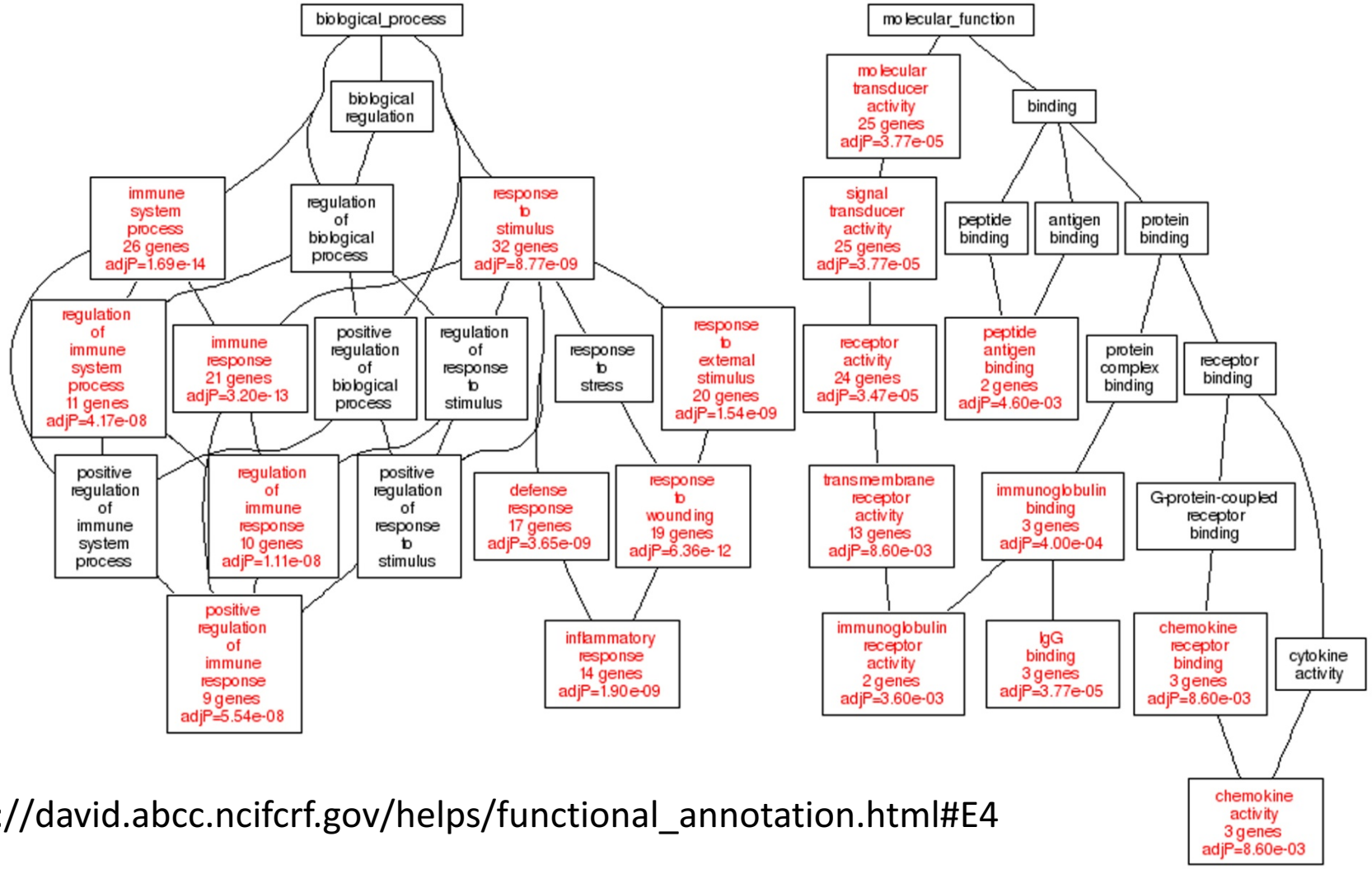
- integrating proteomic information from different organisms;
- assigning functions to protein domains;
- finding functional similarities in genes that are overexpressed or underexpressed in diseases and as we age;
- analysing groups of genes that are co-expressed during development;
- developing automated 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

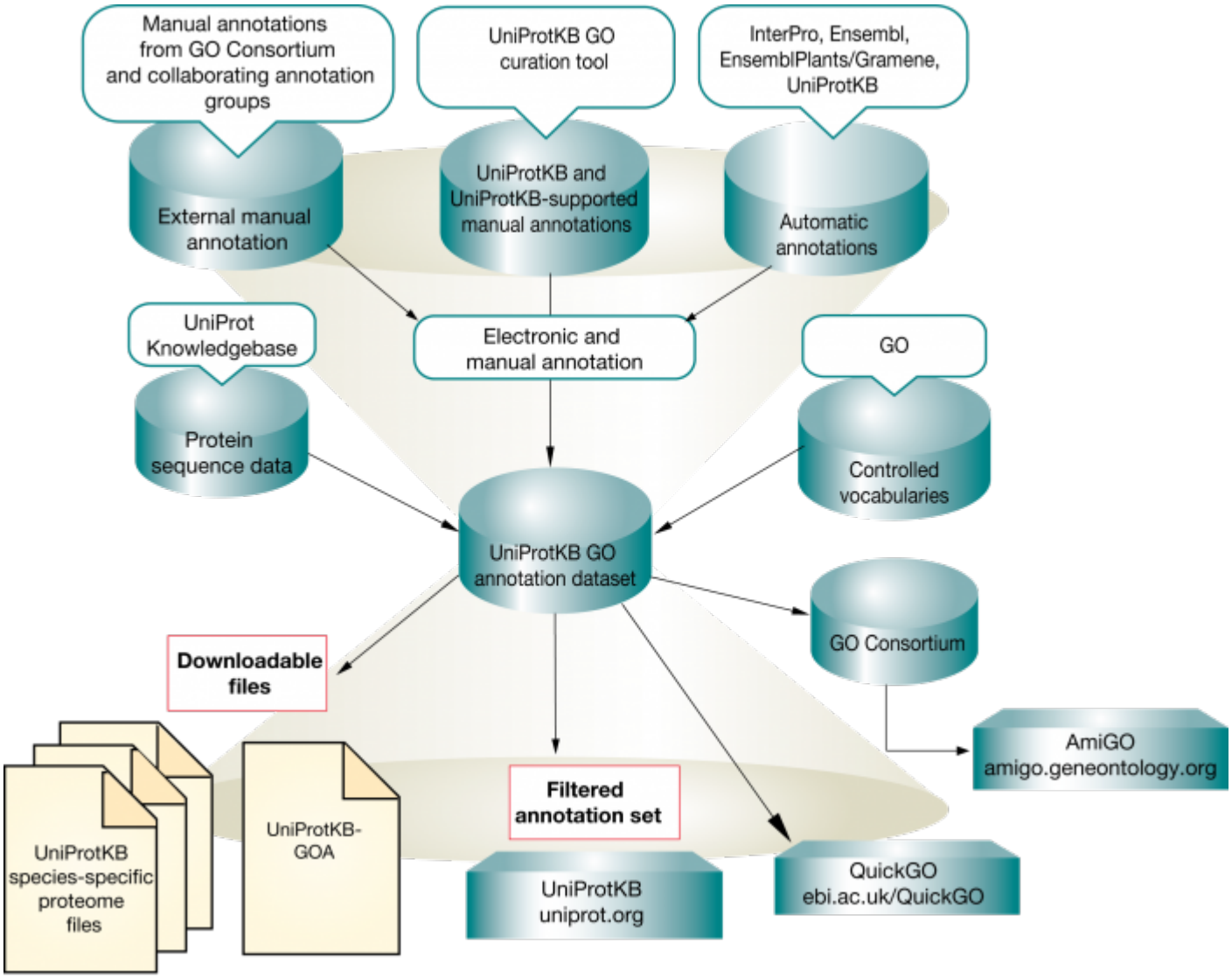
Enrichment analysis: use statistical test e.g. Fisher exact test

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

UniProt-GO annotation (GOA)



http://www.ebi.ac.uk/training/online/course/uniprot-goa-quick-tour/what-uniprot-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

http://geneontology.org/

Gene Ontology Consortium | Home | Documentation | Downloads | User stories | Community | Tools | About | Contact us

Search GO data

terms and gene products

Search

Enrichment analysis (beta)

Your genes here...

biological process

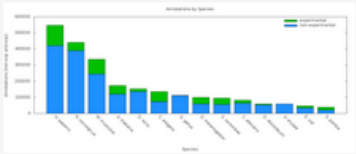
H. sapiens

Submit

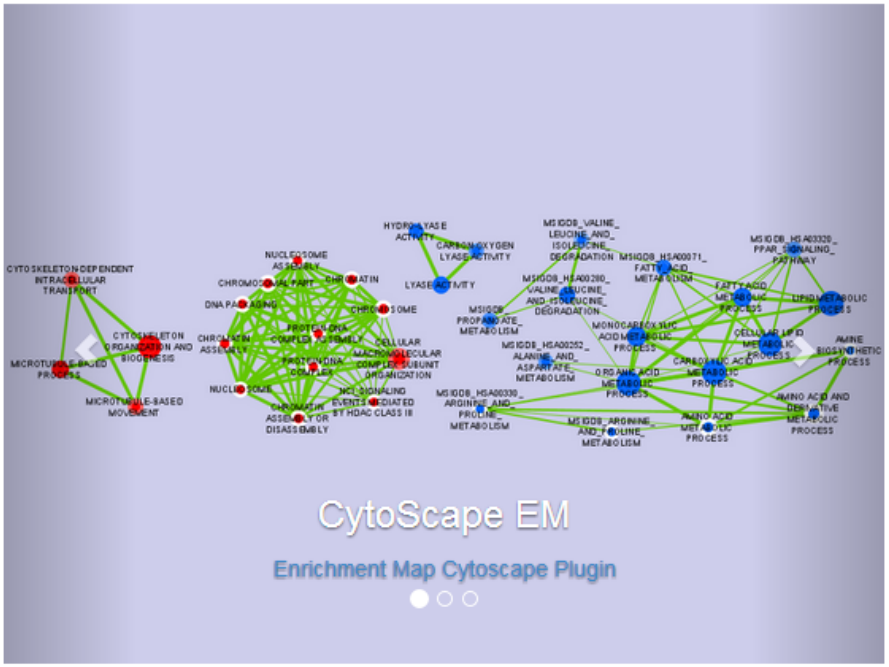
Advanced options

Powered by PANTHER

Statistics



Gene Ontology Consortium



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

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 Gene Ontology and the KEGG Pathway

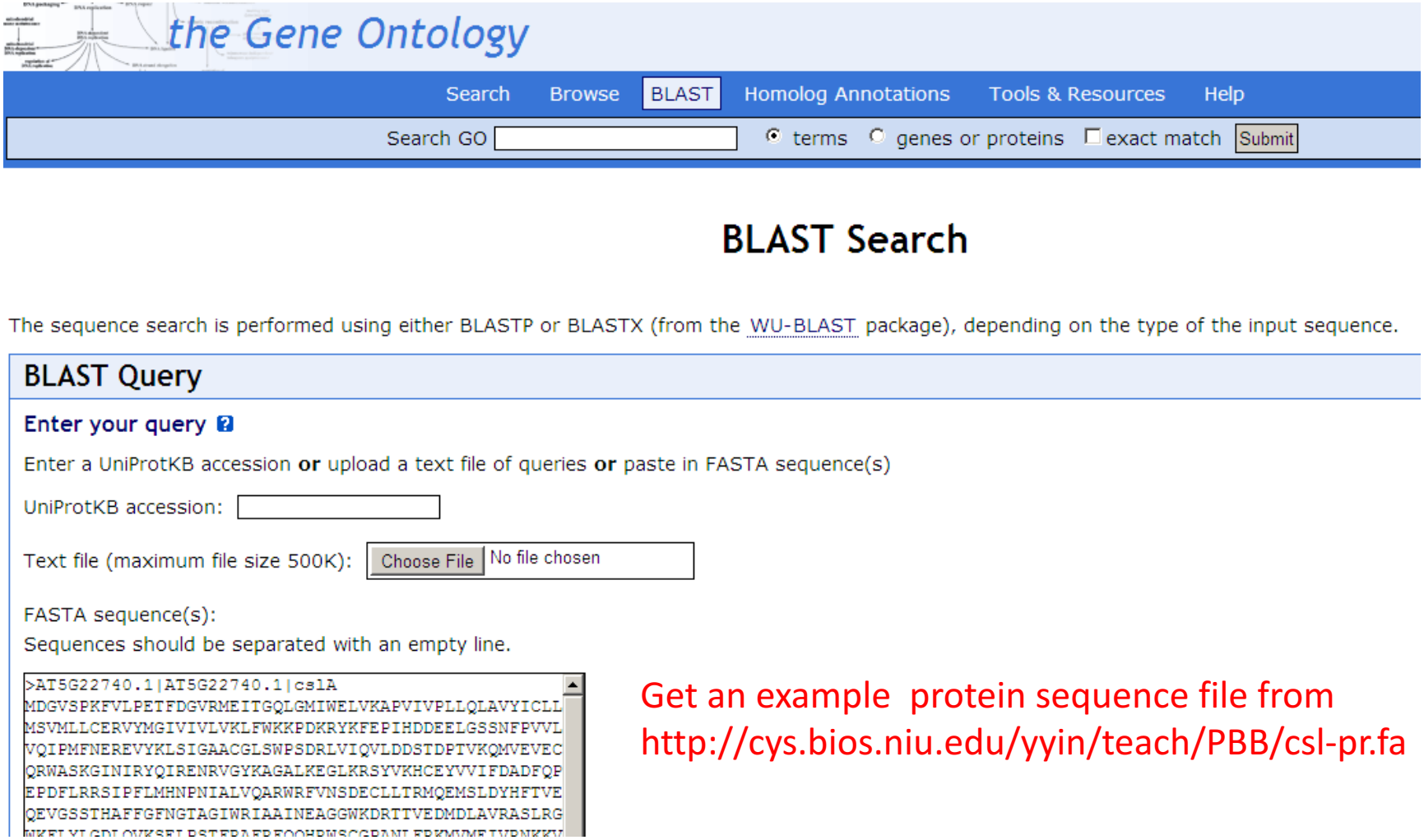
Analysis of Tumor Suppressor Genes Based on Gene Ontology and the KEGG Pathway

An association analysis between psychophysical characteristics and genome-wide gene ...

An association analysis between psychophysical characteristics and genome-wide gene ...

Differentiation of the two rice subspecies indica and japonica

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



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

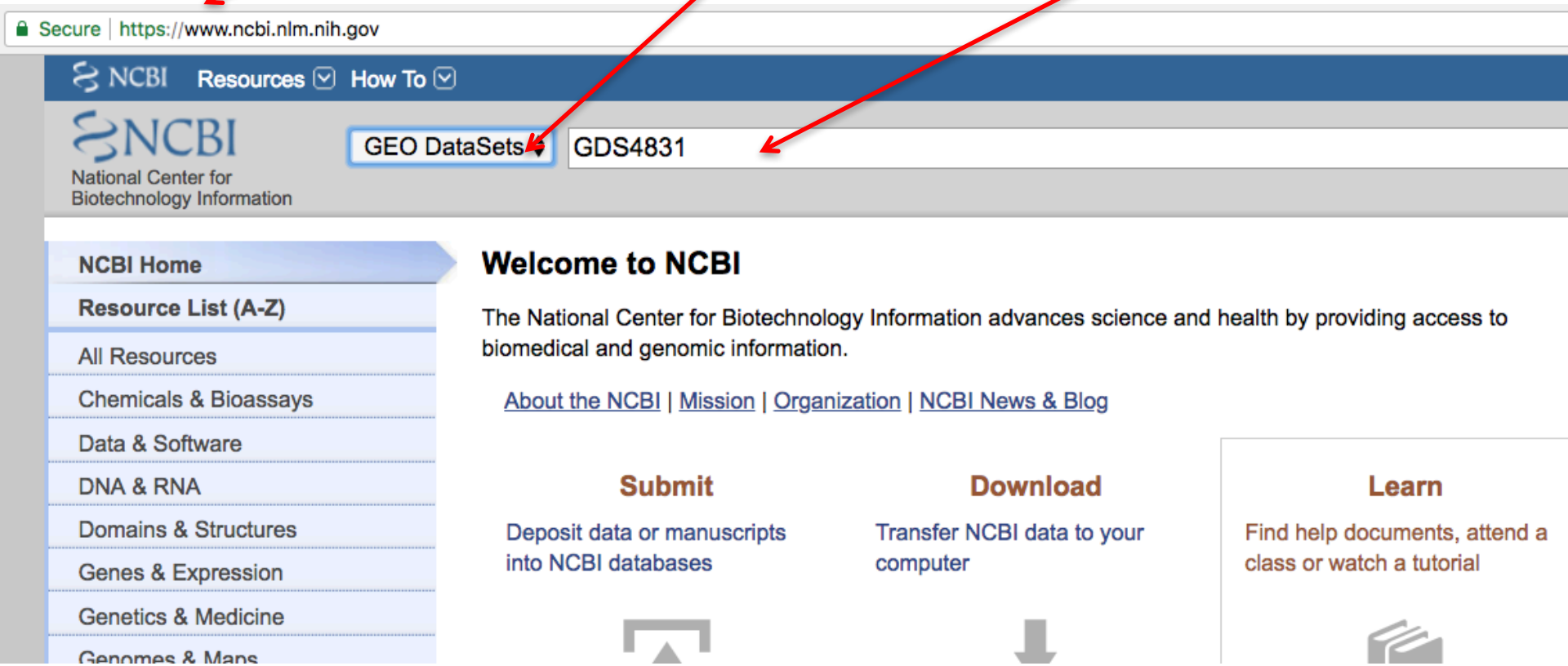
Select all | Clear all | Perform an action with this page's selected gene products... | Go!

Symbol, full name		Information	P value
CSLA02 **cellulose synthase-like A02**	BLAST match view associations BLAST with CSLA02	**protein** from *Arabidopsis thaliana*	3.6e-295
ATCSLA09	BLAST match view associations BLAST with ATCSLA09	**protein** from *Arabidopsis thaliana*	4.5e-217
CSLA03 **cellulose synthase-like A3**	BLAST match view associations BLAST with CSLA03	**protein** from *Arabidopsis thaliana*	4.1e-191
ATCSLA15	BLAST match view associations BLAST with ATCSLA15	**protein** from *Arabidopsis thaliana*	2.9e-183
CSLA07 **cellulose synthase like**	BLAST match view associations BLAST with CSLA07	**protein** from *Arabidopsis thaliana*	1.2e-182
CSLA10 **cellulose synthase-like A10**	BLAST match view associations BLAST with CSLA10	**protein** from *Arabidopsis thaliana*	3.3e-173
CSLA01 **cellulose synthase-like A01**	BLAST match view associations BLAST with CSLA01	**protein** from *Arabidopsis thaliana*	4.0e-170
CSLA14 **cellulose synthase like A14**	BLAST match view associations BLAST with CSLA14	**protein** from *Arabidopsis thaliana*	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



Choose “Compare 2 sets of samples”

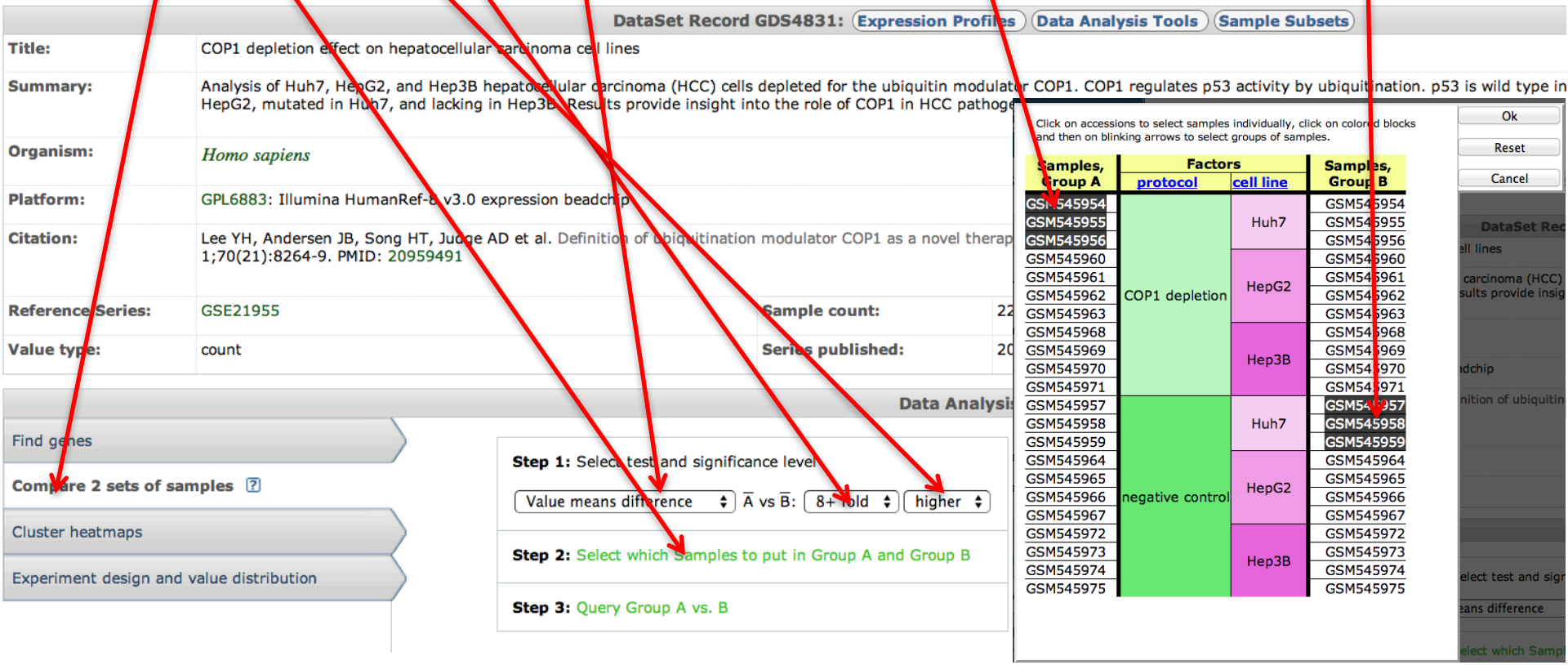
Choose “Value means difference”
Choose “8+ fold”
Choose “higher”

Then go to Step 2

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

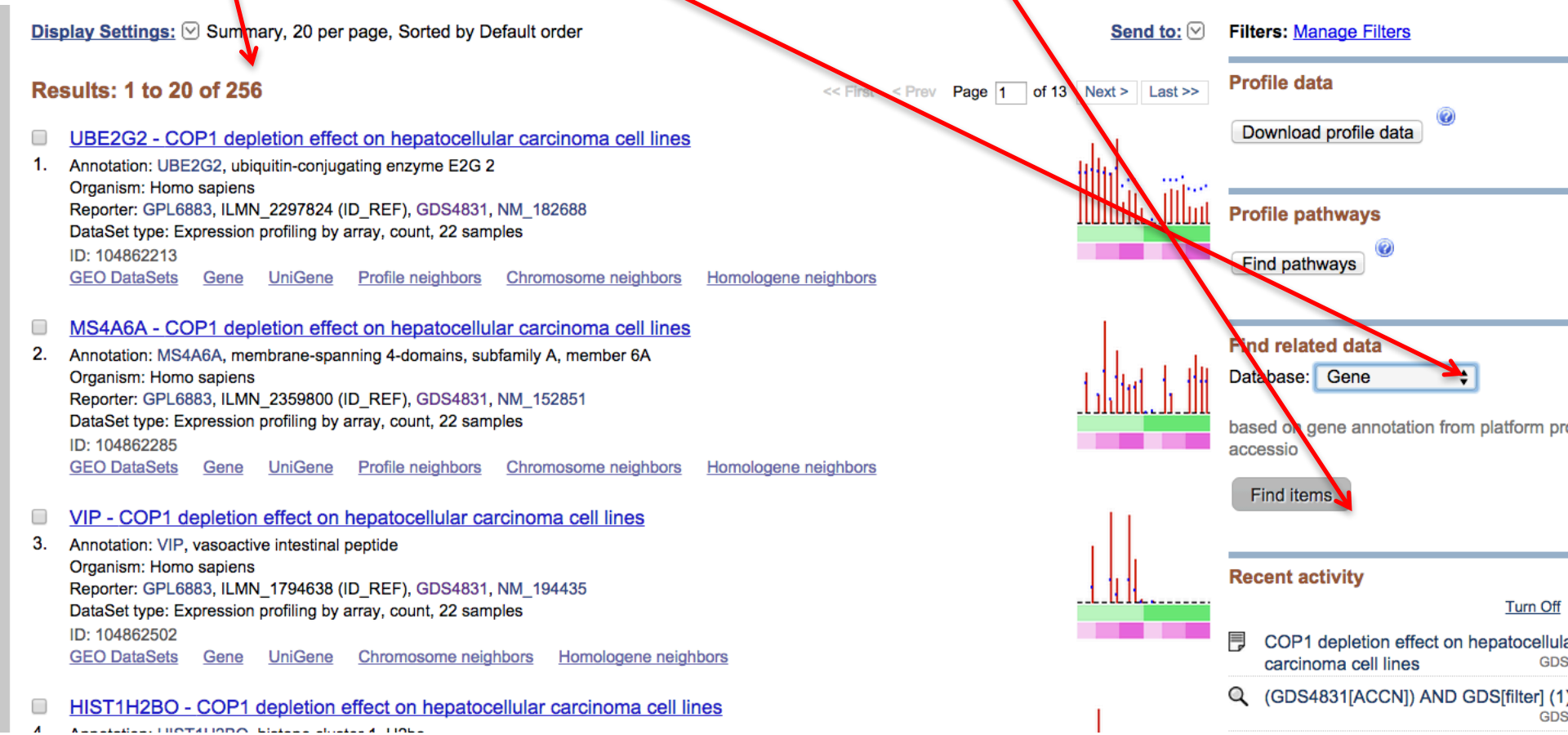
Group B: three samples for negative control and Huh7 cell line

Hit ok, and go to Step 3



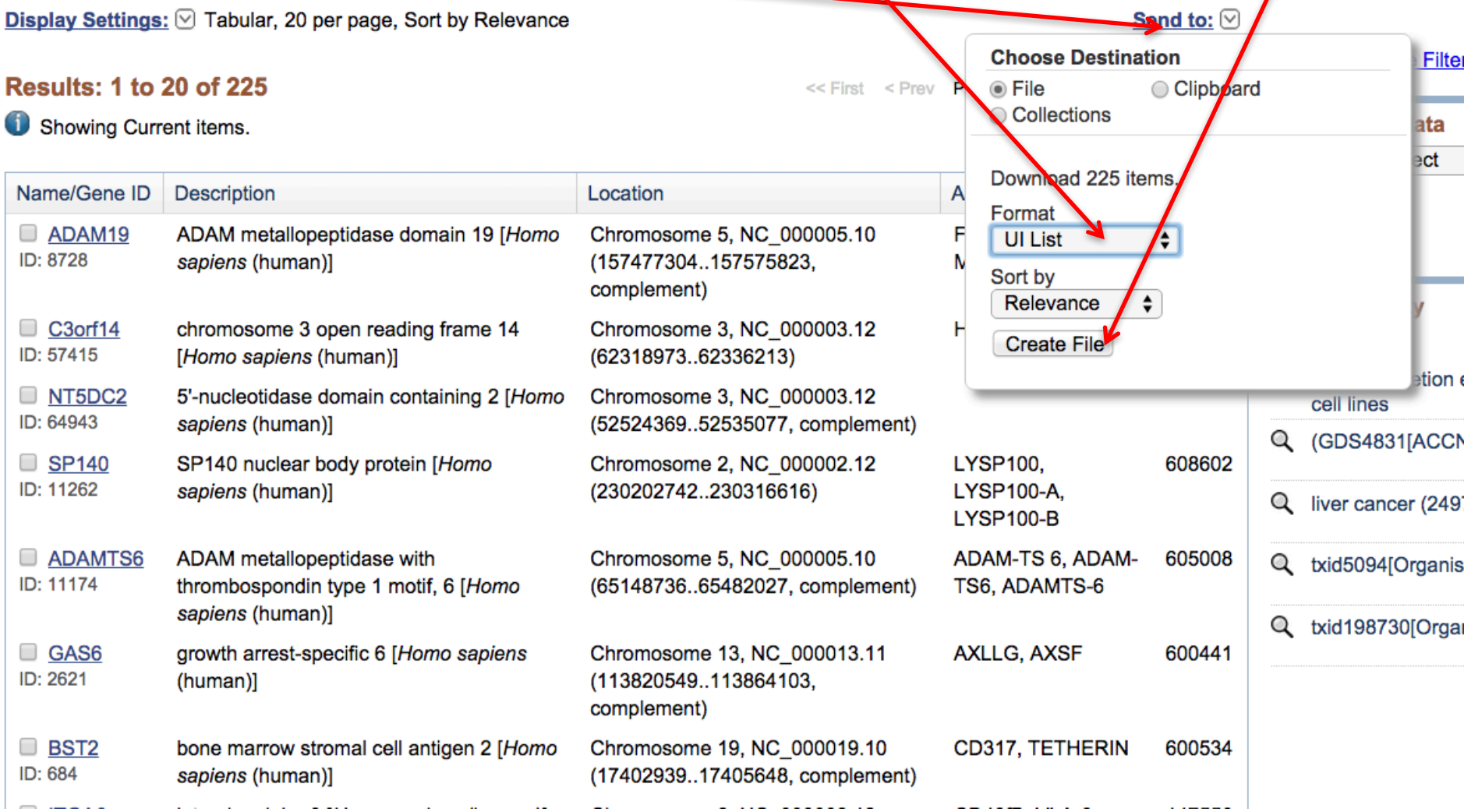
Total 256 gene profiles are found with 8+ fold higher expression in COP 1 depletion than in negative control in Huh7 cell line

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



Total 225 genes correspond to 256 gene profiles

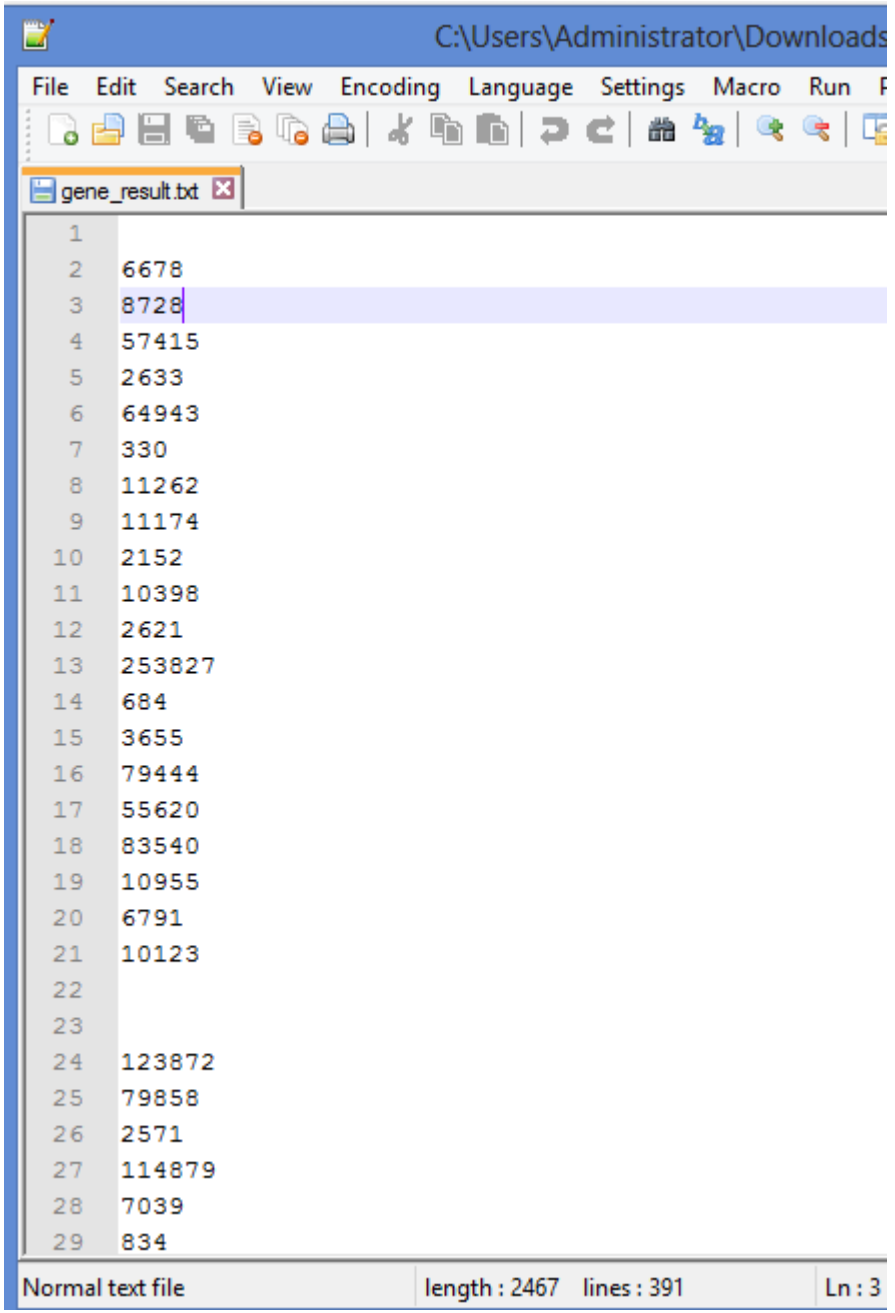
To download the list of Gene IDs, hit Send to, choose UI list as format and hit Create file



A file named “gene_result.txt” will be automatically downloaded to your local computer

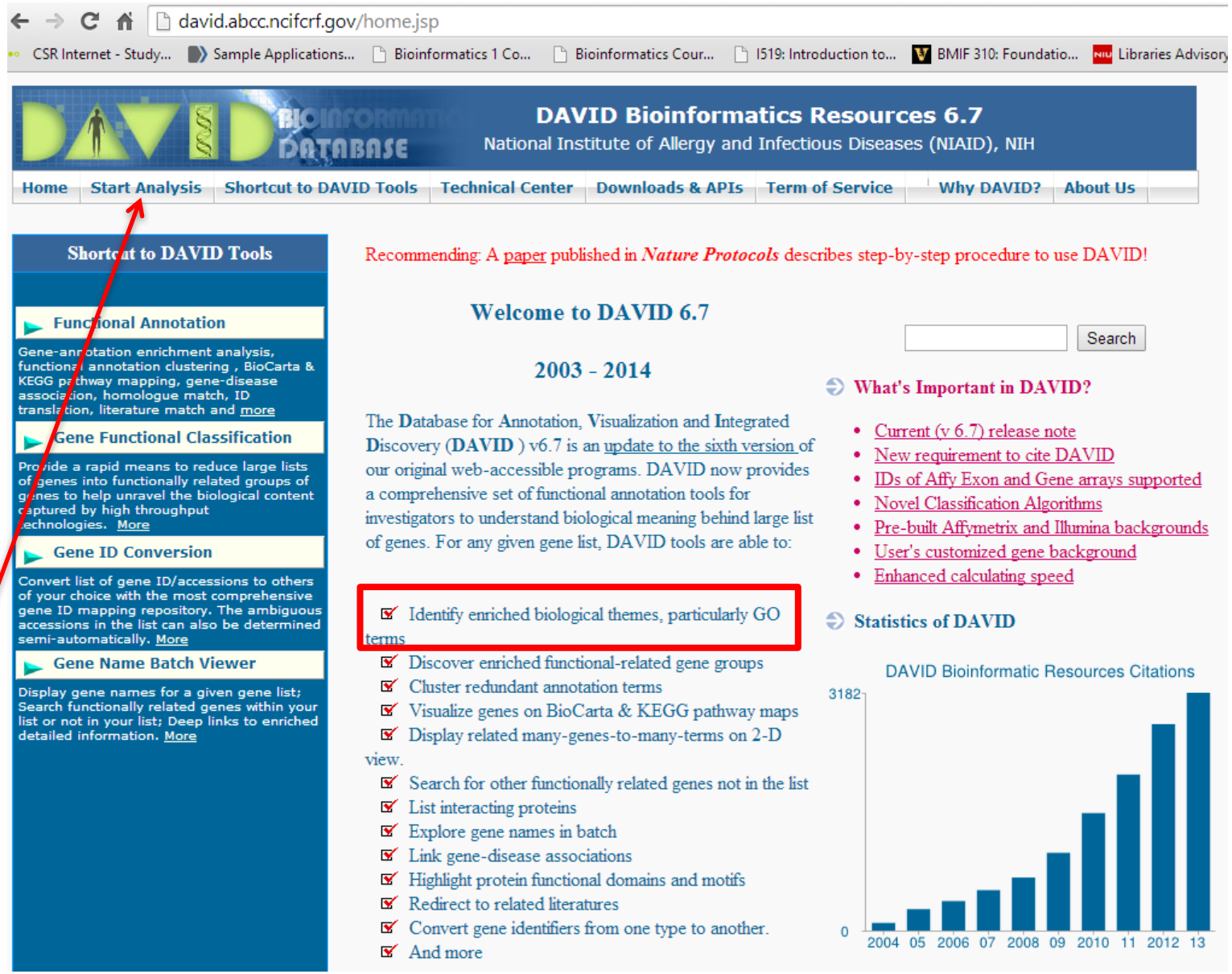
Find out where it is downloaded to, open it using notepad++

View the file using notepad++

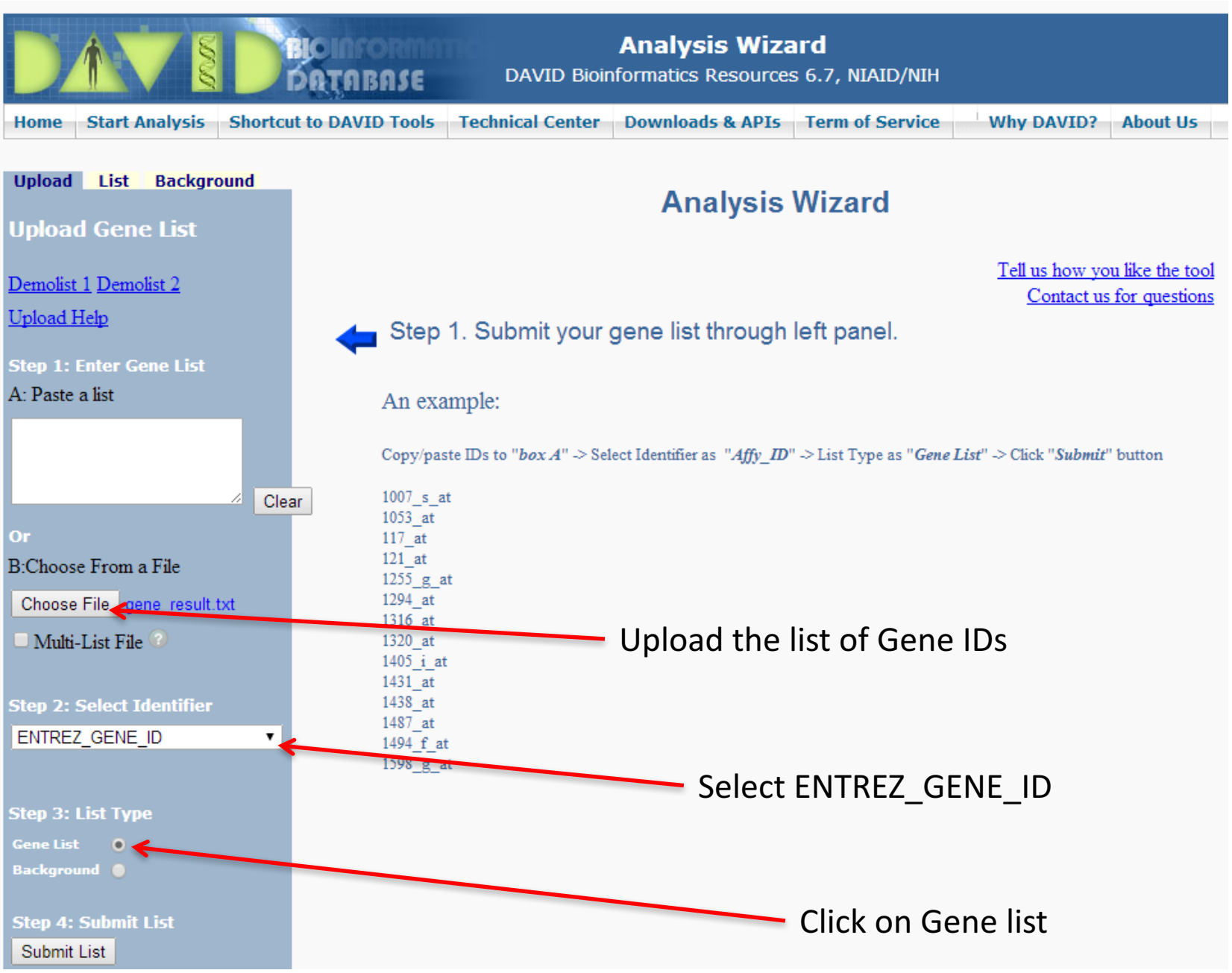


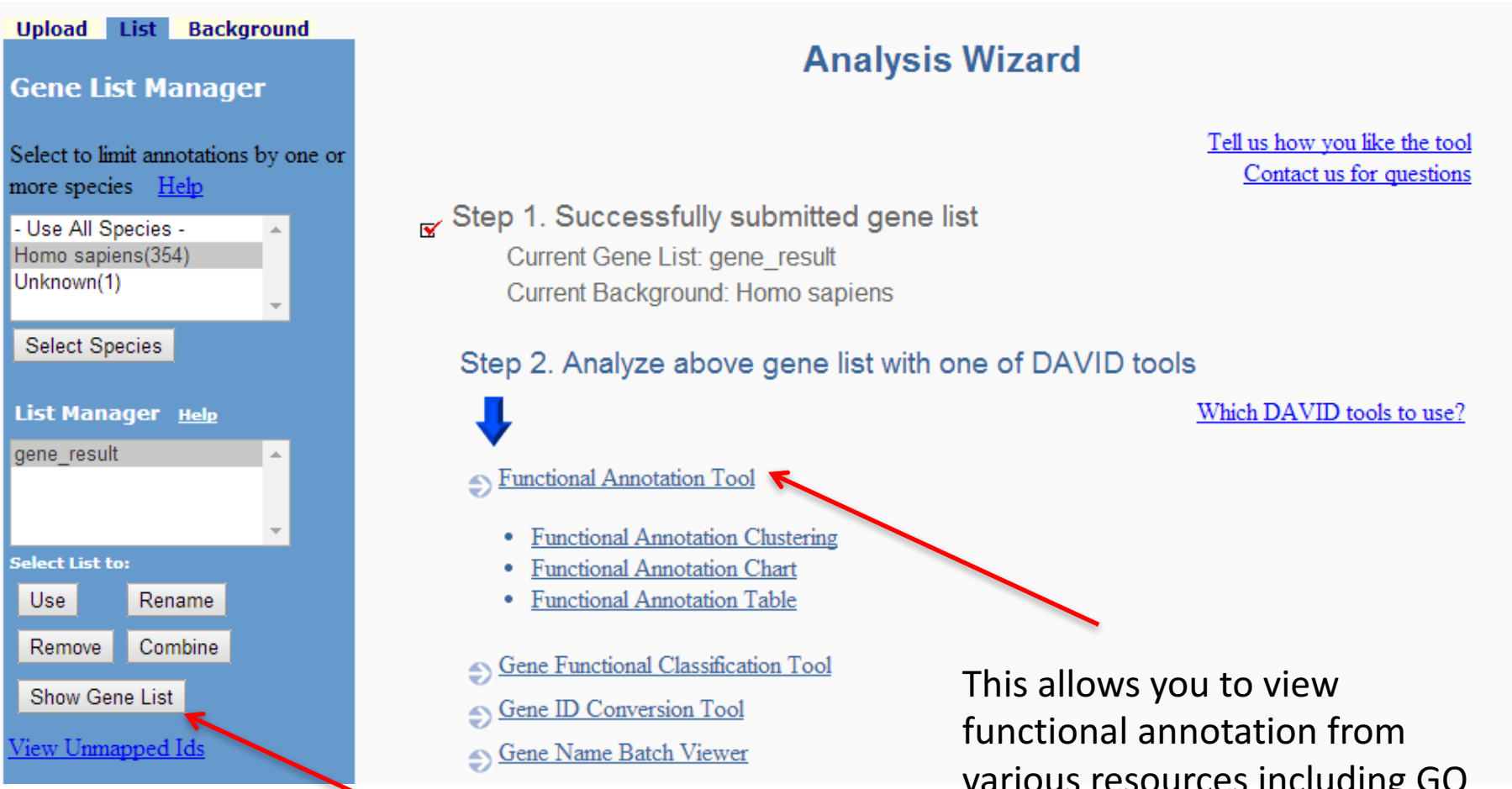
Next we will use DAVID to perform function enrichment analysis

The **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated**D**iscovery (**DAVID**)



Hit start analysis

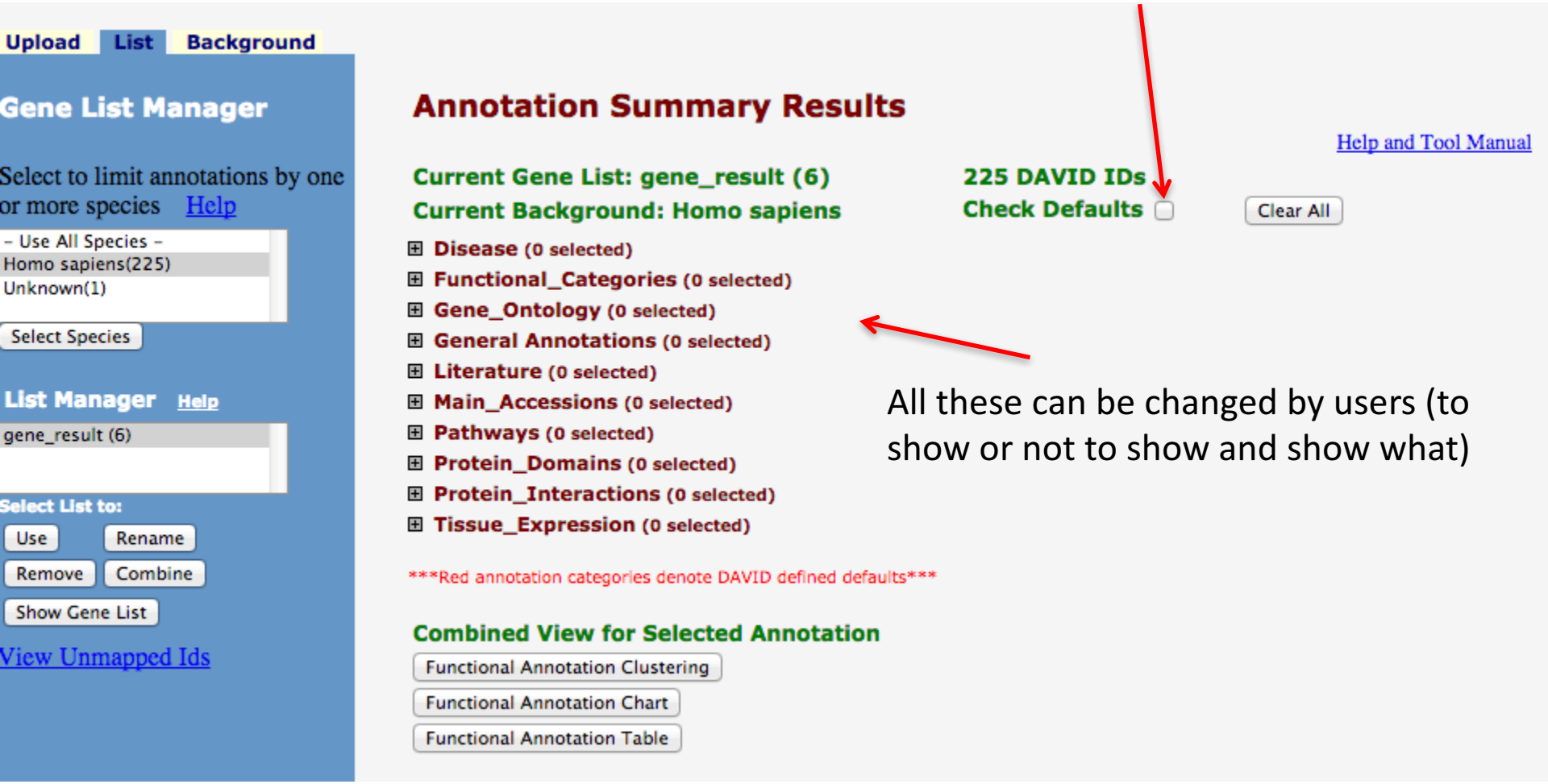

Analysis Wizard
DAVID Bioinformatics Resources 6.7, NIAID/NIH
Home
Start Analysis
Shortcut to DAVID Tools
Technical Center
Downloads & APIs
Term of Service
Why DAVID?
About Us
Upload
List
Background
Upload Gene List
Demolist 1 Demolist 2
Upload Help
Step 1: Enter Gene List
A: Paste a list
Clear
Or
B:Choose 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

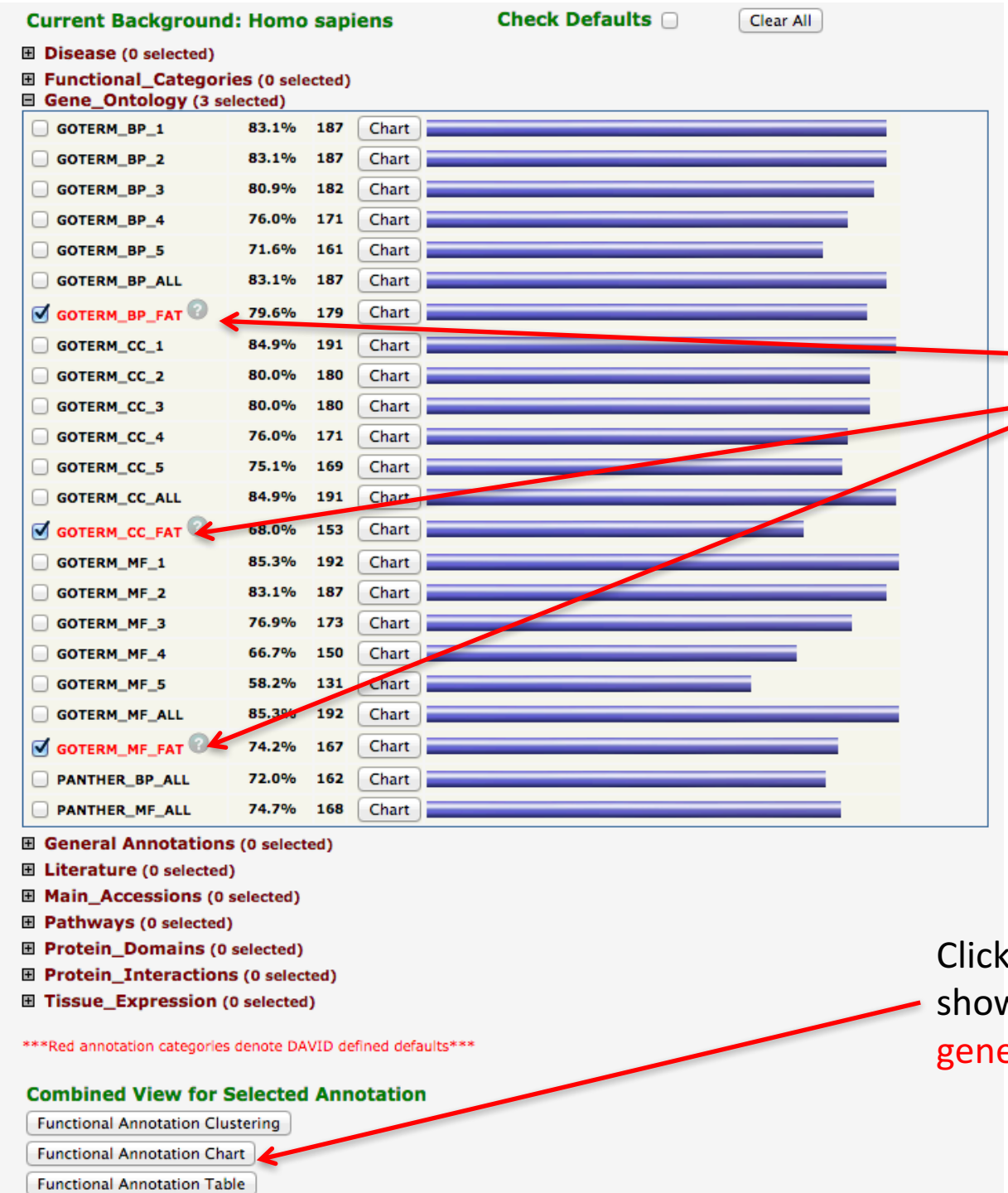


This allows you to view functional annotation from various resources including GO

Check the submitted gene list

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





Select just GO

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

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