EBI web resources I: databases and tools

Yanbin Yin

Fall 2015

Outline

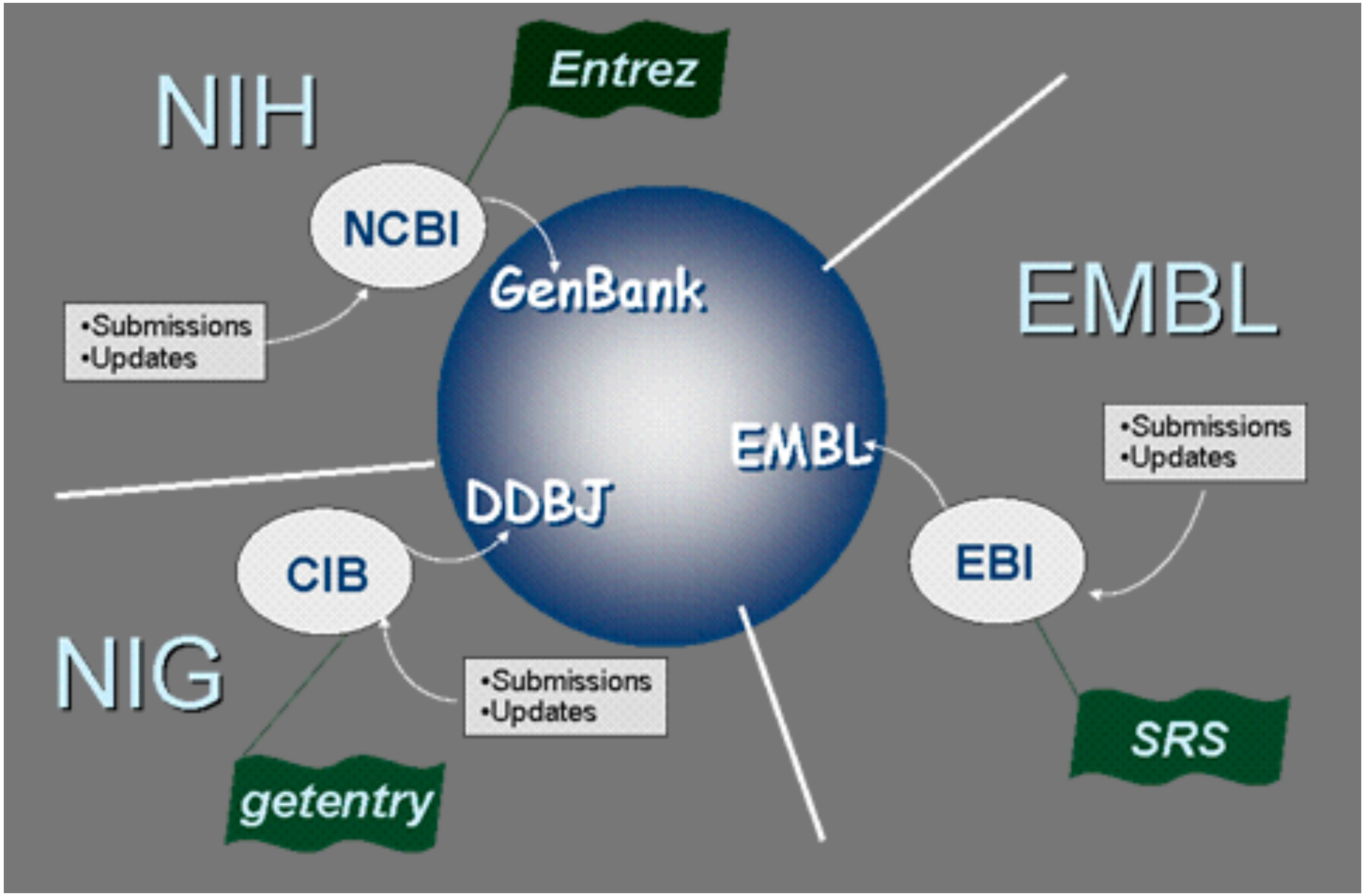
- Intro to EBI

- 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)

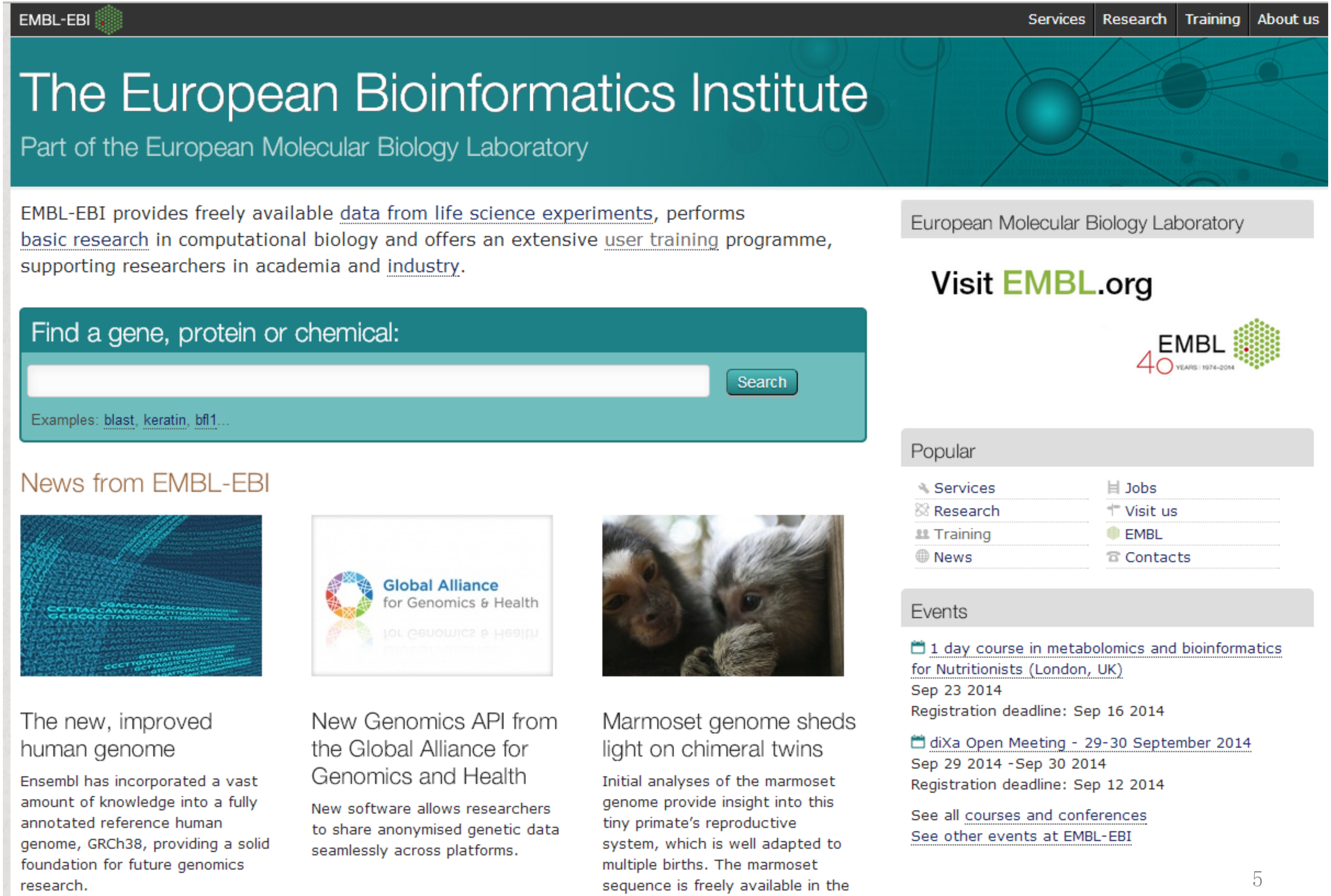
United Kingdom
f Man
Blackpool
Leeds
Manchester
Hull
Liverpool
Sheffield
Stoke-on-Trent
Nottingham
Birmingham
Coventry
A
Haarlem
Nederland
Rotterdam
Bristol
London
Cardiff
Eindhoven
Ewell
Gent
Antwerpen
Southampton
Mechelen
Exeter
Brighton
Kortrijk
België
Bournemouth
Asco
Belgique
Torquay
English Channel
Lëtzebuerg
Octeville
Le Havre
Guernsey
Reims

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/



The European Bioinformatics Institute

Part of the European Molecular Biology Laboratory

EMBL-EBI provides freely available data from life science experiments, performs basic research in computational biology and offers an extensive user training programme, supporting researchers in academia and industry.

Find a gene, protein or chemical:

Search

Examples: blast, keratin, bfl1...

News from EMBL-EBI

The new, improved human genome

Ensembl has incorporated a vast amount of knowledge into a fully annotated reference human genome, GRCh38, providing a solid foundation for future genomics research.

New Genomics API from the Global Alliance for Genomics and Health

New software allows researchers to share anonymised genetic data seamlessly across platforms.

Marmoset genome sheds light on chimeral twins

Initial analyses of the marmoset genome provide insight into this tiny primate's reproductive system, which is well adapted to multiple births. The marmoset sequence is freely available in the

European Molecular Biology Laboratory

Visit EMBL.org

Popular

- Services
- Research
- Training
- News
- Jobs
- Visit us
- EMBL
- Contacts

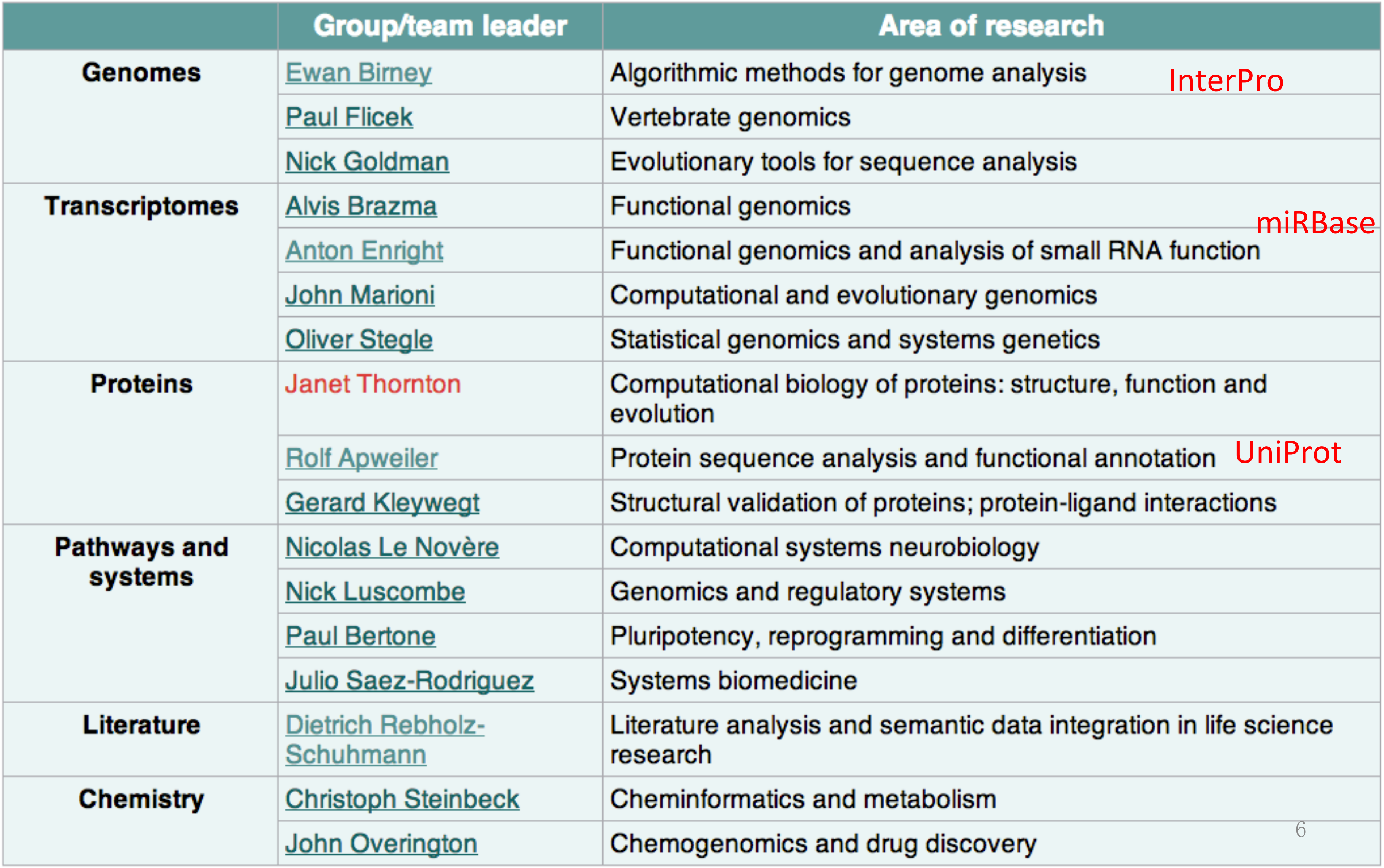
Events

1 day course in metabolomics and bioinformatics for Nutritionists (London, UK)
Sep 23 2014
Registration deadline: Sep 16 2014

diXa Open Meeting - 29-30 September 2014
Sep 29 2014 -Sep 30 2014
Registration deadline: Sep 12 2014

See all courses and conferences
See other events at EMBL-EBI

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

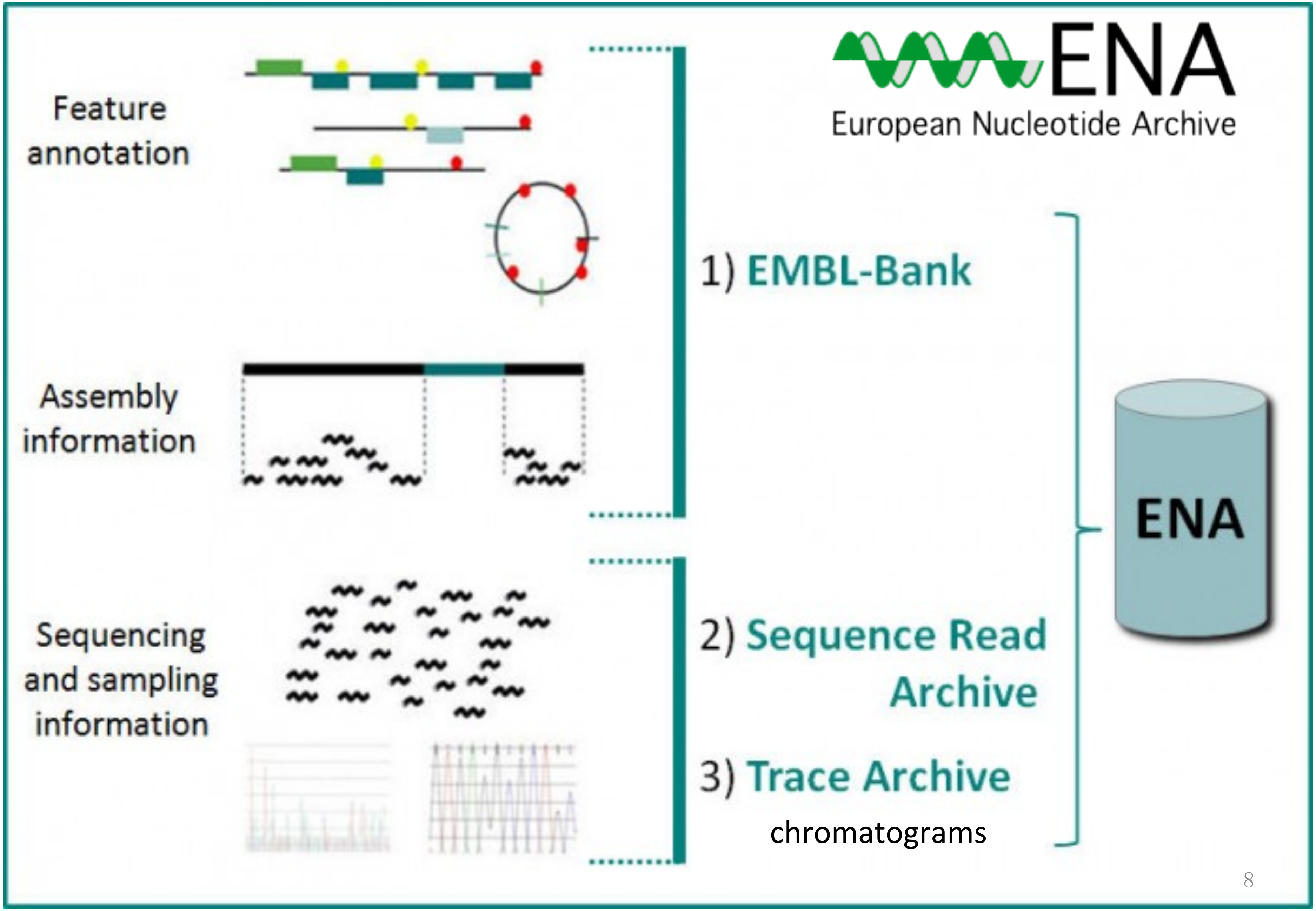
Major databases in EBI

GenBank	EMBL-Bank (DNA and RNA sequences)
Genome MapView	Ensembl (genomes)
GEO	ArrayExpress(microarray-based gene-expression data)
nr (GenPept)	UniProt (protein sequences)
CDD	InterPro(protein families, domains and motifs)
MMDB	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



ENA

European Nucleotide Archive

Feature annotation

Assembly information

Sequencing and sampling information

1) EMBL-Bank

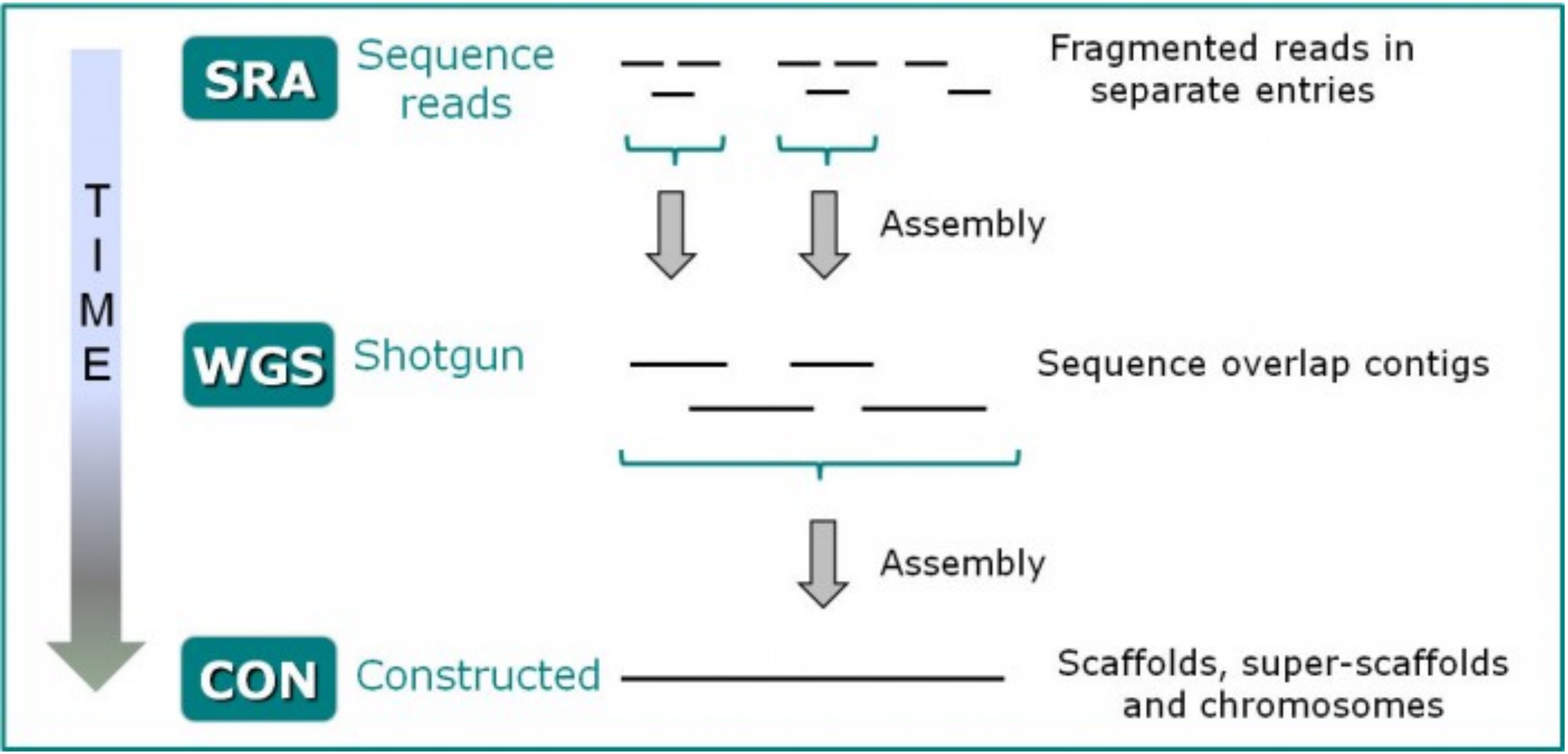
2) Sequence Read Archive

3) Trace Archive

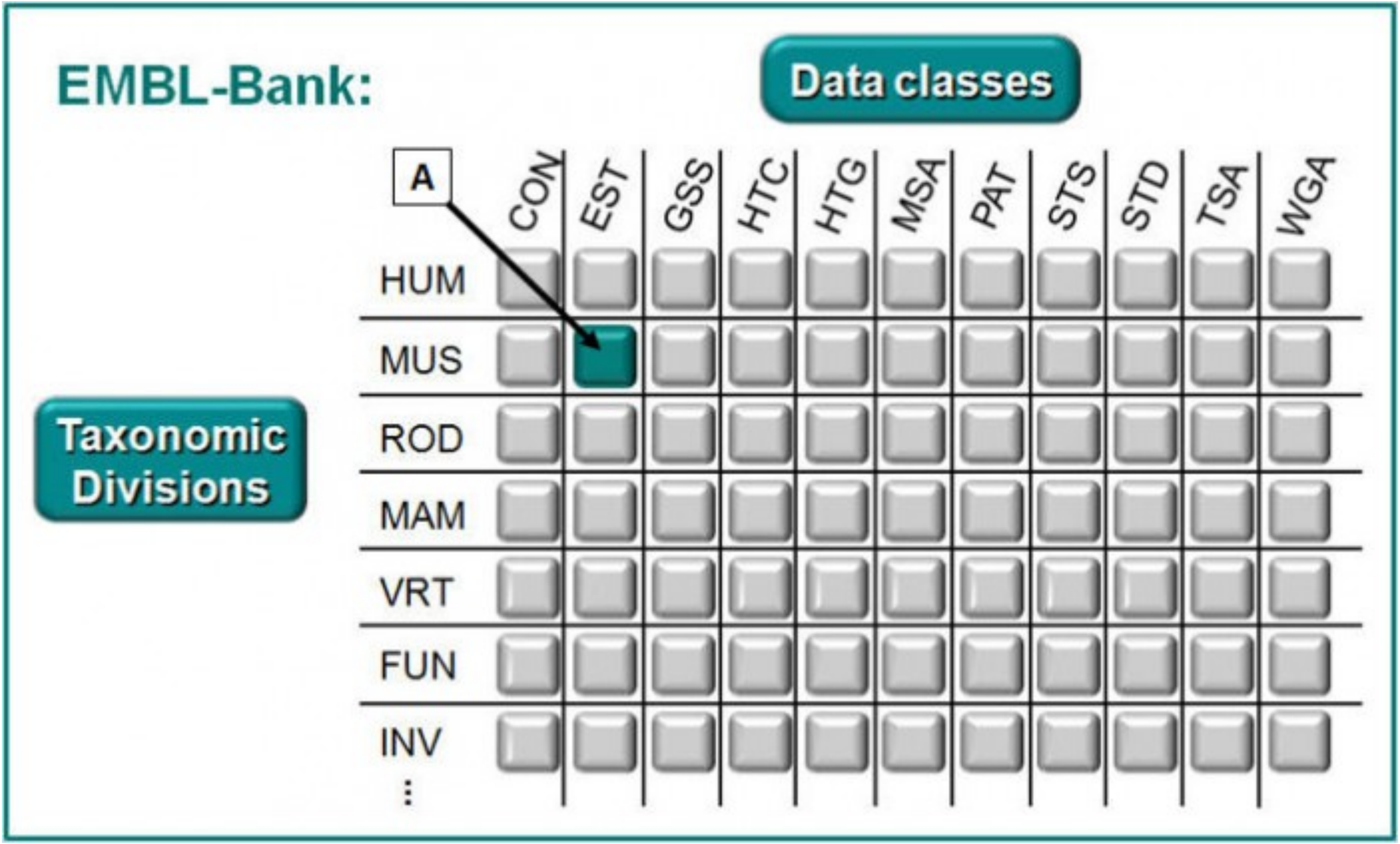
chromatograms

ENA

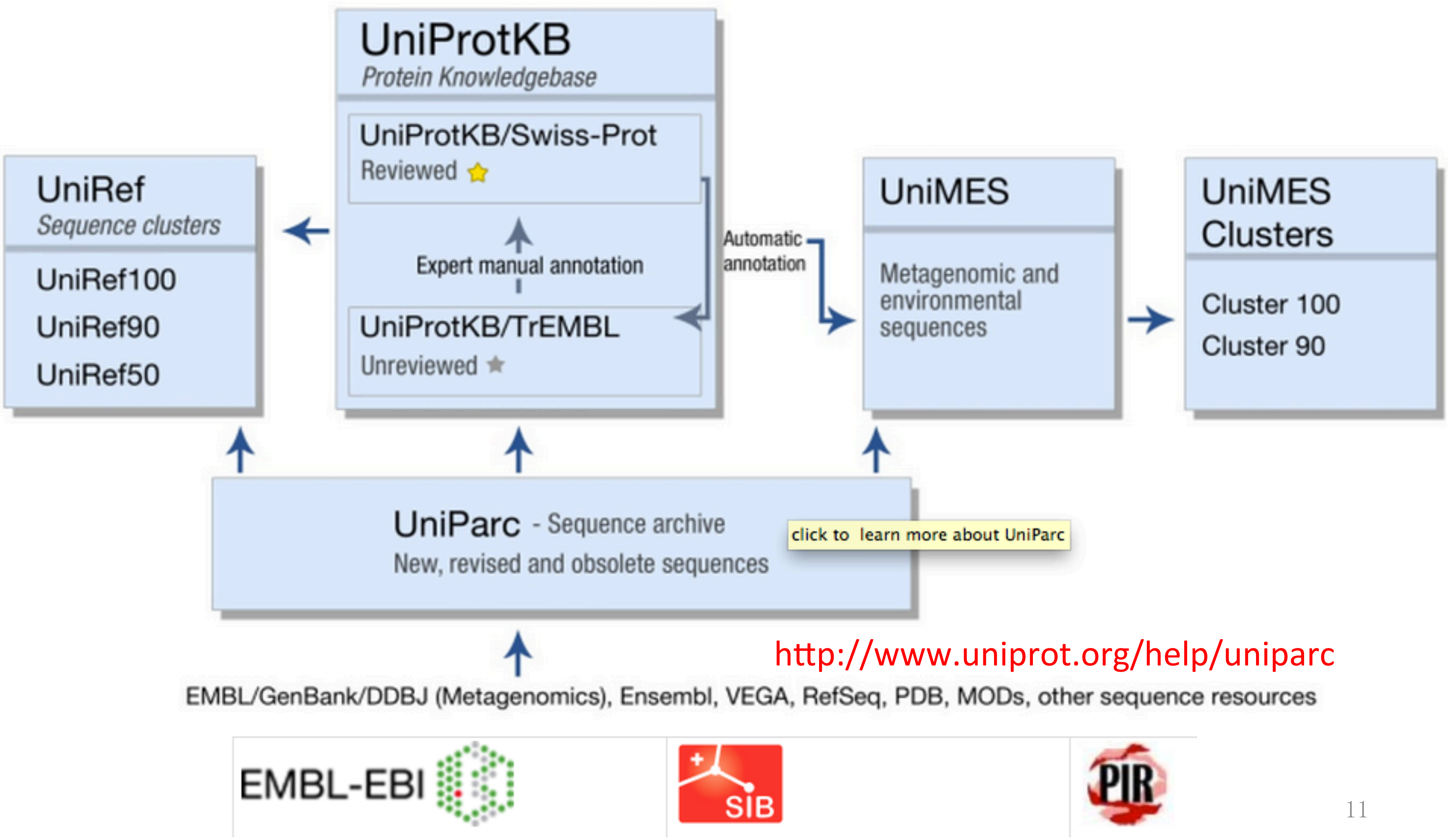
Sequence might first enter ENA as SRA (Sequence Read Archive) fragmented sequence reads; it might be re-submitted as assembled WGS (Whole Genome Shotgun) sequence overlap contigs; it might be re-submitted again with further assembly as CON (Constructed) sequence entries, with the older WGS entries being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



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



UniProt



UniProtKB
Protein Knowledgebase

UniProtKB/Swiss-Prot
Reviewed ★

Expert manual annotation

UniProtKB/TrEMBL
Unreviewed ★

UniRef
Sequence clusters

UniRef100
UniRef90
UniRef50

Automatic annotation

UniMES

Metagenomic and environmental sequences

UniMES Clusters

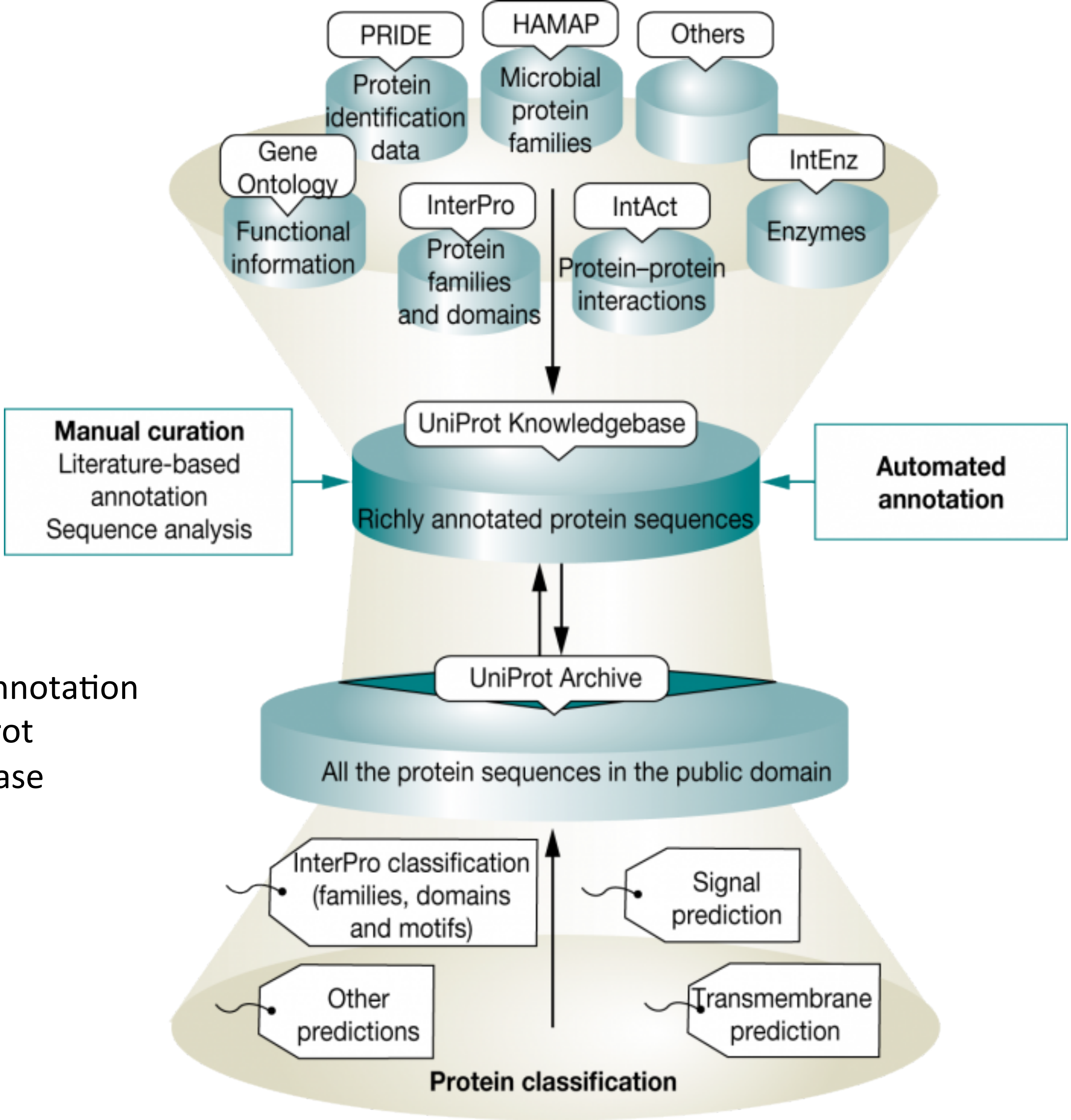
Cluster 100
Cluster 90

UniParc - Sequence archive
New, revised and obsolete sequences

click to learn more about UniParc

EMBL/GenBank/DDBJ (Metagenomics), Ensembl, VEGA, RefSeq, PDB, MODs, other sequence resources

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


PRIDE
Protein identification data
HAMAP
Microbial protein families
Others
Gene Ontology
Functional information
InterPro
Protein families and domains
IntAct
Protein–protein interactions
IntEnz
Enzymes
UniProt Knowledgebase
Richly annotated protein sequences
Manual curation
Literature-based annotation
Sequence analysis
Automated annotation
UniProt Archive
All the protein sequences in the public domain
InterPro classification (families, domains and motifs)
Signal prediction
Other predictions
Transmembrane prediction
Protein classification

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

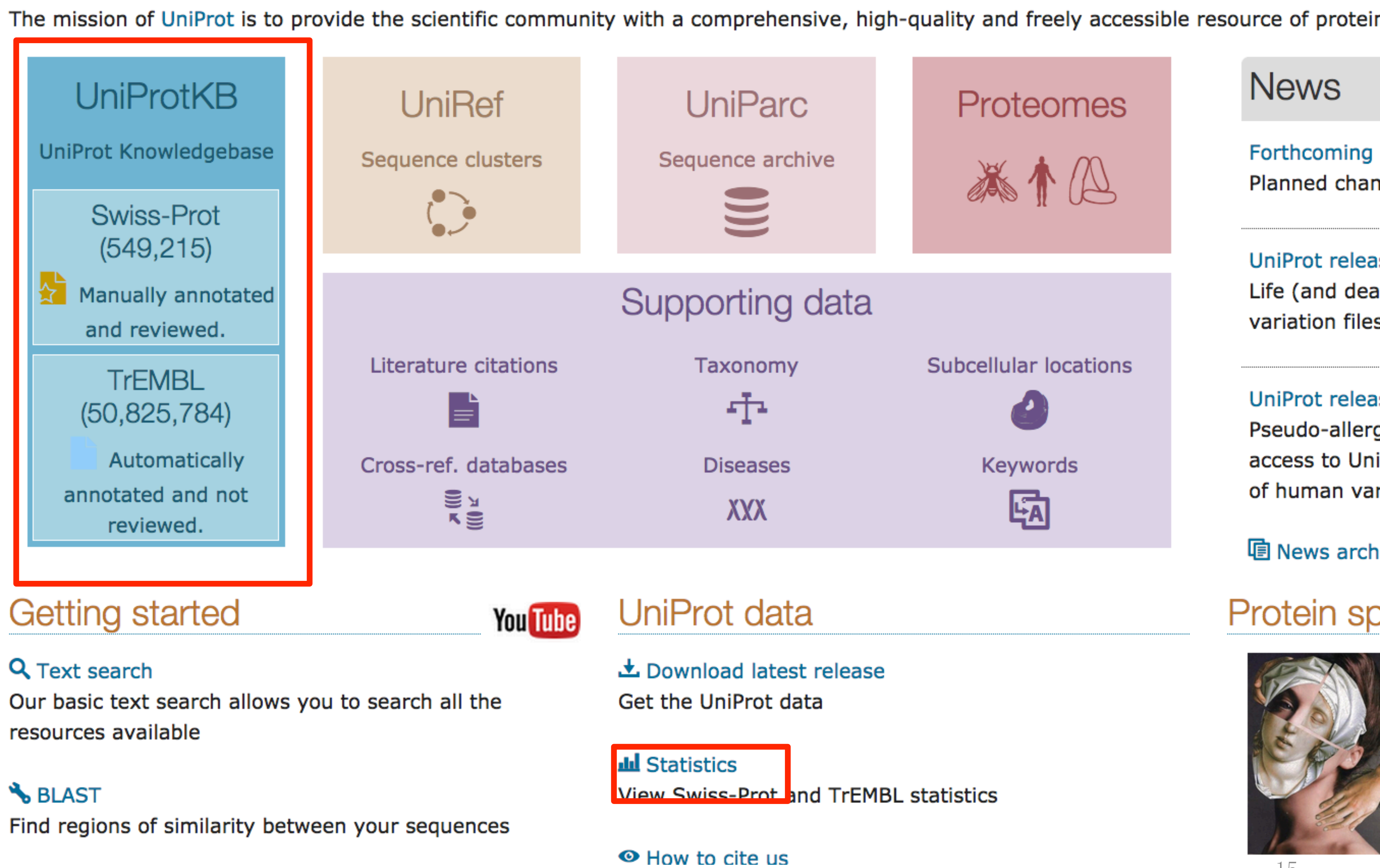
Scientific Database Curator job : Cambridge, United Kingdom
http://www.nature.com/naturejobs/science/jobs/444213-scientific-database-curator

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 protei

UniProtKB

UniProt Knowledgebase

Swiss-Prot (549,215)

Manually annotated and reviewed.

TrEMBL (50,825,784)

Automatically annotated and not reviewed.

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

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

BLAST
Find regions of similarity between your sequences

UniProt data

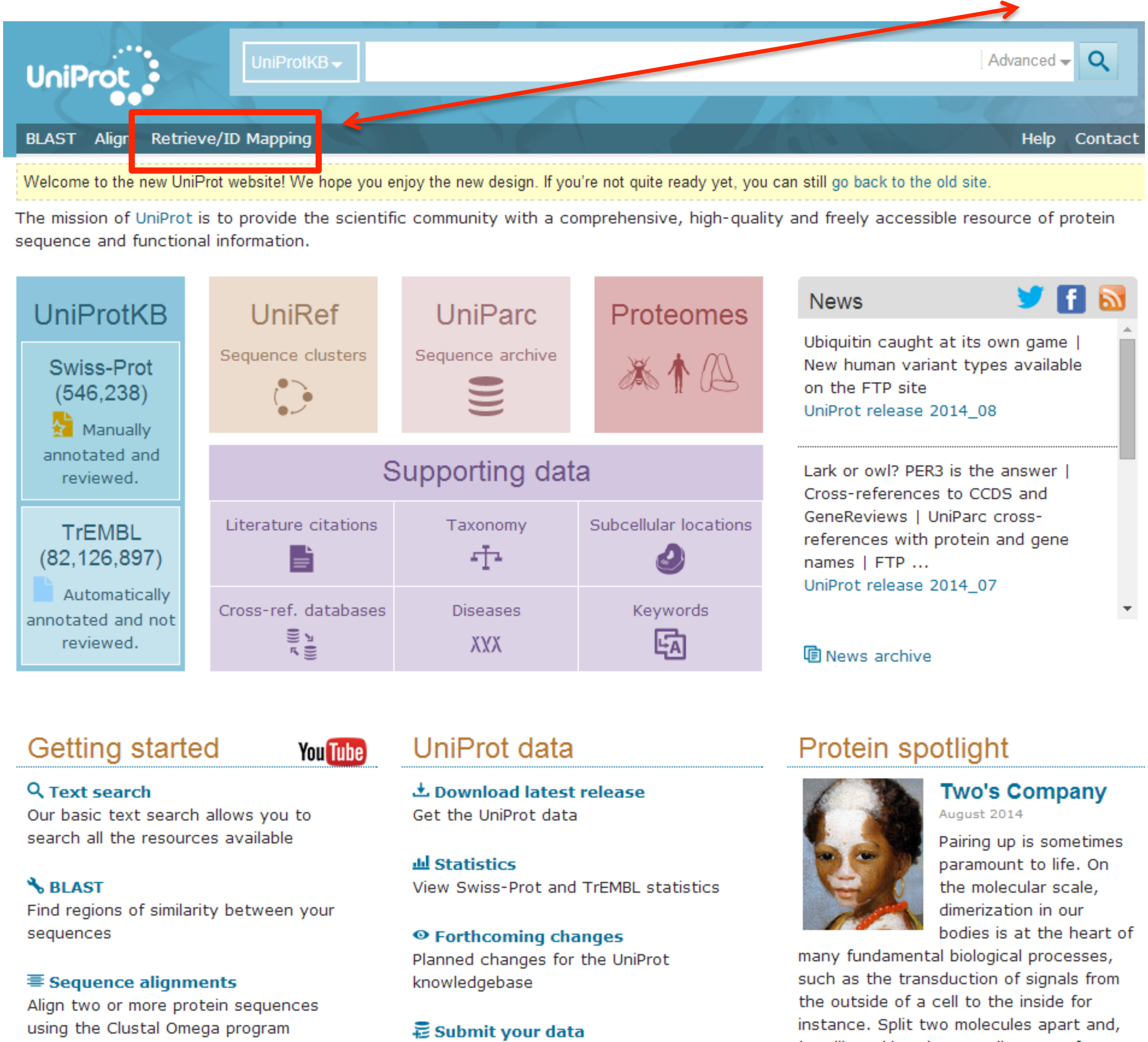
Download latest release
Get the UniProt data

Statistics
View Swiss-Prot and TrEMBL statistics

How to cite us

Protein sp

We are going to do ID mapping



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.

UniProtKB

Swiss-Prot (546,238)

Manually annotated and reviewed.

TrEMBL (82,126,897)

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

Ubiquitin caught at its own game | New human variant types available on the FTP site

UniProt release 2014_08

Lark or owl? PER3 is the answer | Cross-references to CCDS and GeneReviews | UniParc cross-references with protein and gene names | FTP ...

UniProt release 2014_07

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

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

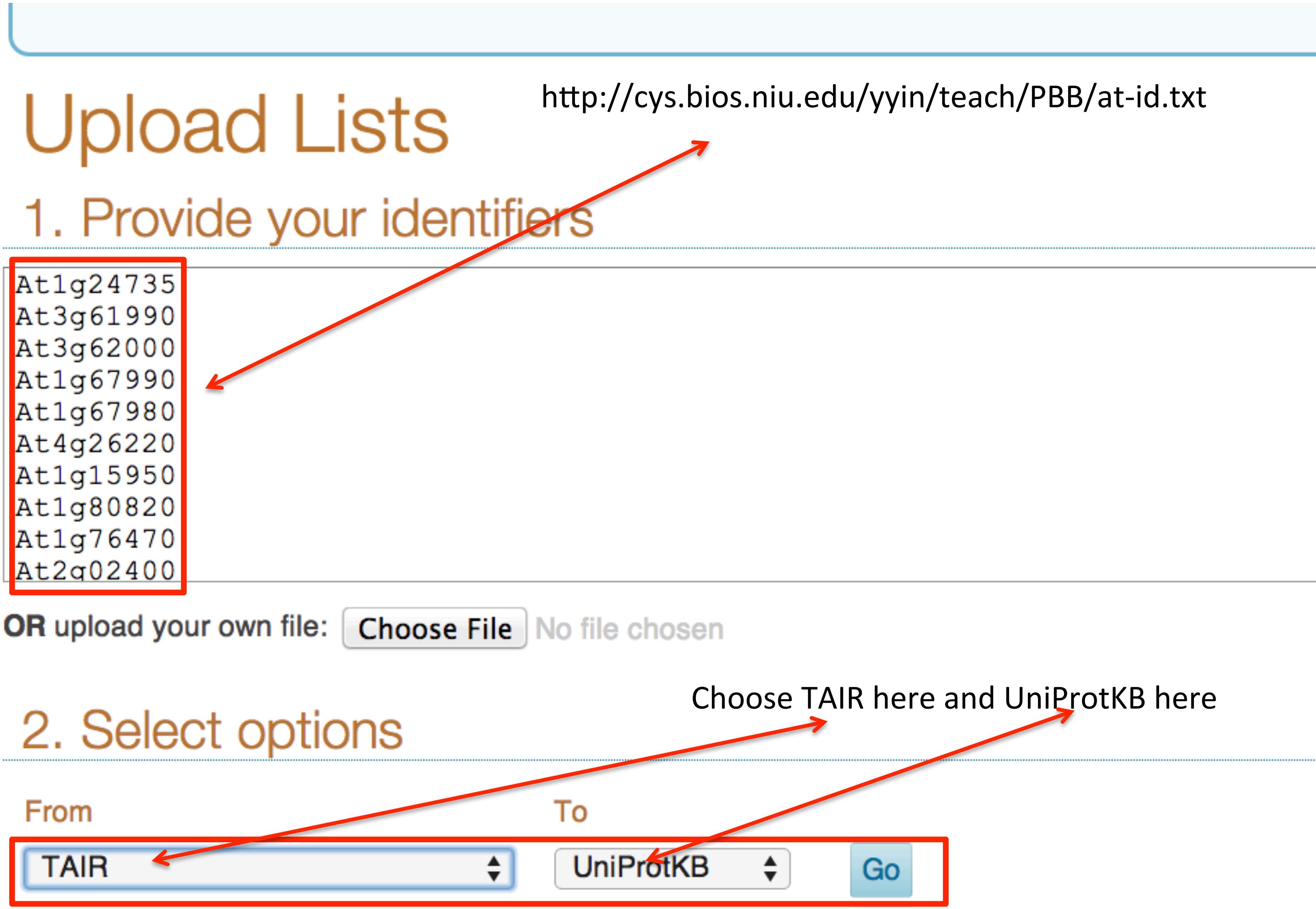
Submit your data

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

Upload Lists

1. Provide your identifiers

```
At1g24735
At3g61990
At3g62000
At1g67990
At1g67980
At4g26220
At1g15950
At1g80820
At1g76470
At2g02400
```

OR upload your own file: Choose File No file chosen

2. Select options

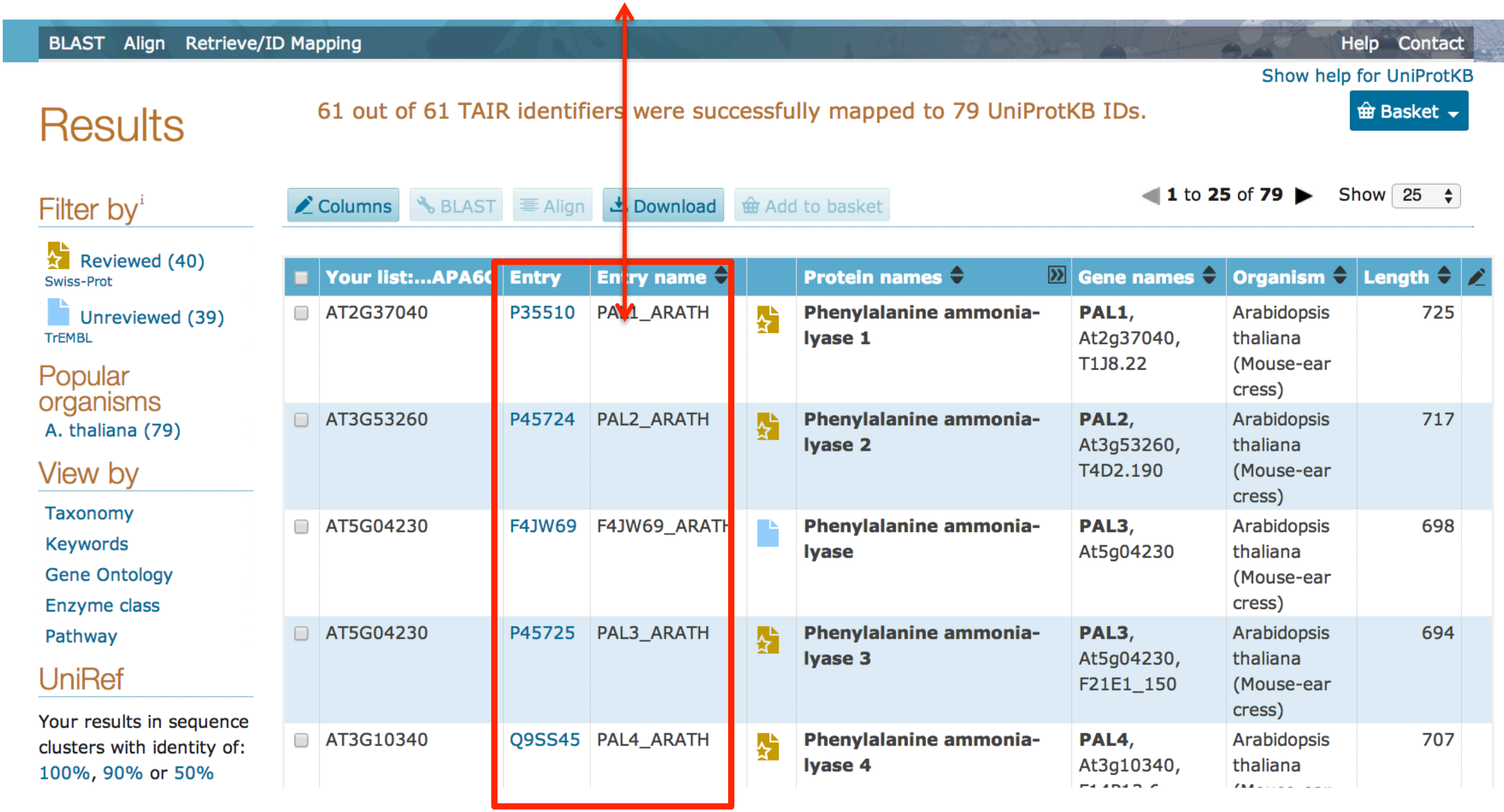
Choose TAIR here and UniProtKB here

From: TAIR

To: UniProtKB

Go

These are UniProt IDs



Results

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

◀ 1 to 25 of 79 ▶

Filter by

★ Reviewed (40)
Swiss-Prot

📄 Unreviewed (39)
TrEMBL

Popular organisms

A. thaliana (79)

View by

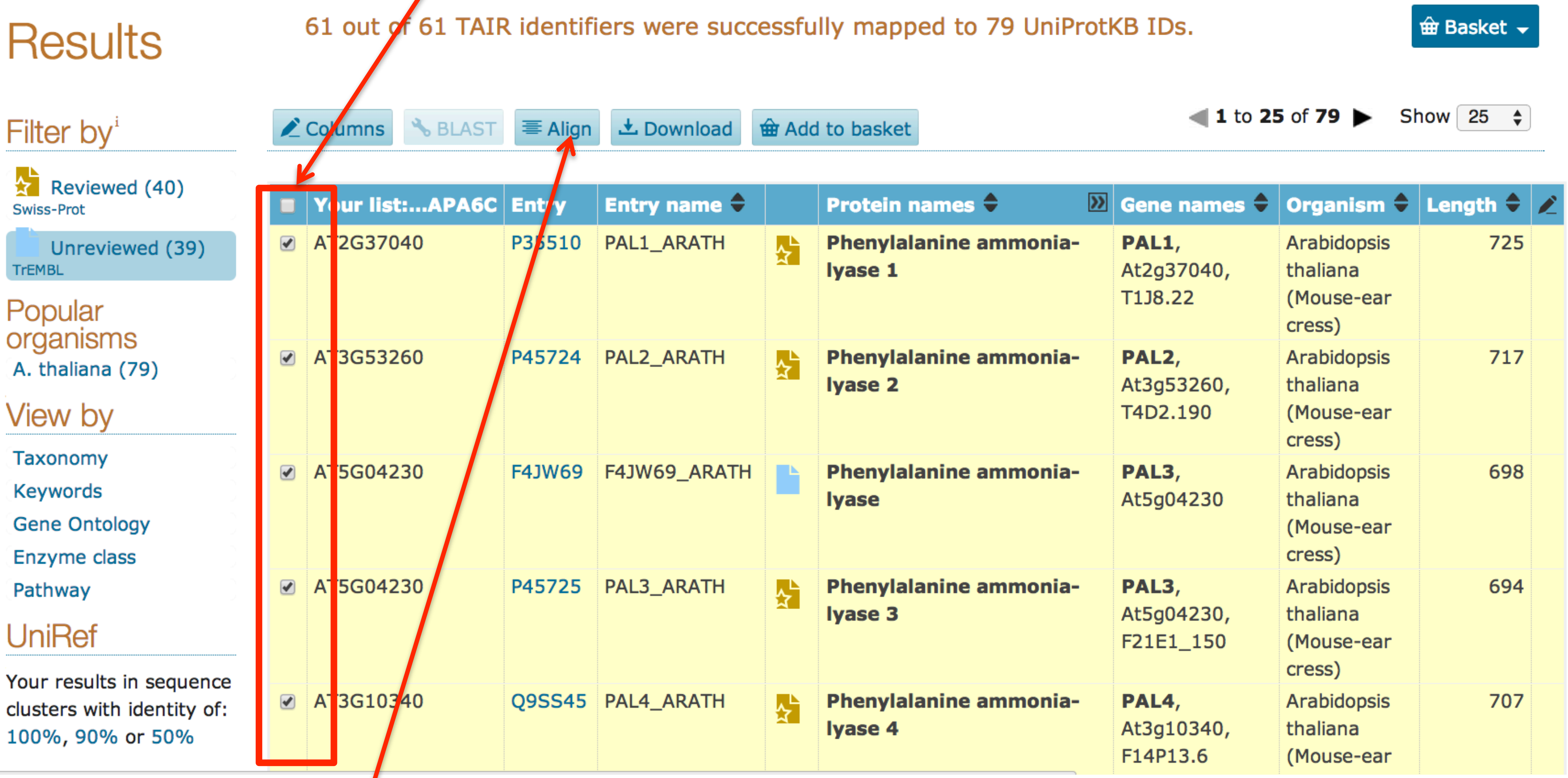
Taxonomy
Keywords
Gene Ontology
Enzyme class
Pathway

UniRef

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

□	Your list:...APA6(	Entry	Entry name		Protein names	Gene names	Organism	Length
□	AT2G37040	P35510	PAL1_ARATH	★	Phenylalanine ammonia-lyase 1	PAL1, At2g37040, T1J8.22	Arabidopsis thaliana (Mouse-ear cress)	725
□	AT3G53260	P45724	PAL2_ARATH	★	Phenylalanine ammonia-lyase 2	PAL2, At3g53260, T4D2.190	Arabidopsis thaliana (Mouse-ear cress)	717
□	AT5G04230	F4JW69	F4JW69_ARATH	📄	Phenylalanine ammonia-lyase	PAL3, At5g04230	Arabidopsis thaliana (Mouse-ear cress)	698
□	AT5G04230	P45725	PAL3_ARATH	★	Phenylalanine ammonia-lyase 3	PAL3, At5g04230, F21E1_150	Arabidopsis thaliana (Mouse-ear cress)	694
□	AT3G10340	Q9SS45	PAL4_ARATH	★	Phenylalanine ammonia-lyase 4	PAL4, At3g10340,	Arabidopsis thaliana	707

Select the PAL proteins and align them



Results

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

Filter by

- Reviewed (40) Swiss-Prot
- Unreviewed (39) TrEMBL

Popular organisms

- A. thaliana (79)

View by

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

UniRef

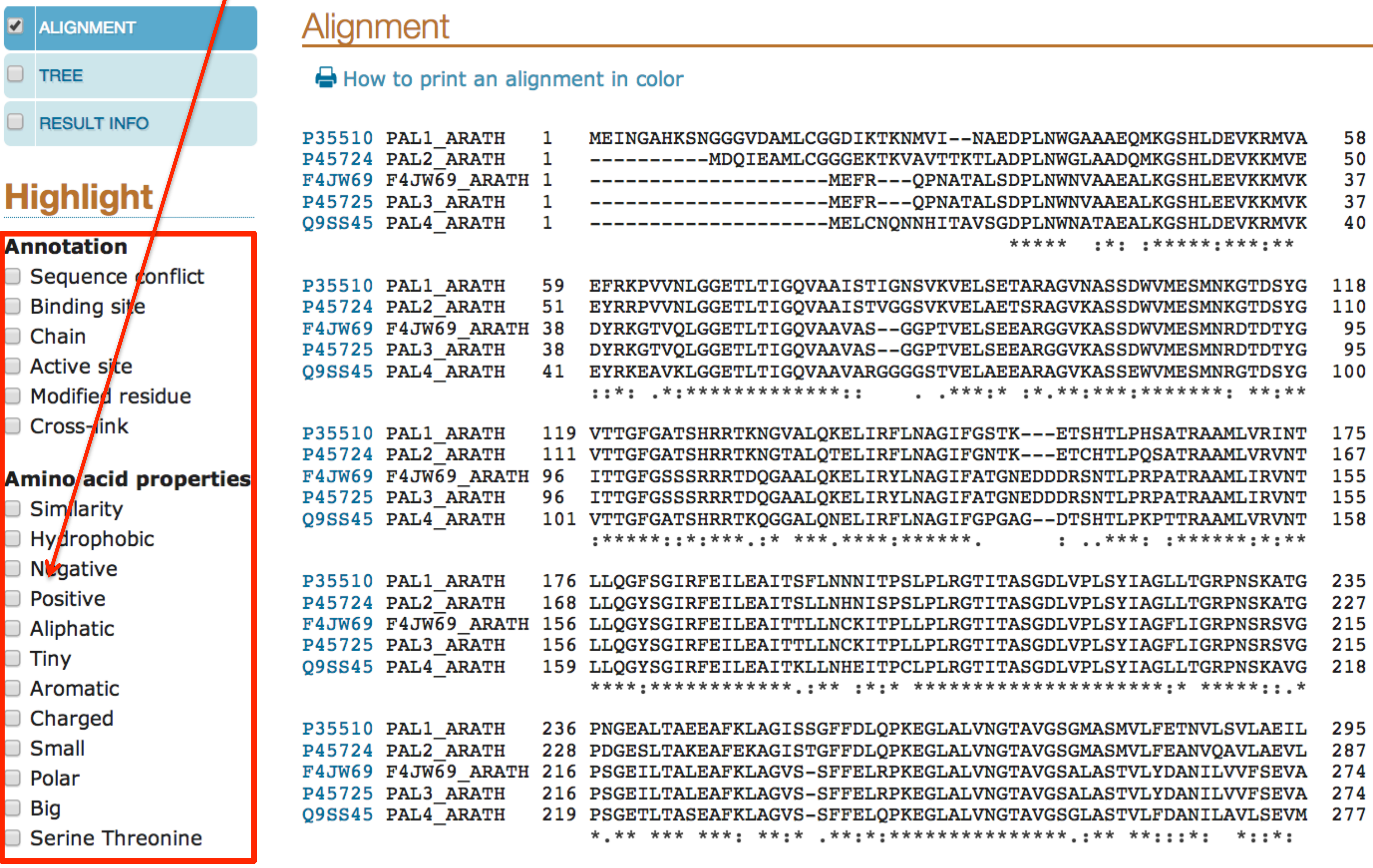
Your results in sequence clusters with identity of: 100%, 90% or 50%

□	Your list:...APA6C	Entry	Entry name		Protein names	Gene names	Organism	Length
☑	AT2G37040	P35510	PAL1_ARATH	★	**Phenylalanine ammonia-lyase 1**	**PAL1**, At2g37040, T1J8.22	Arabidopsis thaliana (Mouse-ear cress)	725
☑	AT3G53260	P45724	PAL2_ARATH	★	**Phenylalanine ammonia-lyase 2**	**PAL2**, At3g53260, T4D2.190	Arabidopsis thaliana (Mouse-ear cress)	717
☑	AT5G04230	F4JW69	F4JW69_ARATH	📄	**Phenylalanine ammonia-lyase**	**PAL3**, At5g04230	Arabidopsis thaliana (Mouse-ear cress)	698
☑	AT5G04230	P45725	PAL3_ARATH	★	**Phenylalanine ammonia-lyase 3**	**PAL3**, At5g04230, F21E1_150	Arabidopsis thaliana (Mouse-ear cress)	694
☑	AT3G10340	Q9SS45	PAL4_ARATH	★	**Phenylalanine ammonia-lyase 4**	**PAL4**, At3g10340, F14P13.6	Arabidopsis thaliana (Mouse-ear	707

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

☐ TREE

☐ RESULT INFO

Highlight

Annotation

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

Amino acid properties

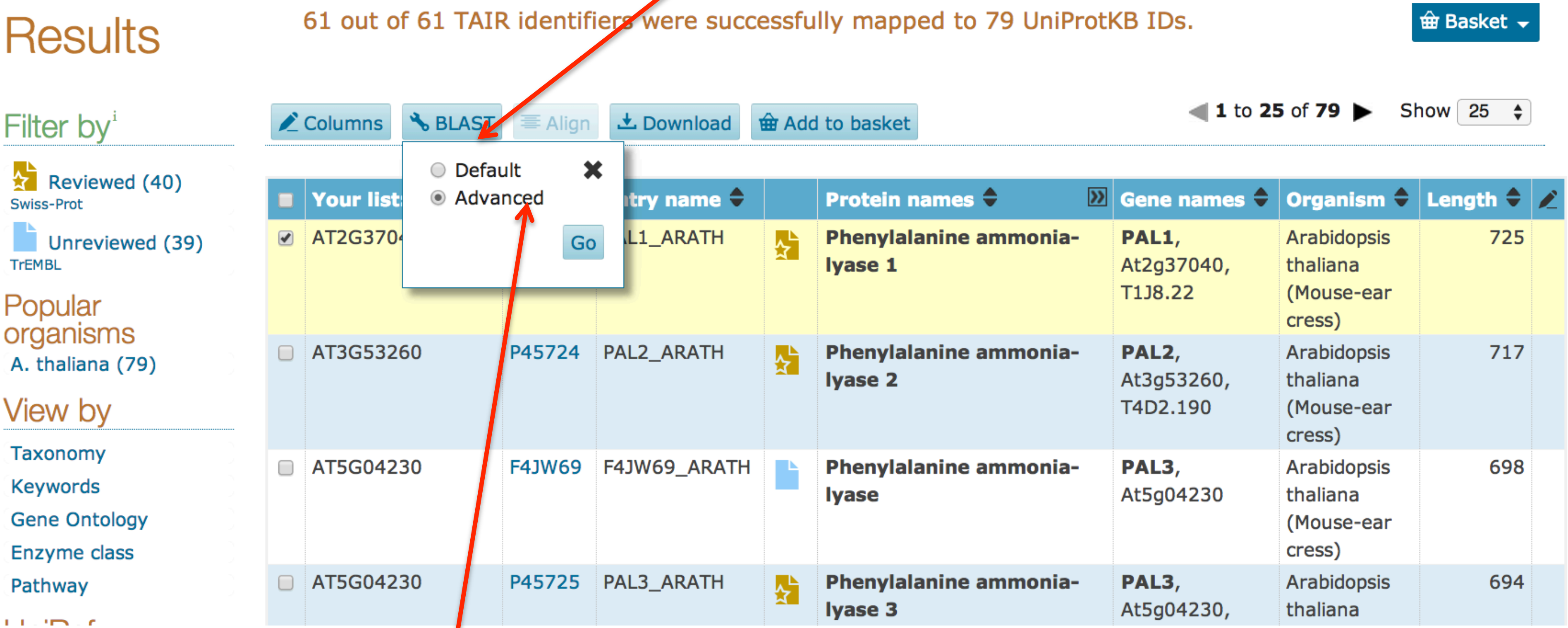
☐ Similarity
☐ Hydrophobic
☐ Negative
☐ Positive
☐ Aliphatic
☐ Tiny
☐ Aromatic
☐ Charged
☐ Small
☐ Polar
☐ Big
☐ Serine Threonine

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



Results

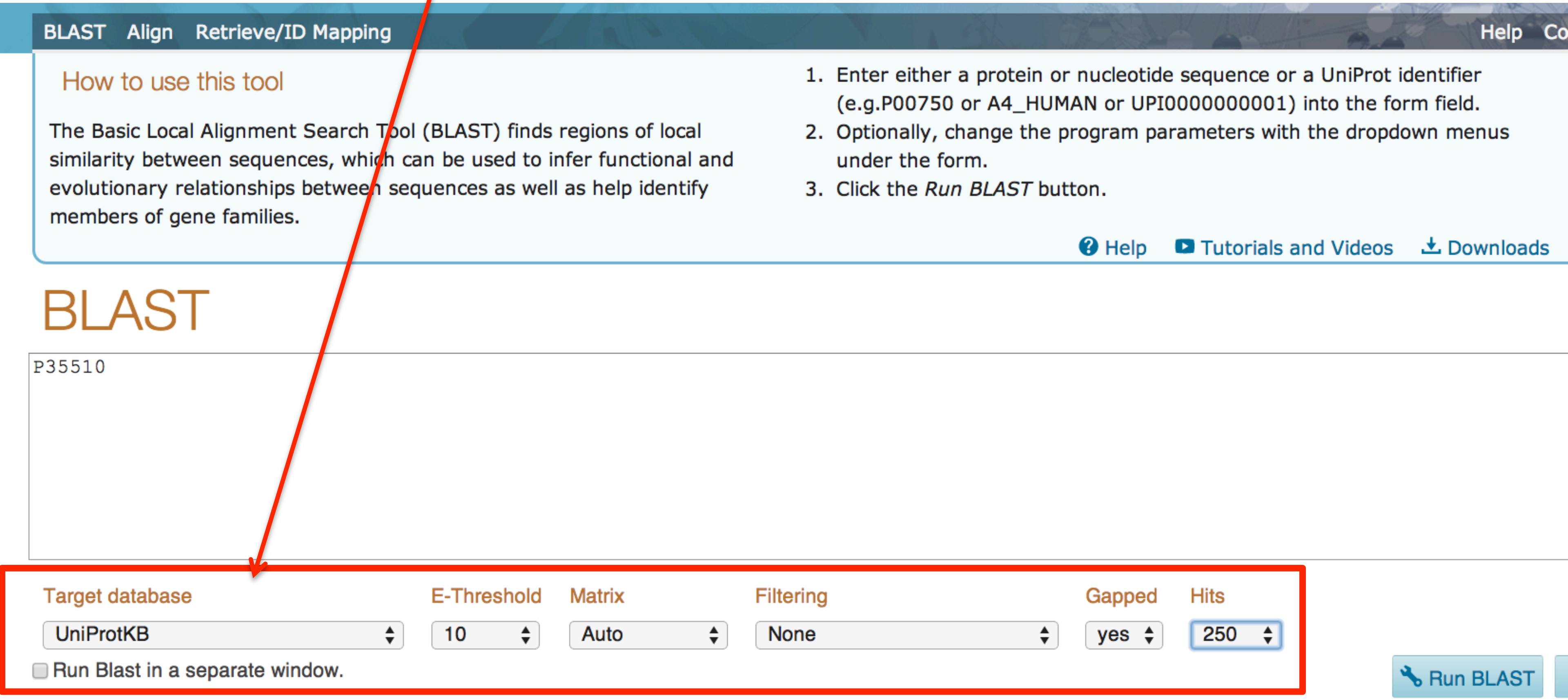
61 out of 61 TAIR identifiers were successfully mapped to 79 UniProtKB IDs.

○ Default ✖
● Advanced
Go

□	Your list:		ntry name		Protein names	Gene names	Organism	Length
☑	AT2G3704		L1_ARATH		**Phenylalanine ammonia-lyase 1**	**PAL1**, At2g37040, T1J8.22	Arabidopsis thaliana (Mouse-ear cress)	725
□	AT3G53260	P45724	PAL2_ARATH		**Phenylalanine ammonia-lyase 2**	**PAL2**, At3g53260, T4D2.190	Arabidopsis thaliana (Mouse-ear cress)	717
□	AT5G04230	F4JW69	F4JW69_ARATH		**Phenylalanine ammonia-lyase**	**PAL3**, At5g04230	Arabidopsis thaliana (Mouse-ear cress)	698
□	AT5G04230	P45725	PAL3_ARATH		**Phenylalanine ammonia-lyase 3**	**PAL3**, 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

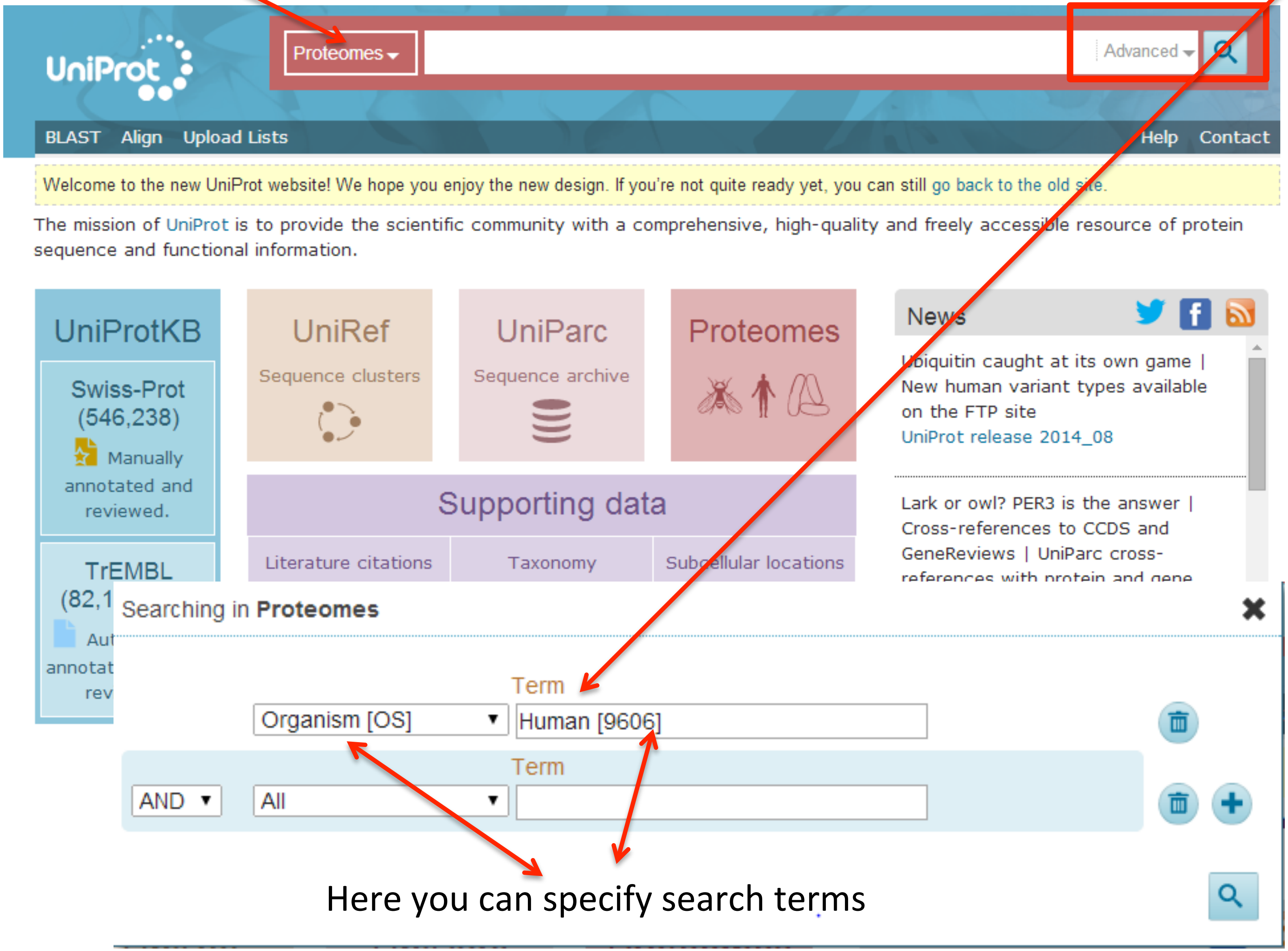
```
P35510
```

Target database	E-Threshold	Matrix	Filtering	Gapped	Hits
UniProtKB	10	Auto	None	yes	250

□ Run Blast in a separate window.

We are going to search UniProt proteomes for human protein set

Click on Advanced you will see a pop-out window



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.

Ubiquitin caught at its own game | New human variant types available on the FTP site

UniProt release 2014_08

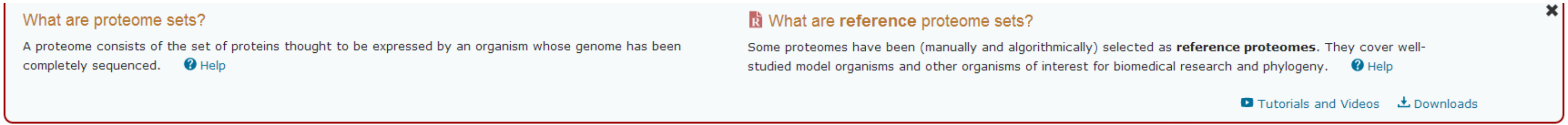
Lark or owl? PER3 is the answer | Cross-references to CCDS and GeneReviews | UniParc cross-references with protein and gene

Searching in **Proteomes**

Term: Organism [OS] ▾ Human [9606]

AND ▾ All ▾ Term:

Here you can specify search terms



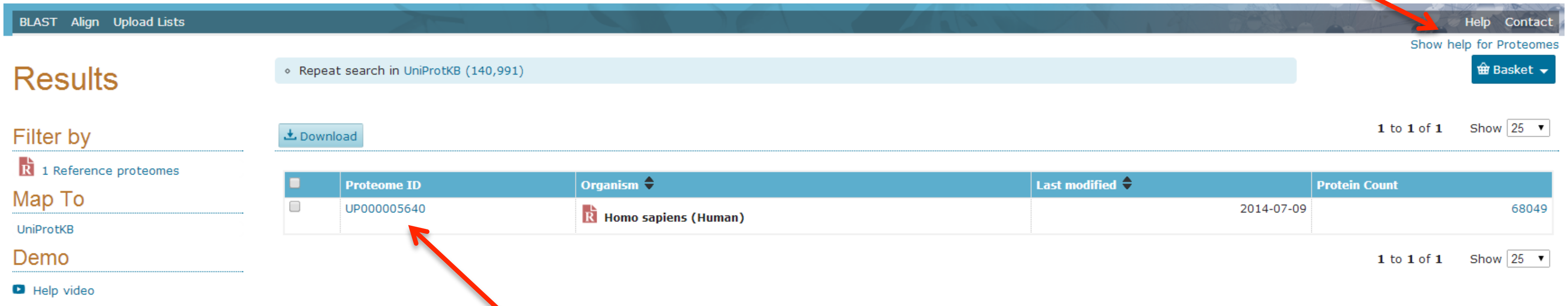
What are proteome sets?

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

Click here to get help



Results

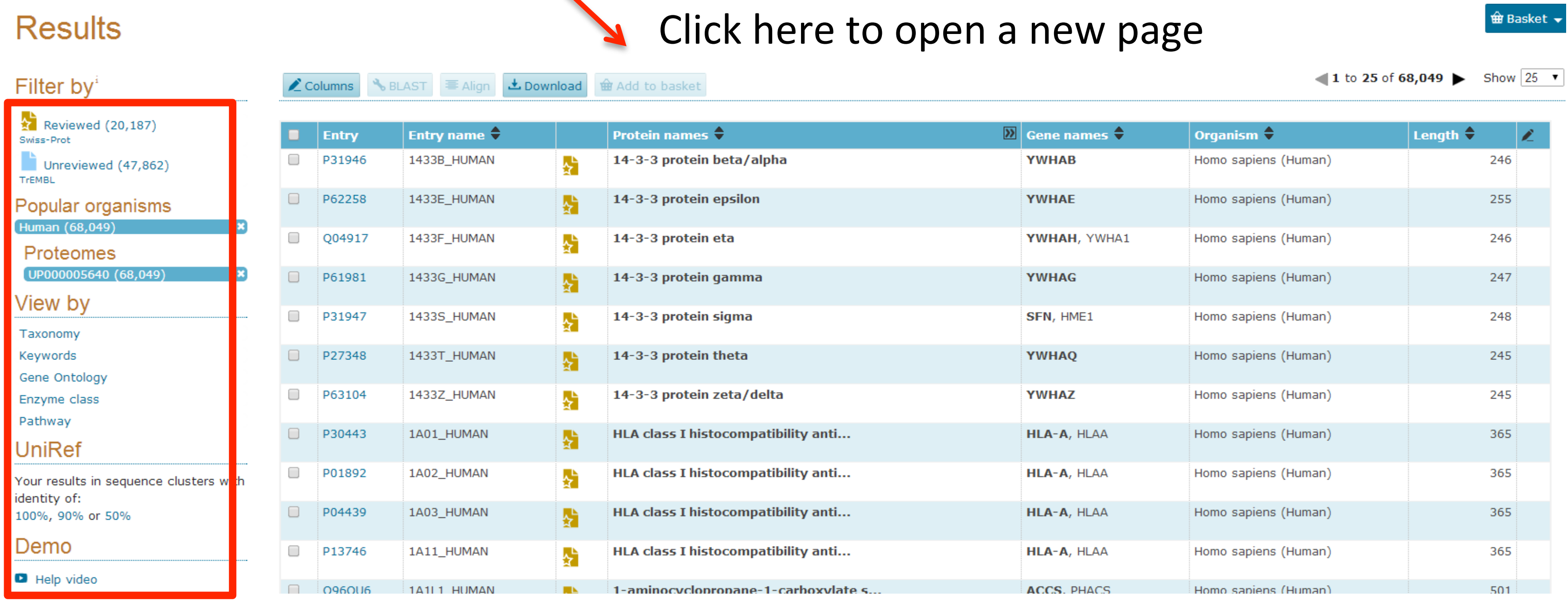
Repeat search in UniProtKB (140,991)

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

Click here to open a new page



Results

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

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
☐	Q96QU6	1A1L1_HUMAN		1-aminocyclopropane-1-carboxylate s...	ACCS, 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;

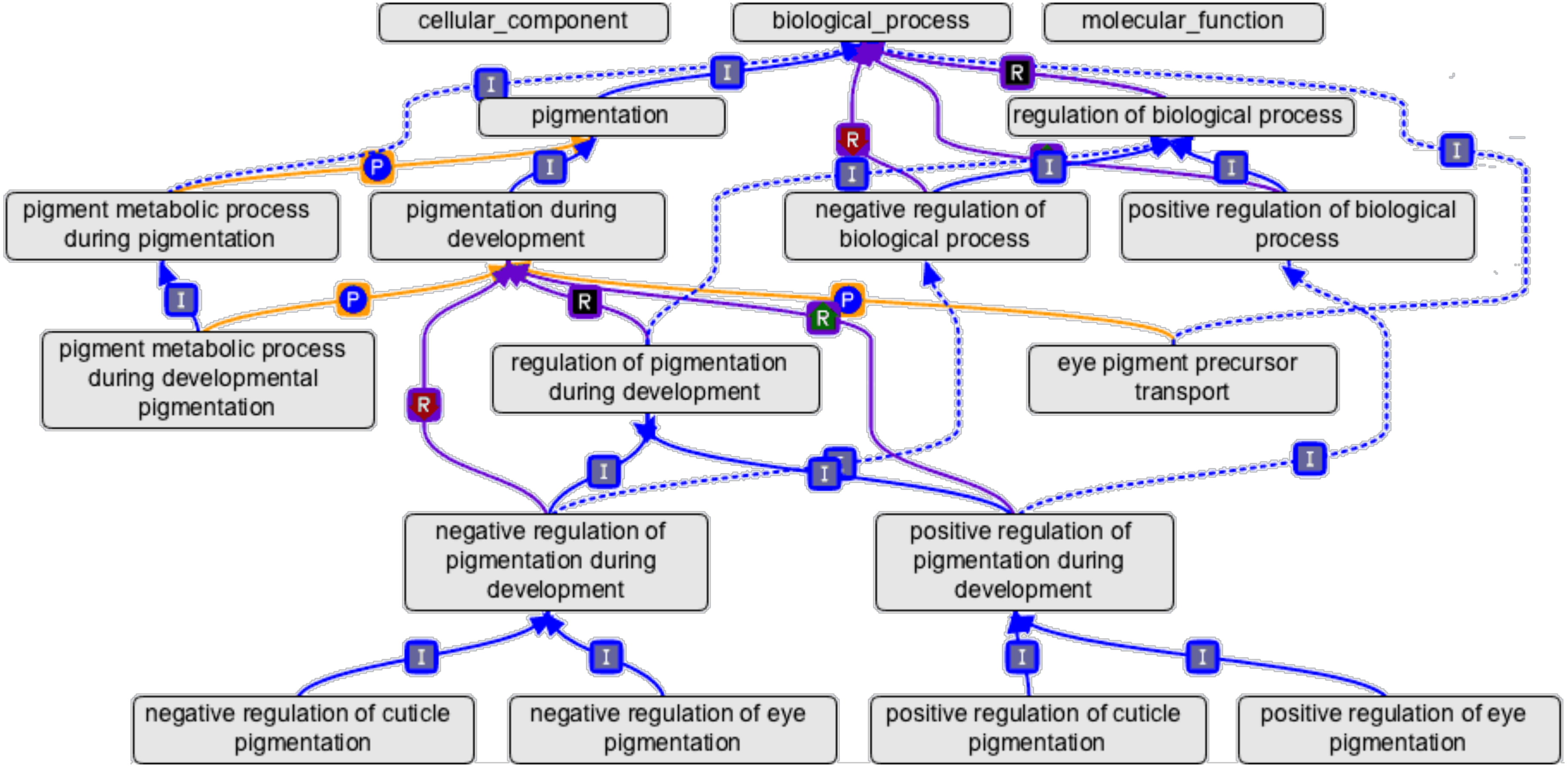
biological process, operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms

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.

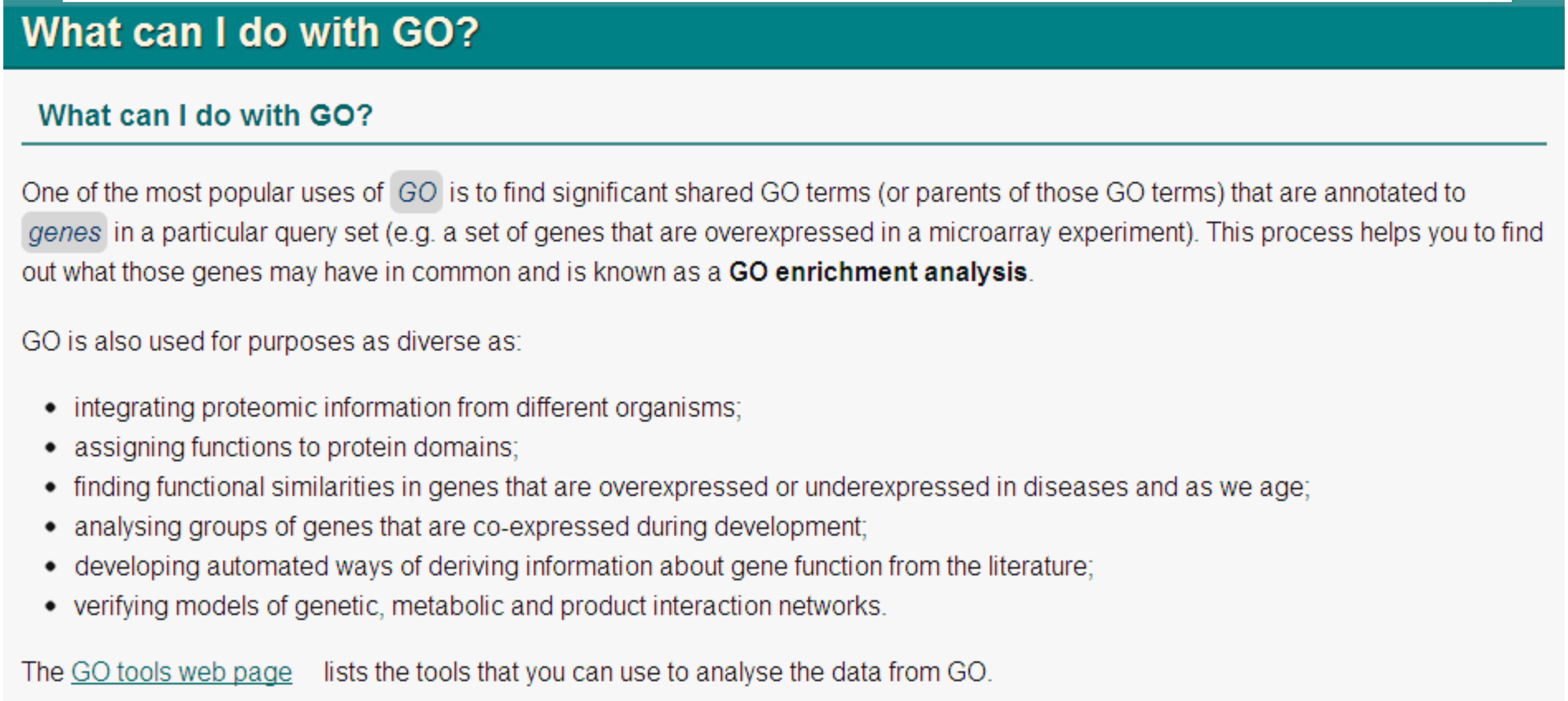
GO is not a way to unify biological databases (i.e. GO is not a 'federated solution'). Sharing vocabulary is a step towards unification, but is not, in itself, sufficient.

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


cellular_component
biological_process
molecular_function
pigmentation
regulation of biological process
pigment metabolic process during pigmentation
pigmentation during development
negative regulation of biological process
positive regulation of biological process
pigment metabolic process during developmental pigmentation
regulation of pigmentation during development
eye pigment precursor transport
negative regulation of pigmentation during development
positive regulation of pigmentation during development
negative regulation of cuticle pigmentation
negative regulation of eye pigmentation
positive regulation of cuticle pigmentation
positive regulation of eye pigmentation
I
P
R

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:

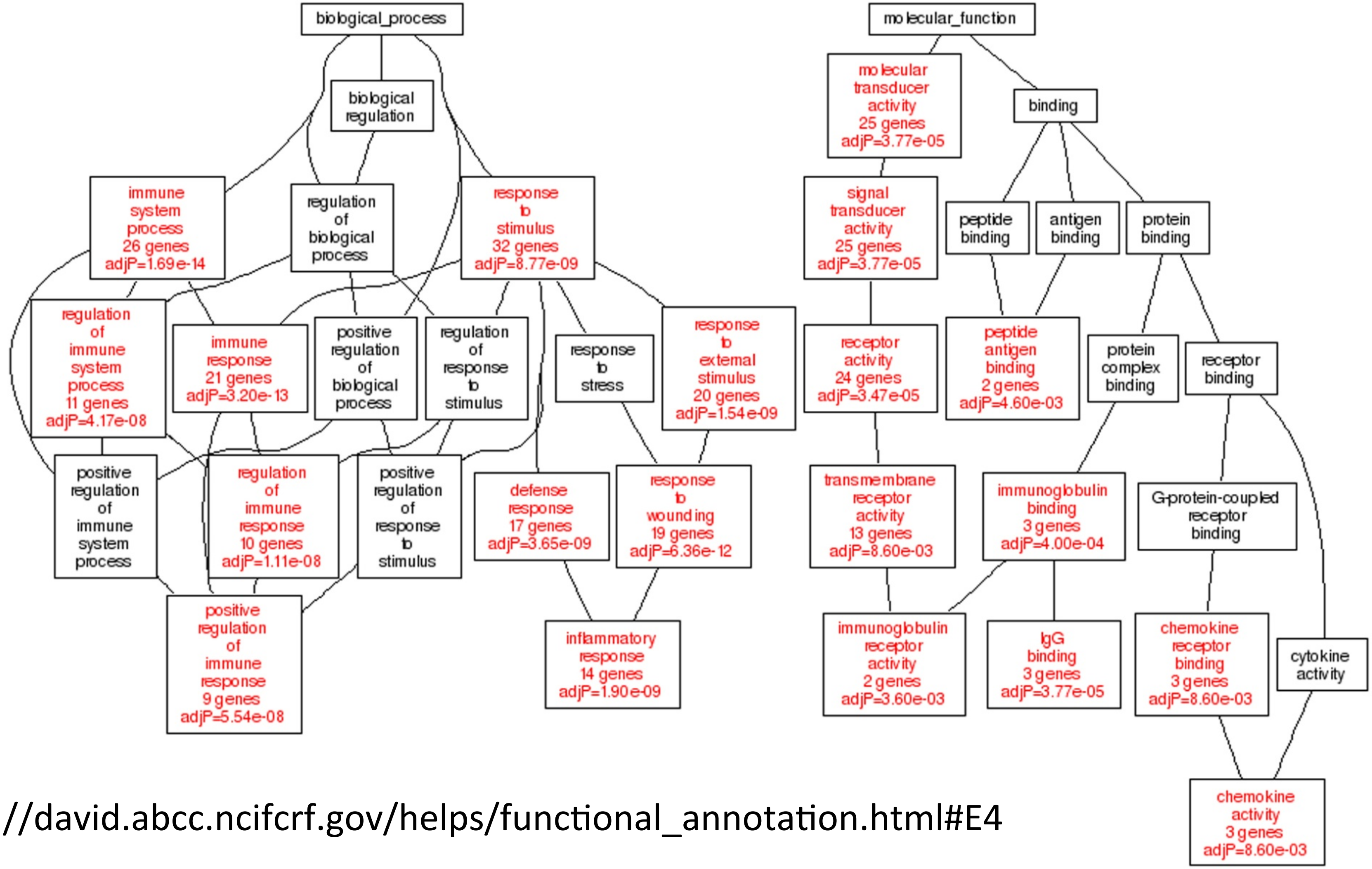
- integrating proteomic information from different organisms;
- assigning functions to protein domains;
- finding functional similarities in genes that are overexpressed or underexpressed in diseases and as we age;
- analysing groups of genes that are co-expressed during development;
- developing automated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



biological_process

biological regulation

immune system process 26 genes adjP=1.69e-14

regulation of biological process

response to stimulus 32 genes adjP=8.77e-09

regulation of immune system process 11 genes adjP=4.17e-08

immune response 21 genes adjP=3.20e-13

positive regulation of biological process

regulation of response to stimulus

response to stress

response to external stimulus 20 genes adjP=1.54e-09

positive regulation of immune system process

regulation of immune response 10 genes adjP=1.11e-08

positive regulation of response to stimulus

defense response 17 genes adjP=3.65e-09

response to wounding 19 genes adjP=6.36e-12

positive regulation of immune response 9 genes adjP=5.54e-08

inflammatory response 14 genes adjP=1.90e-09

molecular_function

molecular transducer activity 25 genes adjP=3.77e-05

binding

signal transducer activity 25 genes adjP=3.77e-05

peptide binding

antigen binding

protein binding

receptor activity 24 genes adjP=3.47e-05

peptide antigen binding 2 genes adjP=4.60e-03

protein complex binding

receptor binding

transmembrane receptor activity 13 genes adjP=8.60e-03

immunoglobulin binding 3 genes adjP=4.00e-04

G-protein-coupled receptor binding

immunoglobulin receptor activity 2 genes adjP=3.60e-03

IgG binding 3 genes adjP=3.77e-05

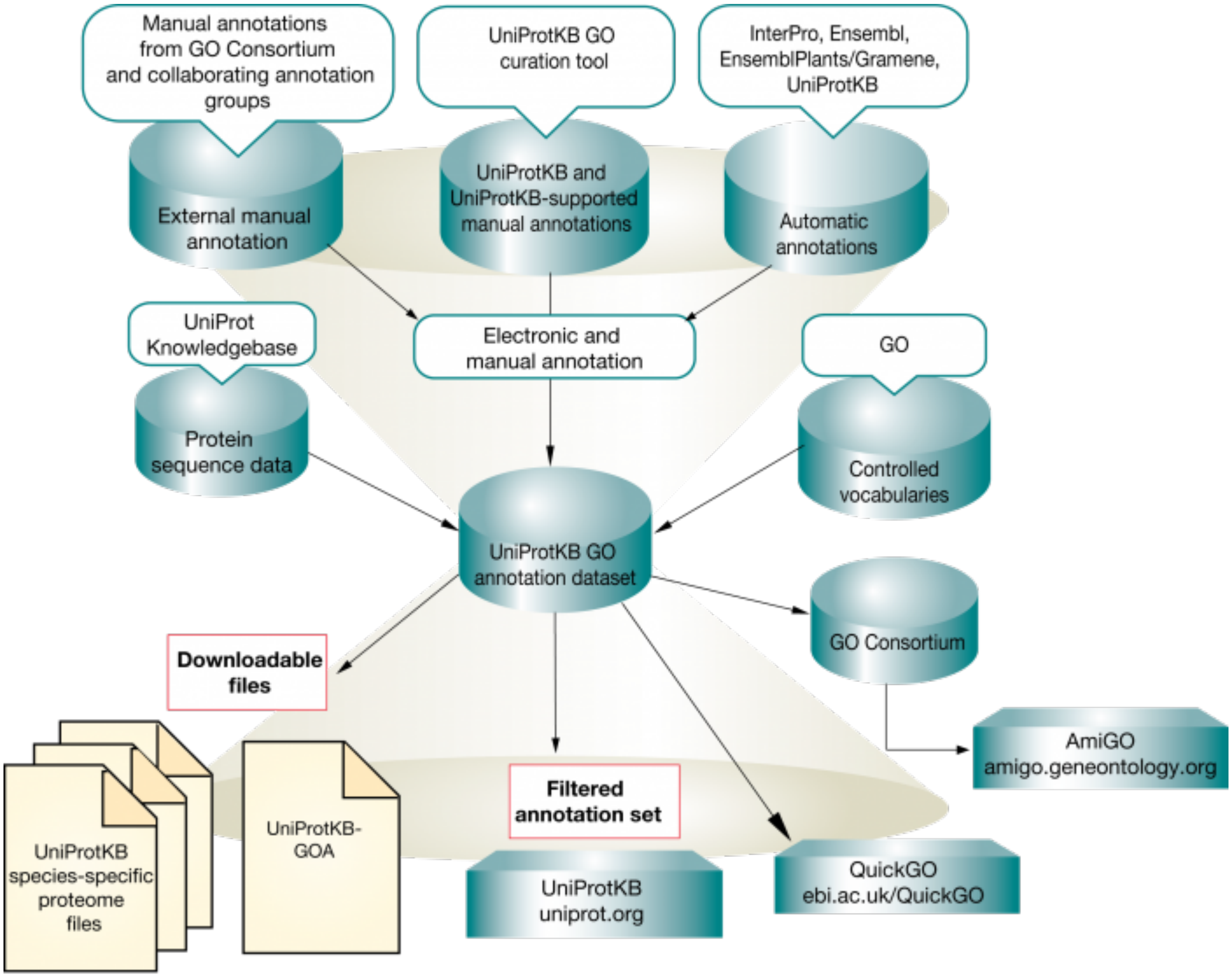
chemokine receptor binding 3 genes adjP=8.60e-03

cytokine activity

chemokine activity 3 genes adjP=8.60e-03

http://david.abcc.ncifcrf.gov/helps/functional_annotation.html#E4

UniProt-GO annotation (GOA)


Manual annotations from GO Consortium and collaborating annotation groups
UniProtKB GO curation tool
InterPro, Ensembl, EnsemblPlants/Gramene, UniProtKB
External manual annotation
UniProtKB and UniProtKB-supported manual annotations
Automatic annotations
UniProt Knowledgebase
Electronic and manual annotation
GO
Protein sequence data
Controlled vocabularies
UniProtKB GO annotation dataset
GO Consortium
Downloadable files
AmiGO amigo.geneontology.org
UniProtKB- GOA
Filtered annotation set
UniProtKB species-specific proteome files
UniProtKB uniprot.org
QuickGO ebi.ac.uk/QuickGO

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?

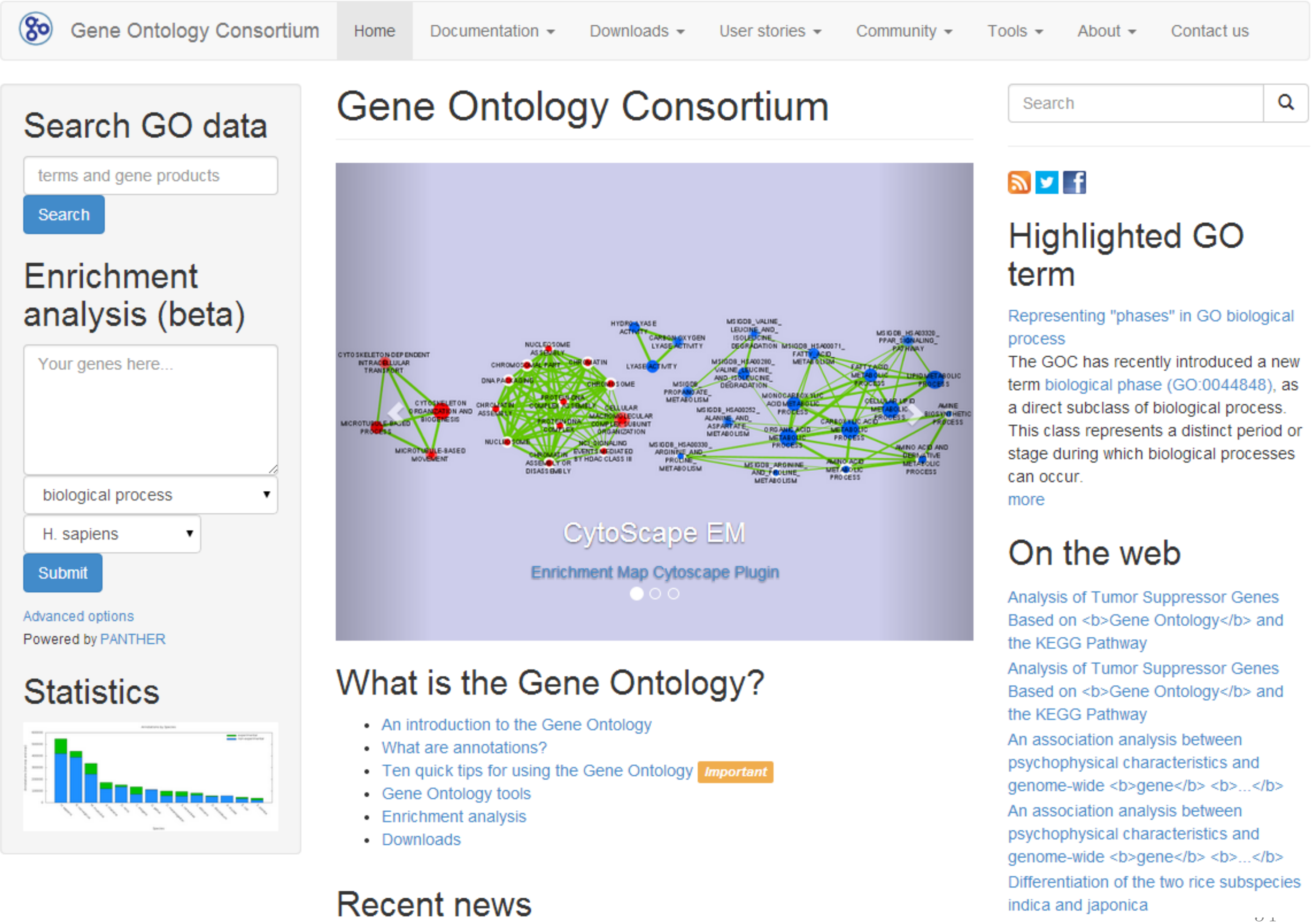
1. Find a closet model organism which has been annotated by GO
2. BLAST your data against this closest organism
3. Transfer the GO annotation of the best match to your query sequences

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

CytoScape EM

Enrichment Map Cytoscape Plugin

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 <b>Gene Ontology</b> and the KEGG Pathway

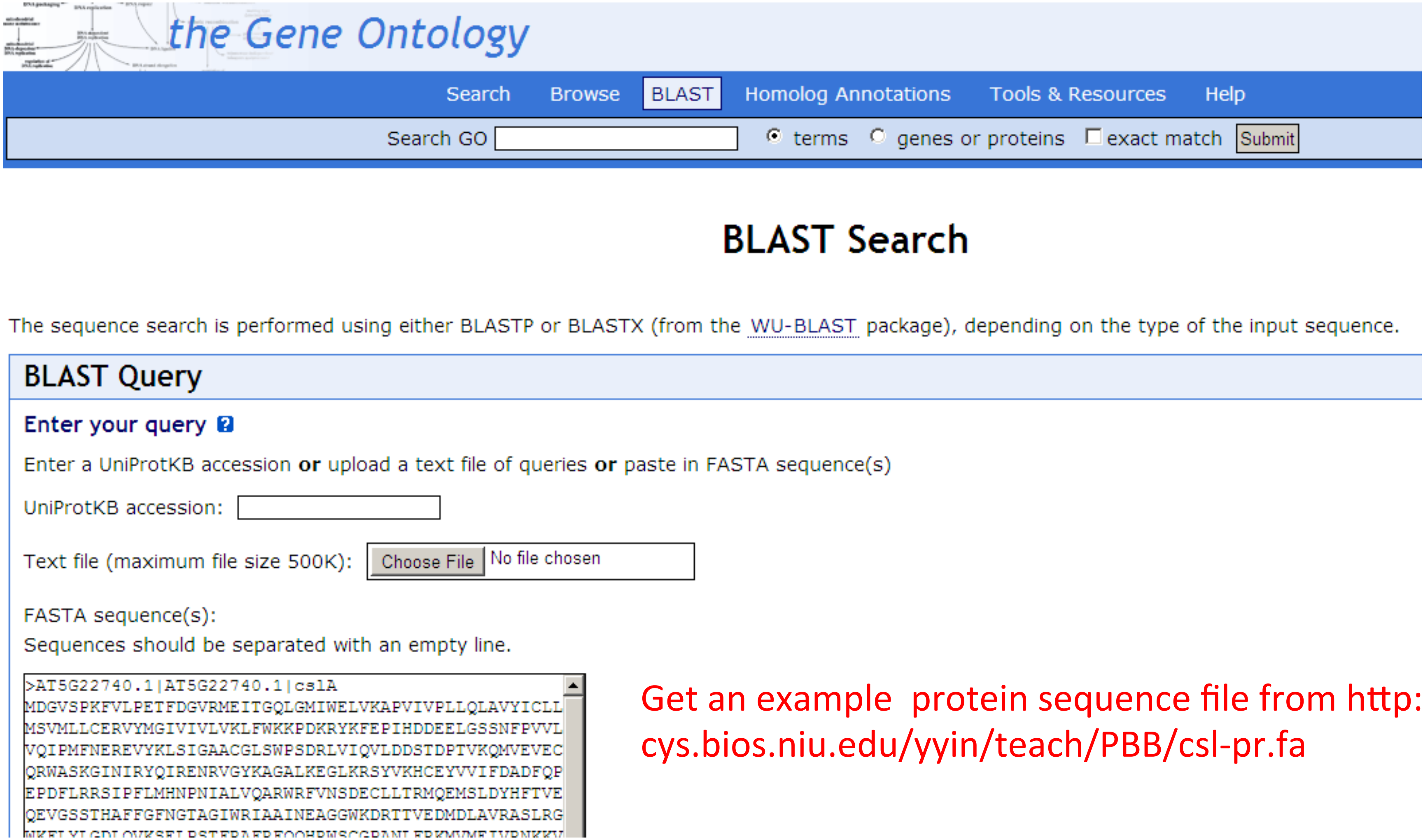
Analysis of Tumor Suppressor Genes Based on <b>Gene Ontology</b> and the KEGG Pathway

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

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:

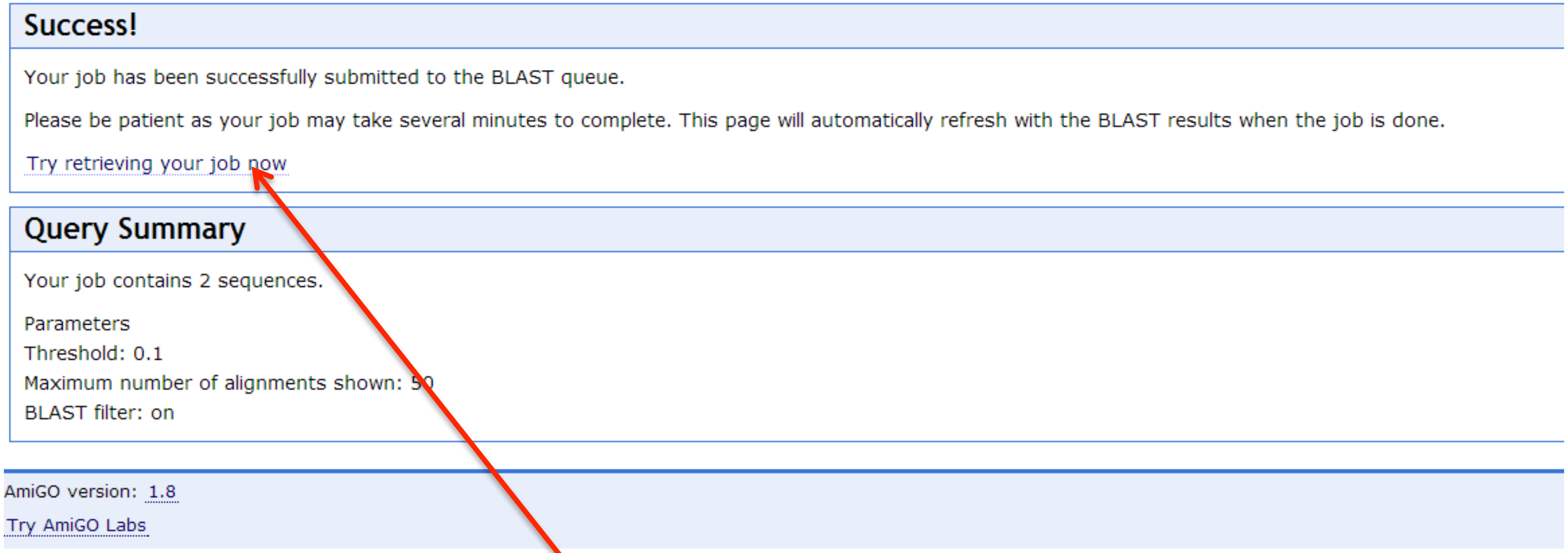
Text file (maximum file size 500K): No file chosen

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

```
>AT5G22740.1|AT5G22740.1|cslA
MDGVSPKFVLPETFDGVRMEITGQLGMIWELVKAPVIVPLLQLAVYICLL
MSVMLLCERVYMGIVIVLVKLFWKKPDKRYKFEPIHDDEELGSSNFPVVL
VQIPMFNEREVYKLSIGAACGLSWPSDRLVIQVLDDSTDPTVKQMVEVEC
QRWASKGINIRYQIRENRVGYKAGALKEGLKRSYVKHCEYVVIFDADFQP
EPDFLRRSIPFLMHNPNIALVQARWRFVNSDECLLTRMQEMSLDYHFTVE
QEVGSSTHAFFGFNGTAGIWRIAAINEAGGWKDRTTVEDMDLAVRASLRG
WKFLYLGDLQVKSELRSTERAERFQQHRWSCGPANLFRKMVMEIVRNKKV
```

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.

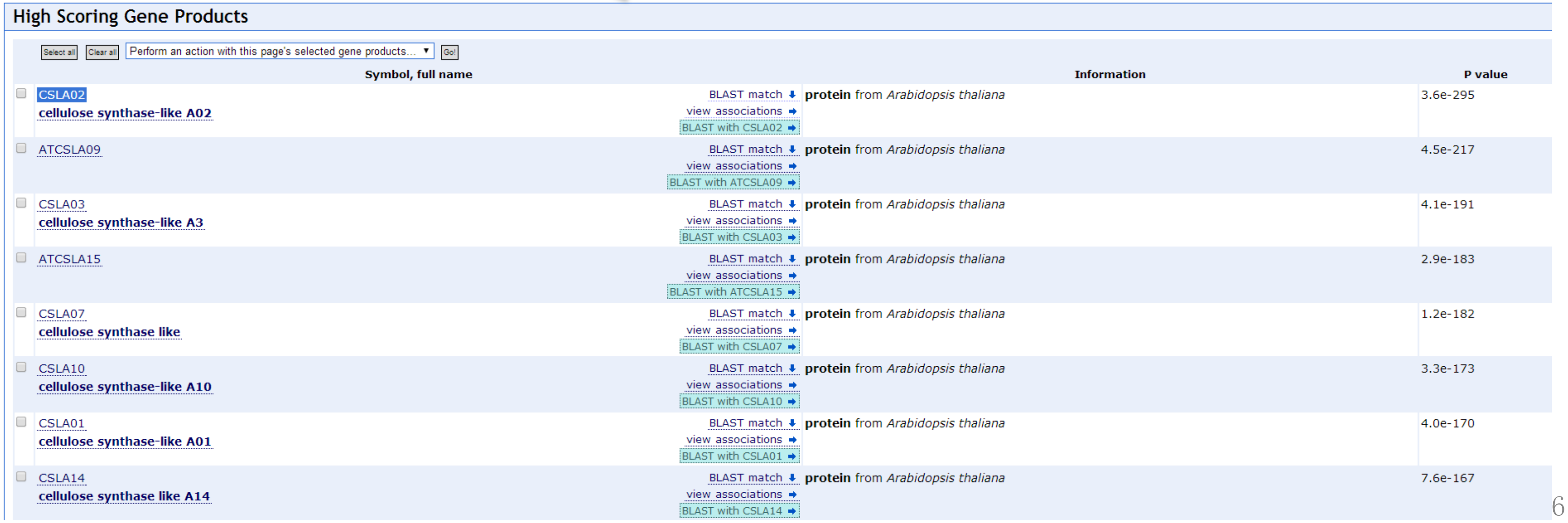
Try retrieving your job now

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

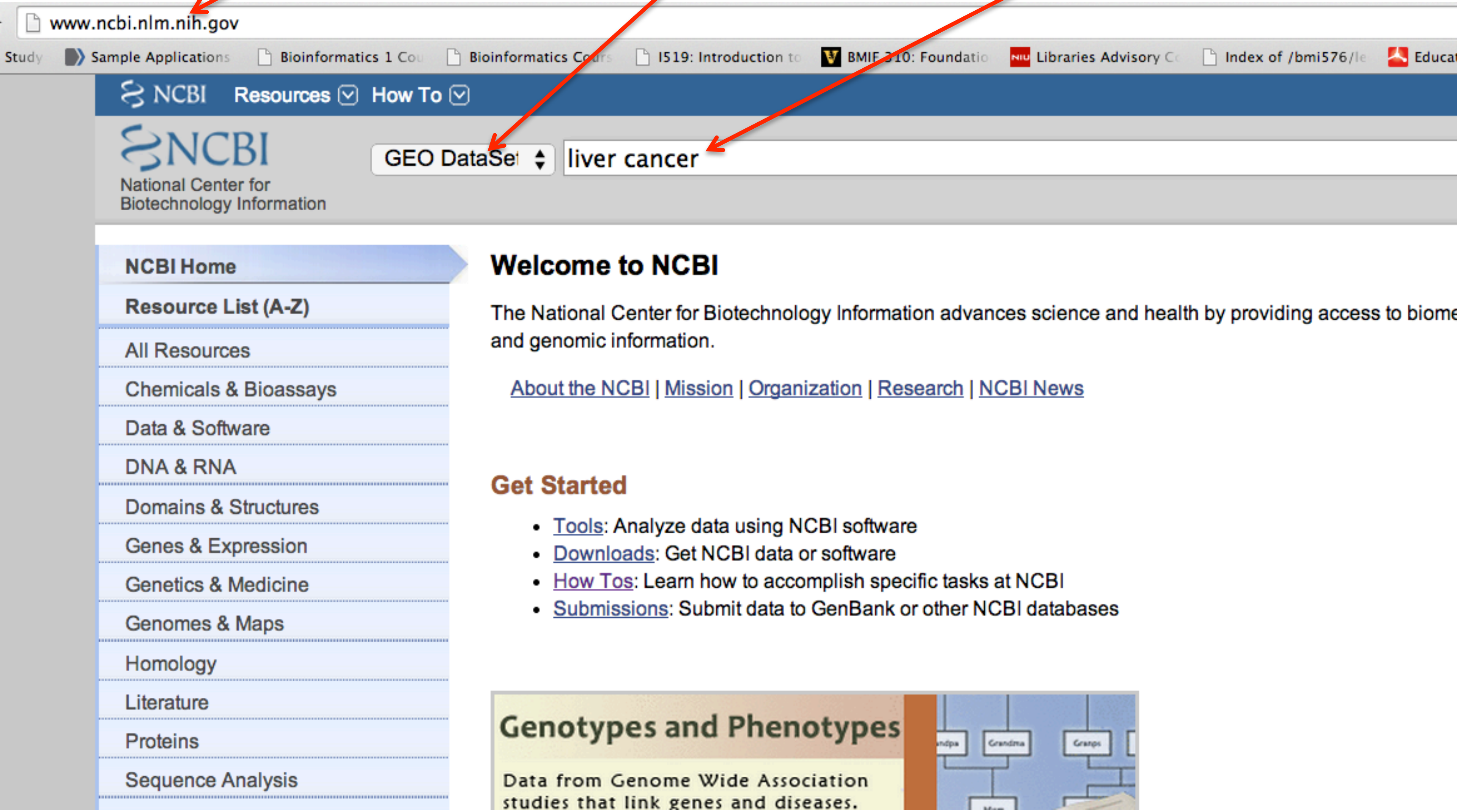
Select all
Clear all
Perform an action with this page's selected gene products...
Go!

	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 “liver cancer”, and hit search



www.ncbi.nlm.nih.gov

GEO DataSet

liver cancer

Welcome to NCBI

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

About the NCBI | Mission | Organization | Research | NCBI News

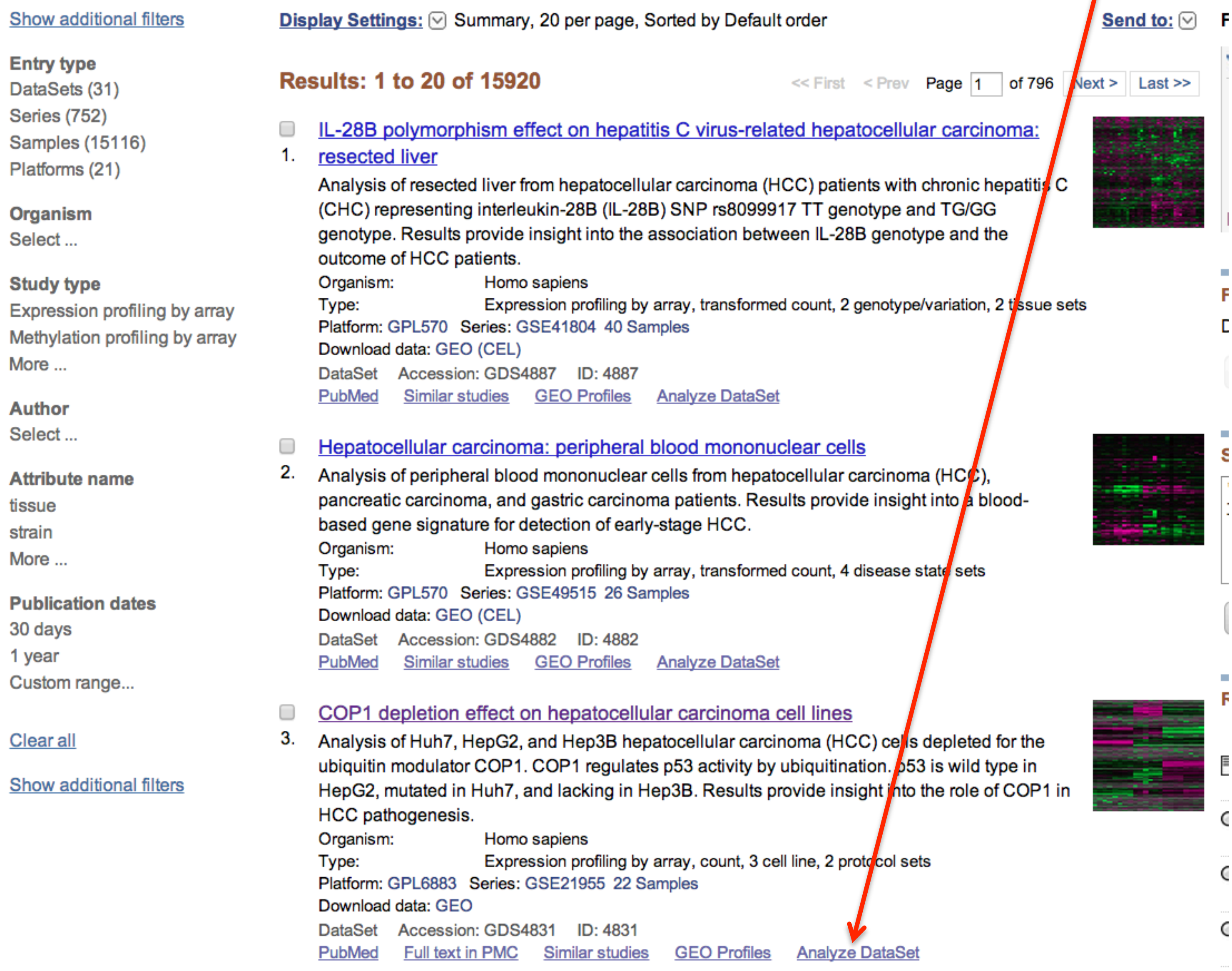
Get Started

- Tools: Analyze data using NCBI software
- Downloads: Get NCBI data or software
- How Tos: Learn how to accomplish specific tasks at NCBI
- Submissions: Submit data to GenBank or other NCBI databases

Genotypes and Phenotypes

Data from Genome Wide Association studies that link genes and diseases.

Top hits are always GEO DataSets, let’s choose the 3rd one, hit Analyze DataSet



Show additional filters

Entry type
DataSets (31)
Series (752)
Samples (15116)
Platforms (21)

Organism
Select ...

Study type
Expression profiling by array
Methylation profiling by array
More ...

Author
Select ...

Attribute name
tissue
strain
More ...

Publication dates
30 days
1 year
Custom range...

Clear all

Show additional filters

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

Send to:

Results: 1 to 20 of 15920

<< First < Prev Page 1 of 796 Next > Last >>

□ 1. IL-28B polymorphism effect on hepatitis C virus-related hepatocellular carcinoma: resected liver

Analysis of resected liver from hepatocellular carcinoma (HCC) patients with chronic hepatitis C (CHC) representing interleukin-28B (IL-28B) SNP rs8099917 TT genotype and TG/GG genotype. Results provide insight into the association between IL-28B genotype and the outcome of HCC patients.
Organism: Homo sapiens
Type: Expression profiling by array, transformed count, 2 genotype/variation, 2 tissue sets
Platform: GPL570 Series: GSE41804 40 Samples
Download data: GEO (CEL)
DataSet Accession: GDS4887 ID: 4887
PubMed Similar studies GEO Profiles Analyze DataSet

□ 2. Hepatocellular carcinoma: peripheral blood mononuclear cells

Analysis of peripheral blood mononuclear cells from hepatocellular carcinoma (HCC), pancreatic carcinoma, and gastric carcinoma patients. Results provide insight into a blood-based gene signature for detection of early-stage HCC.
Organism: Homo sapiens
Type: Expression profiling by array, transformed count, 4 disease state sets
Platform: GPL570 Series: GSE49515 26 Samples
Download data: GEO (CEL)
DataSet Accession: GDS4882 ID: 4882
PubMed Similar studies GEO Profiles Analyze DataSet

□ 3. COP1 depletion effect on hepatocellular carcinoma cell lines

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
Type: Expression profiling by array, count, 3 cell line, 2 protocol sets
Platform: GPL6883 Series: GSE21955 22 Samples
Download data: GEO
DataSet Accession: GDS4831 ID: 4831
PubMed Full text in PMC Similar studies GEO Profiles Analyze DataSet

Choose "Compare 2 sets of samples"

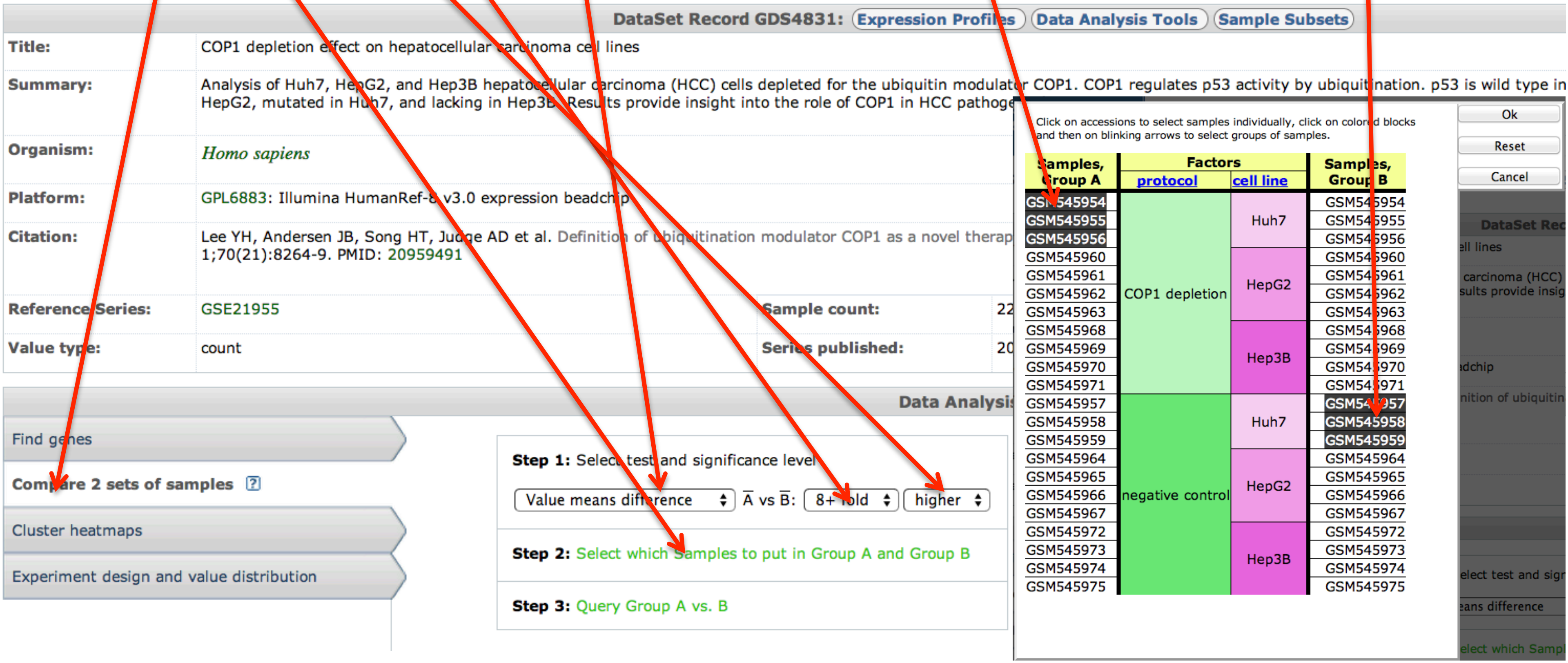
Choose "Value means difference"
Choose "8+ fold"
Choose "higher"

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 pathoge
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 therap 1;70(21):8264-9. PMID: 20959491
Reference Series:	GSE21955 — Sample count: 22
Value type:	count — Series published: 20

Data Analysis

- 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.

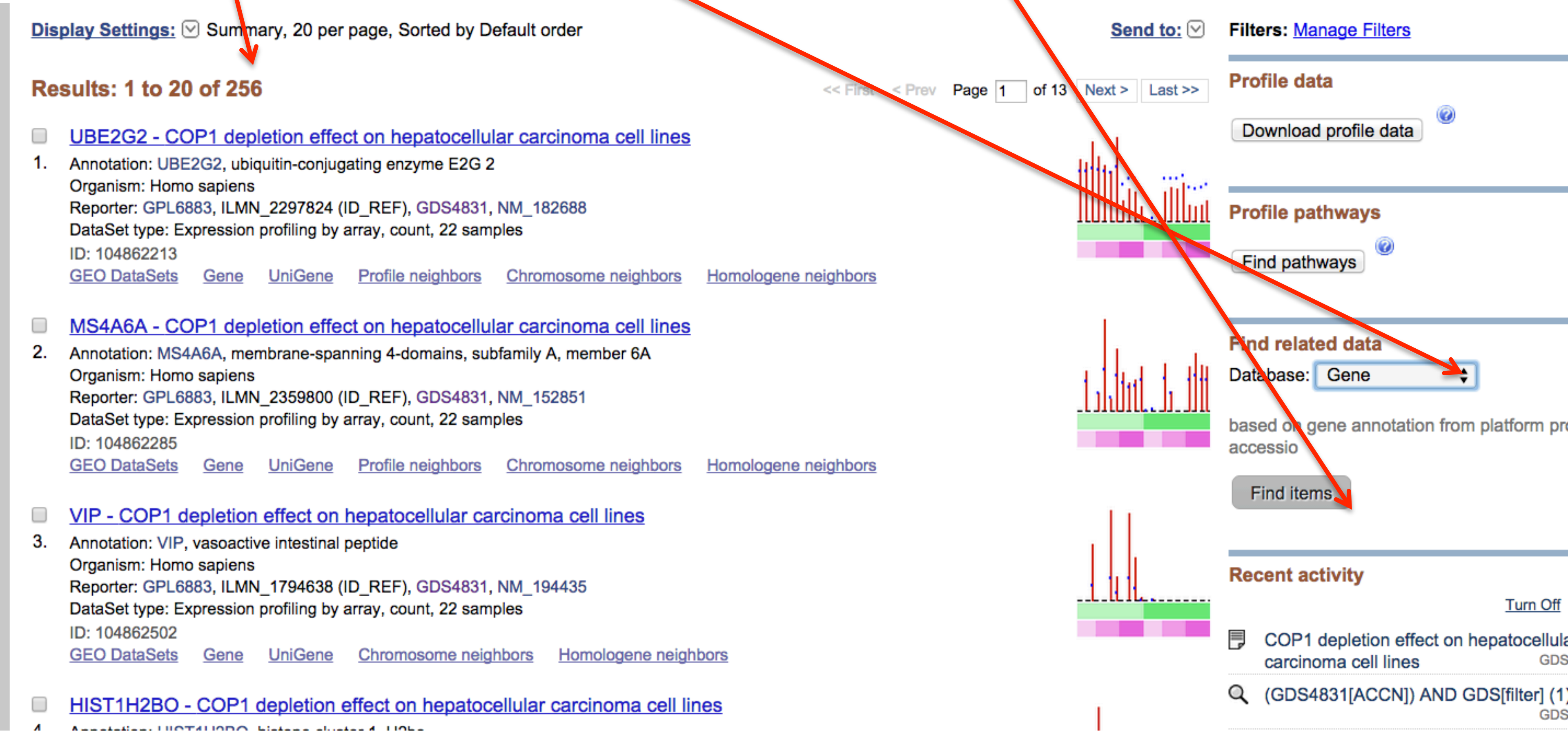
Ok
Reset
Cancel

Samples, Group A	Factors protocol	 cell line	Samples, Group B
GSM545954	COP1 depletion	Huh7	GSM545954
GSM545955			GSM545955
GSM545956			GSM545956
GSM545960		HepG2	GSM545960
GSM545961			GSM545961
GSM545962			GSM545962
GSM545963			GSM545963
GSM545968		Hep3B	GSM545968
GSM545969			GSM545969
GSM545970			GSM545970
GSM545971			GSM545971
GSM545957	negative control	Huh7	GSM545957
GSM545958			GSM545958
GSM545959			GSM545959
GSM545964		HepG2	GSM545964
GSM545965			GSM545965
GSM545966			GSM545966
GSM545967			GSM545967
GSM545972		Hep3B	GSM545972
GSM545973			GSM545973
GSM545974			GSM545974
GSM545975			GSM545975

DataSet Rec
ell lines
carcinoma (HCC)
sults provide insig
dchip
nition of ubiquitin
elect test and sign
eans difference
elect which Sampl

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

Send to:

Filters: Manage Filters

Results: 1 to 20 of 256

<< First < Prev Page 1 of 13 Next > Last >>

□ 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, H2bo

Profile data

Profile pathways

Find related data

Database: Gene

based on gene annotation from platform pr
accessio

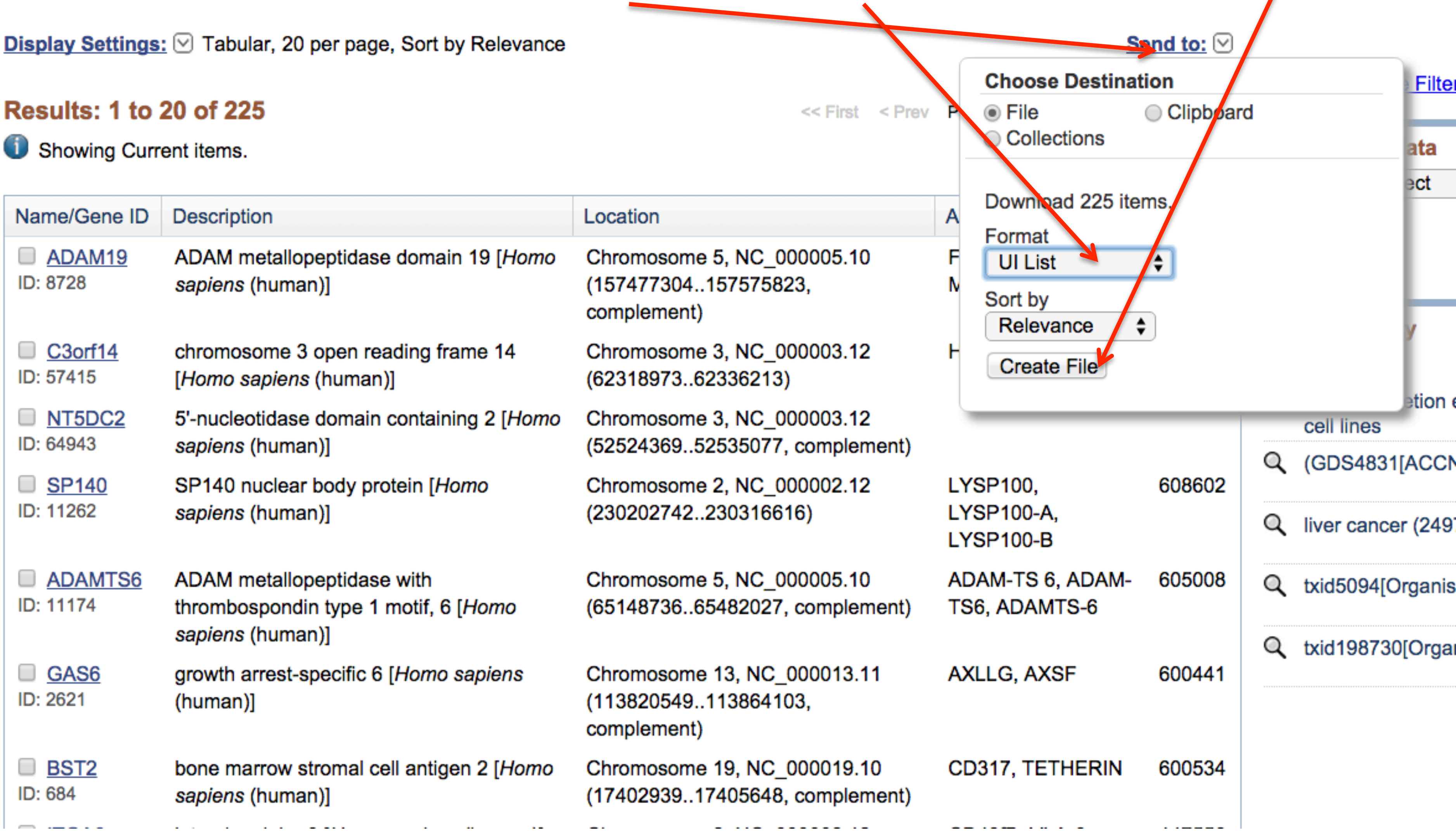
Recent activity

COP1 depletion effect on hepatocellula
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

Send to:

Results: 1 to 20 of 225

<< First < Prev P

Showing Current items.

Choose Destination

● File ○ Clipboard
○ Collections

Download 225 items.

Format: UI List

Sort by: Relevance

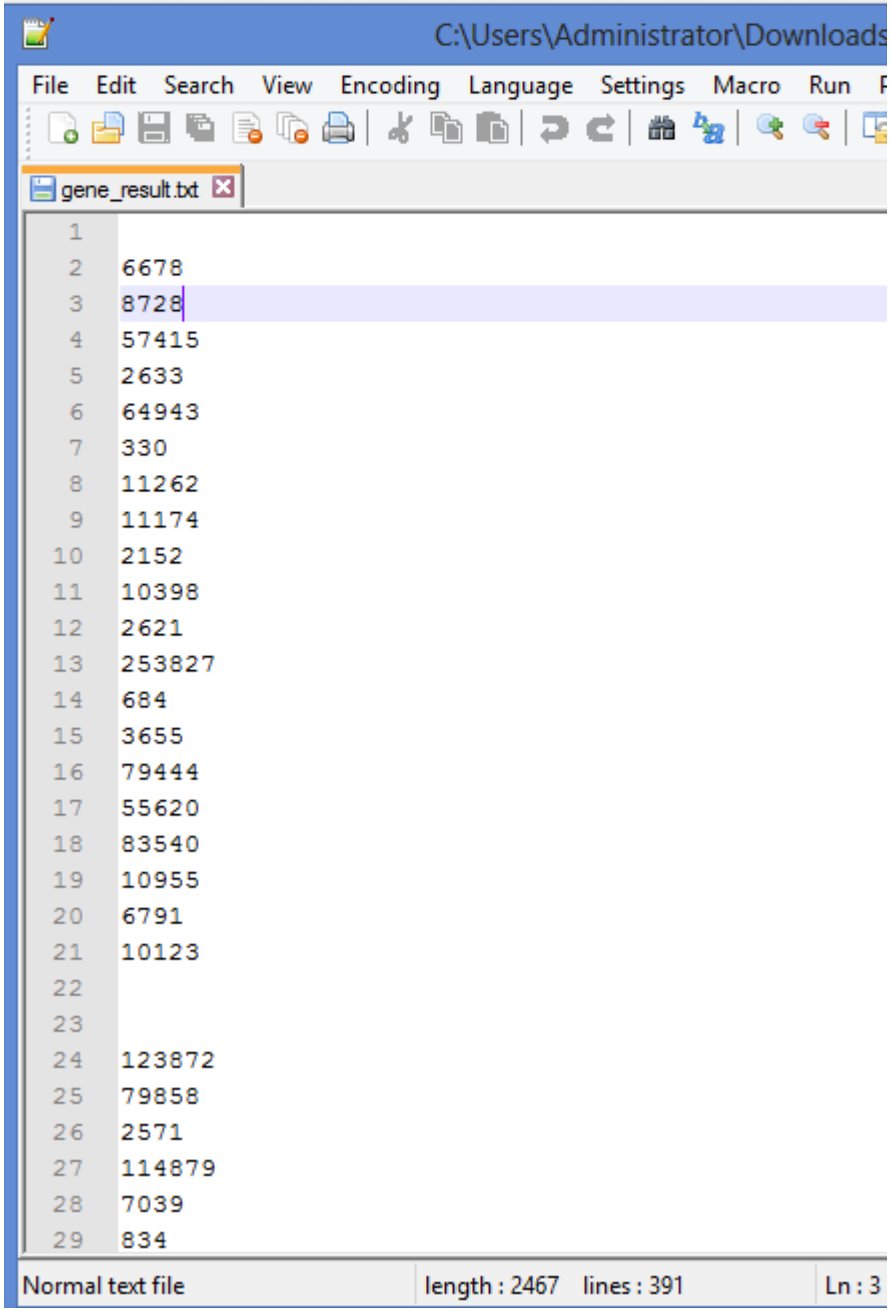
Create File

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

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++

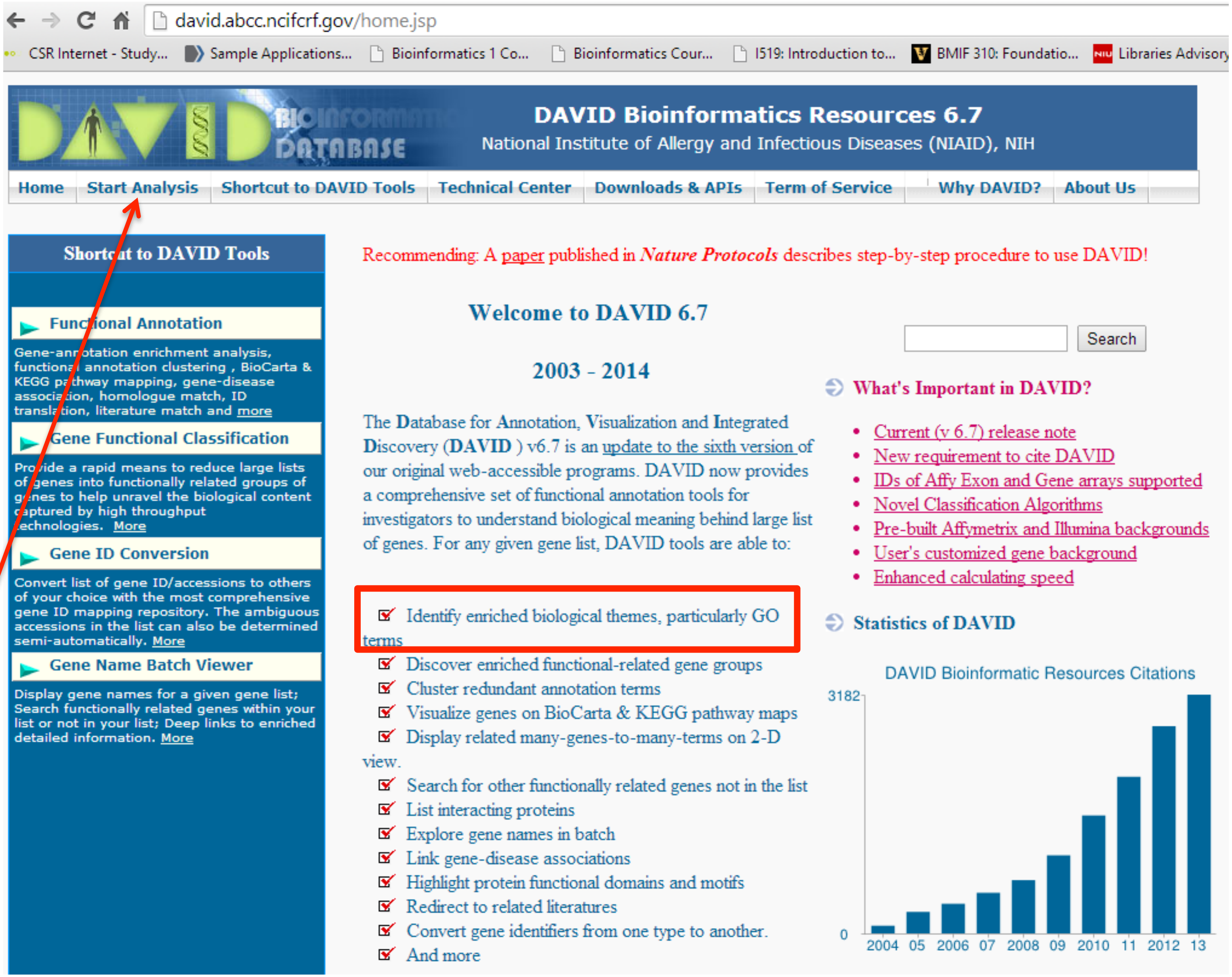


```

6678
8728
57415
2633
64943
330
11262
11174
2152
10398
2621
253827
684
3655
79444
55620
83540
10955
6791
10123


123872
79858
2571
114879
7039
834
```

Normal text file   length : 2467   lines : 391   Ln : 3

Next we will use DAVID to perform function enrichment analysis

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



DAVID Bioinformatics Resources 6.7

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

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 **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery (**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:

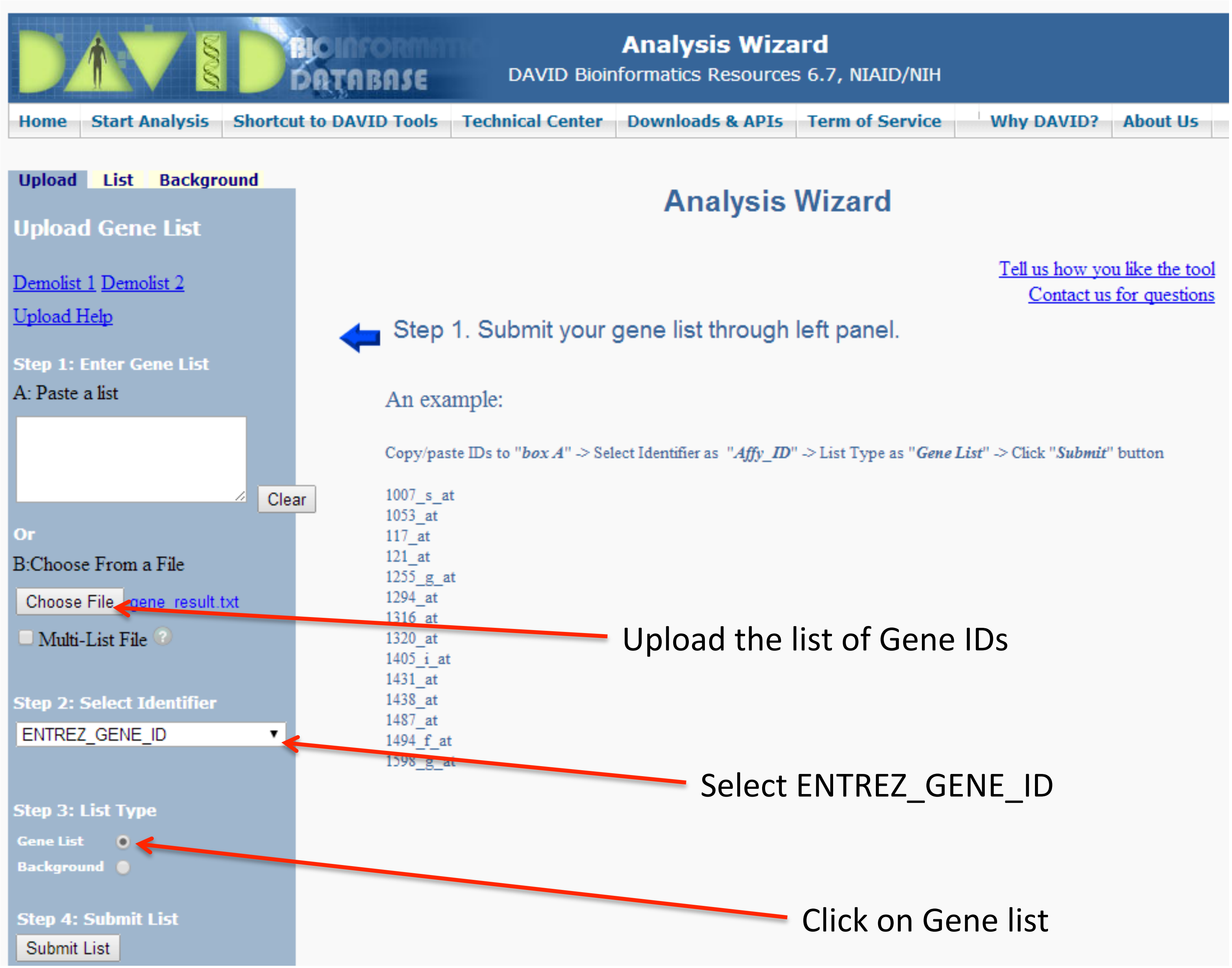
- ☑ Identify enriched biological themes, particularly GO terms
- ☑ Discover enriched functional-related gene groups
- ☑ Cluster redundant annotation terms
- ☑ Visualize genes on BioCarta & KEGG pathway maps
- ☑ Display related many-genes-to-many-terms on 2-D view.
- ☑ Search for other functionally related 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

Hit start analysis



Analysis Wizard

DAVID Bioinformatics Resources 6.7, NIAID/NIH

Upload Gene List

Demolist 1 Demolist 2

Upload Help

Step 1: Enter Gene List

A: Paste a list

Or

B:Choose From a 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

Analysis Wizard

Tell us how you like the tool

Contact us for questions

Step 1. Submit your gene list through left panel.

An example:

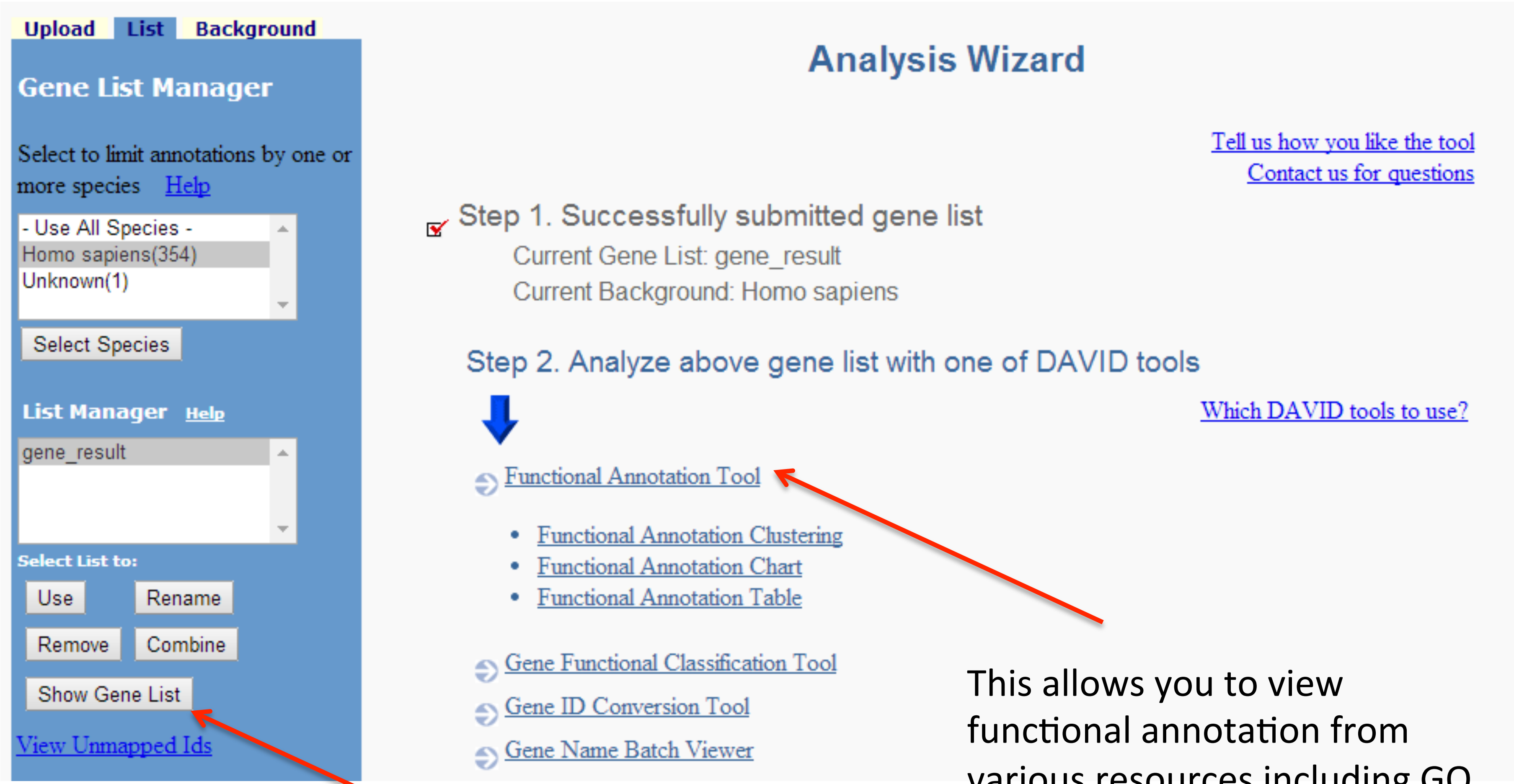
Copy/paste IDs to "*box A*" -> Select Identifier as "*Affy_ID*" -> List Type as "*Gene List*" -> Click "*Submit*" button

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



Select to limit annotations by one or more species

Analysis Wizard

☑ Step 1. Successfully submitted gene list

Current Gene List: gene_result

Current Background: Homo sapiens

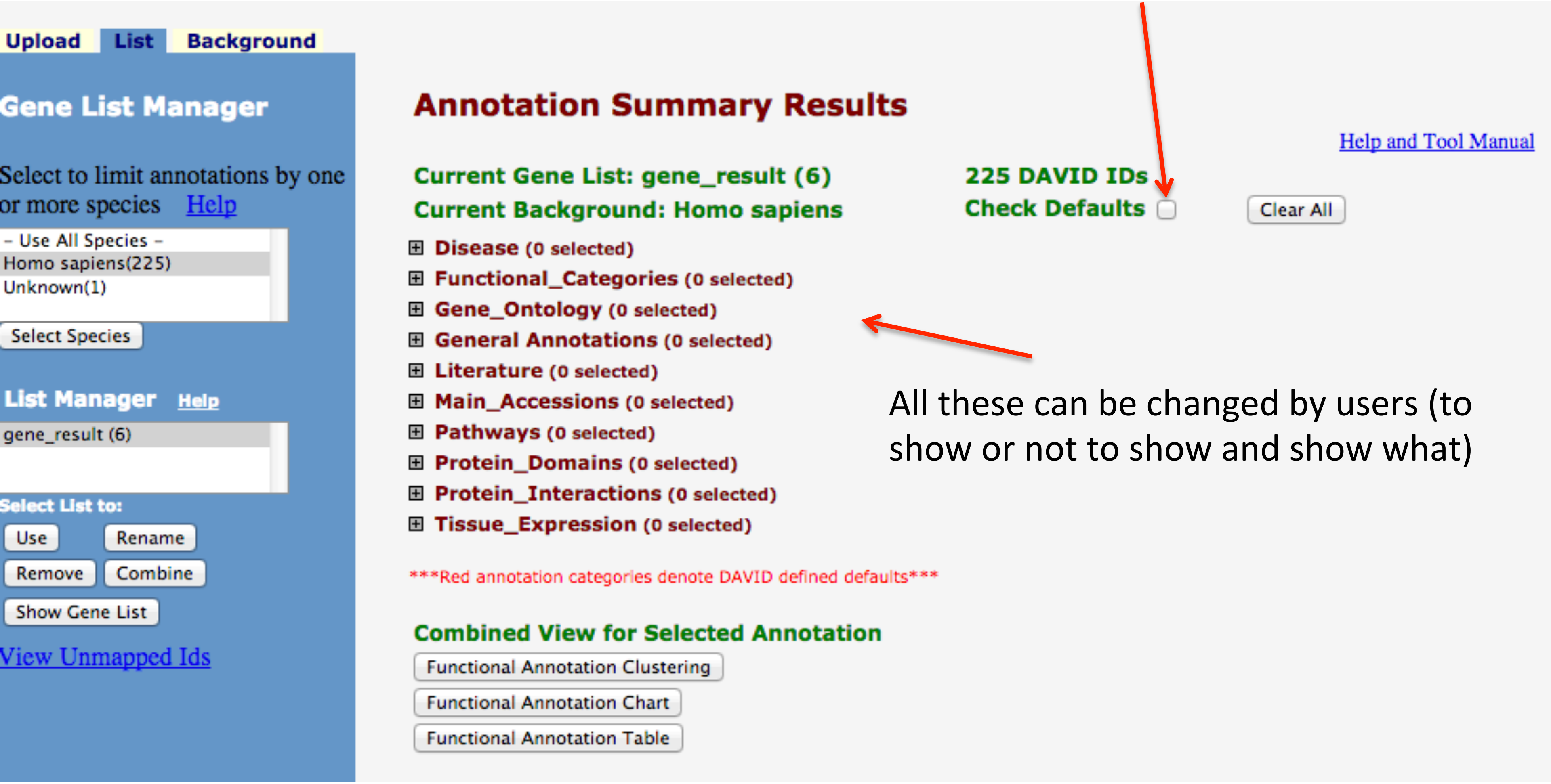
Step 2. Analyze above gene list with one of DAVID tools

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

Uncheck this



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

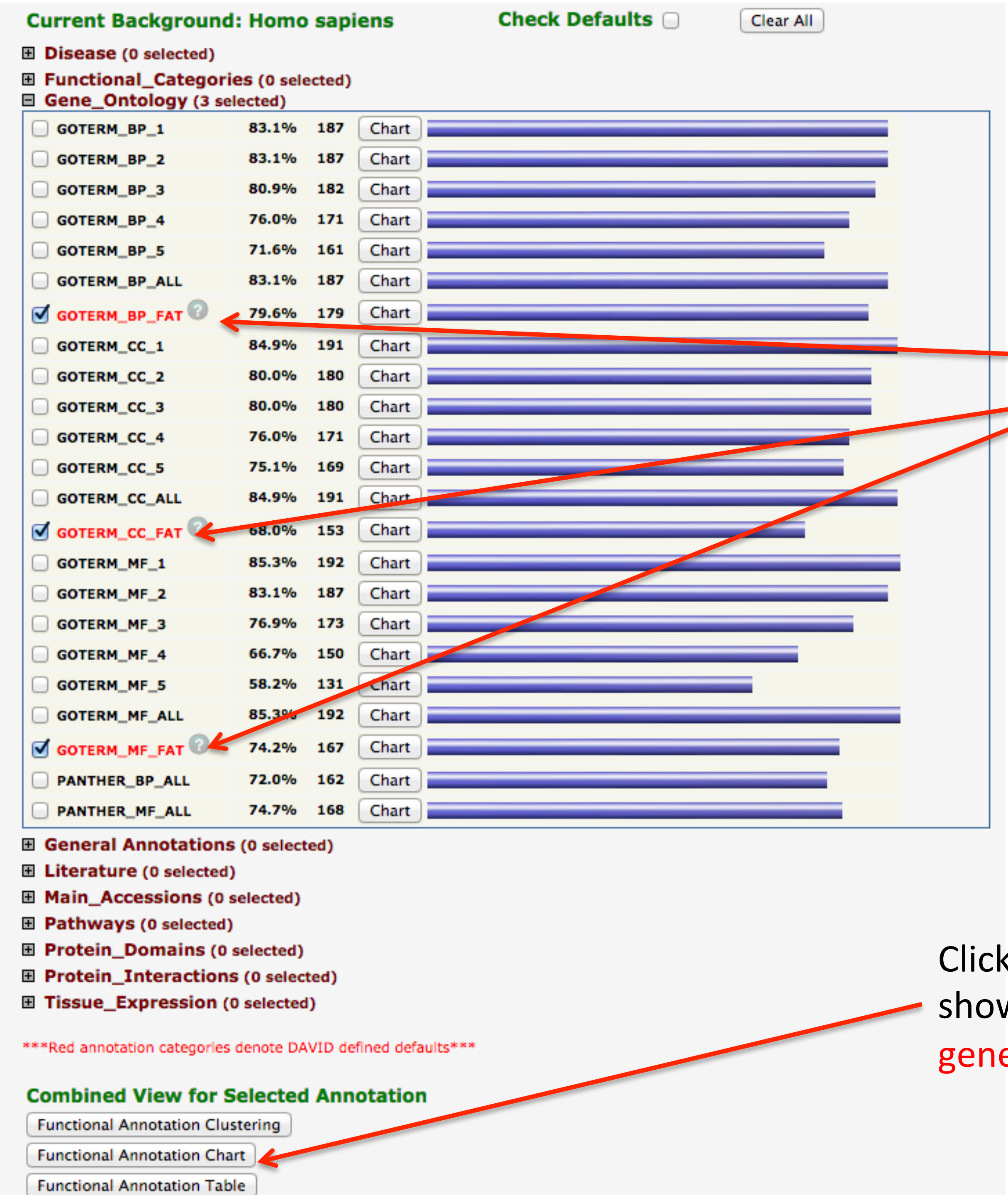
Combined View for Selected Annotation

Functional Annotation Clustering

Functional Annotation Chart

Functional Annotation Table

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



Current Background: Homo sapiens **Check Defaults** □ Clear All

⊞ **Disease** (0 selected)

⊞ **Functional_Categories** (0 selected)

⊟ **Gene_Ontology** (3 selected)

□	GOTERM_BP_1	83.1%	187	Chart
□	GOTERM_BP_2	83.1%	187	Chart
□	GOTERM_BP_3	80.9%	182	Chart
□	GOTERM_BP_4	76.0%	171	Chart
□	GOTERM_BP_5	71.6%	161	Chart
□	GOTERM_BP_ALL	83.1%	187	Chart
☑	GOTERM_BP_FAT	79.6%	179	Chart
□	GOTERM_CC_1	84.9%	191	Chart
□	GOTERM_CC_2	80.0%	180	Chart
□	GOTERM_CC_3	80.0%	180	Chart
□	GOTERM_CC_4	76.0%	171	Chart
□	GOTERM_CC_5	75.1%	169	Chart
□	GOTERM_CC_ALL	84.9%	191	Chart
☑	GOTERM_CC_FAT	68.0%	153	Chart
□	GOTERM_MF_1	85.3%	192	Chart
□	GOTERM_MF_2	83.1%	187	Chart
□	GOTERM_MF_3	76.9%	173	Chart
□	GOTERM_MF_4	66.7%	150	Chart
□	GOTERM_MF_5	58.2%	131	Chart
□	GOTERM_MF_ALL	85.3%	192	Chart
☑	GOTERM_MF_FAT	74.2%	167	Chart
□	PANTHER_BP_ALL	72.0%	162	Chart
□	PANTHER_MF_ALL	74.7%	168	Chart

⊞ **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

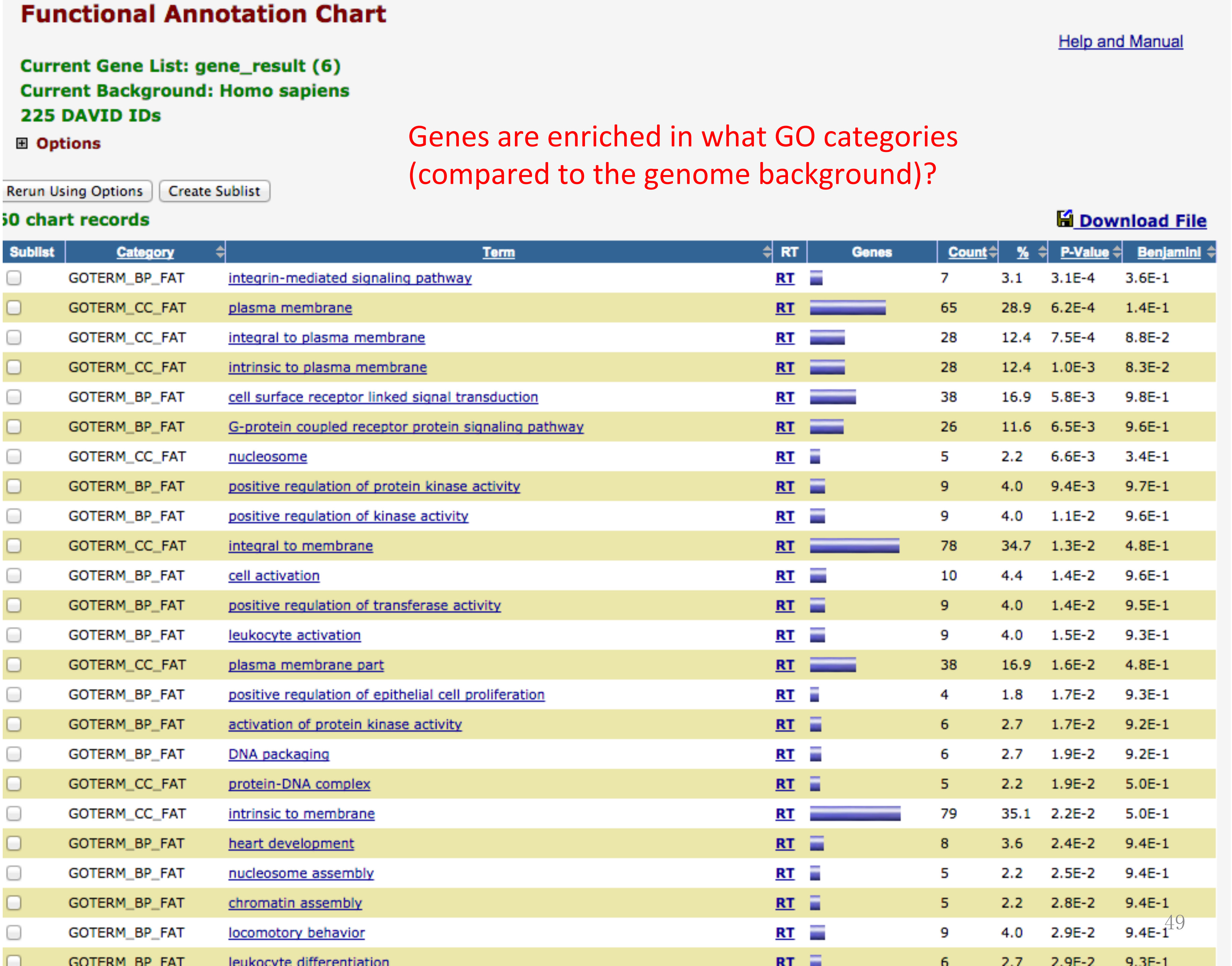
Functional Annotation Clustering

Functional Annotation Chart

Functional Annotation Table

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)`