Introduction to the Gene Ontology and GO annotation resources

Rachael Huntley

UniProt-GOA

EBI

PAGXX

January 2012

EMBL-EBI

UniProt

Why do we need GO?

Reasons for the Gene Ontology

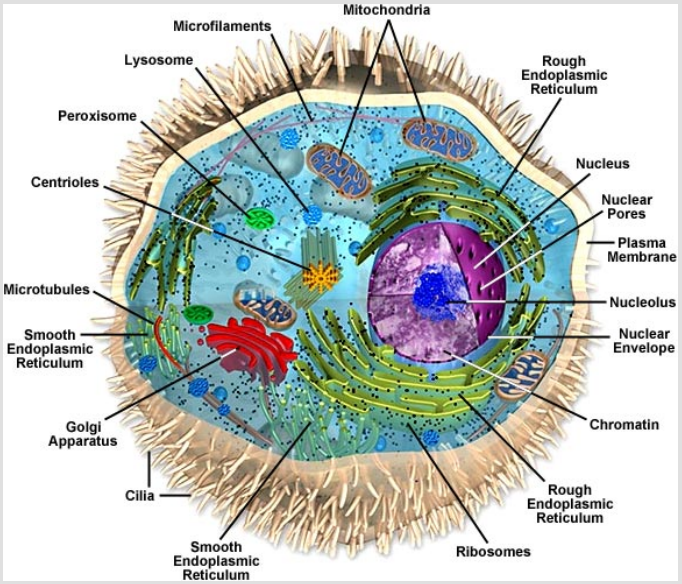
- Inconsistency in English language

Inconsistency in English languauge

- **Same** name for **different** concepts

Cell

or



??

- **Different** names for the **same** concept

Eggplant

Brinjal

Aubergine

Melongene

Same for biological concepts

→ Comparison is difficult – in particular across species or across databases

Just one reason why the Gene Ontology (GO) is is needed…

Reasons for the Gene Ontology

- Inconsistency in English language
- Increasing amounts of biological data available
- Increasing amounts of biological data to come

Increasing amounts of biological data available

NCBI Resources How To

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed | DNA repair | Search

RSS Save search Limits Advanced

Display Settings: Summary, 20 per page, Sorted by Recently Added Send to:

Results: 1 to 20 of 64637

<< First < Prev Page 1 of 3232 Next > Last >>

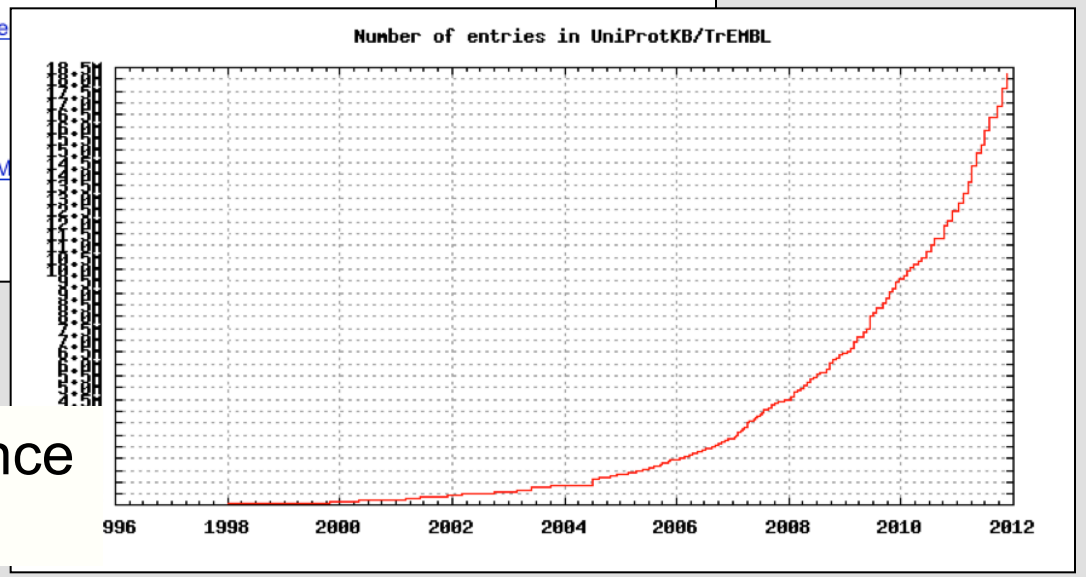
1. Germline Stem Cell Gene PIWIL2 Mediates **DNA Repair** through Relaxation of Chromatin.
   Yin DT, Wang Q, Chen L, Liu MY, Han C, Yan Q, Shen R, He G, Duan W, Li JJ, Wani A, Gao JX.
   PLoS One. 2011;6(11):e27154. Epub 2011 Nov 16.
   PMID: 22110608 [PubMed - in process]

2. An MSI Tumor Specific Frameshift Mutation in a Coding Microsatellite of MSH3 Encodes Epitopes.
   Garbe Y, Maletzki C, Linnebacher M.
   PLoS One. 2011;6(11):e26517. Epub 2011 Nov 14.
   PMID: 22110587 [PubMed - in process]

3. A Review of ERCC1 Gene in Bladder Cancer: Implications for Carcinoge
   Kawashima A, Takayama H, Tsujimura A.
   Adv Urol. 2012;2012:812398. Epub 2011 Oct 26.
   PMID: 22110495 [PubMed - as supplied by publisher]

4. Polymorphisms in **DNA Repair** Gene XRCC1 and Skin Cancer Risk: A M
   Zhang H, Li W, Franklin MJ, Dudek AZ.
   Anticancer Res. 2011 Nov;31(11):3945-52.
   PMID: 22110224 [PubMed - in process]

Search on ‘DNA repair’... get almost 65,000 results



Expansion of sequence information

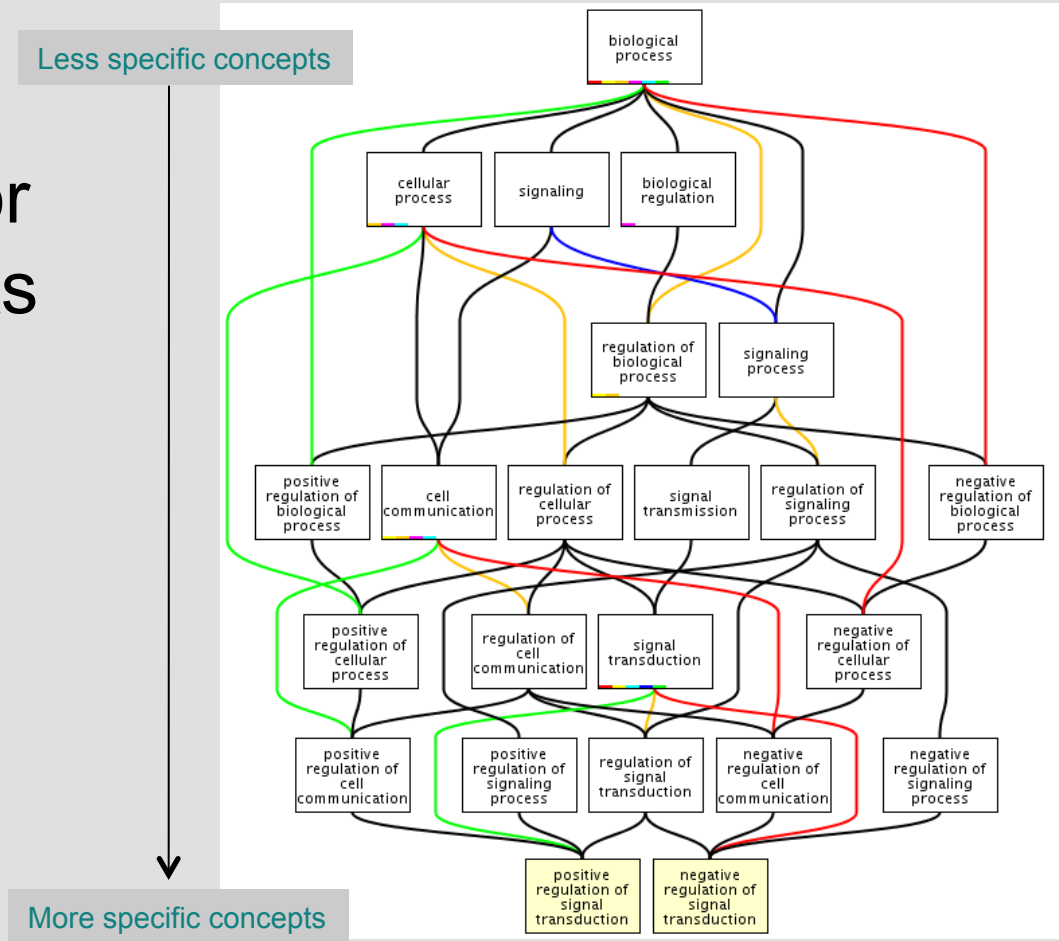
Reasons for the Gene Ontology

- Inconsistency in English language
- Increasing amounts of biological data available
- Increasing amounts of biological data to come
- Large datasets need to be interpreted quickly

The Gene Ontology

- A way to capture biological knowledge for individual gene products in a written and computable form

- A set of concepts and their relationships to each other arranged as a hierarchy

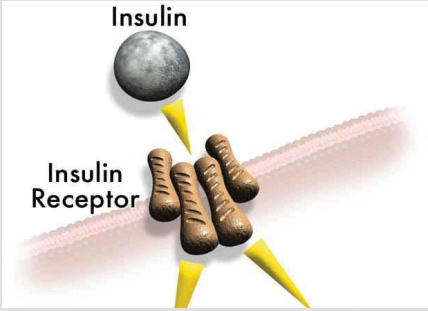


www.ebi.ac.uk/QuickGO

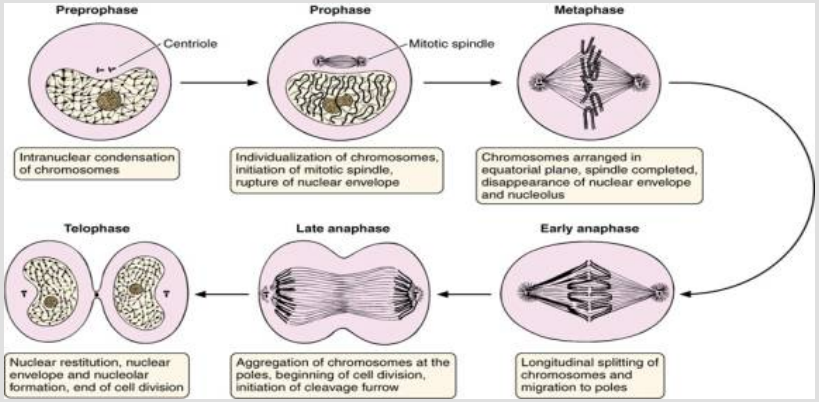
The Concepts in GO

1. Molecular Function

An elemental activity or task or job



- protein kinase activity
- insulin receptor activity



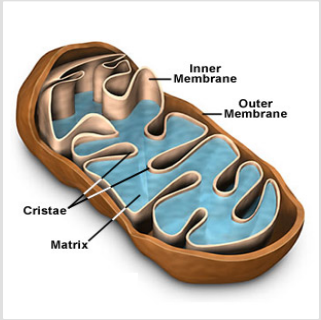
2. Biological Process

A commonly recognised series of events

- cell division

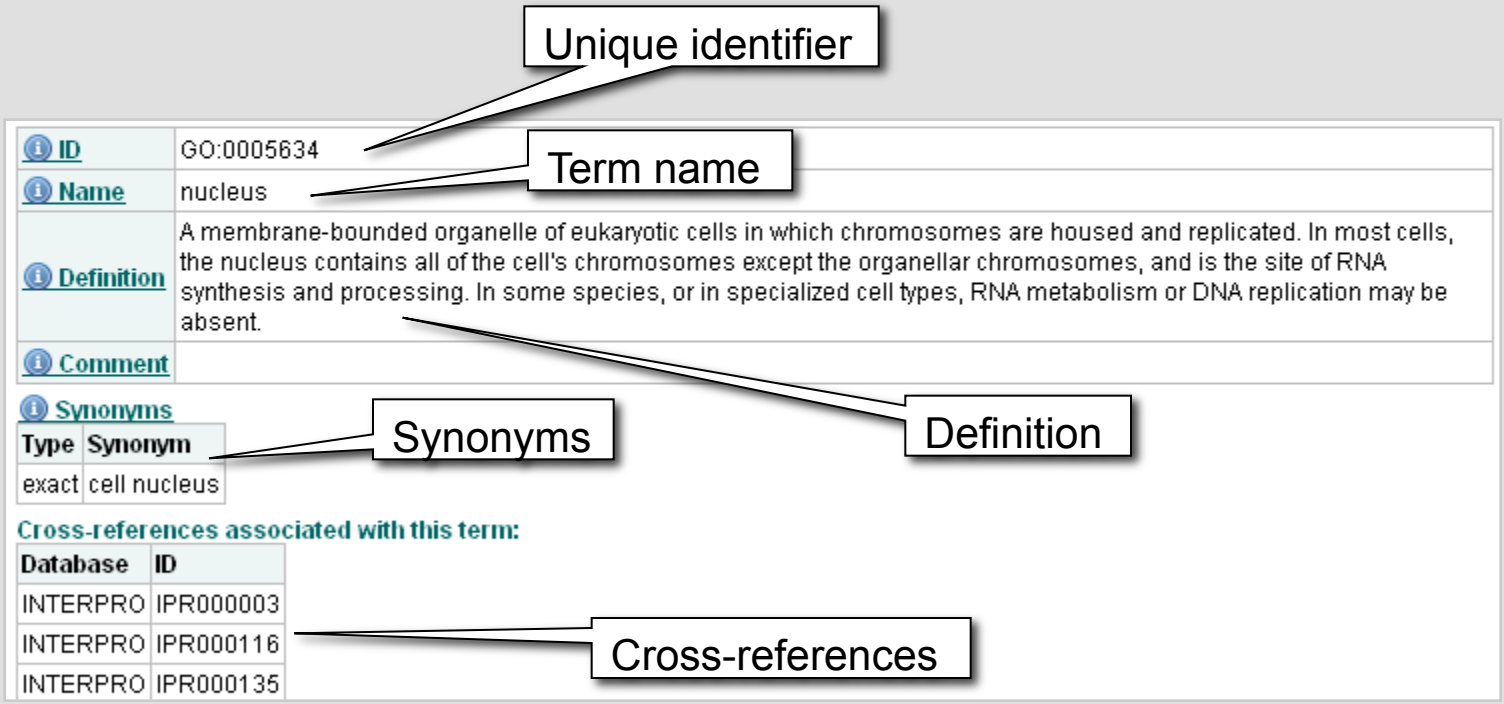
3. Cellular Component

Where a gene product is located



- mitochondrion
- mitochondrial matrix
- mitochondrial inner membrane

Anatomy of a GO term



Ontology structure

- Directed acyclic graph

  Terms can have more than one parent

- Terms are linked by relationships

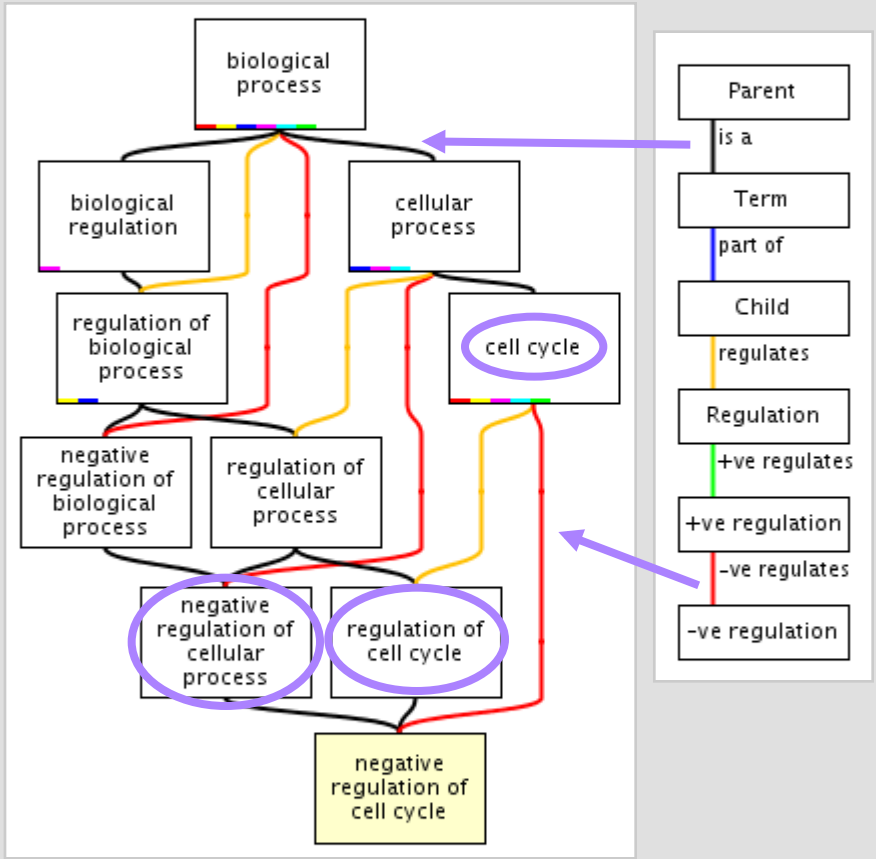
  is_a

  part_of

  regulates (and +/- regulates)

  has_part

  occurs_in



www.ebi.ac.uk/QuickGO

These relationships allow for complex analysis of large datasets

dictyBase

Annotating cardiovascular genes

wellcome trust
sanger
institute
GeneDB

MGI

FlyBase

SGD

GO
GENE ONTOLOGY™ CONSORTIUM
http://www.geneontology.org

EcoliHub

InterPro

WormBase
The Biology and Genome of C. elegans.

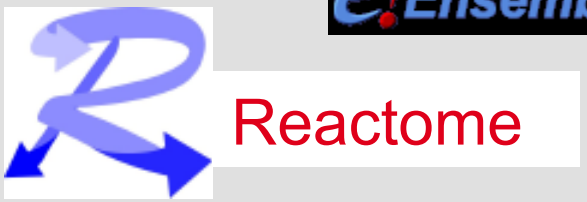
e!EnsemblGenomes

tair

e!Ensembl

BERKELEY LAB

Annotation@EBI


Reactome

AgBase

GRAMENE

University of Maryland School of Medicine
INSTITUTE FOR GENOME SCIENCES

ZFIN

J. Craig Venter
INSTITUTE

RGD

UniProt

EMBL-EBI

Aims of the GO project

- Compile the ontologies

  - currently over 35,000 terms
  - constantly increasing and improving

- Annotate gene products using ontology terms

  - around 30 groups provide annotations

- Provide a public resource of data and tools

  - regular releases of annotations
  - tools for browsing/querying annotations and editing the ontology

GO Annotation

UniProt-Gene Ontology Annotation (UniProt-GOA) database at the EBI

- Largest open-source contributor of annotations to GO
- Provides annotation for more than 397,000 species
- Our priority is to annotate the human proteome

A GO annotation is …

…a statement that a gene product;

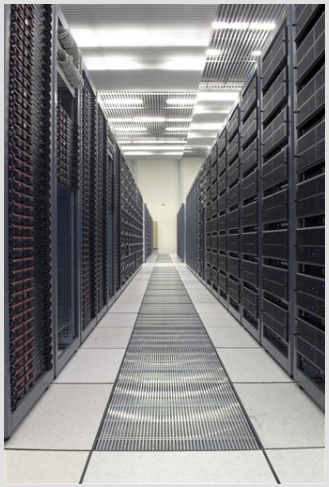
1. has a particular molecular function
   or is involved in a particular biological process
   or is located within a certain cellular component

2. as determined by a particular method

3. as described in a particular reference

Accession	*Name*	*GO ID*	*GO term name*	*Reference*	*Evidence code*
P00505	GOT2	GO:0004069	aspartate transaminase activity	PMID:2731362	IDA

UniProt-GOA incorporates annotations made using two methods



Electronic Annotation

- Quick way of producing large numbers of annotations
- Annotations use less-specific GO terms
- Only source of annotation for many non-model organism species

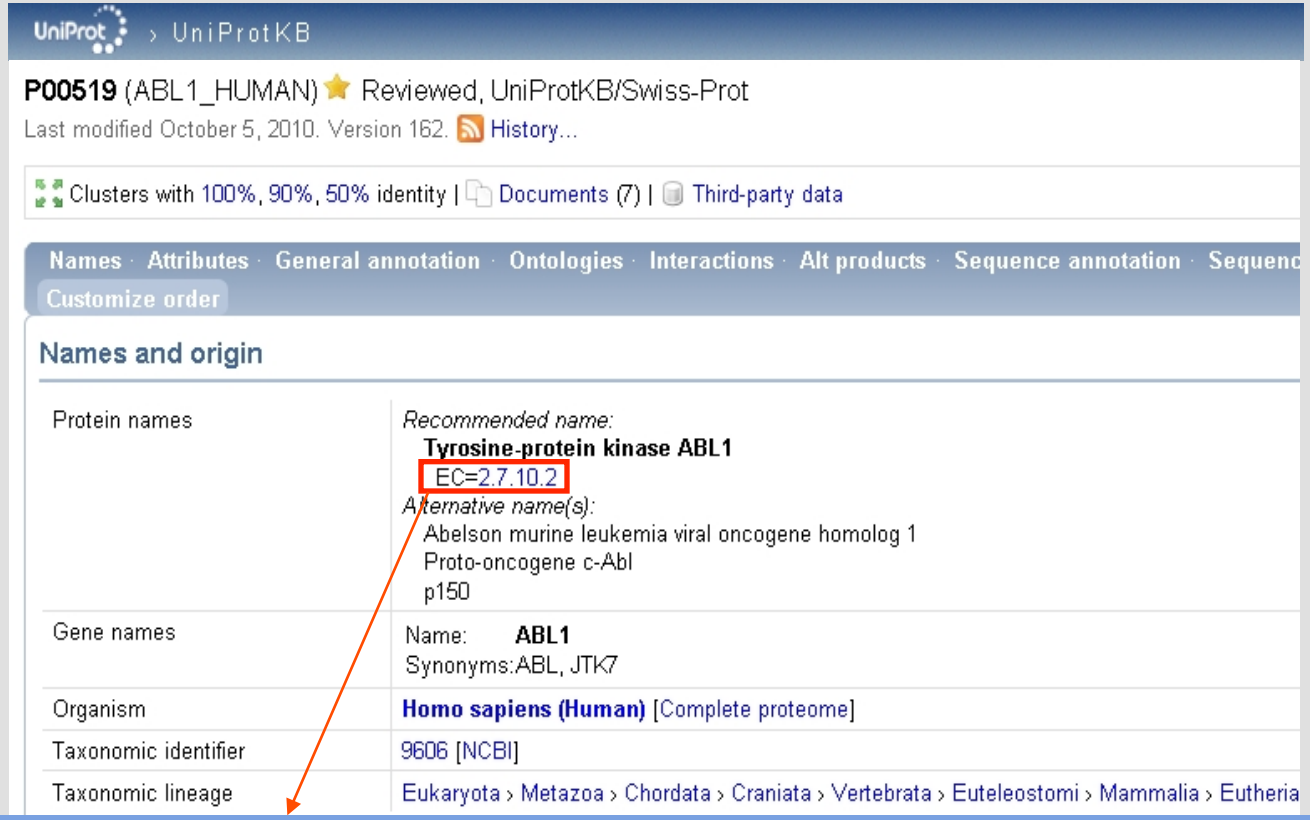
Manual Annotation

- Time-consuming process producing lower numbers of annotations
- Annotations tend to use very specific GO terms

Electronic annotation methods

1. Mapping of external concepts to GO terms

e.g. InterPro2GO, UniProt Keyword2GO, Enzyme Commission2GO



GO:0004715 ; non-membrane spanning protein tyrosine kinase activity

Electronic annotation methods

2. Automatic transfer of manual annotations to orthologs

e.g. Human

Ensembl compara

Macaque

Chimpanzee

Guinea Pig

Rat

...and more

Mouse

Cow

Dog

Chicken

Arabidopsis

Rice

Ensembl compara

Poplar

Maize

Brachypodium

Grape

…and more

Annotations are high-quality and have an explanation of the method (GO_REF)

Manual annotation by GOA

High–quality, specific annotations made using:

- Full text peer-reviewed papers
- A range of evidence codes to categorise the types of evidence found in a paper

  e.g. IDA, IMP, IPI

Number of annotations in UniProt-GOA database

Electronic annotations	110,247,289
Manual annotations*	920,557

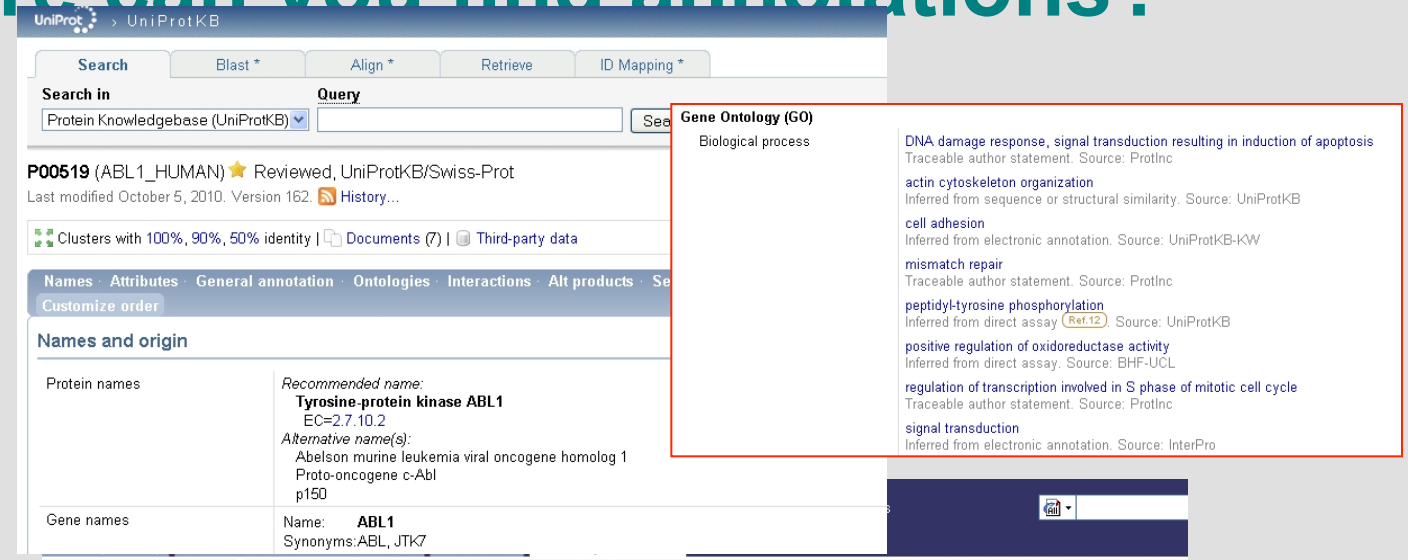
Dec 2011 Statistics

* Includes manual annotations integrated from external model organism and specialist groups

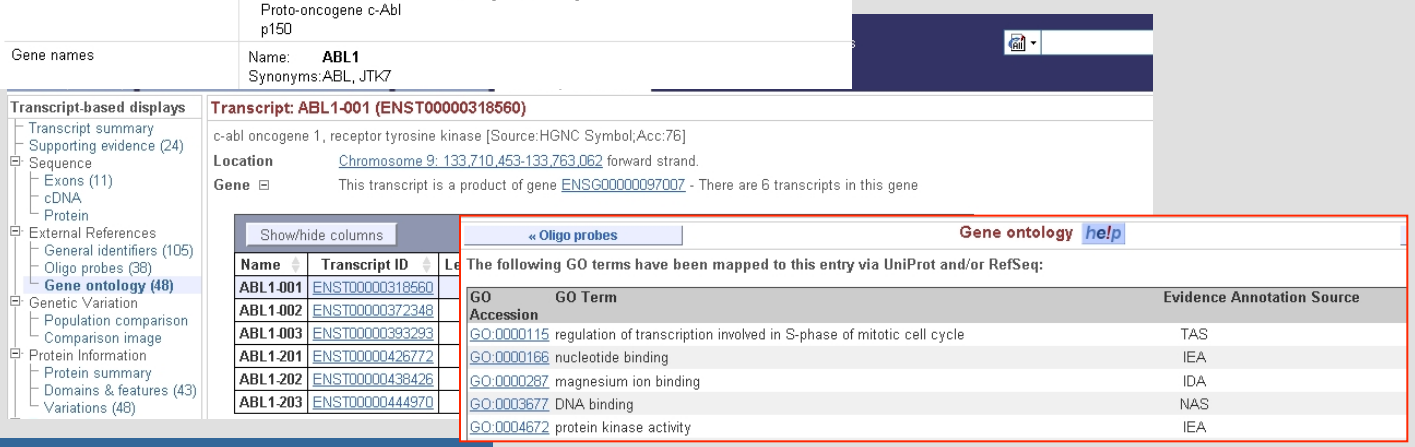
How to access and use GO annotation data

Where can you find annotations?

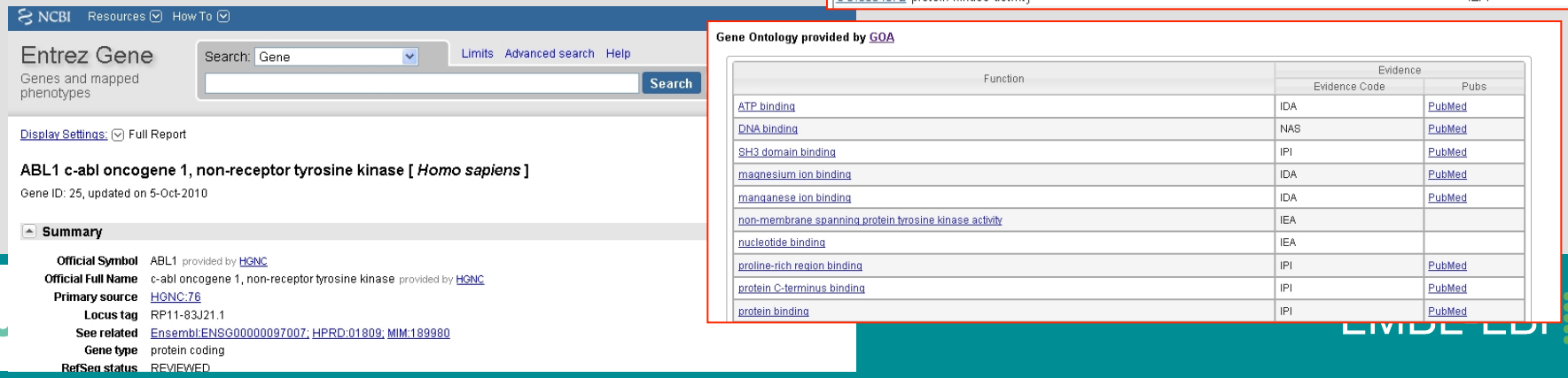
UniProtKB



Ensembl

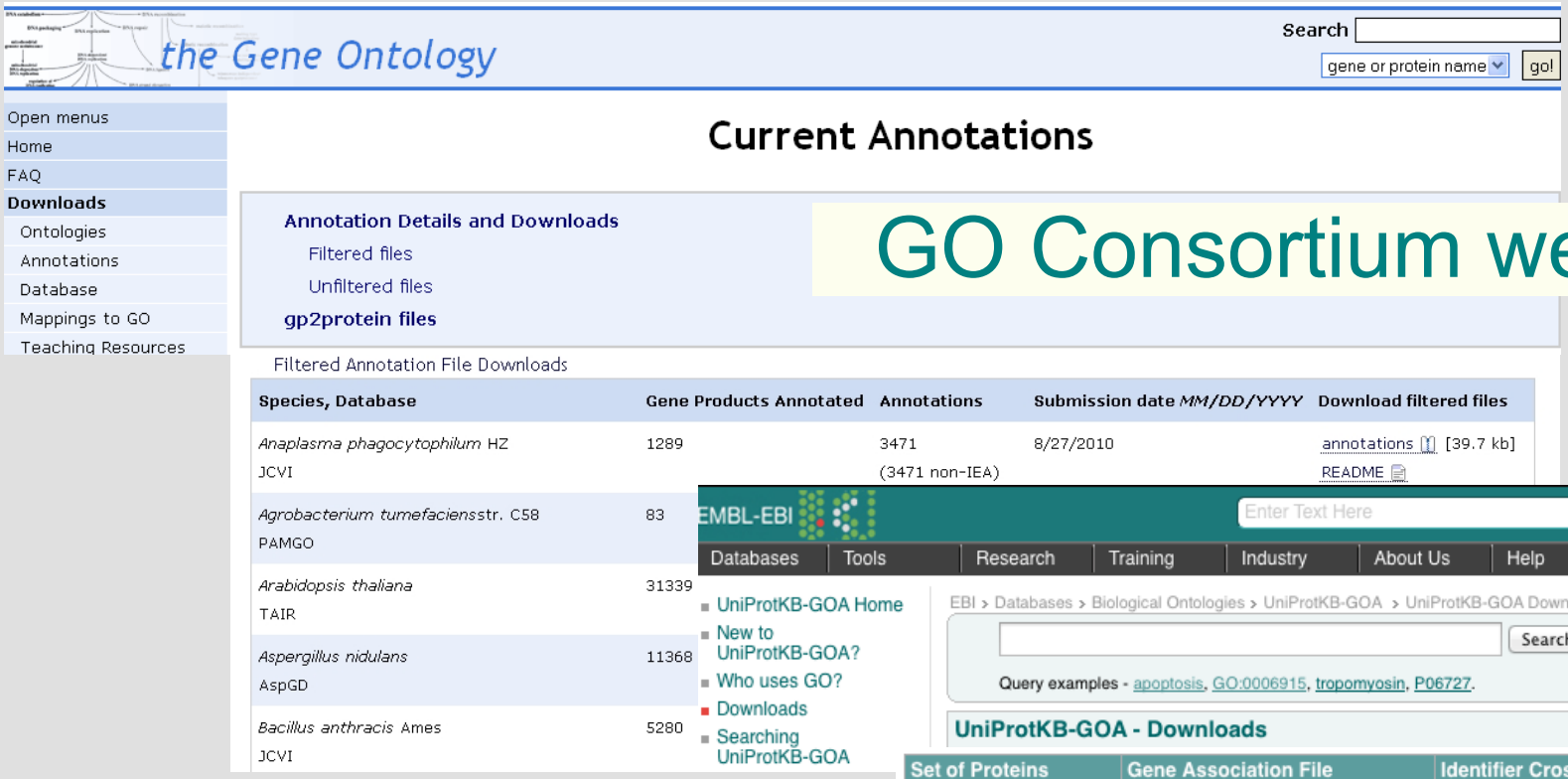


Entrez gene

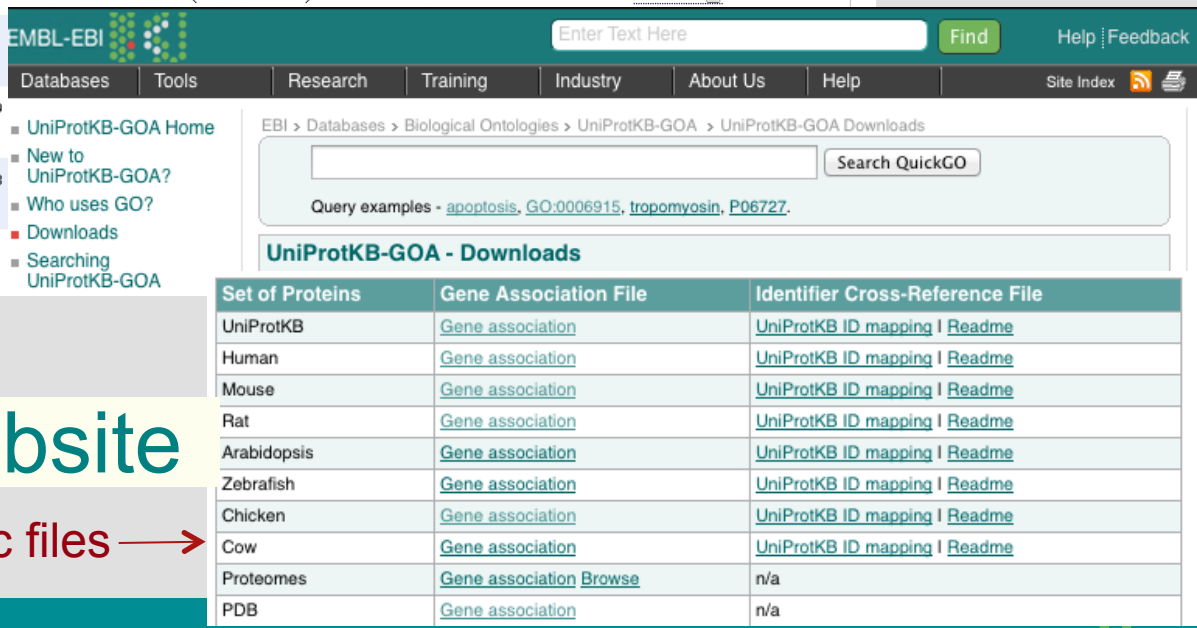


Gene Association Files

17 column files containing all information for each annotation



GO Consortium website



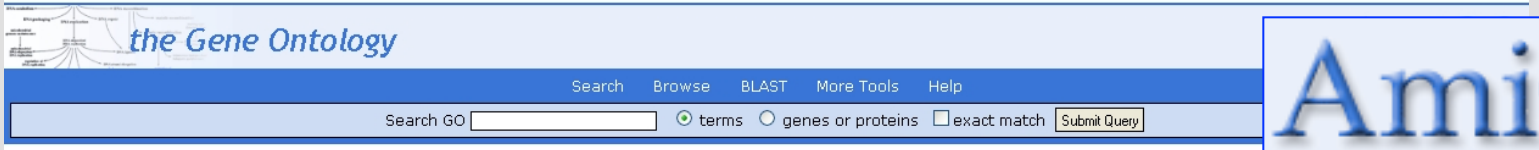
UniProt-GOA website

Numerous species-specific files →

http://www.ebi.ac.uk/GOA/downloads.html

EMBL-EBI

GO browsers



ABL1

Tyrosine-protein kinase ABL1

protein from *Homo sapiens* (human)

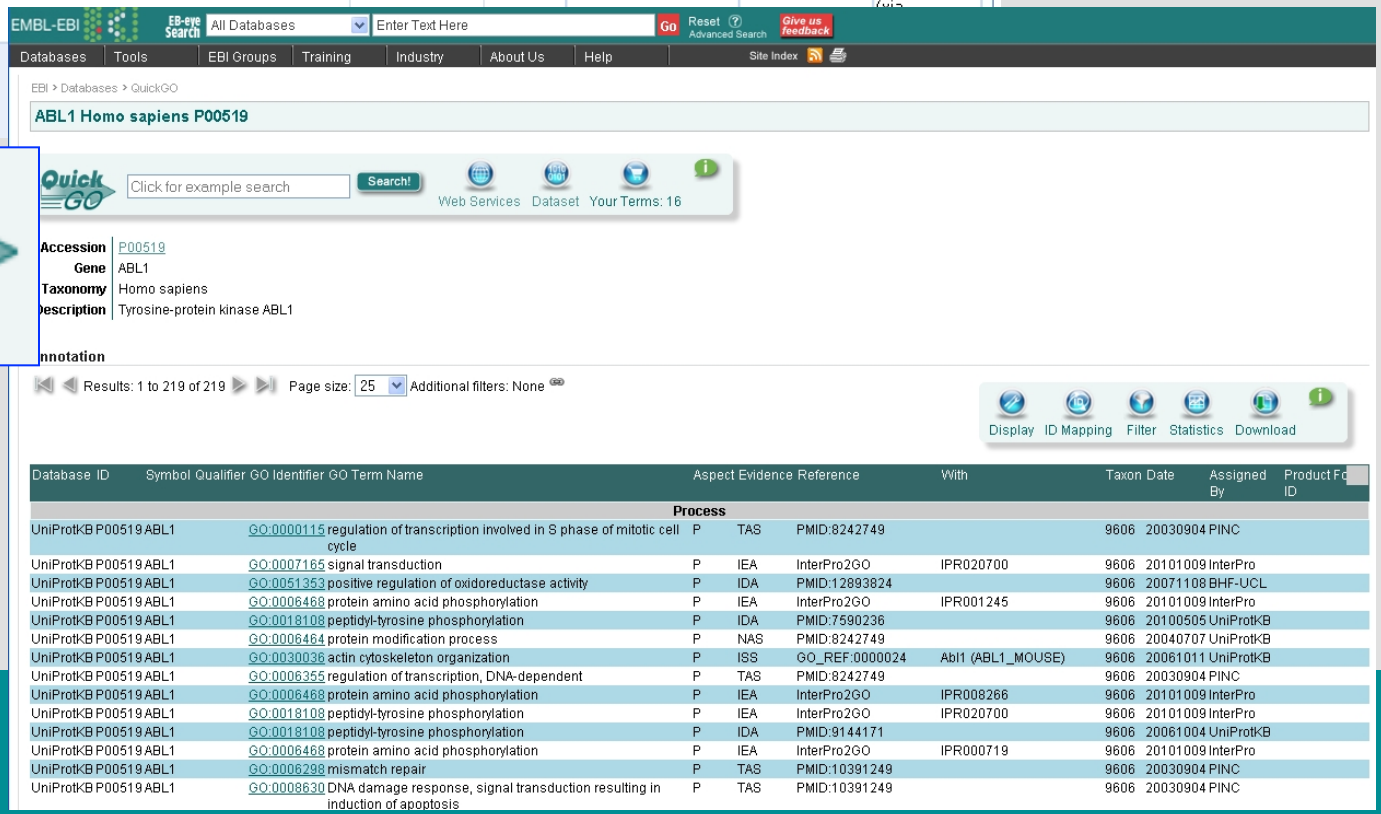
Term associations ⬇ Gene product information ➡ Peptide Sequence ➡ Sequence information ➡

Term Associations

Download all association information in: gene association format RDF-XML

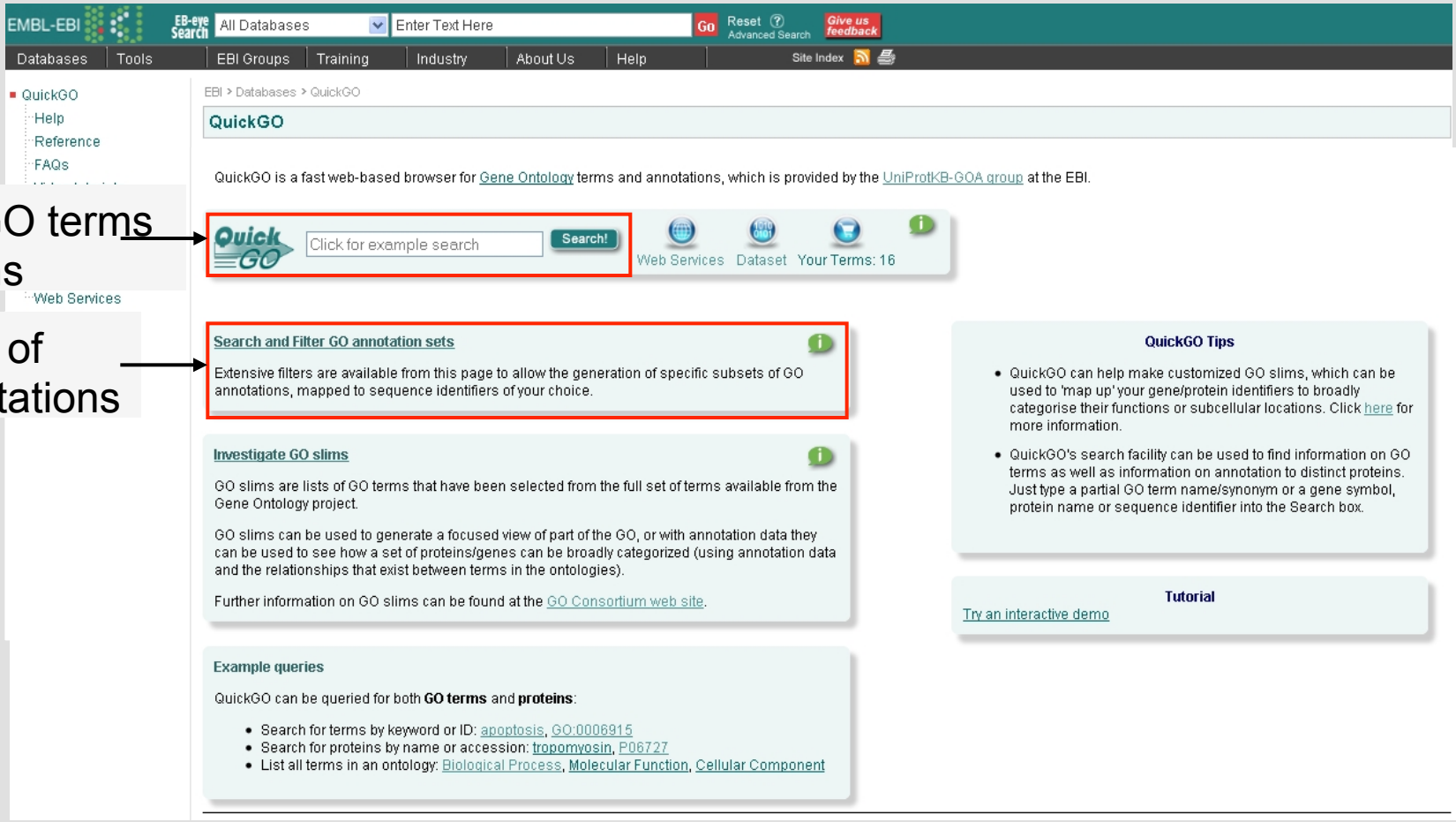
Select all | Clear all | Perform an action with this page's selected terms... | Go!

Accession, Term		Ontology	Qualifier	Evidence	Reference	Assigned by
GO:0030036 : actin cytoskeleton organization	2336 gene products view in tree	biological process		ISS With UniProtKB:P00520	GO REF:0000024	UniProtKB
GO:0007155 : cell adhesion	4735 gene products view in tree	biological process		IEA With SP_KW:KW-0130	GO REF:0000004	UniProtKB
GO:0008630 : DNA damage response, signal transduction resulting in induction of apoptosis	145 gene products view in tree	biological process		TAS	PMID:10391249	Proteome Inc.
GO:0006298 : mismatch repair						



Database ID	Symbol	Qualifier	GO Identifier	GO Term Name	Aspect	Evidence	Reference	With	Taxon	Date	Assigned By	Product Form ID
Process												
UniProtKB	P00519	ABL1		GO:0000115 regulation of transcription involved in S phase of mitotic cell cycle	P	TAS	PMID:8242749		9606	20030904	PINC	
UniProtKB	P00519	ABL1		GO:0007165 signal transduction	P	IEA	InterPro2GO	IPR020700	9606	20101009	InterPro	
UniProtKB	P00519	ABL1		GO:0051353 positive regulation of oxidoreductase activity	P	IDA	PMID:12893824		9606	20071108	BHF-UCL	
UniProtKB	P00519	ABL1		GO:0006468 protein amino acid phosphorylation	P	IEA	InterPro2GO	IPR001245	9606	20101009	InterPro	
UniProtKB	P00519	ABL1		GO:0018108 peptidyl-tyrosine phosphorylation	P	IDA	PMID:7590236		9606	20100505	UniProtKB	
UniProtKB	P00519	ABL1		GO:0006464 protein modification process	P	NAS	PMID:8242749		9606	20040707	UniProtKB	
UniProtKB	P00519	ABL1		GO:0030036 actin cytoskeleton organization	P	ISS	GO_REF:0000024	Abl1 (ABL1_MOUSE)	9606	20061011	UniProtKB	
UniProtKB	P00519	ABL1		GO:0006355 regulation of transcription, DNA-dependent	P	TAS	PMID:8242749		9606	20030904	PINC	
UniProtKB	P00519	ABL1		GO:0006468 protein amino acid phosphorylation	P	IEA	InterPro2GO	IPR008266	9606	20101009	InterPro	
UniProtKB	P00519	ABL1		GO:0018108 peptidyl-tyrosine phosphorylation	P	IEA	InterPro2GO	IPR020700	9606	20101009	InterPro	
UniProtKB	P00519	ABL1		GO:0018108 peptidyl-tyrosine phosphorylation	P	IDA	PMID:9144171		9606	20061004	UniProtKB	
UniProtKB	P00519	ABL1		GO:0006468 protein amino acid phosphorylation	P	IEA	InterPro2GO	IPR000719	9606	20101009	InterPro	
UniProtKB	P00519	ABL1		GO:0006298 mismatch repair	P	TAS	PMID:10391249		9606	20030904	PINC	
UniProtKB	P00519	ABL1		GO:0008630 DNA damage response, signal transduction resulting in induction of apoptosis	P	TAS	PMID:10391249		9606	20030904	PINC	

The EBI's QuickGO browser



Search GO terms or proteins

Find sets of GO annotations

How scientists use the GO

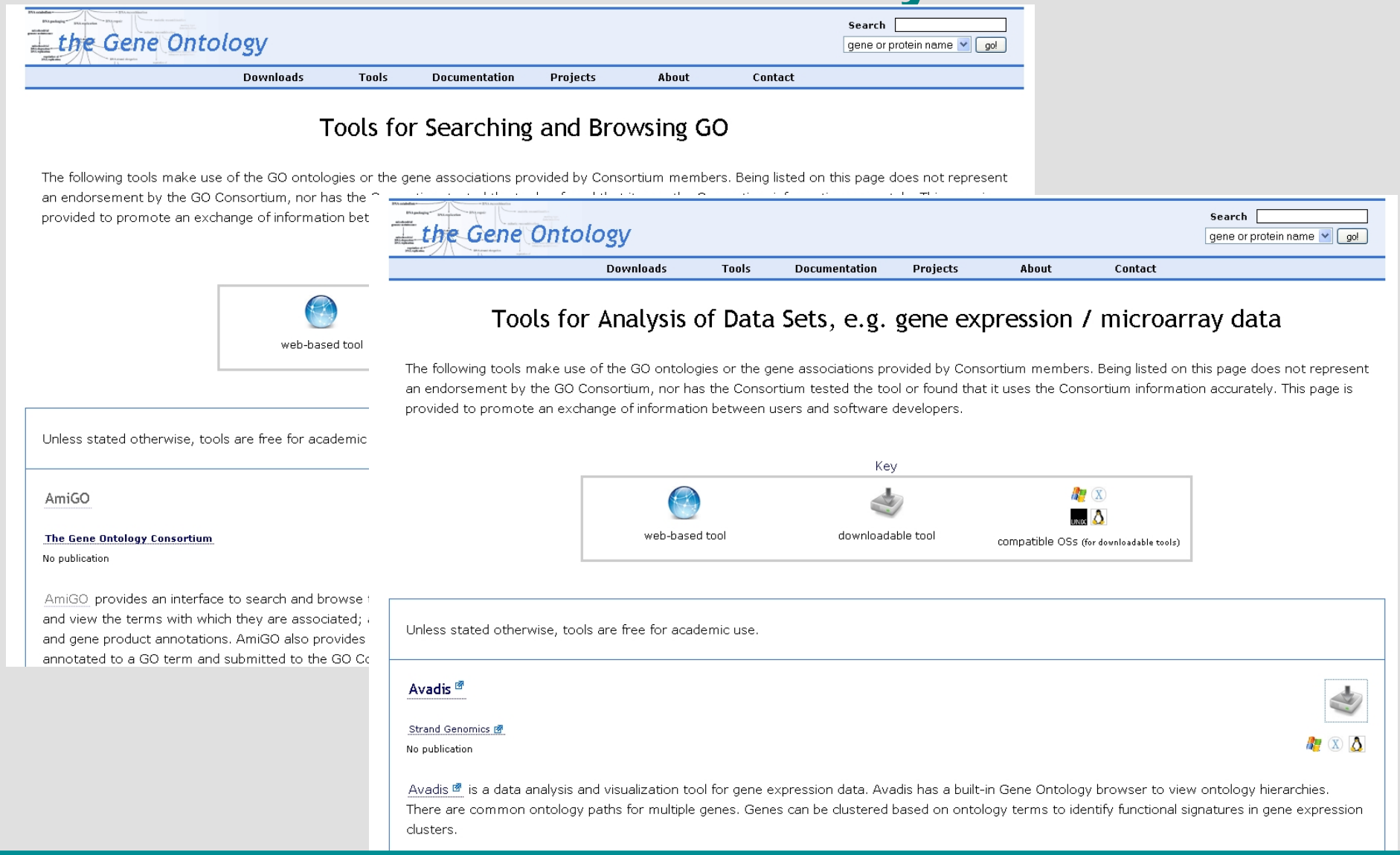
- Access gene product functional information
- Analyse high-throughput genomic or proteomic datasets
- Validation of experimental techniques
- Get a broad overview of a proteome
- Obtain functional information for novel gene products

Some examples…

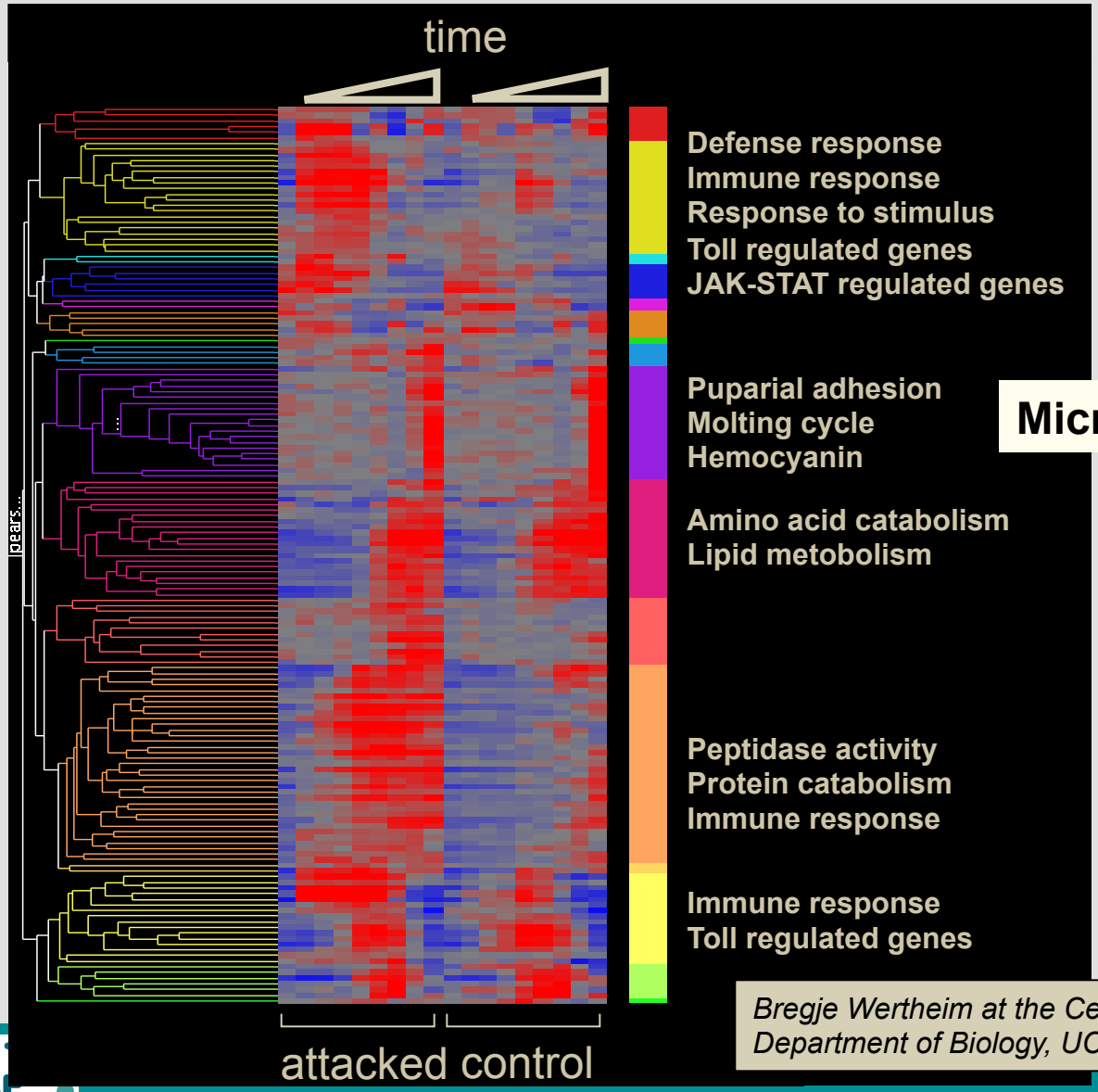
Term enrichment

- Most popular type of GO analysis
- Determines which GO terms are more often associated with a specified list of genes/proteins compared with a control list or rest of genome
- Many tools available to do this analysis
- User must decide which is best for their analysis

Numerous Third Party Tools



Analysis of high-throughput genomic datasets



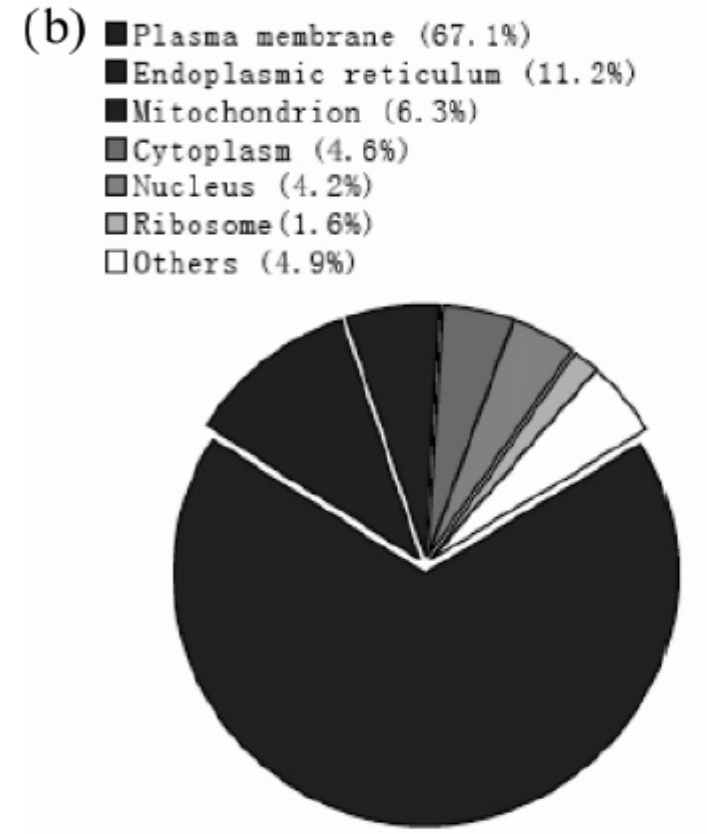
MicroArray data analysis

Bregje Wertheim at the Centre for Evolutionary Genomics, Department of Biology, UCL and Eugene Schuster Group, EBI.

UniProt

EMBL-EBI

Validation of experimental techniques



Rat liver plasma membrane isolation

(Cao *et al*., Journal of Proteome Research 2006)

Annotating novel sequences

- Can use BLAST queries to find similar sequences with GO annotation which can be transferred to the new sequence

- Two tools currently available;

AmiGO BLAST – searches the GO Consortium database

BLAST2GO – searches the NCBI database

Using the GO to provide a functional overview for a large dataset

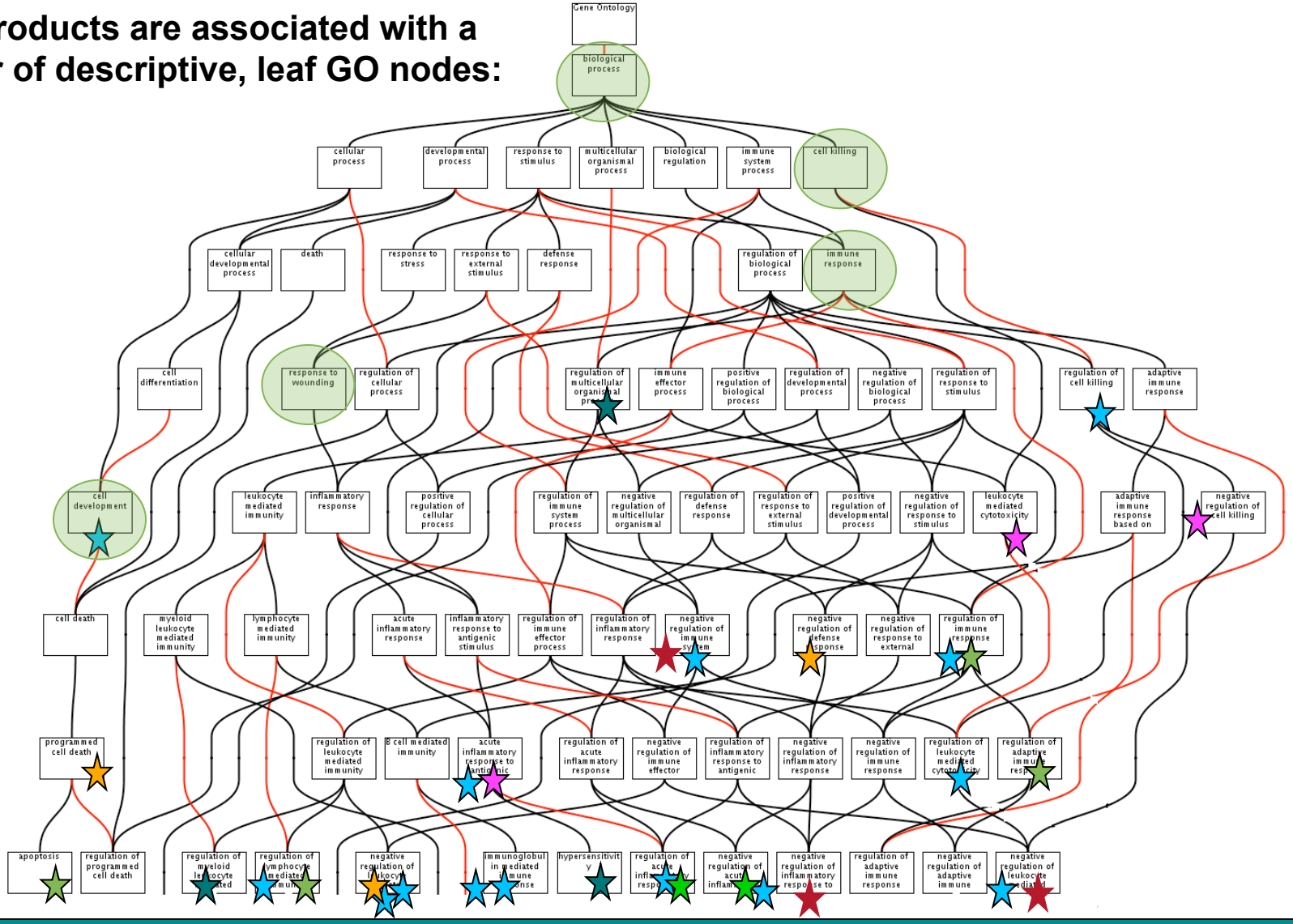
- Many GO analysis tools use GO slims to give a broad overview of the dataset

- GO slims are cut-down versions of the GO and contain a subset of the terms in the whole GO

- GO slims usually contain less-specialised GO terms

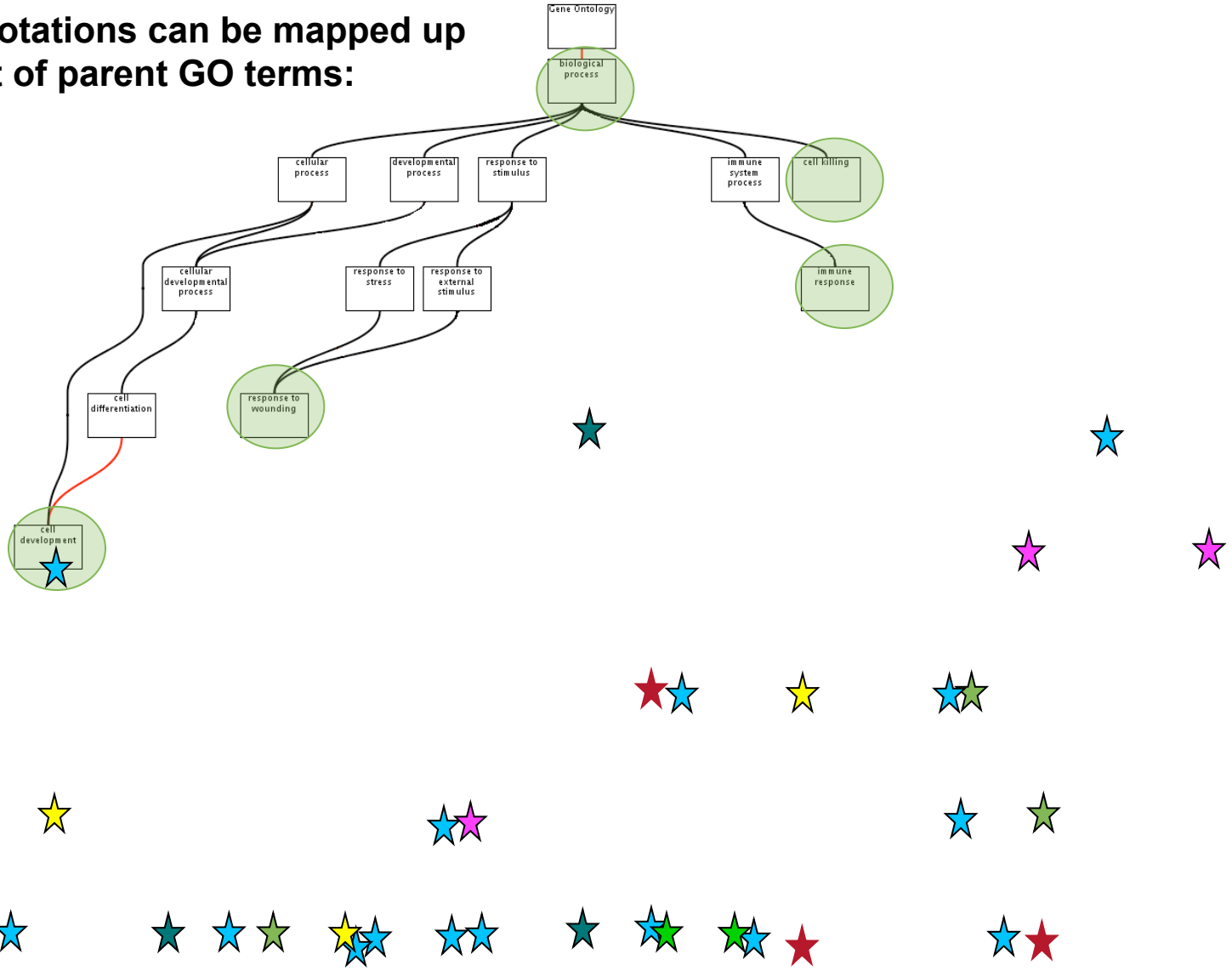
Slimming the GO using the ‘true path rule’

Many gene products are associated with a large number of descriptive, leaf GO nodes:



Slimming the GO using the ‘true path rule’

…however annotations can be mapped up to a smaller set of parent GO terms:



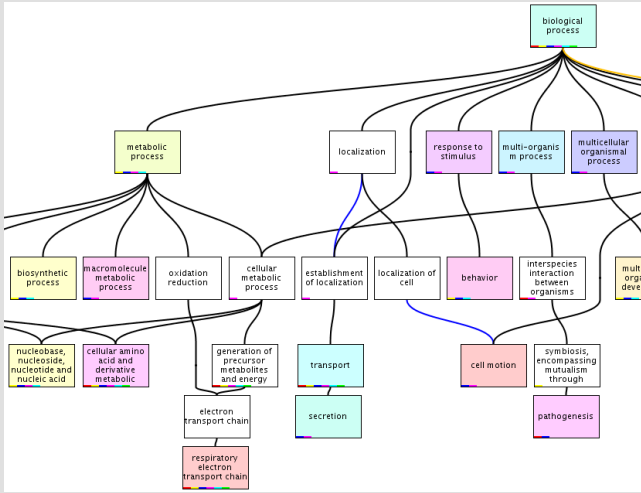
GO slims

Custom slims are available for download;

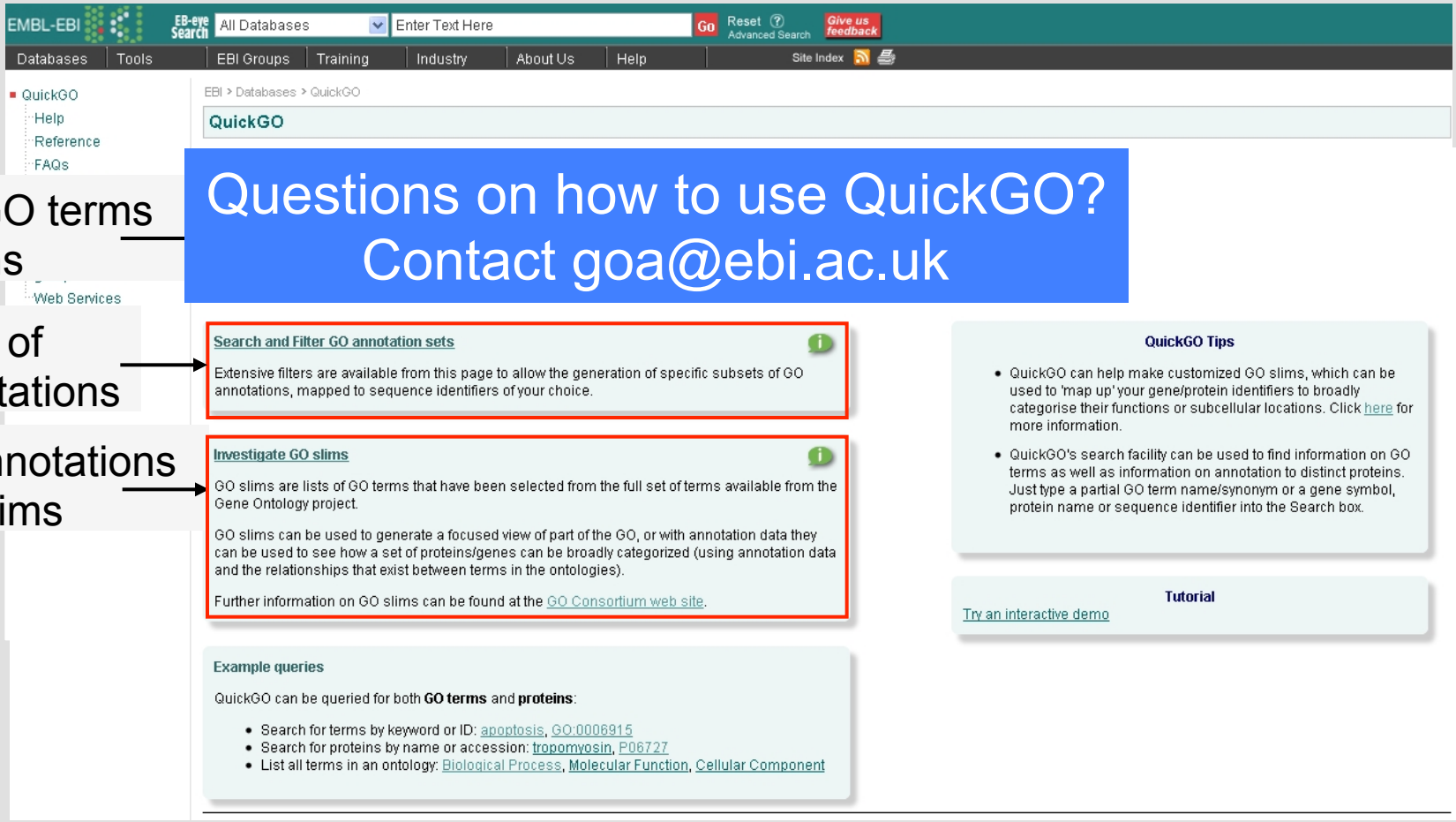
http://www.geneontology.org/GO.slims.shtml

or you can make your own using;

- QuickGO
  http://www.ebi.ac.uk/QuickGO
- AmiGO's GO slimmer
  http://amigo.geneontology.org/cgi-bin/amigo/slimmer



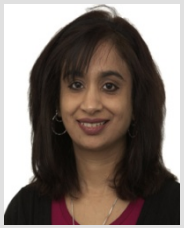
The EBI's QuickGO browser

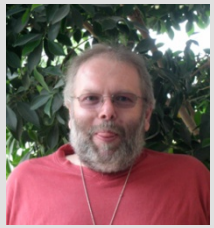


The UniProt-GOA group

Curators:	Emily Dimmer Rachael Huntley Yasmin Alam-Faruque Prudence Mutowo
Software developer:	Tony Sawford
Team leaders:	Rolf Apweiler Claire O'Donovan

Email: *goa@ebi.ac.uk*





Acknowledgements

Members of;

UniProtKB

InterPro

IntAct

HAMAP

Ensembl

Ensembl Genomes

GO Consortium

Funding

National Human Genome Research Institute (NHGRI)
Kidney Research UK
British Heart Foundation
EMBL