How scientists use GO

Prudence Mutowo
UniProt-GOA

Belo Horizonte

September 2014

UniProt

EMBL-EBI

How scientists use the GO

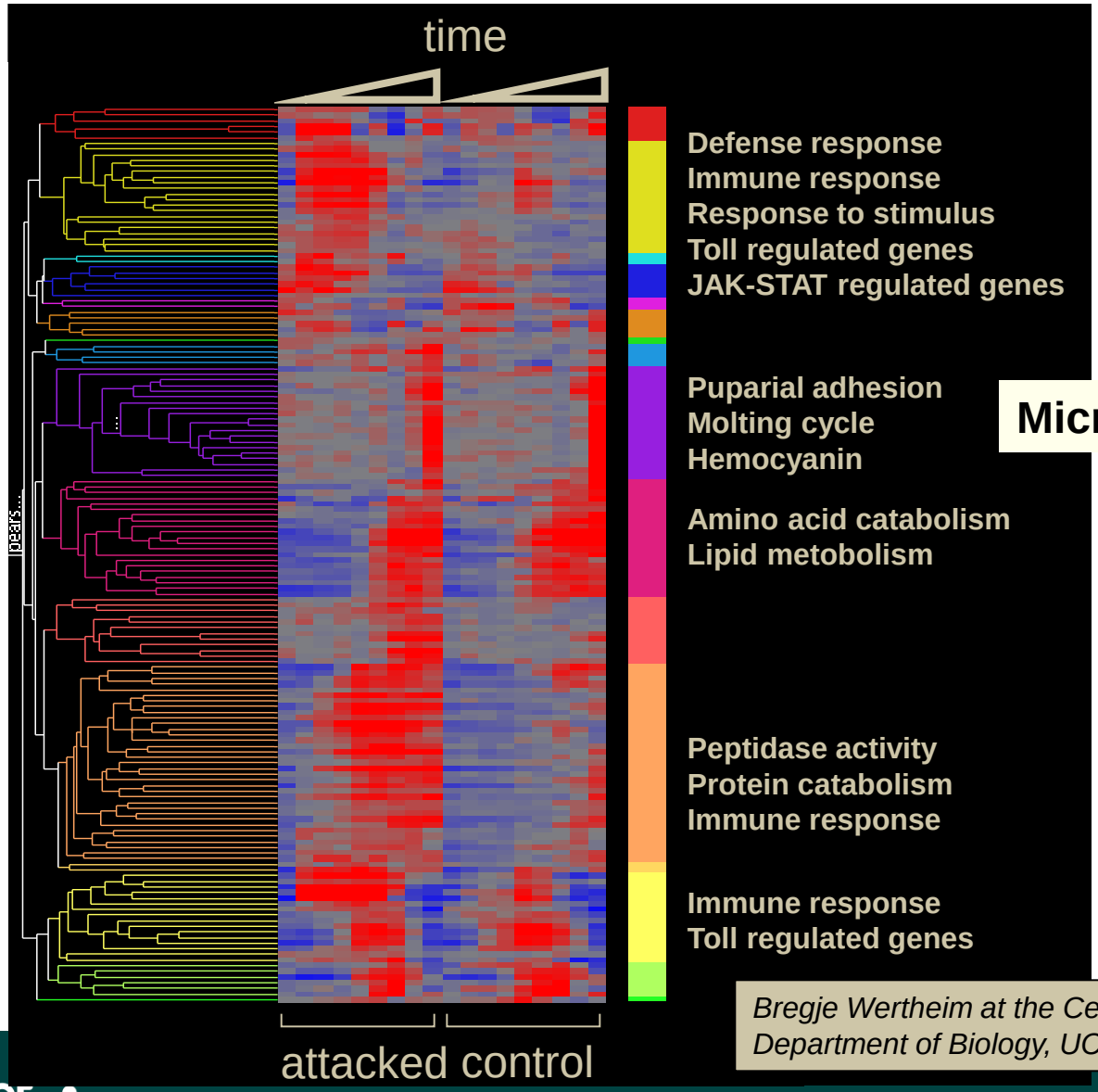
- Access gene product functional information
- Analyse high-throughput genomic or proteomic datasets
- Validation of experimental techniques
- Get a broad overview of a proteome
- Obtain functional information for novel gene products

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

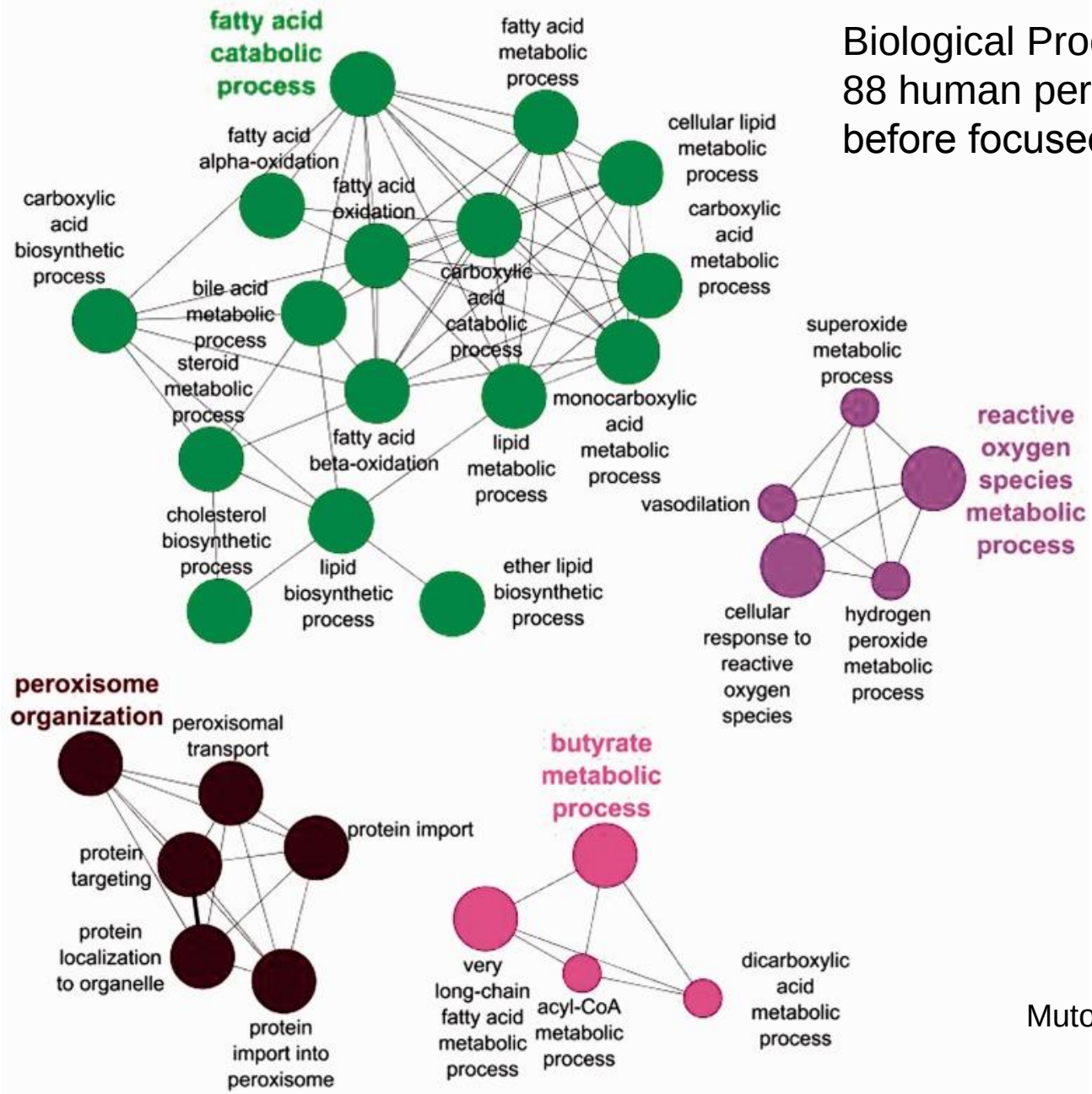
Analysis of high-throughput genomic datasets



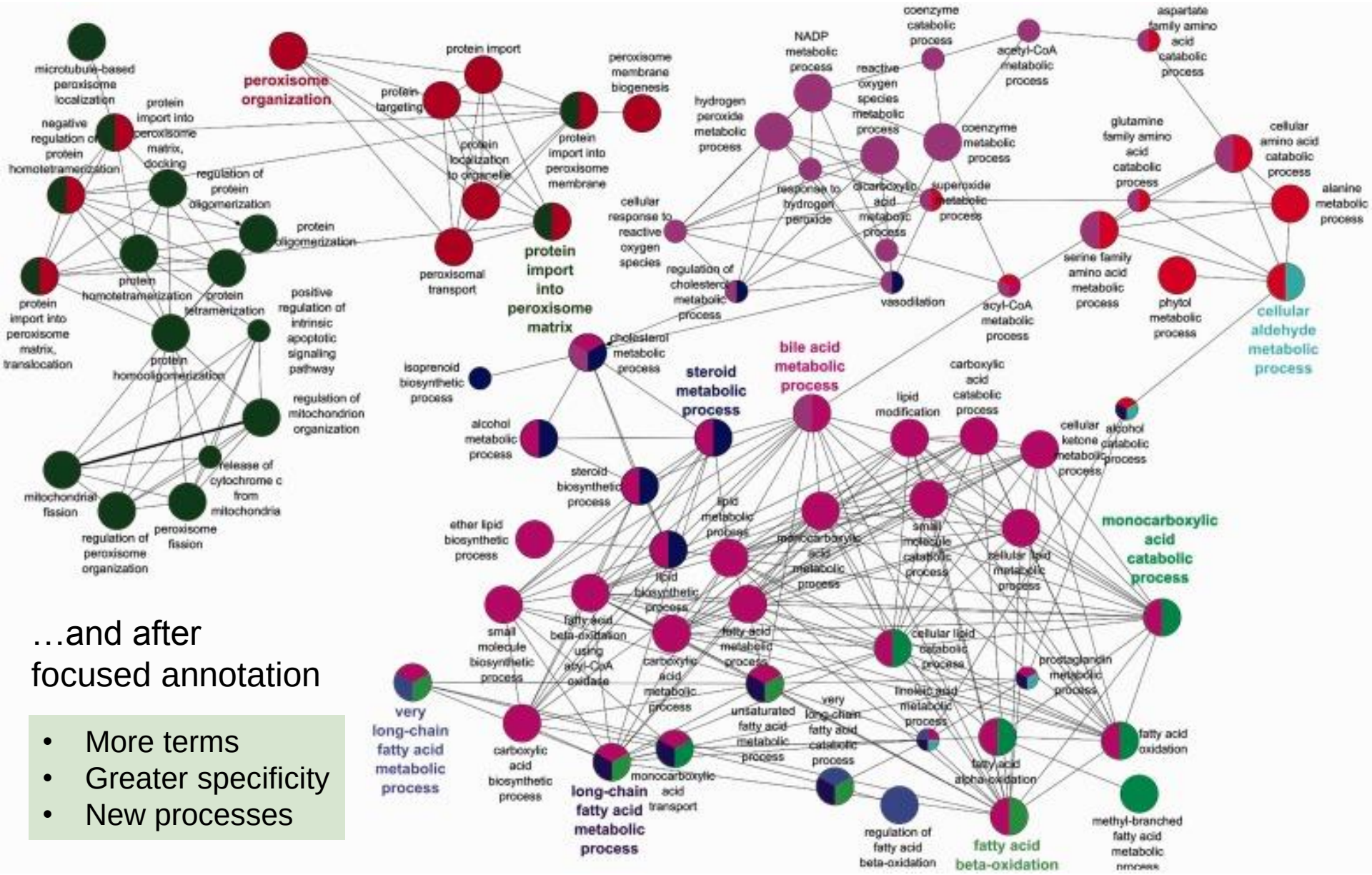
MicroArray data analysis

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

Biological Process GO enrichment of 88 human peroxisome proteins before focused annotation…



Mutowo-Meullenet, Huntley, *et al*. DATABASE 2013

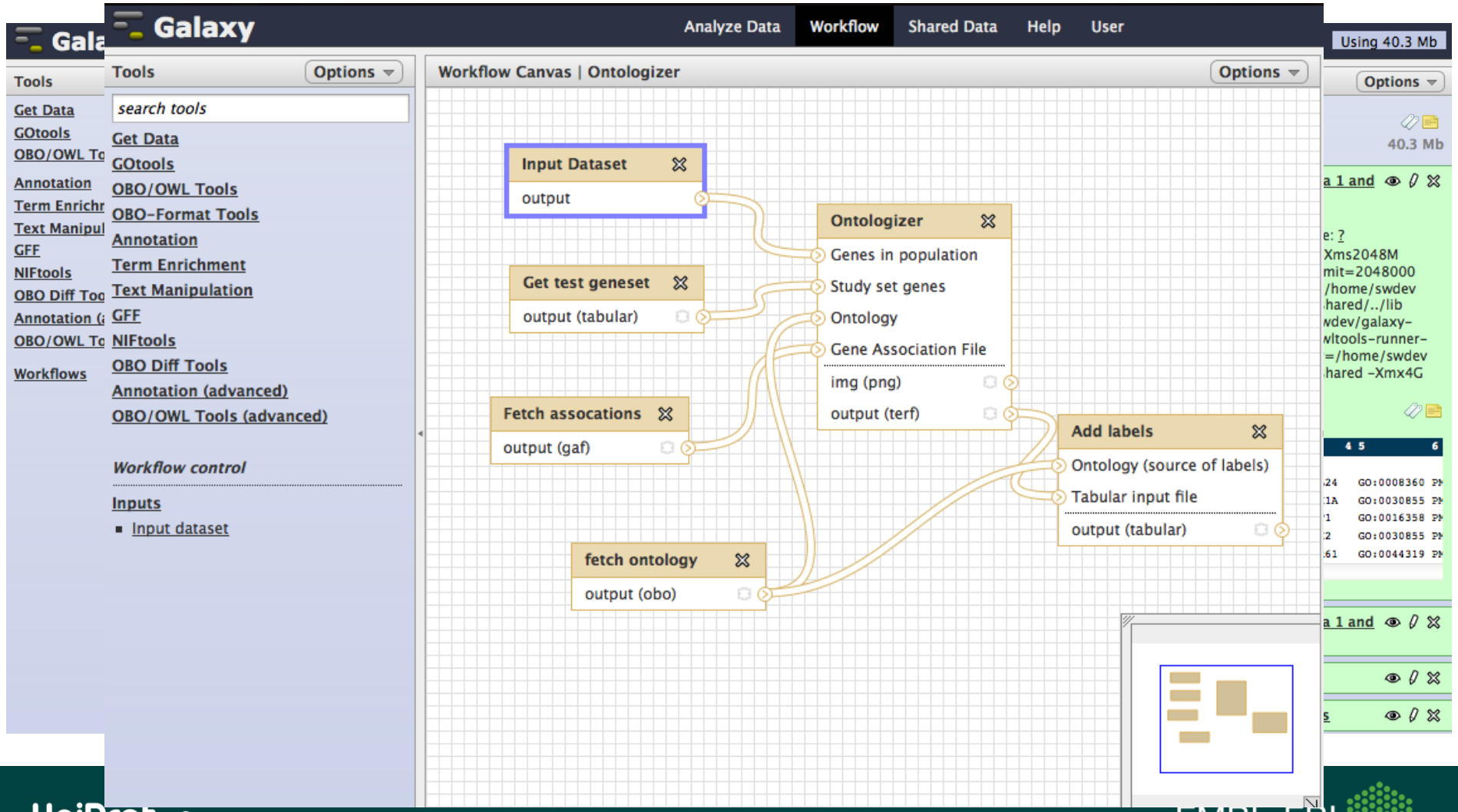


…and after focused annotation

- More terms
- Greater specificity
- New processes

Analysis using GO annotations

GO Galaxy http://galaxy.berkeleybop.org/



Analysis using GO annotations

Many more listed at:

http://neurolex.org/wiki/Category:Resource:Gene_Ontology_Tools

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

Annotating novel sequences

- Can use InterProScan to find GO annotation that is attributed to protein signatures in a submitted protein sequence

EMBL-EBI Services Resear

InterProScan

Input form | Web services | Help & Documentation

Tools > Protein Functional Analysis > InterProScan

InterProScan Sequence Search

This form allows you to scan your sequence for matches against the InterPro collection of protein signature databases.

STEP 1 - Enter your input sequence

Enter or paste a **PROTEIN** sequence in any supported format:

```
MPYVSQSQHIDRVRGAIEGRLPAPGNSSRLVSSWQRSYEQYRLDPGSVIGPRVLTSSELR
DVQGKEEAFLRASGQCLARLHDMIRMADYCVMLTDAHGVTIDYRIDRDRRGDFKHAGLYI
GSCWSEREEGTCGIASVLTDLAPITVHKTDHFRAAFTTLTCSASPIFAPTGELIGVLDAS
AVQSPDNRDSQRLVFQLVRQSAALIEDGYFLNQTAQHWMIFGHASRNFVEAQPEVLIAFD
ECGNIAASNRKAQECIAGLNGPRHVDEIFDTSAVHLHDVARTDTIMPLRLRATGAVLYAR
IRAPLKRVSRSACAVSPSHSGQGTHDAHNDTNLDAISRFLHSRDSRIARNAEVALRIAGK
HLPILILGETGVGKEVFAQALHASGARRAKPFVAVNCGAIPDSLIESELFGYAPGAFTGA
RSRGARGKIAQAHGGTLFLDEIGDMPLNLQTRLLRVLAEGEVLPLGGDAPVRVDIDVICA
```

Or, upload a file: Browse...

STEP 2 - Select the applications to run

Select All Clear All

☑ BlastProDom	☑ FPrintScan	☑ HMMPIR	☑ HMMPfam
☑ HMMTigr	☑ ProfileScan	☑ HAMAP	☑ PatternScan
☑ SignalPHMM	☑ TMHMM	☑ HMMPanther	☑ Gene3D

STEP 3 - Submit your job

Results for job iprscan-I20130326-105632-0639-41160384-oy

Summary Table | Tool Output | Visual Output | Submission Details

IPR002078 RNA polymerase sigma factor 54 interaction domain

Method	Identifier	Description	Matches
PFAM	PF00158	Sigma54_activat	6.699999999999901E-63 [343-507] T
PROFILE	PS50045	SIGMA54_INTERACT_4	0.0 [340-568] T

Parent: IPR003593

Children: No children

Found in: IPR010113 IPR010114 IPR012704 IPR014251 IPR014252 IPR014264 IPR014317 IPR017183

Contains: No entries

GO terms: GO:0005524 ATP binding; GO:0008134 transcription factor binding

IPR002197 DNA binding HTH domain, Fis-type

Method	Identifier	Description	Matches
PRINTS	PR01590	HTHFIS	9.000004079572741E-6 [607-624] T 9.000004079572741E-6 [624-644] T
PFAM	PF02954	HTH_8	9.100000000000008E-12 [605-641] T

Parent: IPR009057

Children: No children

Found in: IPR005412 IPR010113 IPR010114 IPR011785 IPR012704 IPR014264 IPR014317

Contains: No entries

GO terms: GO:0003700 sequence-specific DNA binding transcription factor activity

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

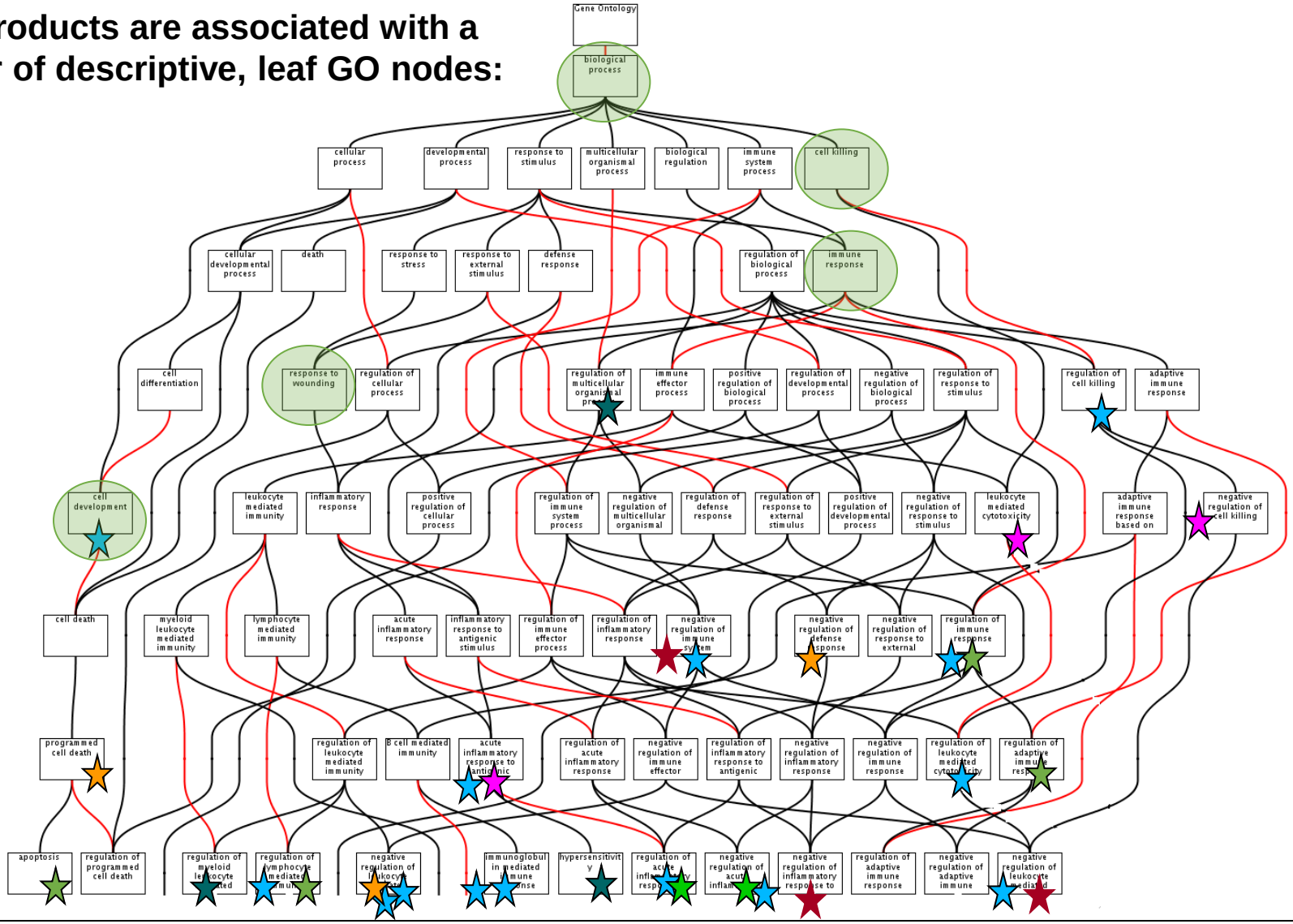
- Many GO analysis tools use GO slims to give a broad overview of the dataset

- 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

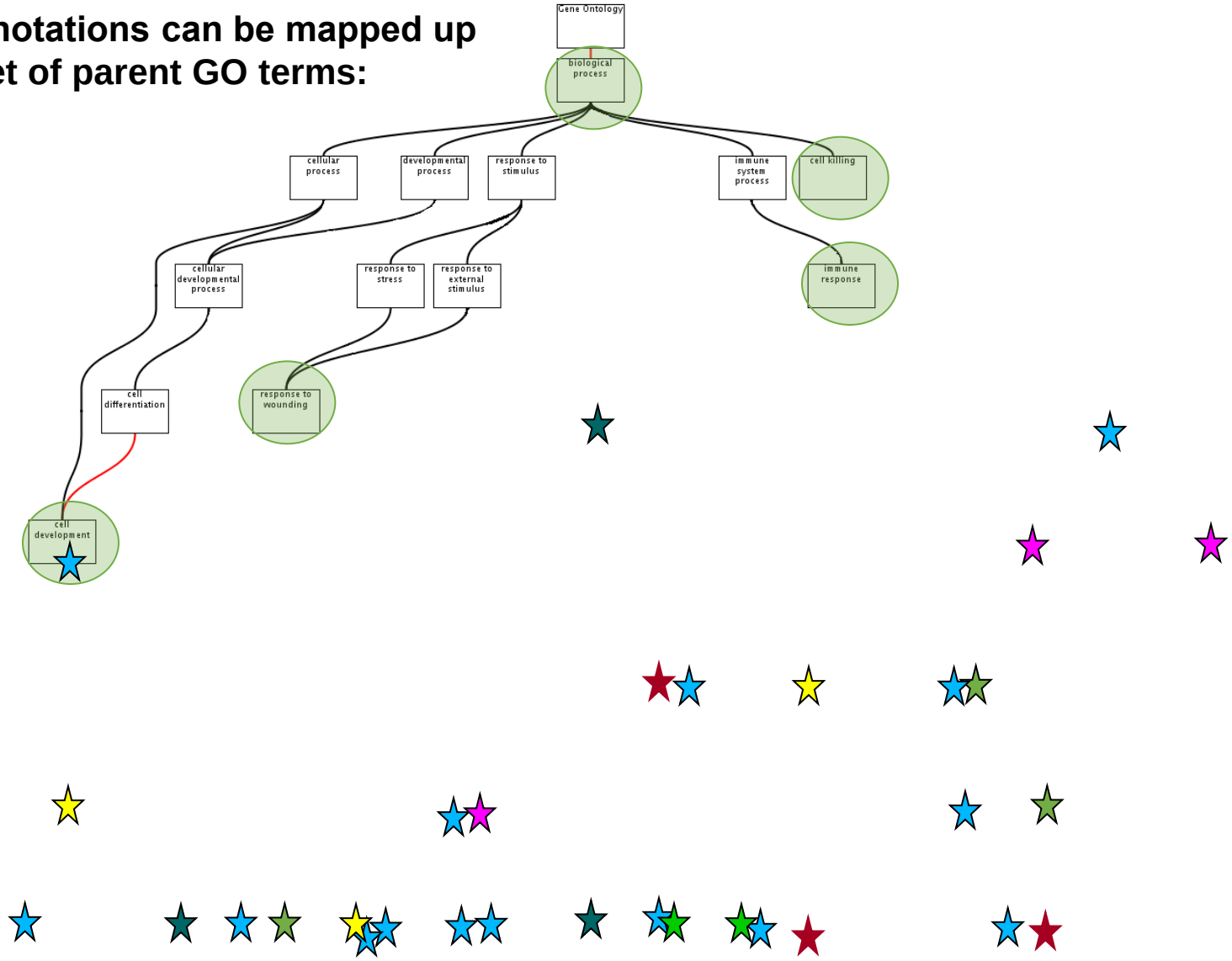
Slimming the GO using the ‘true path rule’

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

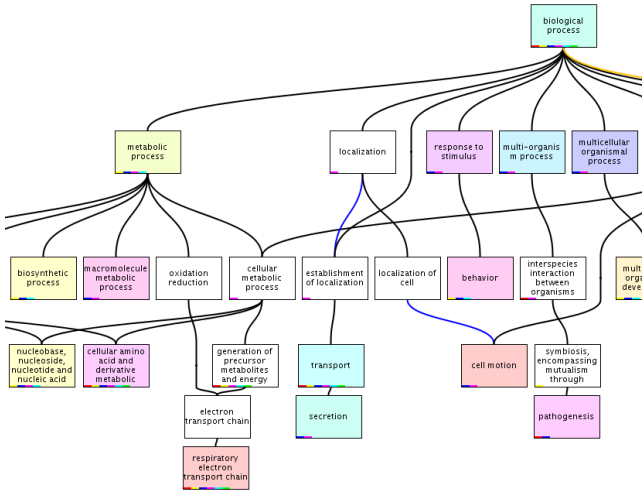
or you can make your own using;

- QuickGO

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

- AmiGO's GO slimmer

  http://amigo.geneontology.org/cgi-bin/amigo/slimmer



The EBI's QuickGO browser

Search GO terms or proteins

Find sets of GO annotations

Map-up annotations with GO slims

EMBL-EBI Services Resear

QuickGO

A fast browser for Gene Ontology terms and annotations.

- QuickGO
  - Help
  - Reference
  - geneontology.org
  - UniProt-GOA project
  - Web Services

EBI > Databases > QuickGO

QuickGO

Click for example search | Search! | Web Services | Dataset | Term Basket: 2

Search and Filter GO annotation sets

Extensive filters are available from this page to allow the generation of specific subsets of GO annotations, mapped to sequence identifiers of your choice.

Investigate GO slims

GO slims are lists of GO terms that have been selected from the full set of terms available from the Gene Ontology project.

GO slims can be used to generate a focused view of part of the GO, or with annotation data they can be used to see how a set of proteins/genes can be broadly categorized (using annotation data and the relationships that exist between terms in the ontologies).

Further information on GO slims can be found at the GO Consortium web site.

View the history of changes to GO

This page allows you to view the changes to GO, optionally filtered by date, term identifier, or type of change.

Exercise

Using GO slims in QuickGO Exercise 1 (pg.27)

Find protein list at:

ftp://ftp.ebi.ac.uk/pub/contrib/goa/Tutorial_Data

Precautions when using GO annotations for analysis

- The Gene Ontology is always changing and GO annotations are continually being created
  - always use a current version of both
  - if publishing your analyses please report the versions/dates you used

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

- Recommended that ‘NOT’ annotations are removed before analysis
  - only ~7000 out of 141 million annotations are ‘NOT’
  - can confuse the analysis

Precautions when using GO annotations for analysis

- Unannotated is **not** unknown

  - where there is no evidence in the literature for a process, function or location the gene product is annotated to the appropriate ontology's root node with an 'ND' evidence code (no biological data), thereby distinguishing between unannotated and unknown

- Pay attention to under-represented GO terms

  - a strong under-representation of a pathway may mean that normal functioning of that pathway is necessary for the given condition