Gene function annotation and gene set analysis

Paul D. Thomas, Ph.D.
University of Southern California
PI, Gene Ontology Consortium

October 25, 2023

1

Outline

- **Goal: using gene function in bioinformatics**
  - Understanding GO and GO annotations so you can use them effectively
- **Gene Ontology:** a computational representation of gene function
  - **Exploring GO**
- **GO annotations**: evidence-based statements about functions of specific genes
- **GO enrichment analysis**
  - Methods and practical considerations

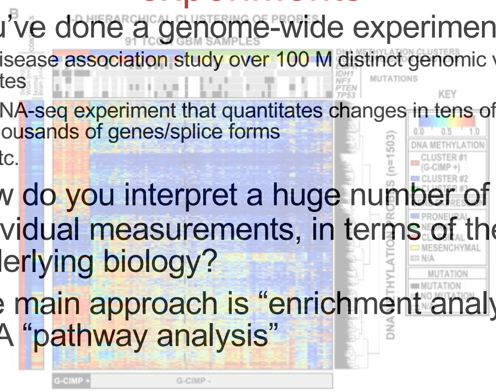
2

Why the Gene Ontology?

- Problem: biology is extremely complex
  - 20,000 human genes, and large numbers in all cellular organisms
  - Millions of publications on gene functions and growing
  - No one person can know it all
- Solution: Encode biological knowledge onto a computer so it can be accessed and used in computational analyses

3

Interpreting large-scale "omics" experiments



- You've done a genome-wide experiment
  - Disease association study over 100 M distinct genomic variant sites
  - RNA-seq experiment that quantitates changes in tens of thousands of genes/splice forms
  - Etc.
- How do you interpret a huge number of individual measurements, in terms of the underlying biology?
- The main approach is "enrichment analysis", AKA "pathway analysis"

4

University of Southern California USC

Enrichment analysis using GO

- Uses ***known information*** about gene function to see if there are any statistical trends in the kinds of FUNCTIONAL CHARACTERISTICS of the genes that are changed in the experiment
- The **Gene Ontology knowledgebase** is the most comprehensive resource on the functions of genes, in a form that can be used in computational analysis

5

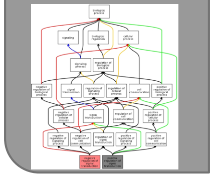
University of Southern California USC

Gene Ontology overview

Ontology

"Universe" of possible function characteristics, and relationships between them:

- Terms
- Relations
- Definitions



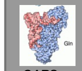
+

Annotations

Statements about the functions of specific **gene products.**

3 aspects:

- Molecular function
- Biological process
- Cellular component



- Glutamine-tRNA ligase activity
- Translation
- Cytoplasm

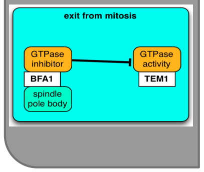
QARS
Gln tRNA synthetase

=

Model of biology

Representation of current knowledge in a manner that is:

- Human understandable
- Machine computable



>40,000 terms
>80,000 relations

~ 750k annotations from expts
in ~175k publications

~7.5M annotations total at GO

13

University of Southern California USC

What is an ontology?

- Ontology: "study of being" originally from Greek philosophy
  - Concepts related to existence
    - "continuants" (material things that persist)
    - "occurrents" (things that happen, develop over time)
  - Categories of concepts (or "terms")
  - Relationships between concepts

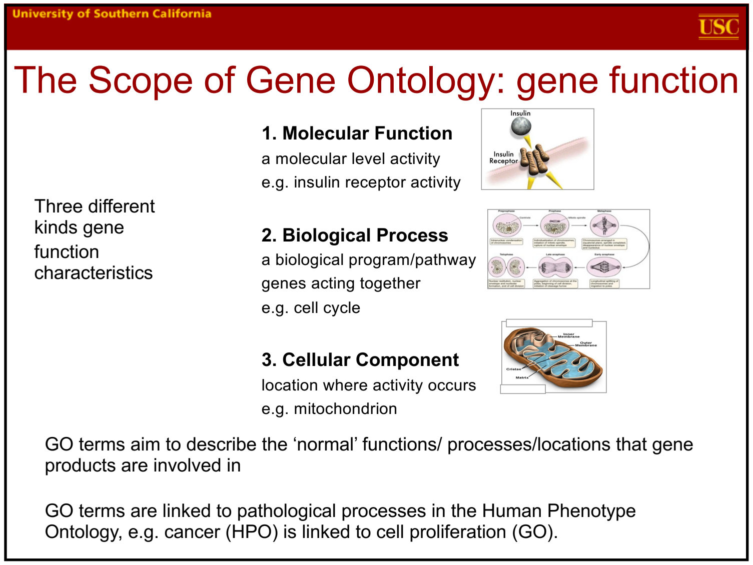
14

University of Southern California USC

Modern definition of ontology

- field of computer science (data science)
- **computational knowledge representation**
- "a formal specification of a shared conceptualization" (Borst, 1997)
  - a **shared conceptualization** is the way we conceive or "model" a particular domain of knowledge
  - a **formal specification** is a formal way of representing (writing out) this model

16



University of Southern California

USC

The Scope of Gene Ontology: gene function

Three different kinds gene function characteristics

1. Molecular Function
a molecular level activity
e.g. insulin receptor activity

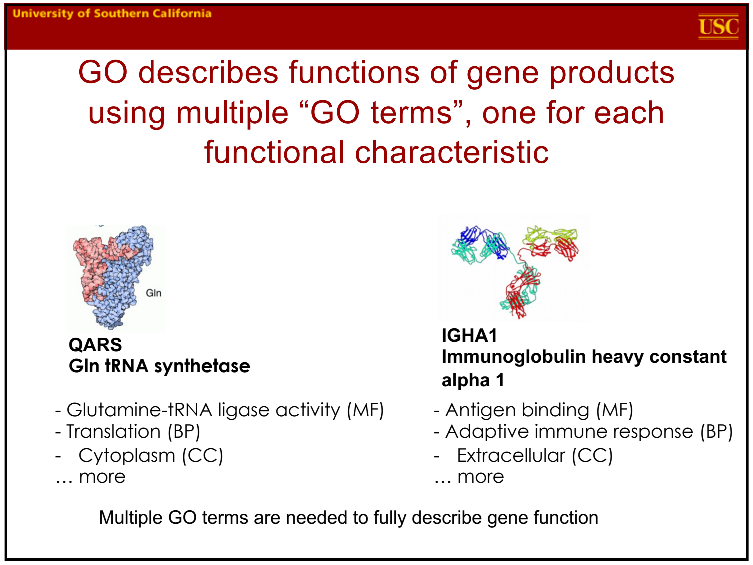
2. Biological Process
a biological program/pathway
genes acting together
e.g. cell cycle

3. Cellular Component
location where activity occurs
e.g. mitochondrion

GO terms aim to describe the ‘normal’ functions/ processes/locations that gene products are involved in

GO terms are linked to pathological processes in the Human Phenotype Ontology, e.g. cancer (HPO) is linked to cell proliferation (GO).

24



University of Southern California

USC

GO describes functions of gene products using multiple “GO terms”, one for each functional characteristic

QARS
Gln tRNA synthetase

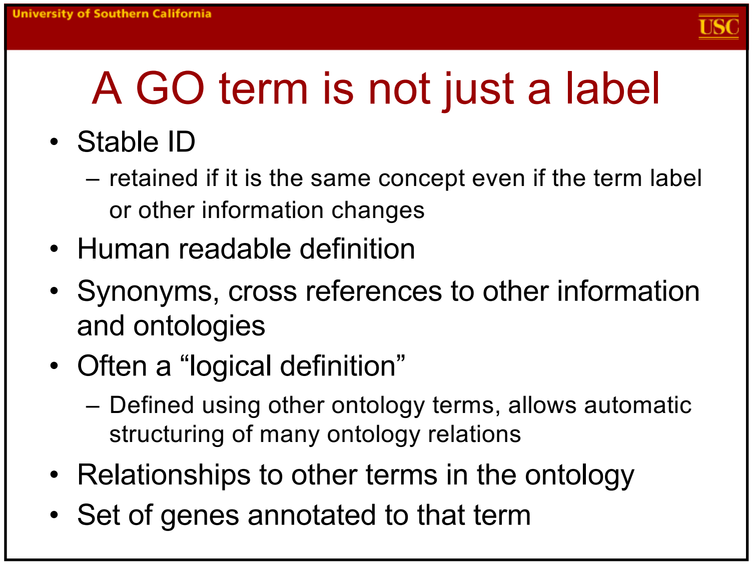
- Glutamine-tRNA ligase activity (MF)
- Translation (BP)
- Cytoplasm (CC)
… more

IGHA1
Immunoglobulin heavy constant alpha 1

- Antigen binding (MF)
- Adaptive immune response (BP)
- Extracellular (CC)
… more

Multiple GO terms are needed to fully describe gene function

25



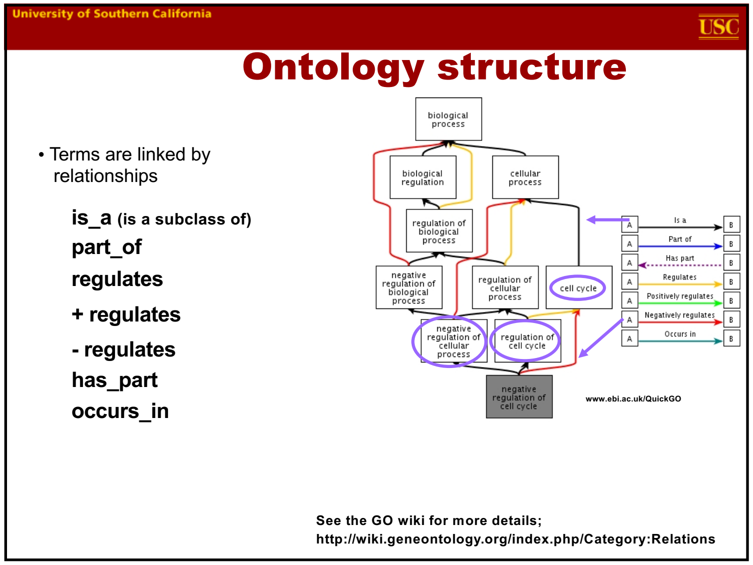
University of Southern California

USC

A GO term is not just a label

- Stable ID
  - retained if it is the same concept even if the term label or other information changes
- Human readable definition
- Synonyms, cross references to other information and ontologies
- Often a “logical definition”
  - Defined using other ontology terms, allows automatic structuring of many ontology relations
- Relationships to other terms in the ontology
- Set of genes annotated to that term

26



University of Southern California

USC

Ontology structure

- Terms are linked by relationships

is_a (is a subclass of)
part_of
regulates
+ regulates
- regulates
has_part
occurs_in

biological process
biological regulation
cellular process
regulation of biological process
negative regulation of biological process
regulation of cellular process
cell cycle
negative regulation of cellular process
regulation of cell cycle
negative regulation of cell cycle
A
Is a
B
Part of
Has part
Regulates
Positively regulates
Negatively regulates
Occurs in

www.ebi.ac.uk/QuickGO

See the GO wiki for more details;
http://wiki.geneontology.org/index.php/Category:Relations

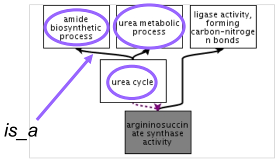
27

Ontology relations

is_a

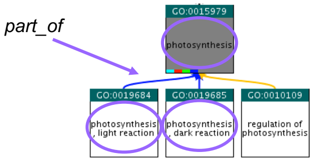
‘urea cycle’ *is_a* type of ‘urea metabolic process’ and *is_a* type of ‘amide biosynthetic process’

Multiple “parents” are allowed: “directed acyclic graph” rather than hierarchy


is_a
amide biosynthetic process
urea metabolic process
ligase activity, forming carbon-nitrogen bonds
urea cycle
argininosuccinate synthase activity

part_of

Photosynthetic dark and light reactions are part_of ‘photosynthesis’


part_of
GO:0015979
photosynthesis
GO:0019684
photosynthesis, light reaction
GO:0019685
photosynthesis, dark reaction
GO:0010109
regulation of photosynthesis

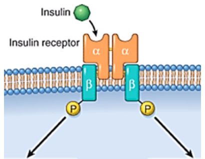
Used for grouping of genes

Hands-on exercise

- Browse the Gene Ontology
  - Go to geneontology.org
  - Click on “browse the ontology”
  - Select a term, read definition, relations
- Explore the Ontology Lookup Service
  - Go to https://www.ebi.ac.uk/ols/
  - Browse the list of available ontologies
  - Browse the Gene Ontology: how is it different here?

GO “annotations”

- An annotation is a statement linking a gene to **one characteristic** of its function (a GO ontology term)



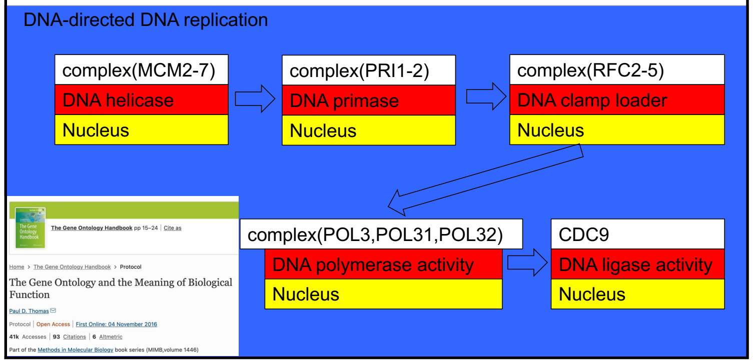
Examples:
Annotation 1: INSR + ‘receptor activity’
Annotation 2: INSR + ‘plasma membrane’
Annotation 3: INSR + ‘insulin receptor signaling pathway’

- Each annotation must be based on *evidence*, which is recorded as part of the annotation
  - **Evidence code** (type of evidence)
  - **Reference** (published journal article)
- Important note: distinct annotations for a given gene may therefore use identical or related GO terms and do not necessarily represent independent characteristics

GO annotations specify a model of biological systems

Tissue regeneration

Cell cycle



The Gene Ontology Handbook pp 15–24 | Cite as

Home › The Gene Ontology Handbook › Protocol

The Gene Ontology and the Meaning of Biological Function

Paul D. Thomas

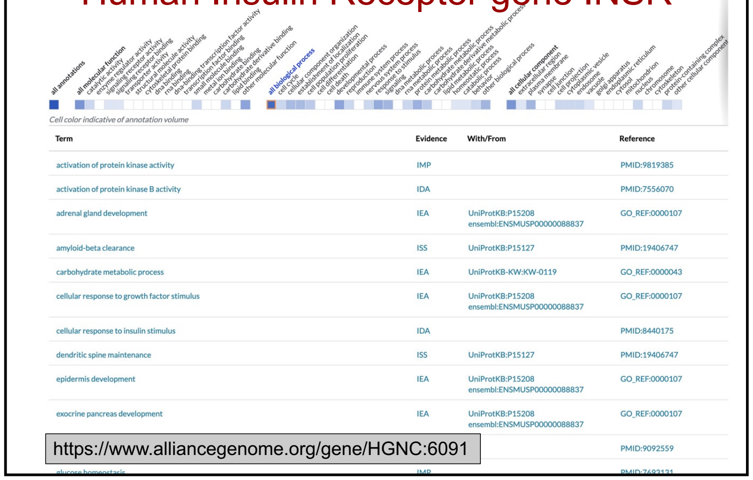
Protocol | Open Access | First Online: 04 November 2016

41k Accesses | **93** Citations | **6** Altmetric

Part of the Methods in Molecular Biology book series (MIMB,volume 1446)

University of Southern California USC

Human Insulin Receptor gene INSR



Cell color indicative of annotation volume

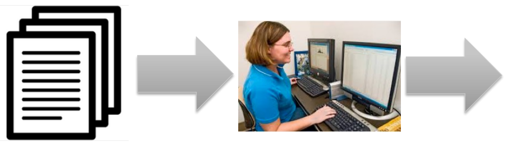
Term	Evidence	With/From	Reference
activation of protein kinase activity	IMP		PMID:9819385
activation of protein kinase B activity	IDA		PMID:7556070
adrenal gland development	IEA	UniProtKB:P15208 ensembl:ENSMUSP00000088837	GO_REF:0000107
amyloid-beta clearance	ISS	UniProtKB:P15127	PMID:19406747
carbohydrate metabolic process	IEA	UniProtKB-KW:KW-0119	GO_REF:0000043
cellular response to growth factor stimulus	IEA	UniProtKB:P15208 ensembl:ENSMUSP00000088837	GO_REF:0000107
cellular response to insulin stimulus	IDA		PMID:8440175
dendritic spine maintenance	ISS	UniProtKB:P15127	PMID:19406747
epidermis development	IEA	UniProtKB:P15208 ensembl:ENSMUSP00000088837	GO_REF:0000107
exocrine pancreas development	IEA	UniProtKB:P15208 ensembl:ENSMUSP00000088837	GO_REF:0000107
			PMID:9092559
glucose homeostasis	IMP		PMID:7693131

https://www.alliancegenome.org/gene/HGNC:6091

32

University of Southern California USC

“Known” functions of genes: Where did this information come from?



MMP2 involved_in collagen catabolic process
ADAMTS2 involved_in collagen catabolic process
ADAMTS3 involved_in collagen catabolic process
…

Published papers

Biocuration

“primary GO annotations”
= gene function characteristics based on direct experimental evidence

33

University of Southern California USC

All GO annotations link to the evidence for that statement about gene function

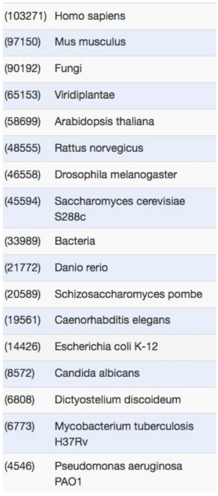
- **Literature evidence (primary & secondary)**
  - Reference provides
    - the experiment demonstrating the function (primary)
    - or the paper with the author assertion (secondary)
- **Homology evidence**
  - Inference from *experimental evidence* for a homologous gene
  - Reference is publication describing inference process

34

University of Southern California USC

Homology inference

- Our knowledge of human genes is limited
  - Only ~25,000 of 150,000 papers used in GO annotations are on human genes
- The GO uses *homology inference* to augment human gene annotations



(103271)	Homo sapiens
(97150)	Mus musculus
(90192)	Fungi
(65153)	Viridiplantae
(58699)	Arabidopsis thaliana
(48555)	Rattus norvegicus
(46558)	Drosophila melanogaster
(45594)	Saccharomyces cerevisiae S288c
(33989)	Bacteria
(21772)	Danio rerio
(20589)	Schizosaccharomyces pombe
(19561)	Caenorhabditis elegans
(14426)	Escherichia coli K-12
(8572)	Candida albicans
(6808)	Dictyostelium discoideum
(6773)	Mycobacterium tuberculosis H37Rv
(4546)	Pseudomonas aeruginosa PAO1

35

Homology-based annotations

- Recommend using annotations based on curated GO assignments
  - Pairwise, individually reviewed: ISS evidence code
  - Phylogeny, individually reviewed: IBA evidence code
  - Family, based on family-level curation and computational assignment to family: InterPro2GO
  - Phylogeny, based on tree curation and computational assignment to a tree branch: PANTHER2GO (TreeGrafter)

36

Two main methods used to annotate by homology inference

- Family-based
  - InterPro2GO, evidence code IEA, reference GO_REF:0000042
- Phylogenetic curation-based
  - PAN-GO, IBA evidence code
  - Extrapolated to proteins that are not in the tree using PANTHER/TreeGrafter, evidence code IEA, reference GO_REF:0000118

38

GO annotation of protein families

- Find functions that are broadly conserved among family members
- Annotate entire family with the corresponding GO terms

Manual GO annotation of predictive protein signatures: the InterPro approach to GO curation

Sarah Burge, Elizabeth Kelly, David Lonsdale, Prudence Mutowo-Muellenet, Craig McAnulla, Alex Mitchell, Amaia Sangrador-Vegas, Siew-Yit Yong, Nicola Mulder, Sarah Hunter

Database, Volume 2012, 1 January 2012, bar068, https://doi.org/10.1093/database/bar068

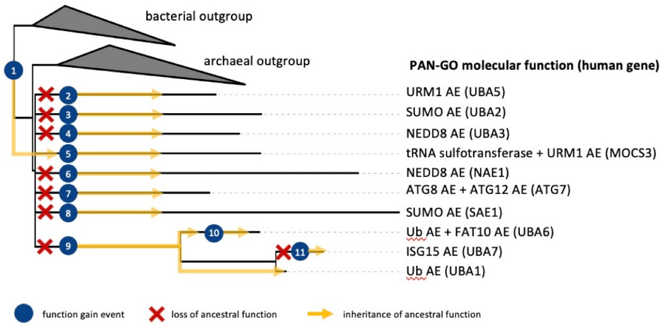
39

InterPro2GO is accurate, but often non-specific

- Inherent limitation of the approach is that the GO terms must apply to all sequences in a family, or with a protein domain
- Many protein families are large and diverse, and have diverse functions

40

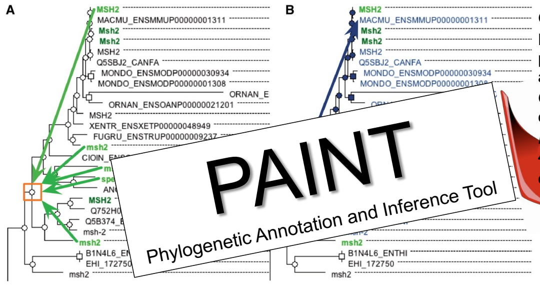
UBL Activating Enzyme Family



41

Phylogenetic Annotation using GO (PAN-GO)

- Manually review experimental GO annotations for related genes, in a gene family tree from PANTHER
- Build a model of branches in the gene tree where specific functions were gained and lost, that explains the distribution of functions found in modern-day genes

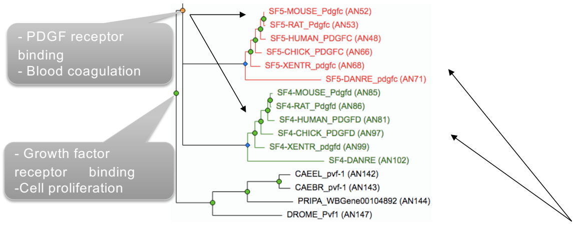


Gaudet, P., et al. (2011). Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Briefings in Bioinformatics*, 12(5), 449–62. doi:10.1093/bib/bbr042

42

Tree with ancestral nodes annotated with GO terms

Annotations are inherited by descendant sequences in the tree; these are individually reviewed and labeled with IBA evidence code

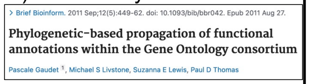


Every node GO annotation must be based on an experimentally supported annotation on a descendant gene: TRACEABLE EVIDENCE

43

Phylogeny-based annotations

- Approach
  - manually curate model of function evolution for each protein family tree using GO terms, based on integrating experimental GO annotations (MOD+human/UniProt) for all family members
- Advantages
  - GO annotations often highly specific
  - All annotations are traceable to experimental evidence



> Brief Bioinform. 2011 Sep;12(5):449-62. doi: 10.1093/bib/bbr042. Epub 2011 Aug 27.

Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium

Pascale Gaudet [1], Michael S Livstone, Suzanna E Lewis, Paul D Thomas

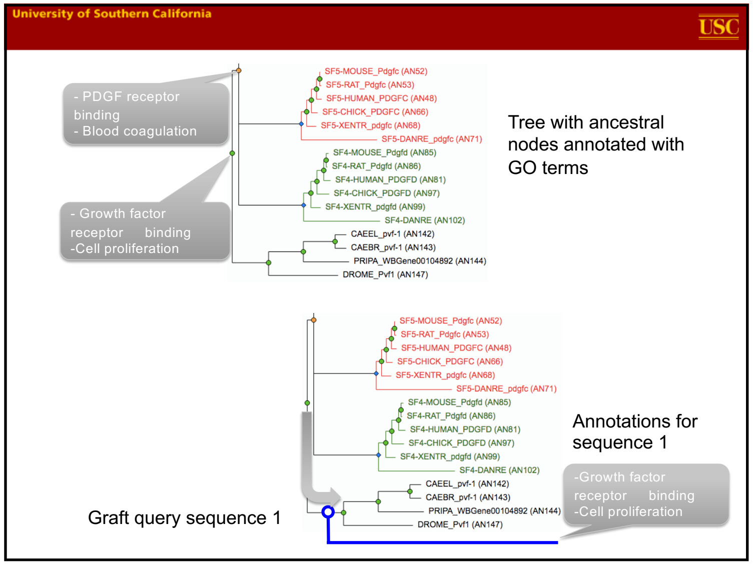
- Additional "query" protein sequences are "grafted" onto tree, to get GO annotations

> Bioinformatics. 2019 Feb 1;35(3):518-520. doi: 10.1093/bioinformatics/bty625.

TreeGrafter: phylogenetic tree-based annotation of proteins with Gene Ontology terms and other annotations

Haiming Tang [1], Robert D Finn [2], Paul D Thomas [1]

44


University of Southern California

USC

- PDGF receptor binding
- Blood coagulation

- Growth factor receptor binding
-Cell proliferation

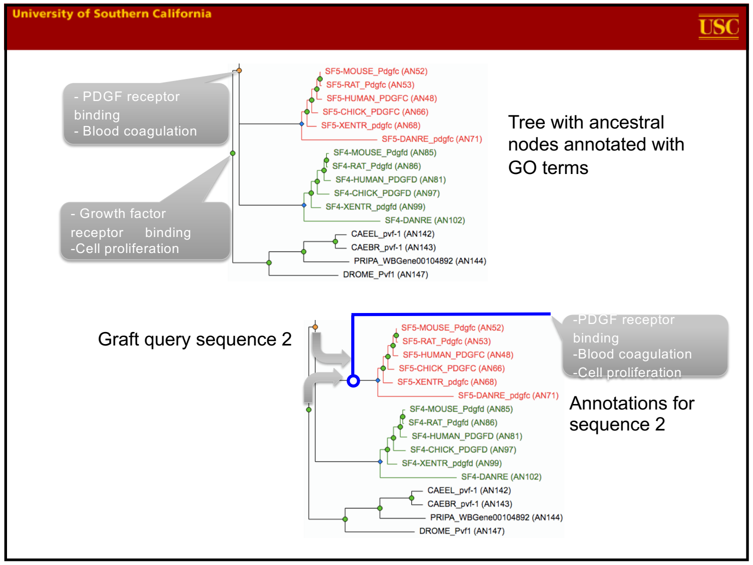
Tree with ancestral nodes annotated with GO terms

Annotations for sequence 1

-Growth factor receptor binding
-Cell proliferation

Graft query sequence 1

45


University of Southern California

USC

- PDGF receptor binding
- Blood coagulation

- Growth factor receptor binding
-Cell proliferation

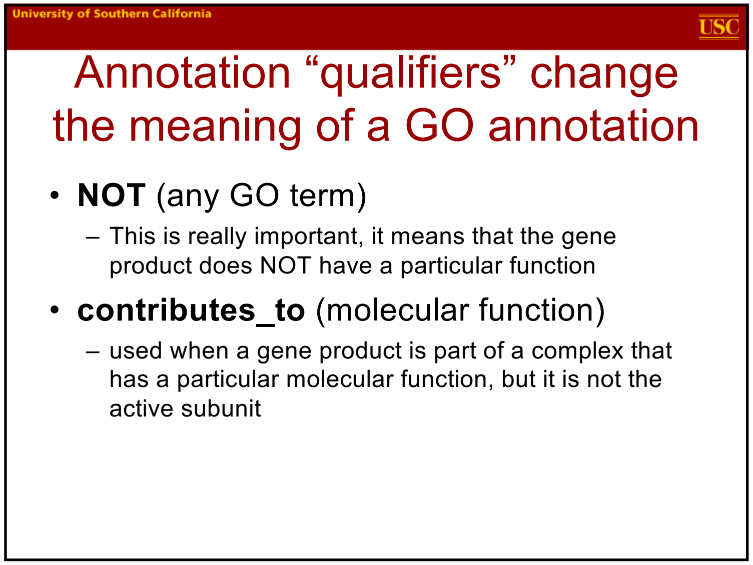
Tree with ancestral nodes annotated with GO terms

Graft query sequence 2

-PDGF receptor binding
-Blood coagulation
-Cell proliferation

Annotations for sequence 2

46

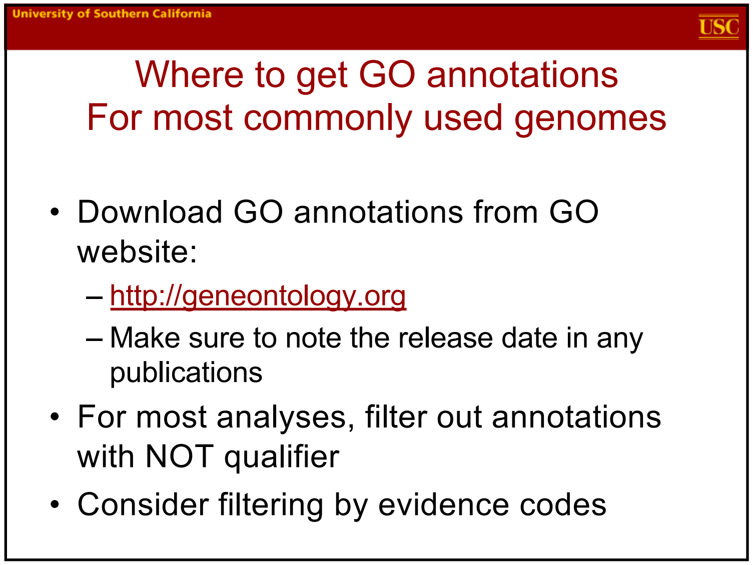

University of Southern California

USC

Annotation “qualifiers” change the meaning of a GO annotation

- **NOT** (any GO term)
  - This is really important, it means that the gene product does NOT have a particular function
- **contributes_to** (molecular function)
  - used when a gene product is part of a complex that has a particular molecular function, but it is not the active subunit

47


University of Southern California

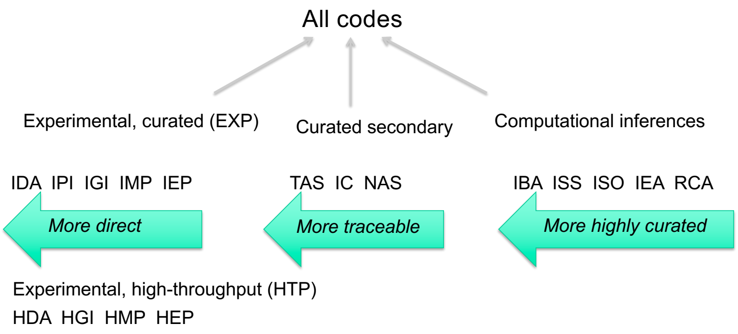
USC

Where to get GO annotations For most commonly used genomes

- Download GO annotations from GO website:
  - http://geneontology.org
  - Make sure to note the release date in any publications
- For most analyses, filter out annotations with NOT qualifier
- Consider filtering by evidence codes

48

GO evidence codes



http://geneontology.org/docs/guide-go-evidence-codes/

General advice for evidence codes

- Filter out less reliable experimental annotations
  - high-throughput evidence codes (HTP*)
  - Large-scale computational predictions (RCA)
  - Expression pattern evidence (IEP)
- Filter to keep only curator-reviewed homology-based annotations
  - ISS, IBA evidence codes
  - If few available for your organism, use IEA with GOREF_0000042, GOREF_0000118

*and more specific HTP codes: HDA, HGI, HMP, HEP

Where to get GO annotations For unannotated genomes

- Use InterProScan
  - https://www.ebi.ac.uk/interpro/download.html
  - Take both sources of annotations
    - Family-based: InterPro2GO
    - Phylogeny-based: PANTHER2GO
- Other computational pipelines for GO annotation are not recommended, as they do not include any manual review

GO enrichment analysis

- Introduction to enrichment analysis
- Annotation sets
- Types of statistical tests
- Overrepresentation analysis using GO/PANTHER
  - Overrepresentation test
  - Enrichment test
  - Both are available on the web and via the PANTHER API

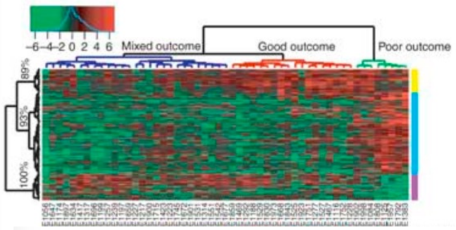
Enrichment analysis

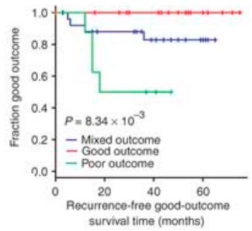
- Uses ***known information*** about gene function
  - are any statistical trends in the kinds of FUNCTIONAL CHARACTERISTICS of the genes that are changed in the experiment?
- For example: genes in the same GO biological process (“module” or “pathway”) tend to be coordinately regulated, or have similar biological effects when perturbed

53

Tip: Use the most up-to-date version of the ontology and annotations

- Analysis using GO annotations in 2008, vs. 2017





Nature Medicine **14**, 518 - 527 (2008)
Published online: 27 April 2008 | doi:10.1038/nm1764

Stromal gene expression predicts clinical outcome in breast cancer

George Guo

54

GO knowledgebase changes over time as we accumulate knowledge

2008

- Good outcome cluster:
  - Upregulation of T-cell mediated immunity processes
- Poor outcome cluster:
  - Several enriched GO terms but no consistency
- Mixed outcome cluster:
  - Several enriched GO terms but no consistency

2017

- Good outcome cluster:
  - Upregulation of T-cell mediated immunity processes
- Poor outcome cluster
  - Upregulation of cell proliferation (rapid growth) and cell motility (metastasis) processes
- Mixed outcome cluster
  - No significant enrichment

55

Common enrichment analysis variations

- Different statistical tests
  - Require different data
- Different “annotation sets”
  - Appropriate sets depend on biological question, but most “omics” data analysis looks for correlated changes across groups of genes that may function together: pathways and GO biological processes
- How do they compare?
  - If there are differences, don’t just choose the one that you’d prefer to be true, examine the results to understand them

56

Two main types of test

- “Overrepresentation”
  - In my list of genes, are any functional classes found more often than expected, compared to a reference list?
- “Enrichment” (e.g. GSEA)
  - No separate reference list. For every gene in a large-scale experiment, a value is measured and computed.
  - Do the genes in a particular functional class have a distribution of values that is different from the expected distribution?

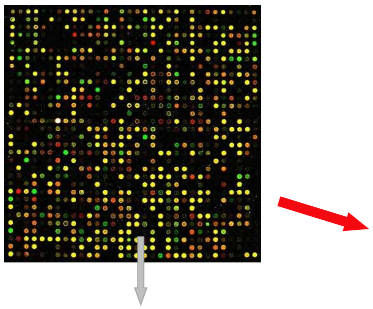
57

Overrepresentation test

- Input
  - A list of genes of interest
  - Optional but recommended: a “reference” list of genes from which the first list was chosen from
    - E.g. all genes with measurable expression in the experiment
- Output
  - Enrichment/depletion: which classes (e.g. pathways) show more (fewer) genes in the list than expected by chance
  - P-value: the probability that the observed enrichment/depletion is significantly different from the null hypothesis of NO ENRICHMENT/DEPLETION

58

Overrepresentation test



Your gene list of interest

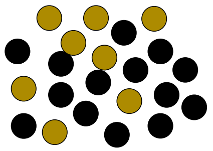
Reference gene list (all the genes you measured)

Need to define:
Gene list(s) of interest
Reference gene list

59

Overrepresentation Test

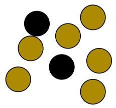
Reference gene list



Genes annotated with a given GO term
Genes not annotated with a given GO term

Is the given annotation class over- or under- represented compared to the reference?

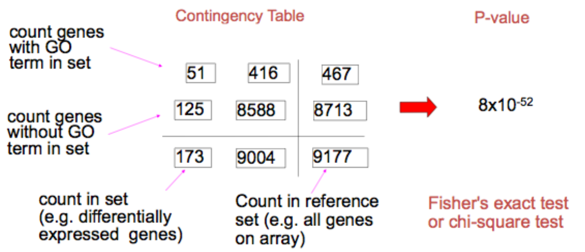
Your gene list of interest



60

University of Southern California

Over (under) representation test example

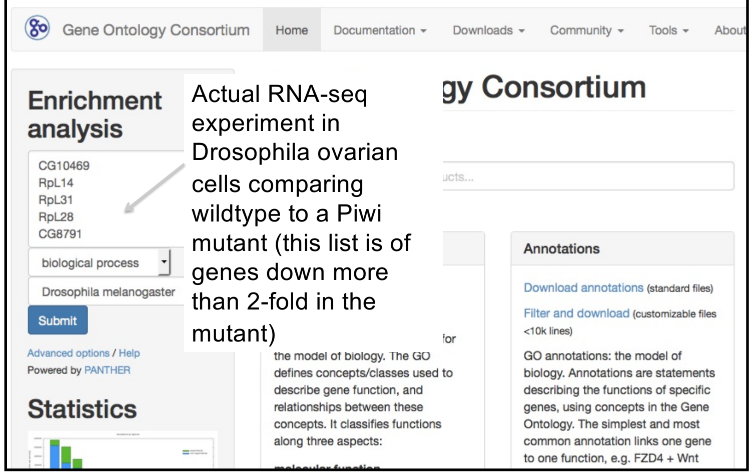

Contingency Table
P-value
count genes with GO term in set
count genes without GO term in set
count in set (e.g. differentially expressed genes)
Count in reference set (e.g. all genes on array)
Fisher's exact test or chi-square test

51	416	467
125	8588	8713
173	9004	9177

8x10^{-52}

62

University of Southern California



Enrichment analysis

CG10469
RpL14
RpL31
RpL28
CG8791

Actual RNA-seq experiment in Drosophila ovarian cells comparing wildtype to a Piwi mutant (this list is of genes down more than 2-fold in the mutant)

for the model of biology. The GO defines concepts/classes used to describe gene function, and relationships between these concepts. It classifies functions along three aspects:

GO annotations: the model of biology. Annotations are statements describing the functions of specific genes, using concepts in the Gene Ontology. The simplest and most common annotation links one gene to one function, e.g. FZD4 + Wnt

63

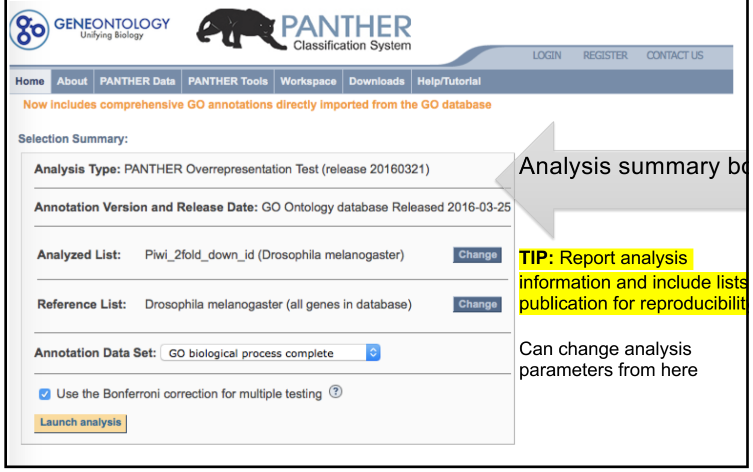
University of Southern California

Exercise 2: GO enrichment analysis

- Download the files at http://data.pantherdb.org/ftp/tools/samples/
- They are from the publication https://www.ncbi.nlm.nih.gov/pubmed/26780607

64

University of Southern California



Now includes comprehensive GO annotations directly imported from the GO database

Selection Summary:

Analysis Type: PANTHER Overrepresentation Test (release 20160321)

Annotation Version and Release Date: GO Ontology database Released 2016-03-25

Analyzed List: Piwi_2fold_down_id (Drosophila melanogaster)

Reference List: Drosophila melanogaster (all genes in database)

Annotation Data Set: GO biological process complete

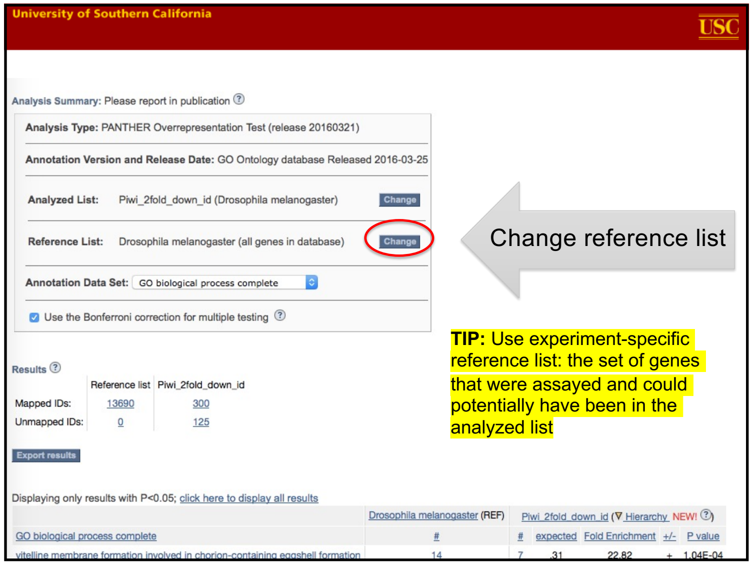
☑ Use the Bonferroni correction for multiple testing

Analysis summary bo

TIP: Report analysis information and include lists publication for reproducibilit

Can change analysis parameters from here

65



University of Southern California

USC

Analysis Summary: Please report in publication ?

Analysis Type: PANTHER Overrepresentation Test (release 20160321)

Annotation Version and Release Date: GO Ontology database Released 2016-03-25

Analyzed List: Piwi_2fold_down_id (Drosophila melanogaster) [Change]

Reference List: Drosophila melanogaster (all genes in database) [Change]

Annotation Data Set: GO biological process complete

☑ Use the Bonferroni correction for multiple testing ?

Change reference list

TIP: Use experiment-specific reference list: the set of genes that were assayed and could potentially have been in the analyzed list

Results ?

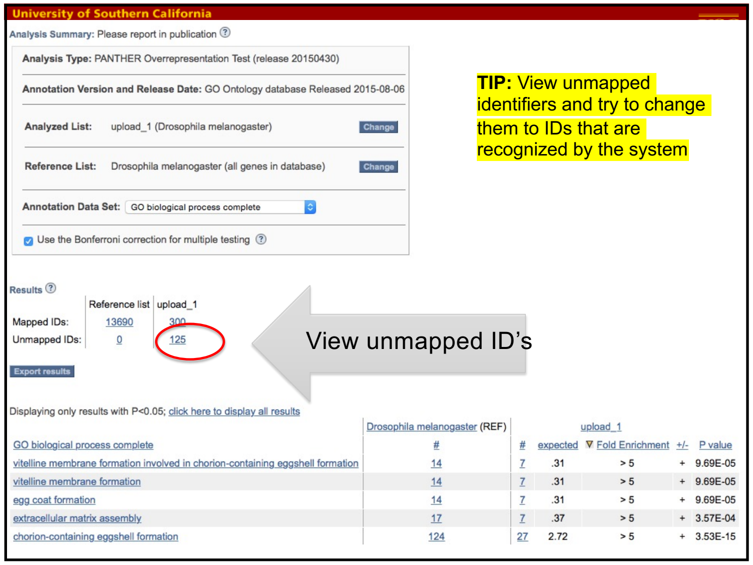
	Reference list	Piwi_2fold_down_id
Mapped IDs:	13690	300
Unmapped IDs:	0	125

[Export results]

Displaying only results with P<0.05; click here to display all results

	Drosophila melanogaster (REF)	Piwi_2fold_down_id (▼ Hierarchy NEW! ?)				
GO biological process complete	#	#	expected	Fold Enrichment	+/-	P value
vitelline membrane formation involved in chorion-containing eggshell formation	14	7	.31	22.82	+	1.04E-04

66



University of Southern California

Analysis Summary: Please report in publication ?

Analysis Type: PANTHER Overrepresentation Test (release 20150430)

Annotation Version and Release Date: GO Ontology database Released 2015-08-06

Analyzed List: upload_1 (Drosophila melanogaster) [Change]

Reference List: Drosophila melanogaster (all genes in database) [Change]

Annotation Data Set: GO biological process complete

☑ Use the Bonferroni correction for multiple testing ?

TIP: View unmapped identifiers and try to change them to IDs that are recognized by the system

Results ?

	Reference list	upload_1
Mapped IDs:	13690	300
Unmapped IDs:	0	125

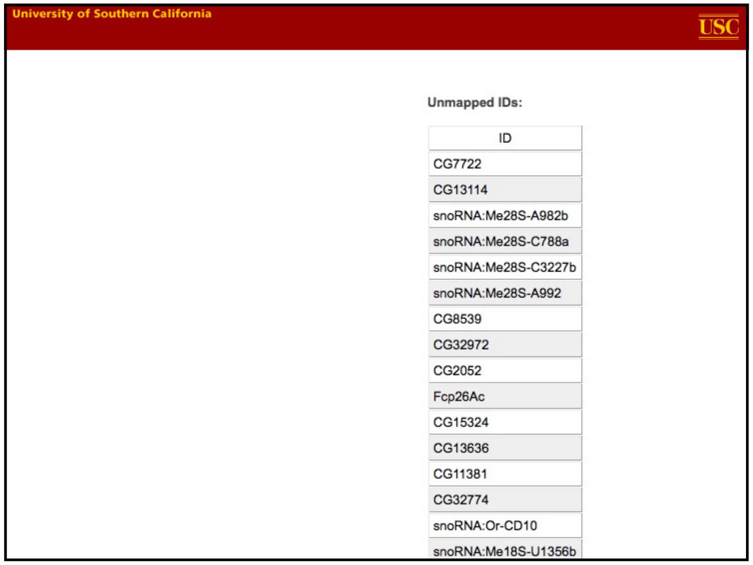
View unmapped ID's

[Export results]

Displaying only results with P<0.05; click here to display all results

	Drosophila melanogaster (REF)	upload_1				
GO biological process complete	#	#	expected	▼ Fold Enrichment	+/-	P value
vitelline membrane formation involved in chorion-containing eggshell formation	14	7	.31	> 5	+	9.69E-05
vitelline membrane formation	14	7	.31	> 5	+	9.69E-05
egg coat formation	14	7	.31	> 5	+	9.69E-05
extracellular matrix assembly	17	7	.37	> 5	+	3.57E-04
chorion-containing eggshell formation	124	27	2.72	> 5	+	3.53E-15

67



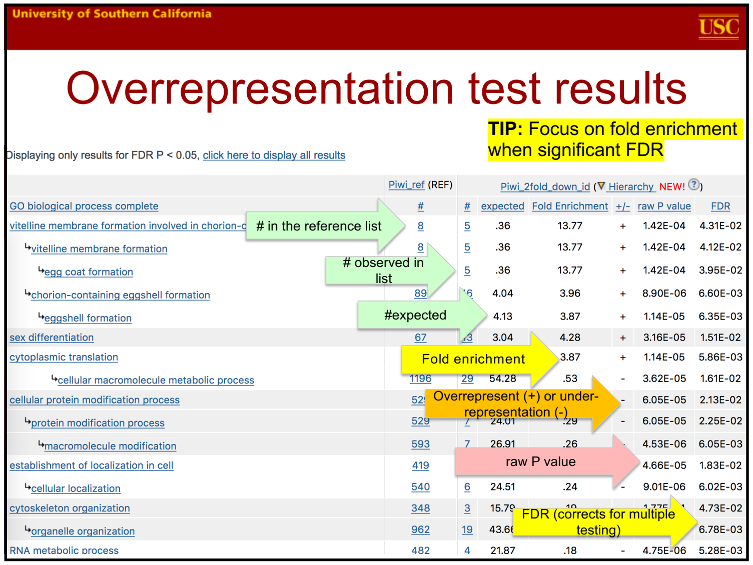
University of Southern California

USC

Unmapped IDs:

ID
CG7722
CG13114
snoRNA:Me28S-A982b
snoRNA:Me28S-C788a
snoRNA:Me28S-C3227b
snoRNA:Me28S-A992
CG8539
CG32972
CG2052
Fcp26Ac
CG15324
CG13636
CG11381
CG32774
snoRNA:Or-CD10
snoRNA:Me18S-U1356b

68



University of Southern California

USC

Overrepresentation test results

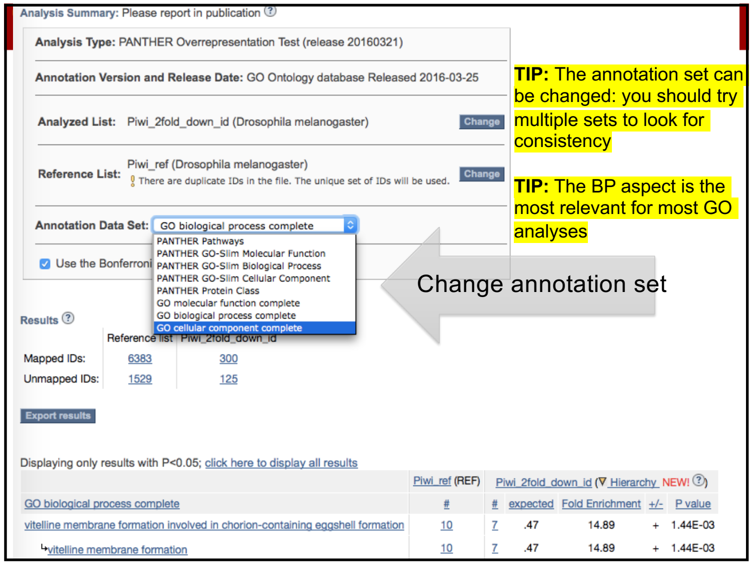
TIP: Focus on fold enrichment when significant FDR

Displaying only results for FDR P < 0.05, click here to display all results

	Piwi_ref (REF)	Piwi_2fold_down_id (▼ Hierarchy NEW! ?)					
GO biological process complete	#	#	expected	Fold Enrichment	+/-	raw P value	FDR
vitelline membrane formation involved in chorion-c	8	5	.36	13.77	+	1.42E-04	4.31E-02
↳vitelline membrane formation	8	5	.36	13.77	+	1.42E-04	4.12E-02
↳egg coat formation		5	.36	13.77	+	1.42E-04	3.95E-02
↳chorion-containing eggshell formation	89	16	4.04	3.96	+	8.90E-06	6.60E-03
↳eggshell formation	[illegible]	[illegible]	4.13	3.87	+	1.14E-05	6.35E-03
sex differentiation	67	13	3.04	4.28	+	3.16E-05	1.51E-02
cytoplasmic translation	[illegible]	[illegible]	[illegible]	3.87	+	1.14E-05	5.86E-03
↳cellular macromolecule metabolic process	1196	29	54.28	.53	-	3.62E-05	1.61E-02
cellular protein modification process	529	[illegible]	[illegible]	[illegible]	-	6.05E-05	2.13E-02
↳protein modification process	529	7	24.01	.29	-	6.05E-05	2.25E-02
↳macromolecule modification	593	7	26.91	.26	-	4.53E-06	6.05E-03
establishment of localization in cell	419	[illegible]	[illegible]	[illegible]	[illegible]	4.66E-05	1.83E-02
↳cellular localization	540	6	24.51	.24	-	9.01E-06	6.02E-03
cytoskeleton organization	348	3	15.79	.19	[illegible]	1.77E-[illegible]	4.73E-02
↳organelle organization	962	19	43.6[illegible]	[illegible]	[illegible]	[illegible]	6.78E-03
RNA metabolic process	482	4	21.87	.18	-	4.75E-06	5.28E-03

# in the reference list
# observed in list
#expected
Fold enrichment
Overrepresent (+) or under-representation (-)
raw P value
FDR (corrects for multiple testing)

69



Analysis Summary: Please report in publication

Analysis Type: PANTHER Overrepresentation Test (release 20160321)

Annotation Version and Release Date: GO Ontology database Released 2016-03-25

Analyzed List: Piwi_2fold_down_id (Drosophila melanogaster) Change

Reference List: Piwi_ref (Drosophila melanogaster)
There are duplicate IDs in the file. The unique set of IDs will be used. Change

Annotation Data Set: GO biological process complete

PANTHER Pathways
PANTHER GO-Slim Molecular Function
PANTHER GO-Slim Biological Process
PANTHER GO-Slim Cellular Component
PANTHER Protein Class
GO molecular function complete
GO biological process complete
GO cellular component complete

☑ Use the Bonferroni

Results

	Reference list	Piwi_2fold_down_id
Mapped IDs:	6383	300
Unmapped IDs:	1529	125

Export results

Displaying only results with P<0.05; click here to display all results

	Piwi_ref (REF)	Piwi_2fold_down_id (Hierarchy NEW!)				
GO biological process complete	#	#	expected	Fold Enrichment	+/-	P value
vitelline membrane formation involved in chorion-containing eggshell formation	10	7	.47	14.89	+	1.44E-03
↳vitelline membrane formation	10	7	.47	14.89	+	1.44E-03

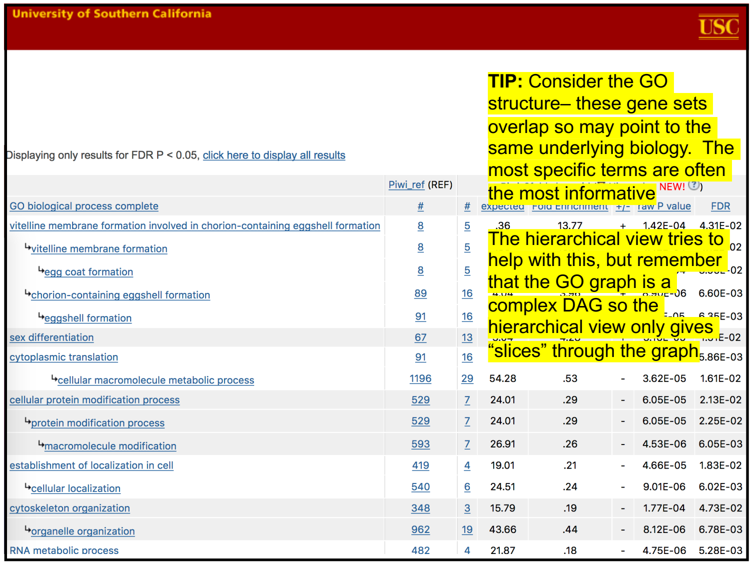
TIP: The annotation set can be changed: you should try multiple sets to look for consistency

TIP: The BP aspect is the most relevant for most GO analyses

Change annotation set

70

University of Southern California



Displaying only results for FDR P < 0.05, click here to display all results

	Piwi_ref (REF)	NEW!					
GO biological process complete	#	#	expected	Fold Enrichment	+/-	raw P value	FDR
vitelline membrane formation involved in chorion-containing eggshell formation	8	5	.36	13.77	+	1.42E-04	4.31E-02
↳vitelline membrane formation	8	5	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
↳egg coat formation	8	5	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
↳chorion-containing eggshell formation	89	16	[illegible]	[illegible]	+	[illegible]	6.60E-03
↳eggshell formation	91	16	[illegible]	[illegible]	[illegible]	[illegible]	6.35E-03
sex differentiation	67	13	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
cytoplasmic translation	91	16	[illegible]	[illegible]	[illegible]	[illegible]	5.86E-03
↳cellular macromolecule metabolic process	1196	29	54.28	.53	-	3.62E-05	1.61E-02
cellular protein modification process	529	7	24.01	.29	-	6.05E-05	2.13E-02
↳protein modification process	529	7	24.01	.29	-	6.05E-05	2.25E-02
↳macromolecule modification	593	7	26.91	.26	-	4.53E-06	6.05E-03
establishment of localization in cell	419	4	19.01	.21	-	4.66E-05	1.83E-02
↳cellular localization	540	6	24.51	.24	-	9.01E-06	6.02E-03
cytoskeleton organization	348	3	15.79	.19	-	1.77E-04	4.73E-02
↳organelle organization	962	19	43.66	.44	-	8.12E-06	6.78E-03
RNA metabolic process	482	4	21.87	.18	-	4.75E-06	5.28E-03

TIP: Consider the GO structure– these gene sets overlap so may point to the same underlying biology. The most specific terms are often the most informative

The hierarchical view tries to help with this, but remember that the GO graph is a complex DAG so the hierarchical view only gives "slices" through the graph

72

University of Southern California

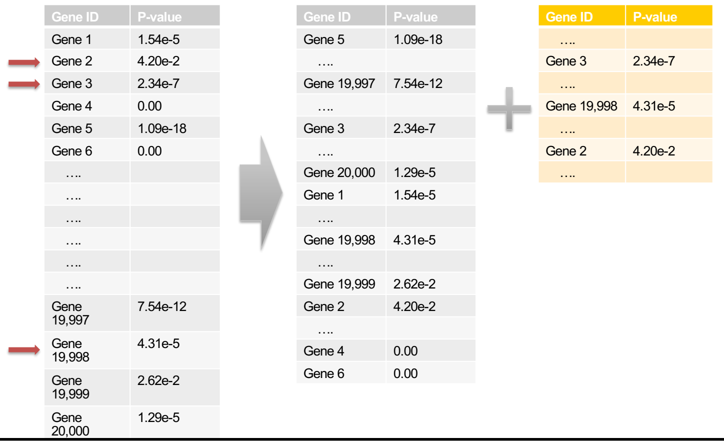
Enrichment test

- Input
  - A list of genes (as many as possible, to get good statistics!) **and a quantitative value for each gene** (e.g. fold change)
- Output
  - The probability that the distribution of values for the **genes in a given GO class** was drawn randomly from the distribution of values for **all genes**

73

University of Southern California

Gene set enrichment



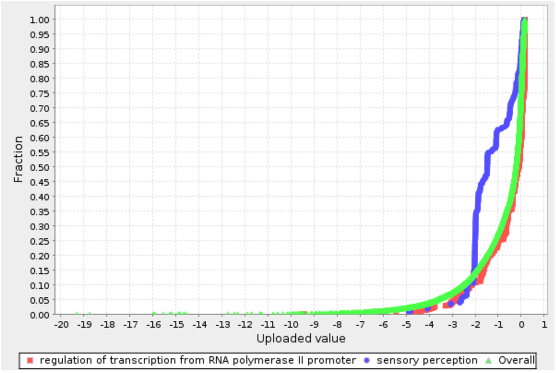
Gene ID	P-value
Gene 1	1.54e-5
Gene 2	4.20e-2
Gene 3	2.34e-7
Gene 4	0.00
Gene 5	1.09e-18
Gene 6	0.00
….	
….	
….	
….	
….	
….	
Gene 19,997	7.54e-12
Gene 19,998	4.31e-5
Gene 19,999	2.62e-2
Gene 20,000	1.29e-5

Gene ID	P-value
Gene 5	1.09e-18
….	
Gene 19,997	7.54e-12
….	
Gene 3	2.34e-7
….	
Gene 20,000	1.29e-5
Gene 1	1.54e-5
….	
Gene 19,998	4.31e-5
….	
Gene 19,999	2.62e-2
Gene 2	4.20e-2
….	
Gene 4	0.00
Gene 6	0.00

Gene ID	P-value
….	
Gene 3	2.34e-7
….	
Gene 19,998	4.31e-5
….	
Gene 2	4.20e-2
….	

74

Gene set enrichment

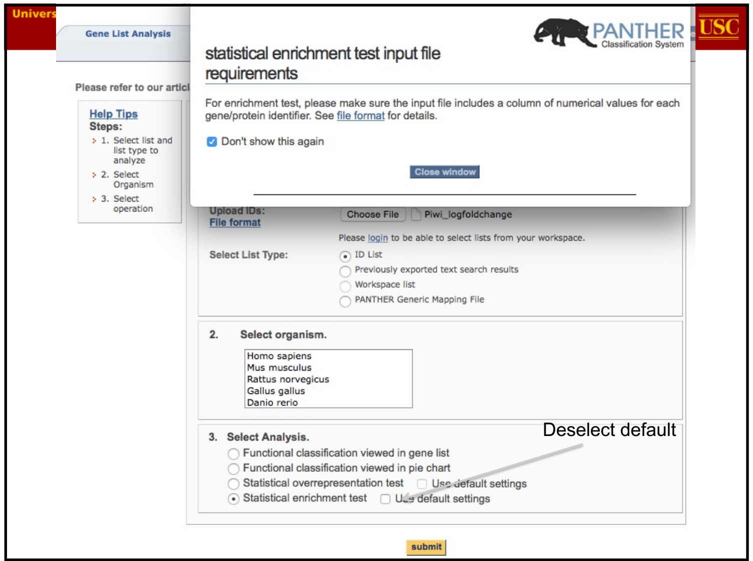


Statistically compares distribution of values for genes in a given annotation class, with distribution for all genes

PANTHER uses Mann-Whitney U Test

GSEA uses Kolmogorov-Smirnov with permutation

75



Gene List Analysis

Please refer to our articl

Help Tips
Steps:
1. Select list and list type to analyze
2. Select Organism
3. Select operation

statistical enrichment test input file requirements

PANTHER Classification System

For enrichment test, please make sure the input file includes a column of numerical values for each gene/protein identifier. See file format for details.

☑ Don't show this again

Close window

Upload IDs: File format — Choose File Piwi_logfoldchange

Please login to be able to select lists from your workspace.

Select List Type:
- ● ID List
- ○ Previously exported text search results
- ○ Workspace list
- ○ PANTHER Generic Mapping File

2. Select organism.

Homo sapiens
Mus musculus
Rattus norvegicus
Gallus gallus
Danio rerio

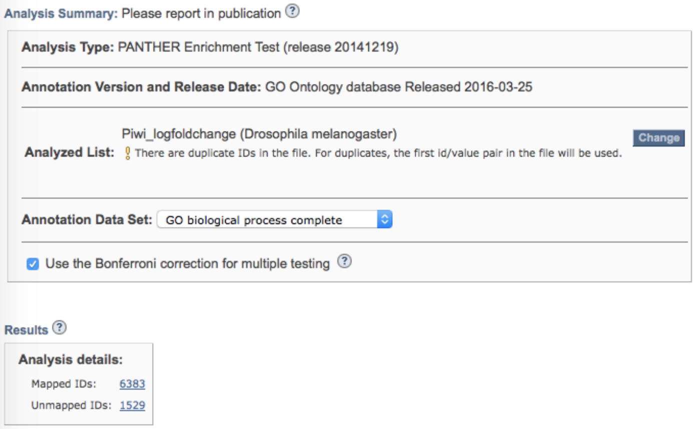
3. Select Analysis.
- ○ Functional classification viewed in gene list
- ○ Functional classification viewed in pie chart
- ○ Statistical overrepresentation test □ Use default settings
- ● Statistical enrichment test □ Use default settings

submit

Deselect default

76

Enrichment test summary



Analysis Summary: Please report in publication

Analysis Type: PANTHER Enrichment Test (release 20141219)

Annotation Version and Release Date: GO Ontology database Released 2016-03-25

Analyzed List: Piwi_logfoldchange (Drosophila melanogaster) Change
There are duplicate IDs in the file. For duplicates, the first id/value pair in the file will be used.

Annotation Data Set: GO biological process complete

☑ Use the Bonferroni correction for multiple testing

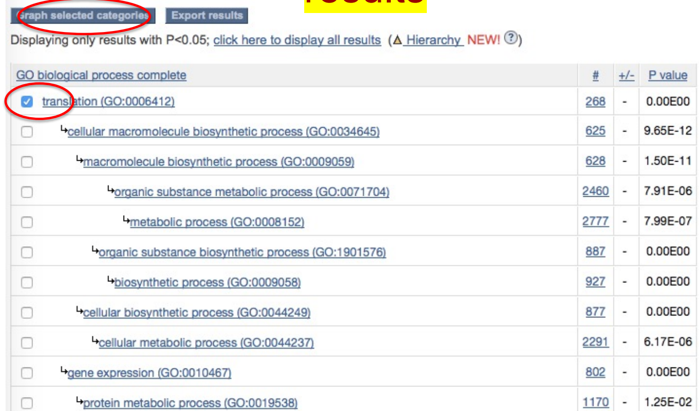
Results

Analysis details:

Mapped IDs:	6383
Unmapped IDs:	1529

77

TIP: Graphing distribution for different classes helps interpret results



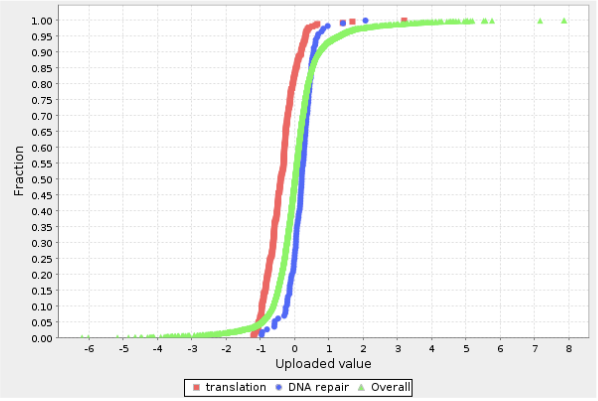
Graph selected categories   Export results

Displaying only results with P<0.05; click here to display all results (⚠ Hierarchy NEW! ?)

GO biological process complete	#	+/-	P value
☑ translation (GO:0006412)	268	-	0.00E00
□ ↳cellular macromolecule biosynthetic process (GO:0034645)	625	-	9.65E-12
□ ↳macromolecule biosynthetic process (GO:0009059)	628	-	1.50E-11
□ ↳organic substance metabolic process (GO:0071704)	2460	-	7.91E-06
□ ↳metabolic process (GO:0008152)	2777	-	7.99E-07
□ ↳organic substance biosynthetic process (GO:1901576)	887	-	0.00E00
□ ↳biosynthetic process (GO:0009058)	927	-	0.00E00
□ ↳cellular biosynthetic process (GO:0044249)	877	-	0.00E00
□ ↳cellular metabolic process (GO:0044237)	2291	-	6.17E-06
□ ↳gene expression (GO:0010467)	802	-	0.00E00
□ ↳protein metabolic process (GO:0019538)	1170	-	1.25E-02

78

Graphing distribution for different classes helps interpret results



Summary of best practices General

- Enable others to reproduce your results
  - Report version of data, and tool
  - And provide data, of course
- Improving analysis (general)
  - Make sure GO annotations are up-to-date
  - For most tools, analysis is gene-centric– ensure that your data are also for individual genes (not splice forms, etc)
    - Example retracted paper https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4169929/
  - Check input identifiers that did not map to the database
    - Can these be fixed using alternative identifiers?
  - Are enriched classes related? (consider GO structure)
  - Consider ALL results, not just the ones you want to see
    - Explore the genes in enriched classes that are unexpected

Best practices: For overrepresentation tests

- Use appropriate reference list (what could have been observed)
- Fold enrichment can be more informative than P-value, as long as the P-value is significant
  - P-value can depend on size of the gene set

Best practices: For enrichment tests

- Upload quantitative values for as many genes as possible
- Graph distributions for enriched classes to help interpretation