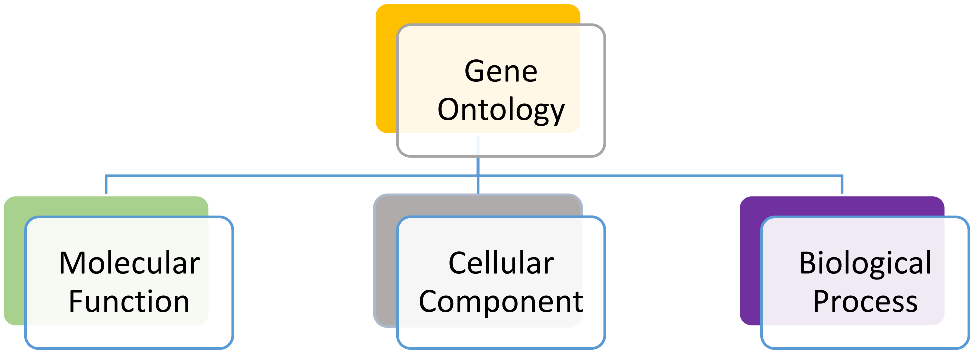
Gene Ontology (GO) Enrichment

Learning objectives:

- Run a GO enrichment analysis
- Explore GO enrichment results

Background:

The gene ontology describes the knowledge of biological sciences and divides this knowledge into three broad categories: Molecular function, cellular component and biological process.



Activities at the molecular level performed by gene products, e.g. Toxin activity, catalytic activity of transporter activity

Where a gene product performs its function, e.g. Cilium Mitochondrion, plastid, Golgi etc…

Processes accomplished by multiple activities, e.g. pyrimidine biosynthesis

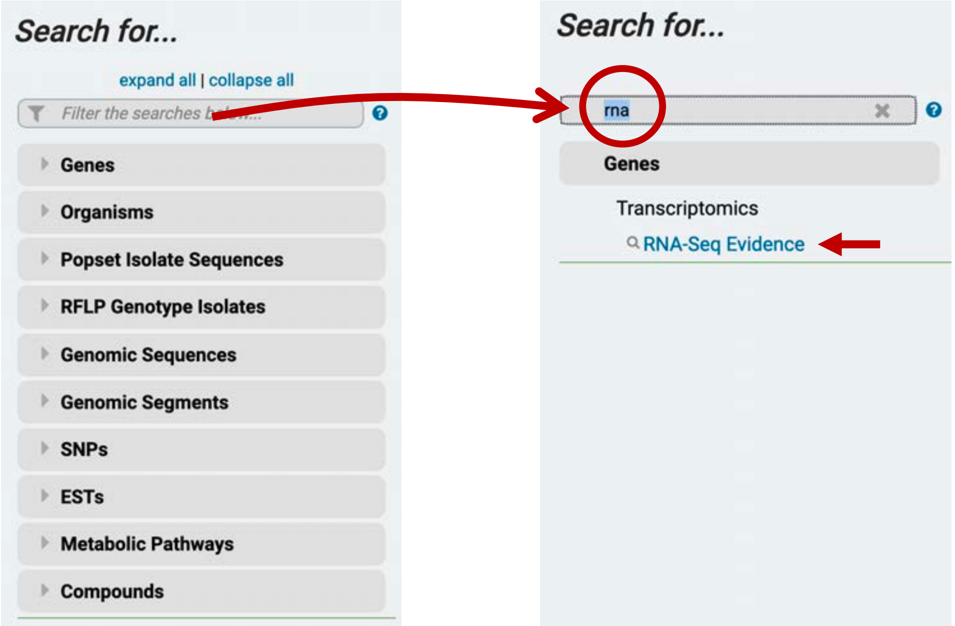
To learn more about Gene ontology please visit: http://geneontology.org/docs/ontology-documentation/

Genes can be assigned a GO term either manually or computationally based on transfer by similarity, by domain association or by many other computational methods. GO terms can be used in enrichment analysis!
For example: Does my list of genes have an over-representation of specific GO terms compared to the rest of the genome?

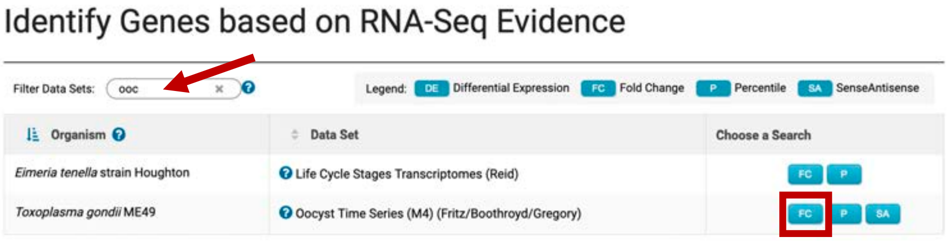
A standard enrichment method is Fisher’s exact test which is a statistical test used when analyzing contingency tables. Typically used when you have a small sample size. But when you are doing enrichment analysis on a list of genes with the background being the whole genome, your sample size is not small. As a result, the P- value you get from a Fisher’s exact test might be misleading.

With a small sample size, a P-value of less than 0.05 is considered significant (5% chance of being wrong/random). But if you are doing an enrichment analysis with all genes in the genome then each gene can be considered a test, so the chances of a type one error becomes higher. As a result, you should correct for this which can be done in different ways including Benjamini-Hochberg false discovery rate (FDR) or Bonferroni adjusted p-value

1. In order to run a GO enrichment analysis, we need a list of genes to test. This can be a list of gene IDs from your results that you can upload using the ID search or a gene list resulting from a search you conducted in the database. For this example, in ToxoDB, we will identify genes that are differentially regulated over time.
    a. Navigate to the RNA-Seq searches and find the data set called "**Oocyst Time Series (M4)**" from Fritz *et al.* A fast way of getting to the RNA-Seq searches is type 'rna' in the filter box on the left of the home page then click on the RNA-Seq Evidence link. See image below.



    b. The RNA-Seq evidence page include a list of all the data sets that are loaded in the database. To quickly find a dataset you can start typing key words in the "Filter Data Sets" box. For example, start typing the word "oocyst".



Identify Genes based on RNA-Seq Evidence

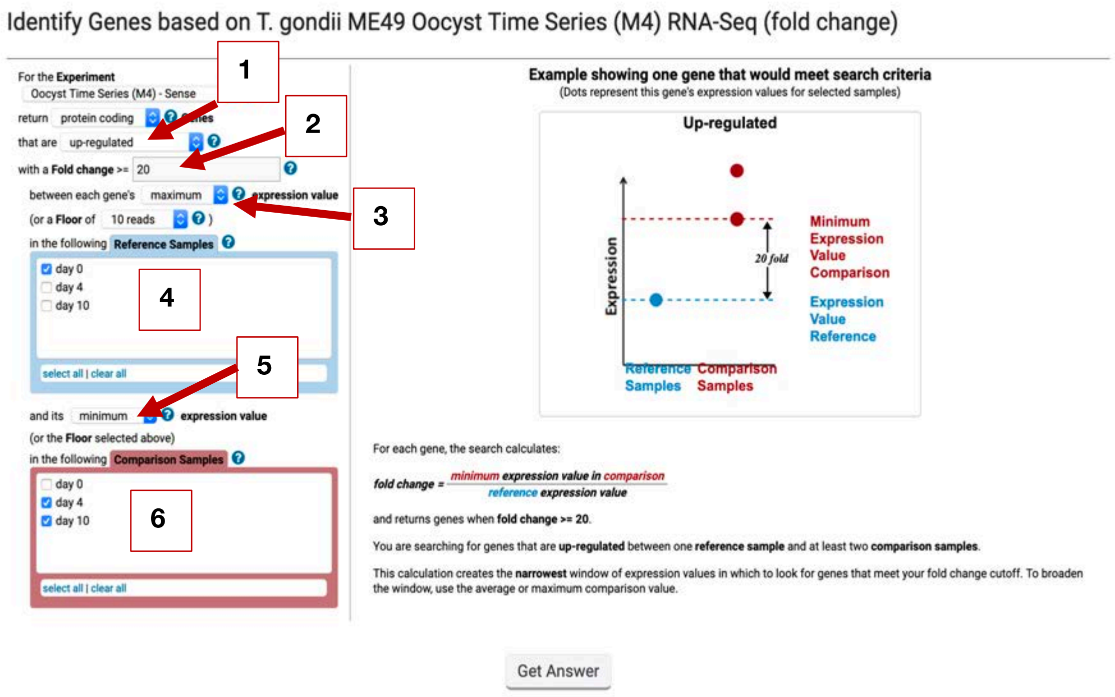
Filter Data Sets: ooc

Legend: DE Differential Expression FC Fold Change P Percentile SA SenseAntisense

Organism	Data Set	Choose a Search
Eimeria tenella strain Houghton	Life Cycle Stages Transcriptomes (Reid)	FC P
Toxoplasma gondii ME49	Oocyst Time Series (M4) (Fritz/Boothroyd/Gregory)	FC P SA

    c. Once you find the data set of interest click on the fold change option. This will make available to you all the parameters that you can manipulate to search this data set. For this exercise identify genes that are upregulated by 20-fold in the day 4 and day 10 time points compared to the day 0 time point. Parameters to set:

1. Up-regulated
2. 20-fold
3. Maximum
4. Day 0
5. Minimum
6. Day 4 and 10



Identify Genes based on T. gondii ME49 Oocyst Time Series (M4) RNA-Seq (fold change)

For the **Experiment**
Oocyst Time Series (M4) - Sense
return protein coding genes
that are up-regulated
with a **Fold change** >= 20
between each gene's maximum **expression value**
(or a **Floor** of 10 reads)
in the following **Reference Samples**

☑ day 0
☐ day 4
☐ day 10

select all | clear all

and its minimum **expression value**
(or the **Floor** selected above)
in the following **Comparison Samples**

☐ day 0
☑ day 4
☑ day 10

select all | clear all

Example showing one gene that would meet search criteria
(Dots represent this gene's expression values for selected samples)

For each gene, the search calculates:

$$\textit{fold change} = \frac{\textit{minimum expression value in comparison}}{\textit{reference expression value}}$$

and returns genes when **fold change >= 20**.

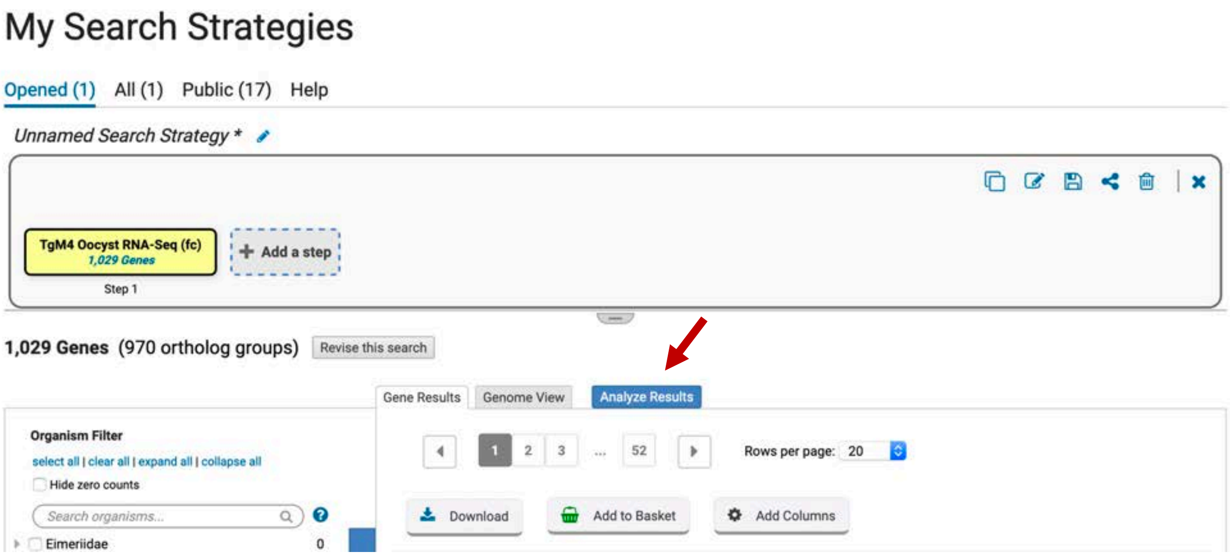
You are searching for genes that are **up-regulated** between one **reference sample** and at least two **comparison samples**.

This calculation creates the **narrowest** window of expression values in which to look for genes that meet your fold change cutoff. To broaden the window, use the average or maximum comparison value.

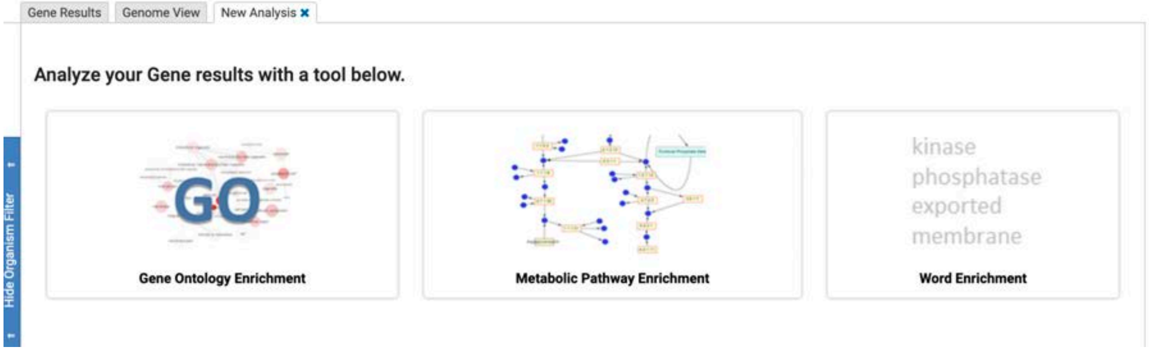
d. Once you have set the parameters you can click on the “Get Answer” button at the bottom of the search. This will return a one-step search strategy. How many genes did you get?

2. To run a GO enrichment analysis on these results, do the following:

a. Click on the Analyze Results tab right above the list of genes (arrow in image below).



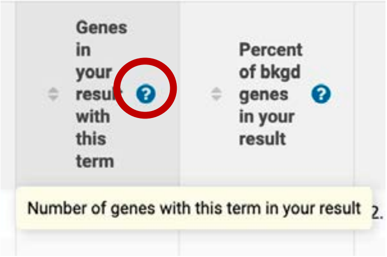
b. Clicking on the “Analyze Results” tab will reveal the different analyses that you can run on your results. Besides GO enrichment what other analyses are available?



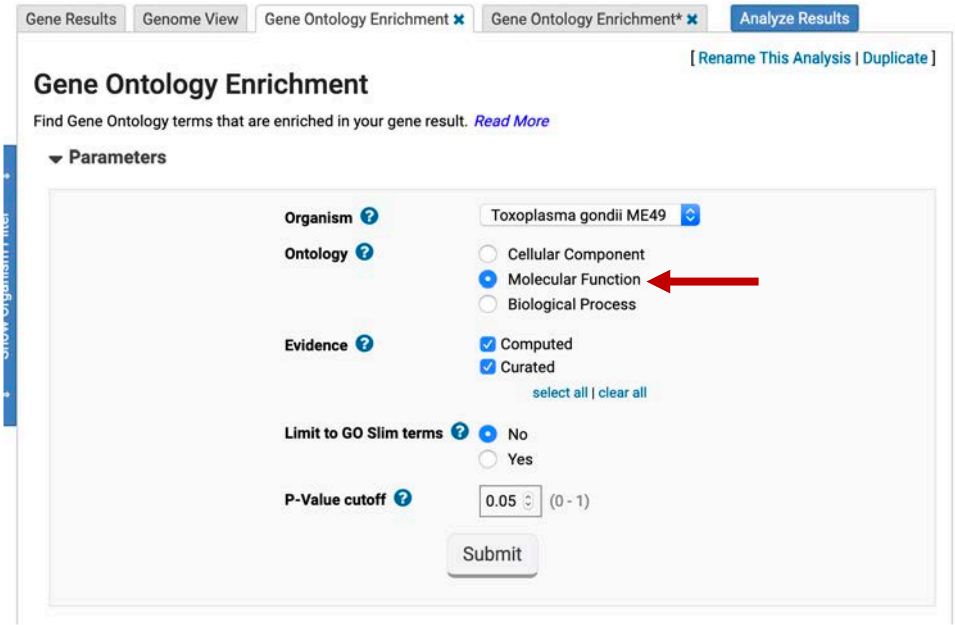
c. Click on the GO enrichment option. This will reveal the parameters that you can modify. For the purpose of this exercise, keep all the defaults and click on “Submit”.

d. What is the top enriched GO term from this analysis?

e. What do each of the columns in the analysis table represent? (hint: move your mouse over the question mark next to each column header to get more information)


Genes in your result with this term
Percent of bkgd genes in your result
Number of genes with this term in your result
2.

f. Try rerunning the GO enrichment analysis but this time select the Molecular Function ontology. What is the top enriched GO term?


Gene Results
Genome View
Gene Ontology Enrichment ✖
Gene Ontology Enrichment* ✖
Analyze Results
[Rename This Analysis | Duplicate]
Show Organism Filter

Gene Ontology Enrichment

Find Gene Ontology terms that are enriched in your gene result. *Read More*

▾ Parameters

Organism ? Toxoplasma gondii ME49

Ontology ?
○ Cellular Component
● Molecular Function
○ Biological Process

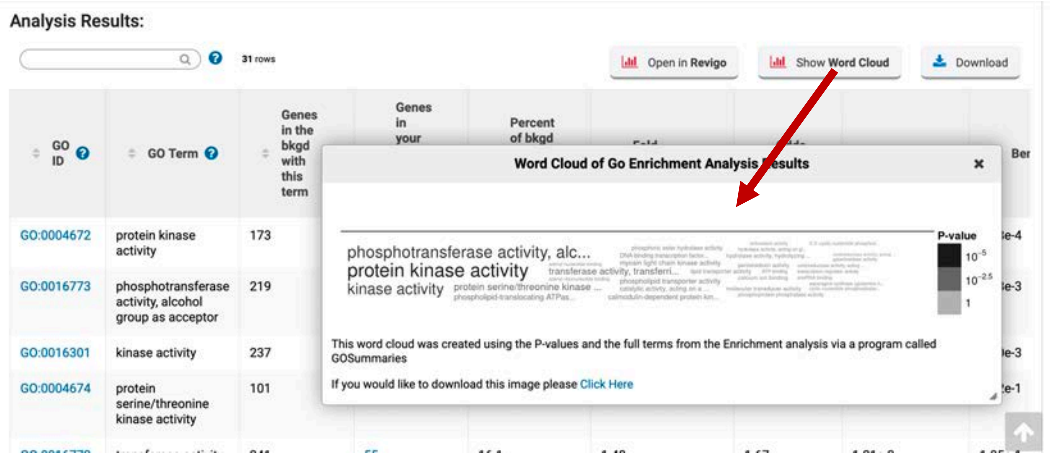
Evidence ?
☑ Computed
☑ Curated
select all | clear all

Limit to GO Slim terms ?
● No
○ Yes

P-Value cutoff ? 0.05 (0 - 1)

Submit

g. Click on the “Word Cloud” button above the analysis results. What does this do? (See image below).


Open in Revigo
Show Word Cloud
Download
Word Cloud of Go Enrichment Analysis Results
phosphotransferase activity, alc...
protein kinase activity
kinase activity
transferase activity, transferri...
protein serine/threonine kinase ...
phospholipid-translocating ATPas...
phospholipid transporter activity
calmodulin-dependent protein kin...
P-value
10^{-5}
$10^{-2.5}$
1

Analysis Results:

31 rows

GO ID	GO Term	Genes in the bkgd with this term
GO:0004672	protein kinase activity	173
GO:0016773	phosphotransferase activity, alcohol group as acceptor	219
GO:0016301	kinase activity	237
GO:0004674	protein serine/threonine kinase activity	101

This word cloud was created using the P-values and the full terms from the Enrichment analysis via a program called GOSummaries

If you would like to download this image please Click Here

Additional resources:
Gene Ontology:
http://geneontology.org/docs/ontology-documentation/
Enzyme Commission numbers:
https://www.qmul.ac.uk/sbcs/iubmb/enzyme/
More info on Fischer’s exact test:

http://www.biostathandbook.com/fishers.html
Fisher's Exact Test and the Hypergeometric Distribution (the M&M example):
https://youtu.be/udyAvvaMjfM
Some more info about Odds ratios:
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2938757/
False discovery rates and P value correction:
http://brainder.org/2011/09/05/fdr-corrected-fdr-adjusted-p-values/
GO Slim:
http://www-legacy.geneontology.org/GO.slims.shtml
REVIGO:
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0021800