Gene Ontology (GO) Enrichment

Learning objectives:

- Run a GO enrichment analysis
- Explore GO enrichment results
- Port GO enrichment results to Revigo

Background:

Ontologies are a controlled vocabulary of terms and concepts with relationships between them. The Gene Ontology describes the knowledge of biological sciences and divides this knowledge into three broad categories: Molecular function, cellular component and biological process.

Gene Ontology

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

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

Biological Process: Processes accomplished by multiple activities, e.g. pyrimidine biosynthesis, gluconeogenesis

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

A gene can be assigned a GO term either manually (by an annotator or curator when they evaluate experimental evidence from a publication) or computationally (based on the GO terms of genes that share sequence or functional domains). The origin of the assignment is documented; some researchers believe that manually assigned functional annotations are more accurate than those that are electronically transferred since a researcher has reviewed the manually annotated assignments. GO terms can be used to test whether your set of genes are enriched for a molecular function, cellular component, or biological process.

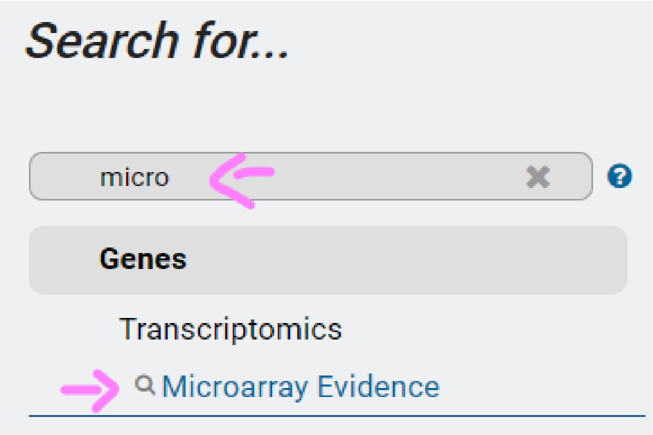
For example: A researcher performs a proteomics experiment on a protein fraction collected during an antimalarial treatment and identifies 100 proteins in total. When they examine the GO terms assigned to the gene set corresponding to the proteome, they see that 25 genes are assigned GO:0016301, kinase activity. Out of 5000 genes in the genome, only 100 are assigned GO:0016301. There is an overrepresentation of

GO:0016301 in the researcher's proteome which is 'enriched' for kinase activity.

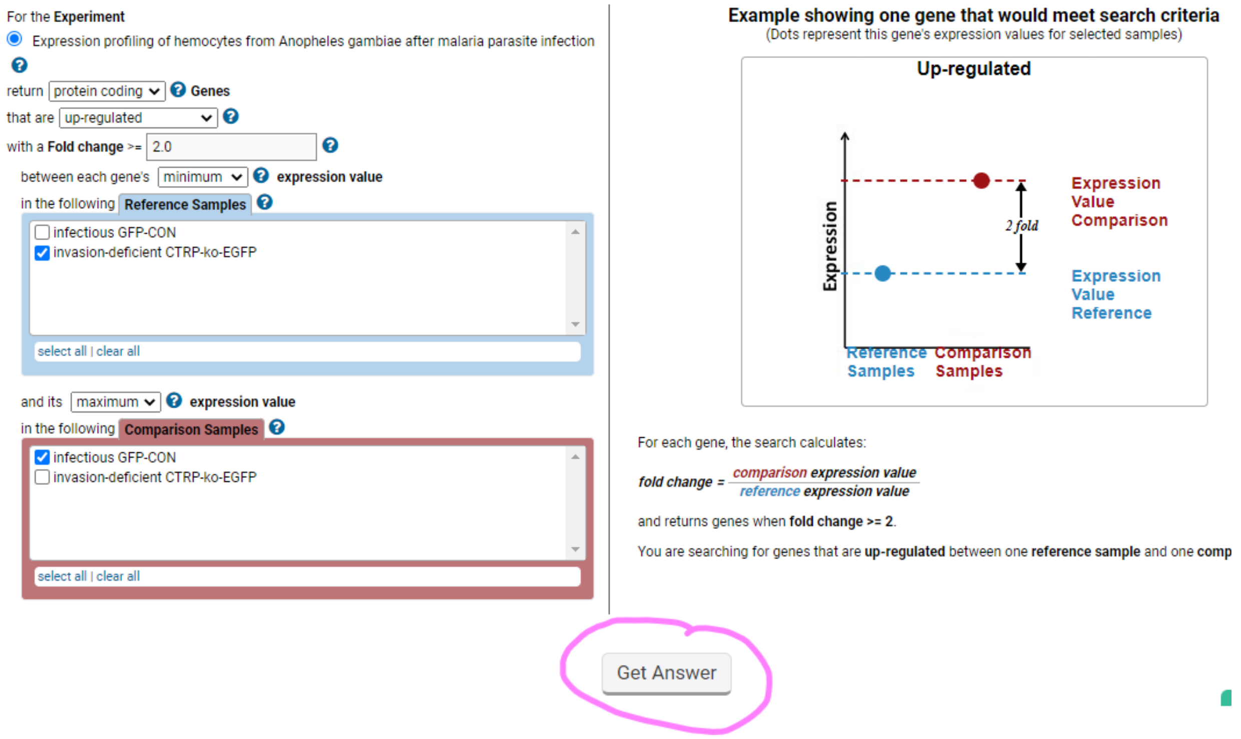
A standard enrichment determination method employs Fisher's exact test, a statistical test that evaluates a 2x2 contingency table (in this case, number of genes in my set *versus* number of genes from genome not in my set, and number of genes with GO term Z *versus* number of genes without term Z). This test produces a p-value between 0 and 1, where $p \leq 0.05$ is considered significant (that is, less than 5% probability that the enrichment is due to chance). However, the test is performed for each of the 100s of GO terms, increasing the chances that a GO term will be incorrectly considered enriched (a false positive, or type I, error). Thus, the original p-value must be adjusted for so-called multiple hypothesis testing, resulting in an adjusted p-value such as the Benjamini-Hochberg false discovery rate (FDR) or Bonferroni adjusted p-value.

In order to run a GO enrichment analysis, you need a list of genes to test. This can be a list of gene IDs from your experimental results (upload them with the ID search) or a gene list resulting from a search you conducted on a VEuPathDB website.

1. For this example, in VectorBase, we will use the microarray data "Expression profiling of hemocytes from *Anopheles gambiae* after malaria parasite infection (Pinto et al.)", to find genes that are upregulated in the mosquito hemocyte cells after infected with *Plasmodium.* Select the fold change search.



- Set the following parameters:



For the **Experiment**

● Expression profiling of hemocytes from Anopheles gambiae after malaria parasite infection

return protein coding **Genes**

that are up-regulated

with a **Fold change** >= 2.0

between each gene's minimum **expression value**

in the following **Reference Samples**

□ infectious GFP-CON
☑ invasion-deficient CTRP-ko-EGFP

select all | clear all

and its maximum **expression value**

in the following **Comparison Samples**

☑ infectious GFP-CON
□ invasion-deficient CTRP-ko-EGFP

select all | clear all

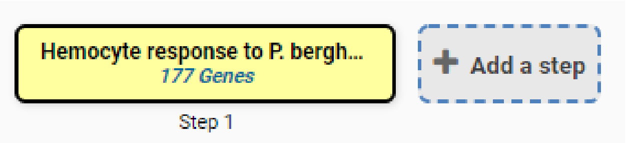
Example showing one gene that would meet search criteria
(Dots represent this gene's expression values for selected samples)

For each gene, the search calculates:

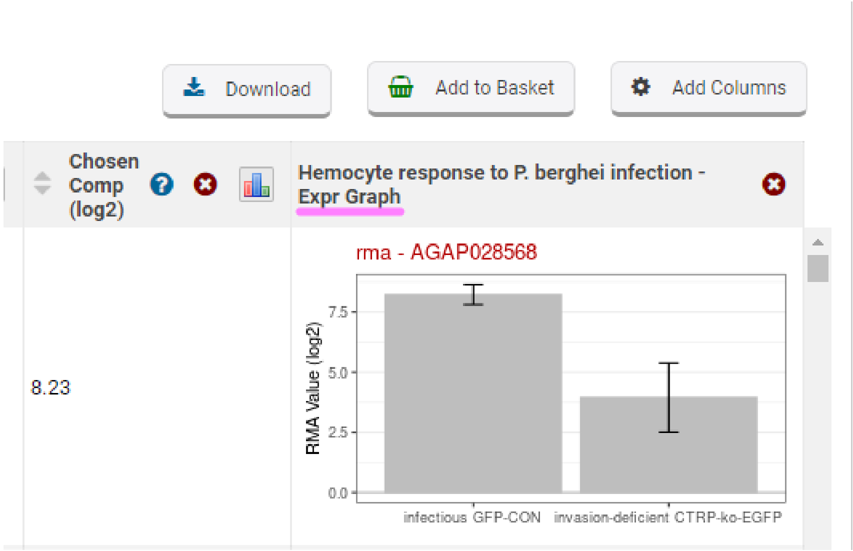
$$\textit{fold change} = \frac{\textit{comparison expression value}}{\textit{reference expression value}}$$

and returns genes when **fold change >= 2**.

You are searching for genes that are **up-regulated** between one **reference sample** and one **comp**

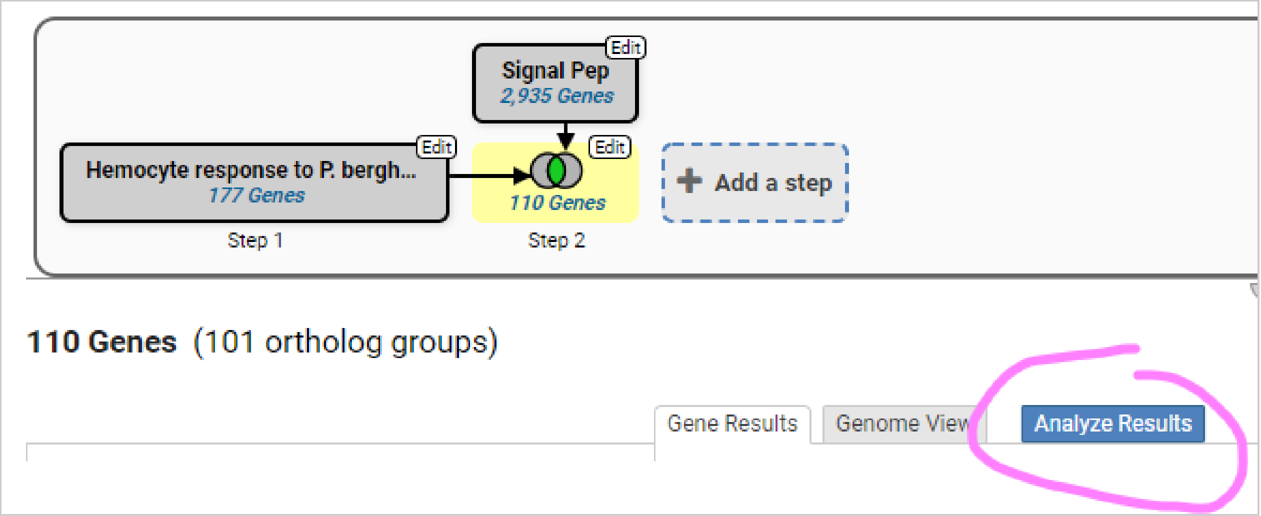


- Scroll down to explore the obtained results. Notice there is a column with the expression graphs for each gene

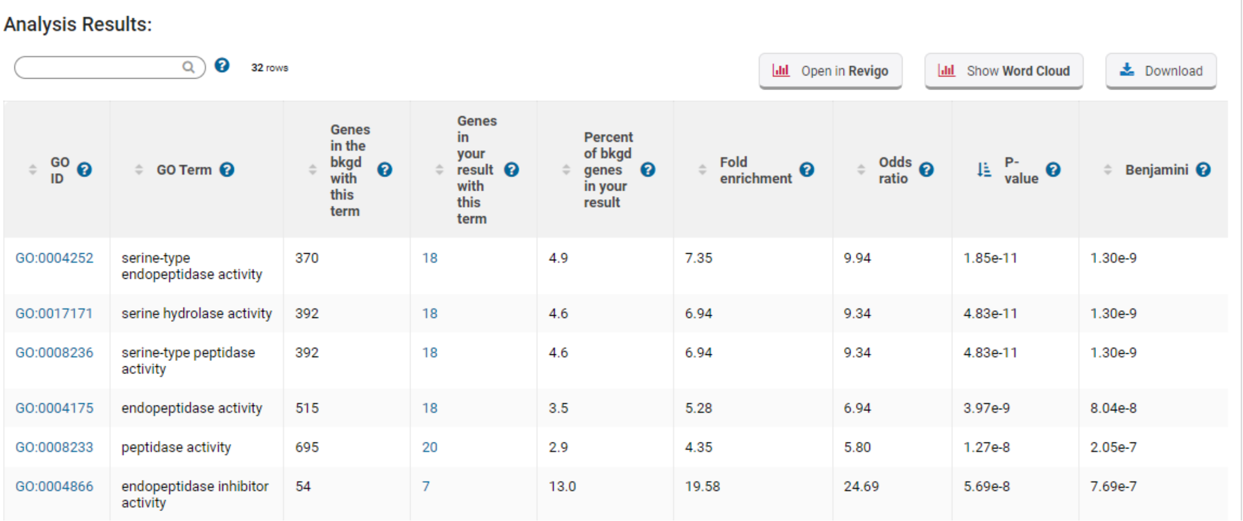


Chosen Comp (log2)	Hemocyte response to P. berghei infection - Expr Graph
8.23	

- Add a step to look for which of these genes are potentially secreted (signal peptide search). Now let´s run an enrichment analysis: Analyze Results > GO Enrichment > use default parameters for Molecular Function



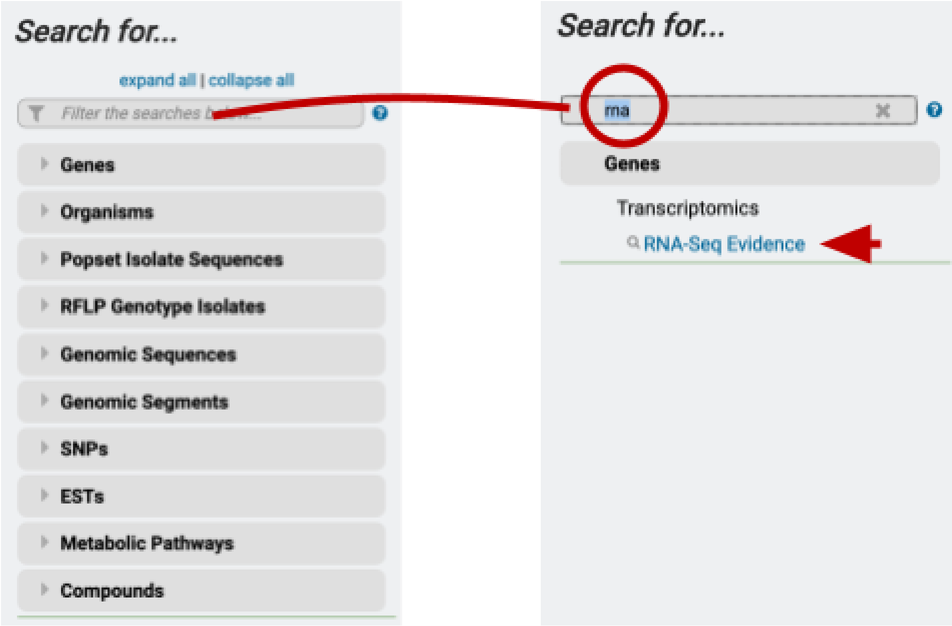
- Are the top hits the molecular functions that you would expect to find?



GO ID	GO Term	Genes in the bkgd with this term	Genes in your result with this term	Percent of bkgd genes in your result	Fold enrichment	Odds ratio	P-value	Benjamini
GO:0004252	serine-type endopeptidase activity	370	18	4.9	7.35	9.94	1.85e-11	1.30e-9
GO:0017171	serine hydrolase activity	392	18	4.6	6.94	9.34	4.83e-11	1.30e-9
GO:0008236	serine-type peptidase activity	392	18	4.6	6.94	9.34	4.83e-11	1.30e-9
GO:0004175	endopeptidase activity	515	18	3.5	5.28	6.94	3.97e-9	8.04e-8
GO:0008233	peptidase activity	695	20	2.9	4.35	5.80	1.27e-8	2.05e-7
GO:0004866	endopeptidase inhibitor activity	54	7	13.0	19.58	24.69	5.69e-8	7.69e-7

2. 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 quick way of getting to the RNA-Seq searches is to type ‘rna’ in the filter box on the left of the home page and click on the RNA Seq Evidence link. See image below.



Search for...

expand all | collapse all

Filter the searches below...

- Genes
- Organisms
- Popset Isolate Sequences
- RFLP Genotype Isolates
- Genomic Sequences
- Genomic Segments
- SNPs
- ESTs
- Metabolic Pathways
- Compounds

Search for...

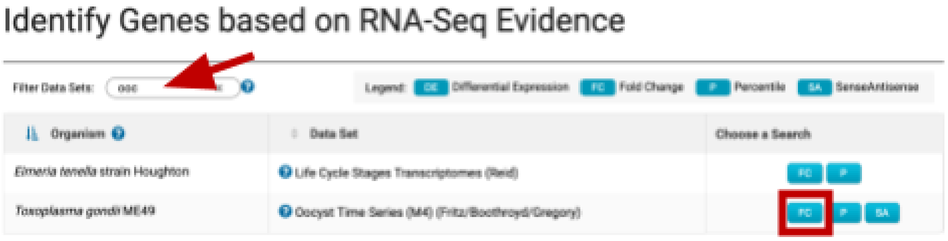
rna

Genes

Transcriptomics

RNA-Seq Evidence

b. The RNA-Seq evidence page includes a list of all data sets that are loaded in the website. 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

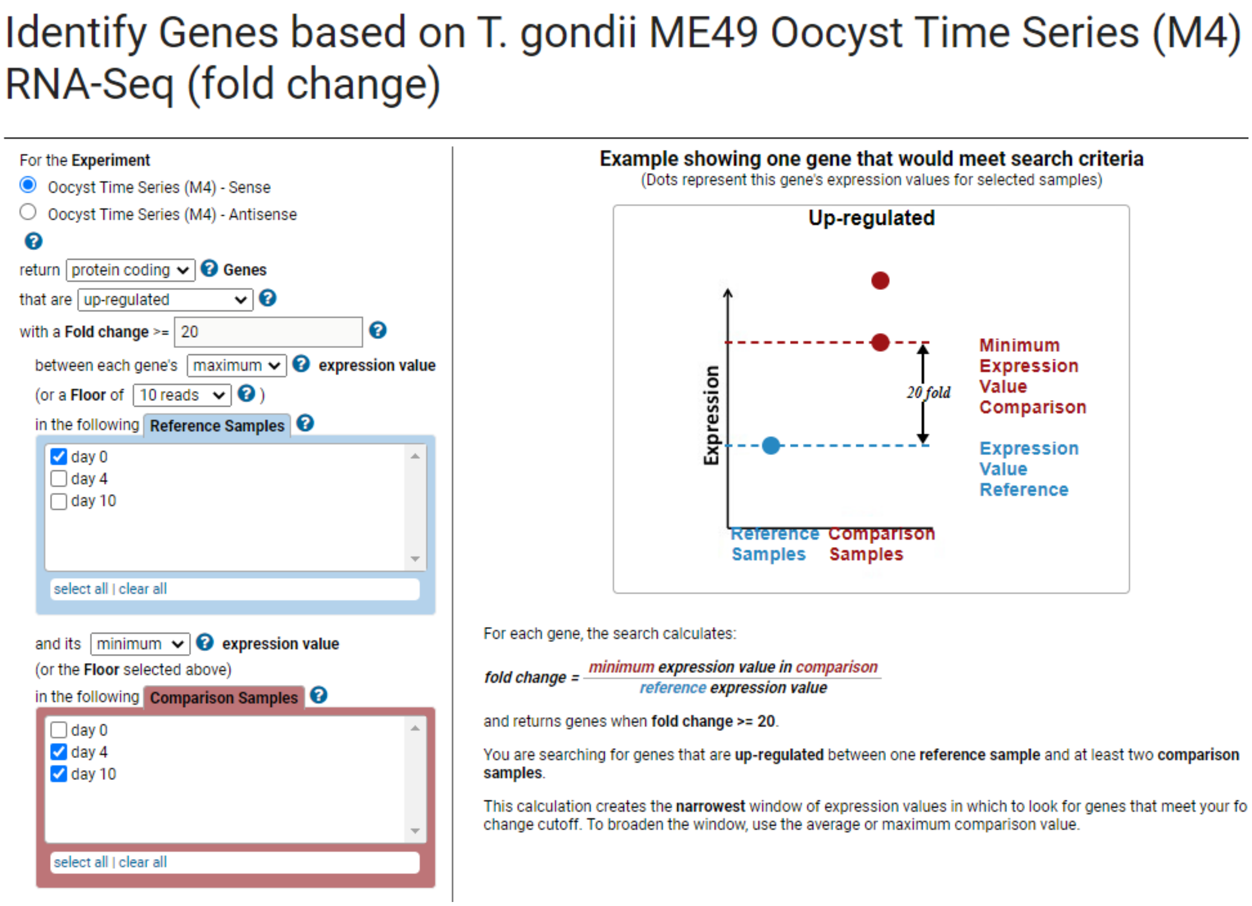
Legend: DE Differential Expression FC Fold Change P Percentile SA SenseAntisense

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, choose the fold-change (FC) search. For this exercise, identify genes that are upregulated by 20-fold in days 4 and 10 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

○ Oocyst Time Series (M4) - Antisense

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:

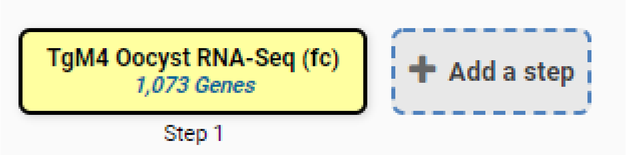
$$\textit{fold change} = \frac{\textit{minimum expression value in comparison}}{\textit{reference expression value}}$$

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 fo change cutoff. To broaden the window, use the average or maximum comparison value.

d. Click "Get Answer" to initiate the search. This will return a one-step search strategy. How many genes did you get?

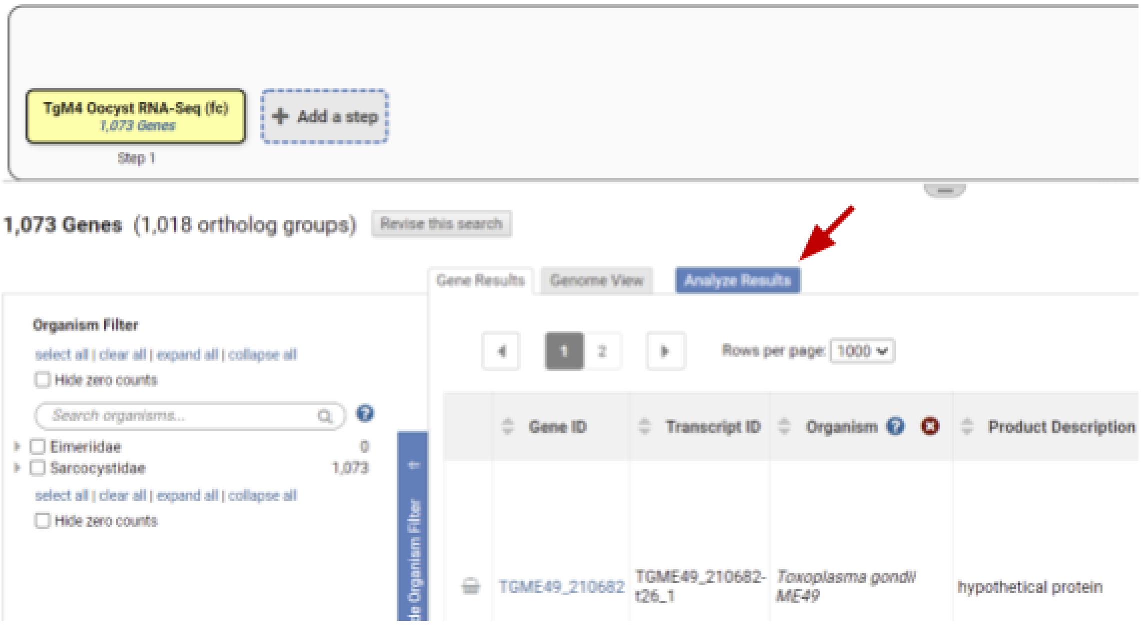


TgM4 Oocyst RNA-Seq (fc)
1,073 Genes

Step 1

2. To run a GO enrichment analysis on these results, do the following: a. Click on the Analyze Results tab just above the list of genes (arrow in image below) to

open the enrichment tools. Besides GO enrichment, what other analyses are available?



TgM4 Oocyst RNA-Seq (fc)
1,073 Genes
Step 1

➕ Add a step

1,073 Genes (1,018 ortholog groups) Revise this search

Gene Results | Genome View | Analyze Results

Organism Filter
select all | clear all | expand all | collapse all
□ Hide zero counts

Search organisms...

□ Eimeriidae	0
□ Sarcocystidae	1,073

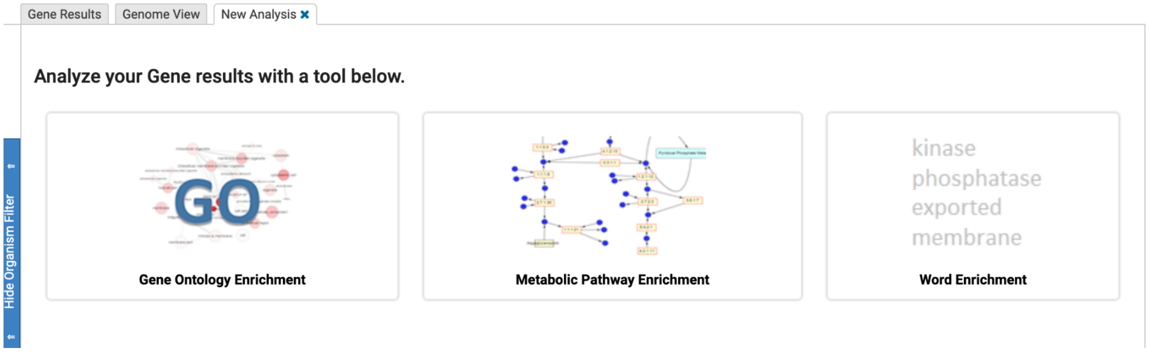
select all | clear all | expand all | collapse all
□ Hide zero counts

Hide Organism Filter

Rows per page: 1000

Gene ID	Transcript ID	Organism	Product Description
TGME49_210682	TGME49_210682-t26_1	*Toxoplasma gondii ME49*	hypothetical protein

b. Click on the GO enrichment option. This will reveal the parameters that you can modify. For the purpose of this exercise, set the following parameters and click on “Submit”.
Organism = T. gondii ME49
Ontology = Cellular Component
Evidence = Computed and Curated
Limit to GO Slim terms? = NO



Gene Results | Genome View | New Analysis ✖

Analyze your Gene results with a tool below.

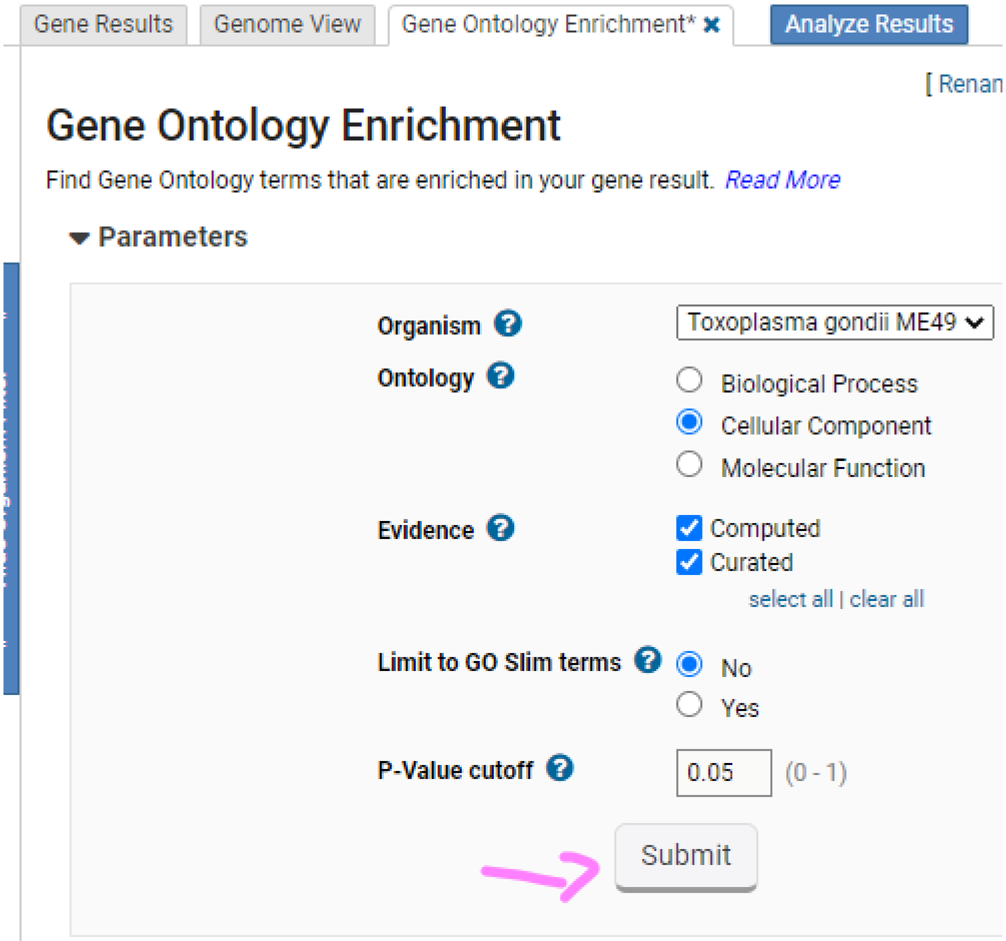
Gene Ontology Enrichment

Metabolic Pathway Enrichment

kinase
phosphatase
exported
membrane

Word Enrichment

Hide Organism Filter



Gene Ontology Enrichment

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

▾ **Parameters**

Organism ? Toxoplasma gondii ME49

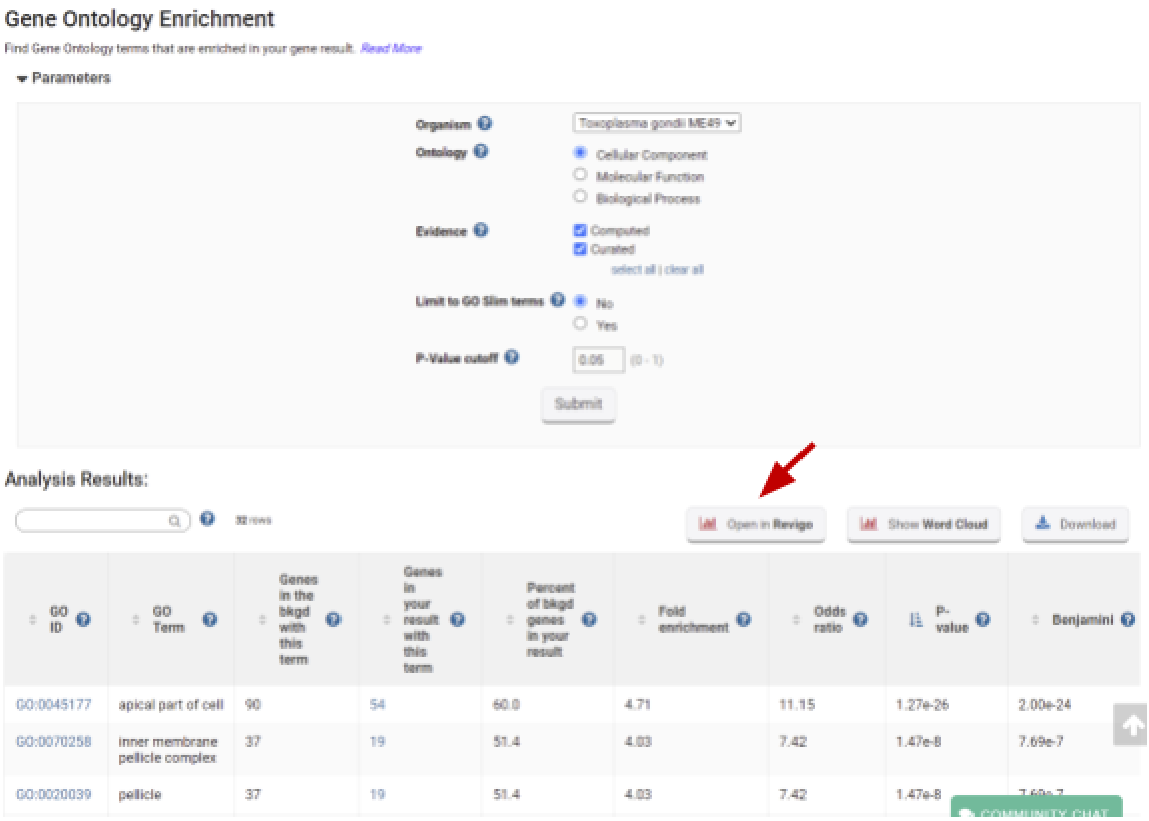
Ontology ? ○ Biological Process ● Cellular Component ○ Molecular Function

Evidence ? ☑ Computed ☑ Curated select all | clear all

Limit to GO Slim terms ? ● No ○ Yes

P-Value cutoff ? 0.05 (0 - 1)

c. What is the top enriched GO term from this analysis? Does this make sense for an enrichment analysis of the cellular component of your Oocyst expressed genes? Notice that the p-value is a rather low, 10^{-24}.



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

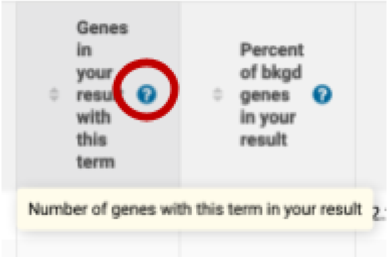
P-Value cutoff: 0.05 (0 - 1)

Analysis Results:

GO ID	GO Term	Genes in the bkgd with this term	Genes in your result with this term	Percent of bkgd genes in your result	Fold enrichment	Odds ratio	P-value	Benjamini
GO:0045177	apical part of cell	90	54	60.0	4.71	11.15	1.27e-26	2.00e-24
GO:0070258	inner membrane pellicle complex	37	19	51.4	4.03	7.42	1.47e-8	7.69e-7
GO:0020039	pellicle	37	19	51.4	4.03	7.42	1.47e-8	7.69e-7

d. What do each of the columns in the analysis table represent? (Hint: move your mouse over the question mark next to each column header)

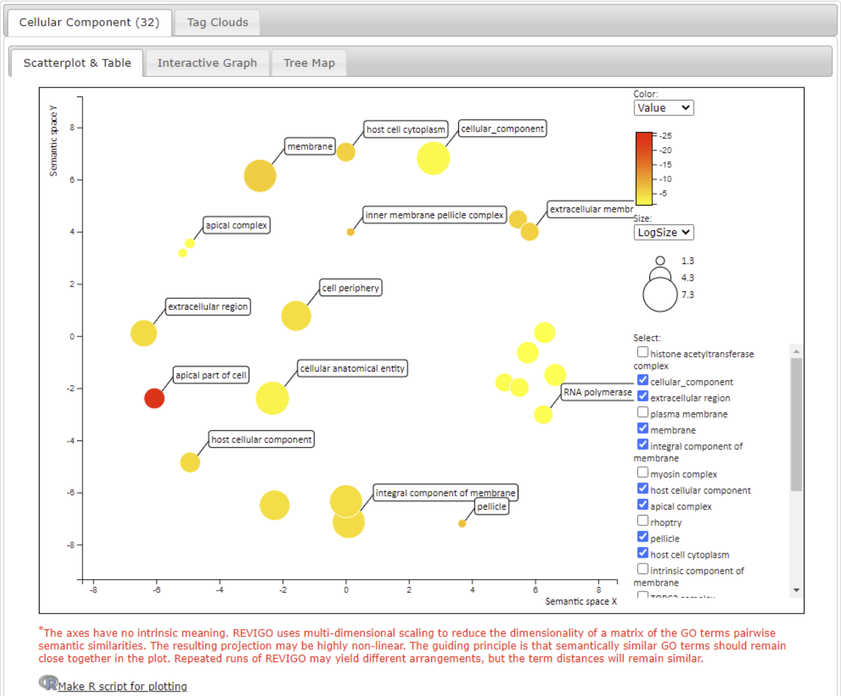
- Fold enrichment -The ratio of the proportion of genes in the list of interest with a specific GO term over the proportion of genes in the background with that term
- Odds ratio -The odds of the GO term appearing in the gene list are the same as that for the background list
- P-value –The null hypothesis or the probability of getting a result that is equal or greater than what was observed
- Benjamini-Hochburg false discovery rate – A method for controlling false discovery rates for type 1 errors
- Bonferroni adjusted P-values -A method for correcting significance based on multiple comparisons



Genes in your result with this term
Percent of bkgd genes in your result

Number of genes with this term in your result

e. Click the Open in Revigo button to port the results to Revigo, the Reduce and Visualize Ontology tool. Once at Revigo, you may need to scroll down to click Start Revigo to run the analysis with default paramters. Revigo provides a scatterplot and table, an integrative map and a tree map to supplement the table provided in the VEuPathDB site. [Revigo publication]



Color: Value

Size: LogSize

Select:

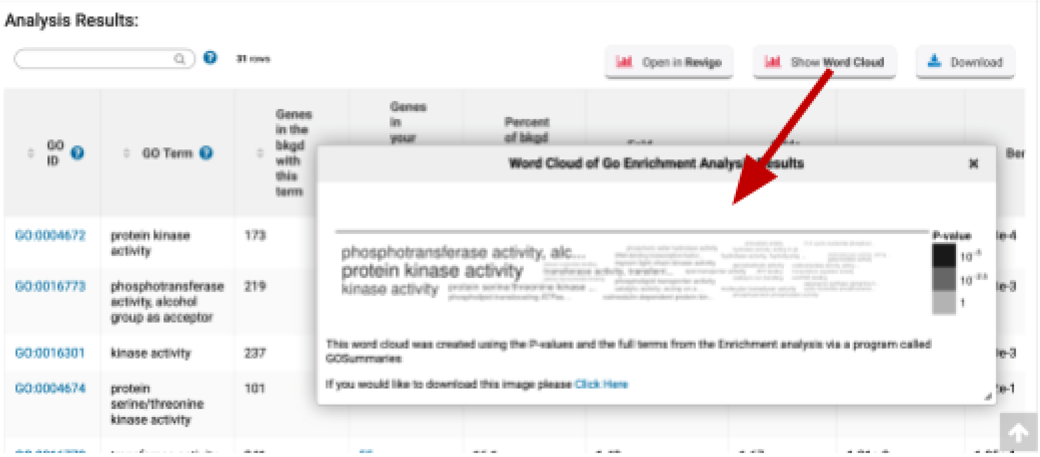
☐ histone acetyltransferase complex
☑ cellular_component
☑ extracellular region
☐ plasma membrane
☑ membrane
☑ integral component of membrane
☐ myosin complex
☑ host cellular component
☑ apical complex
☐ rhoptry
☑ pellicle
☑ host cell cytoplasm
☐ intrinsic component of membrane
☐ TORC2 complex

*The axes have no intrinsic meaning. REVIGO uses multi-dimensional scaling to reduce the dimensionality of a matrix of the GO terms pairwise semantic similarities. The resulting projection may be highly non-linear. The guiding principle is that semantically similar GO terms should remain close together in the plot. Repeated runs of REVIGO may yield different arrangements, but the term distances will remain similar.

Make R script for plotting

f. Try rerunning the GO enrichment analysis, but this time select the Molecular Function ontology. What is the top enriched GO term? What is the p-value for the enrichment? Do you have more or less confidence than in 2c that this function is enriched in your gene set?

g. Click on the "Word Cloud" button above the analysis results. What type of analysis is this? What information can you (See image below).



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

Word Cloud of Go Enrichment Analysis Results

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