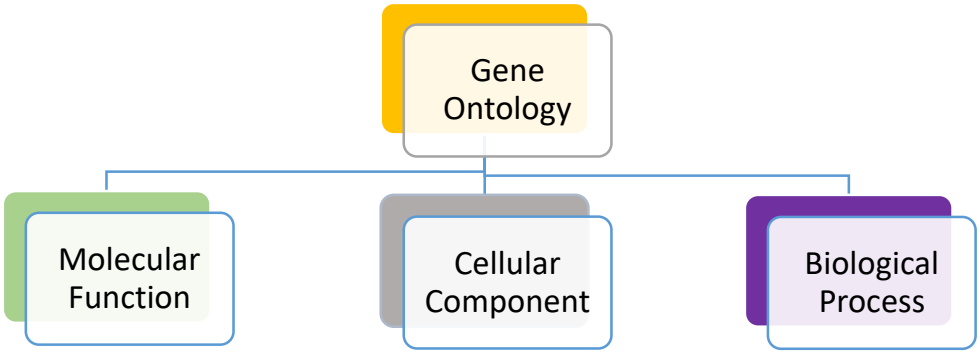
# Gene Ontology (GO) Enrichment

**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.**




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, 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.

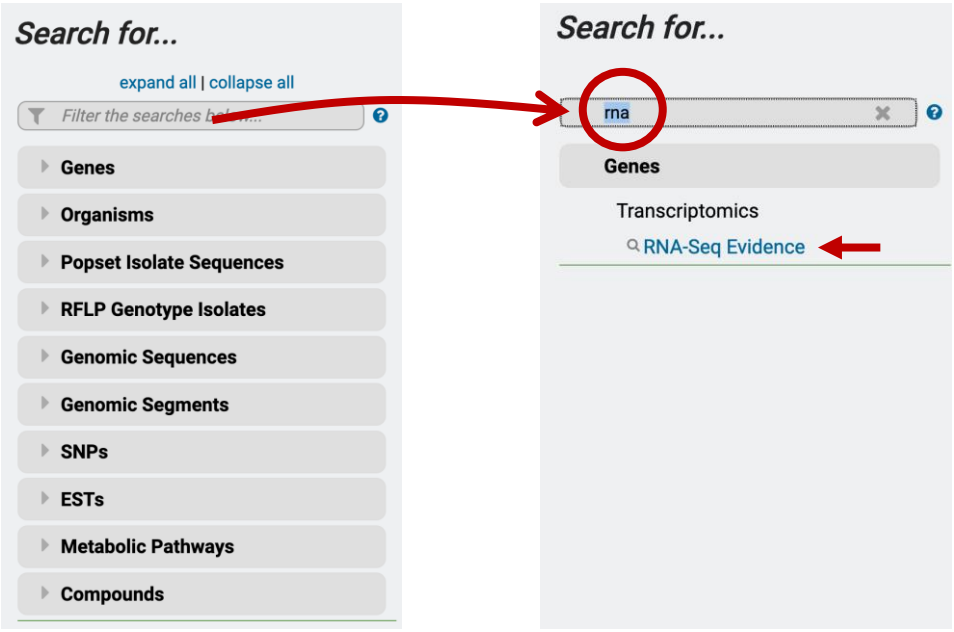
**For example:** A researcher performs a proteomics experiment on a protein fraction collected during an antimalarial treatment and identifies 100 proteins in total. When

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.

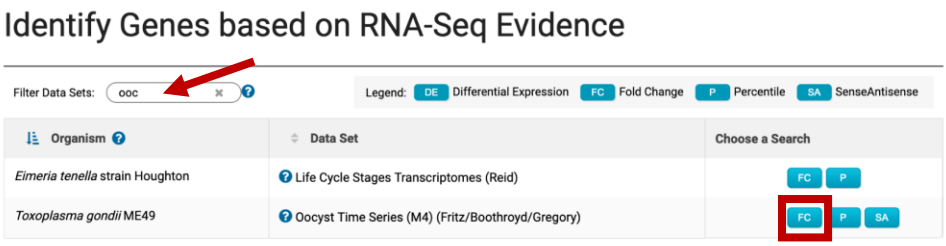
A standard enrichment determination method employs Fisher's exact test, a statistical test that evaluates a 2x2 contingency table (in this case, number of genes in my set *versus* number of genes from genome not in my set, and number of genes with GO term Z *versus* number of genes without term Z). This 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.

1. 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. 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.

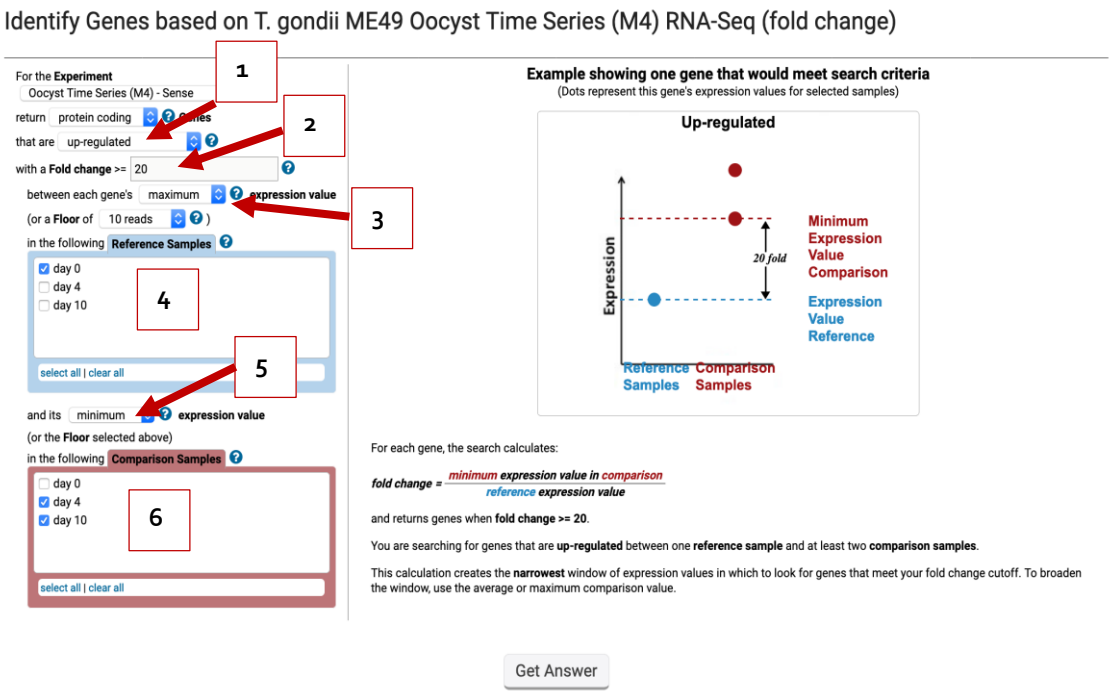


b. The RNA-Seq evidence page includes a list of all data sets that are loaded in the website. To quickly find a dataset, you can start typing key words in the “Filter Data Sets” box. For example, start typing the word “oocyst”.

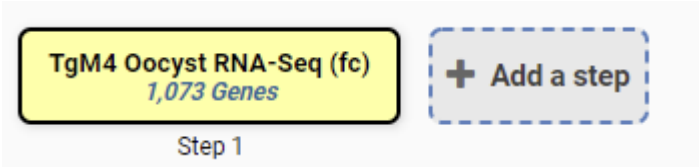



c. Once you find the data set of interest, choose the fold-change (FC) search. For this exercise, identify genes that are upregulated by 20-fold in days 4 and 10 compared to the day 0 time point. Parameters to set:

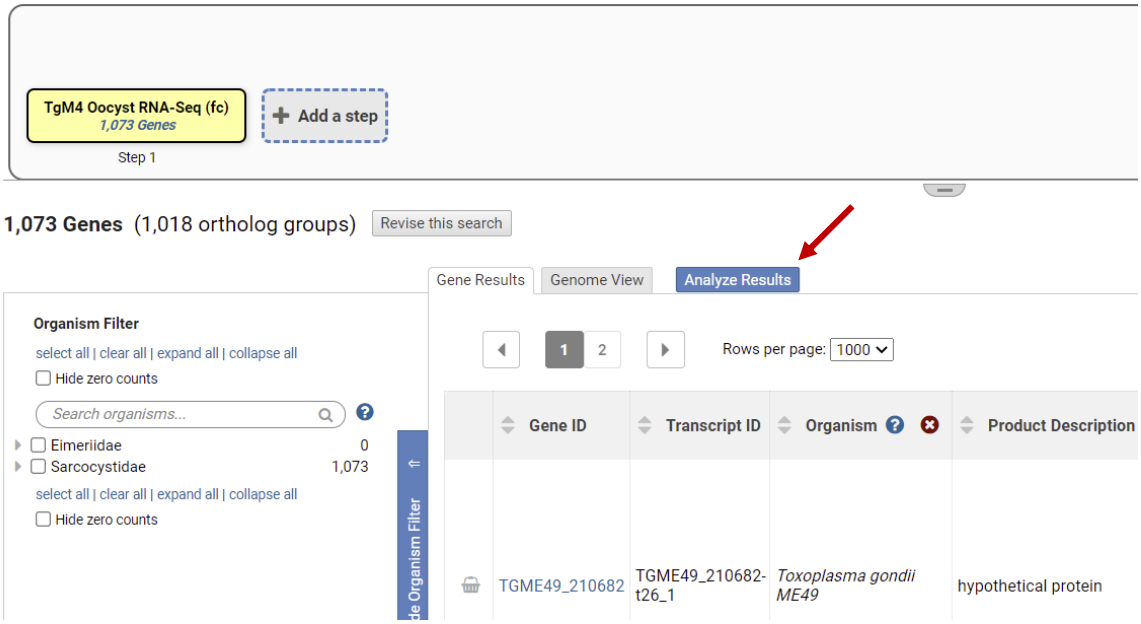
1. Up-regulated
2. 20-fold
3. Maximum
4. Day 0
5. Minimum
6. Day 4 and 10



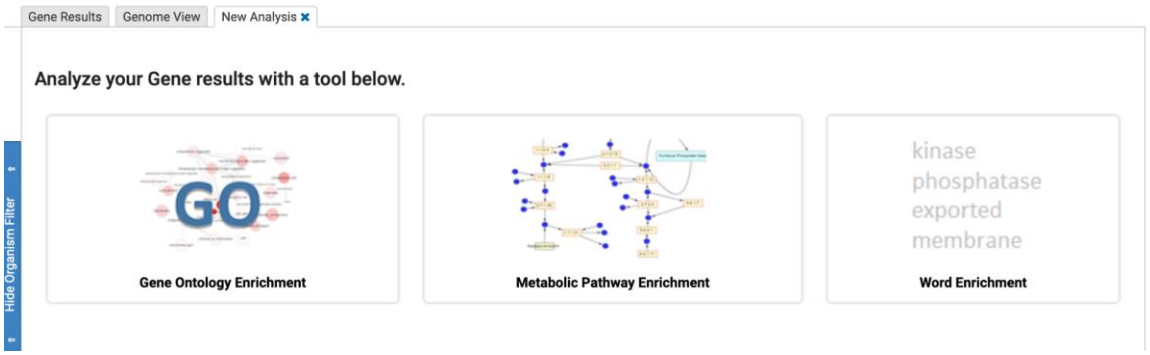

d. Click “Get Answer” to initiate the search. This will return a one-step search strategy. How many genes did you get?



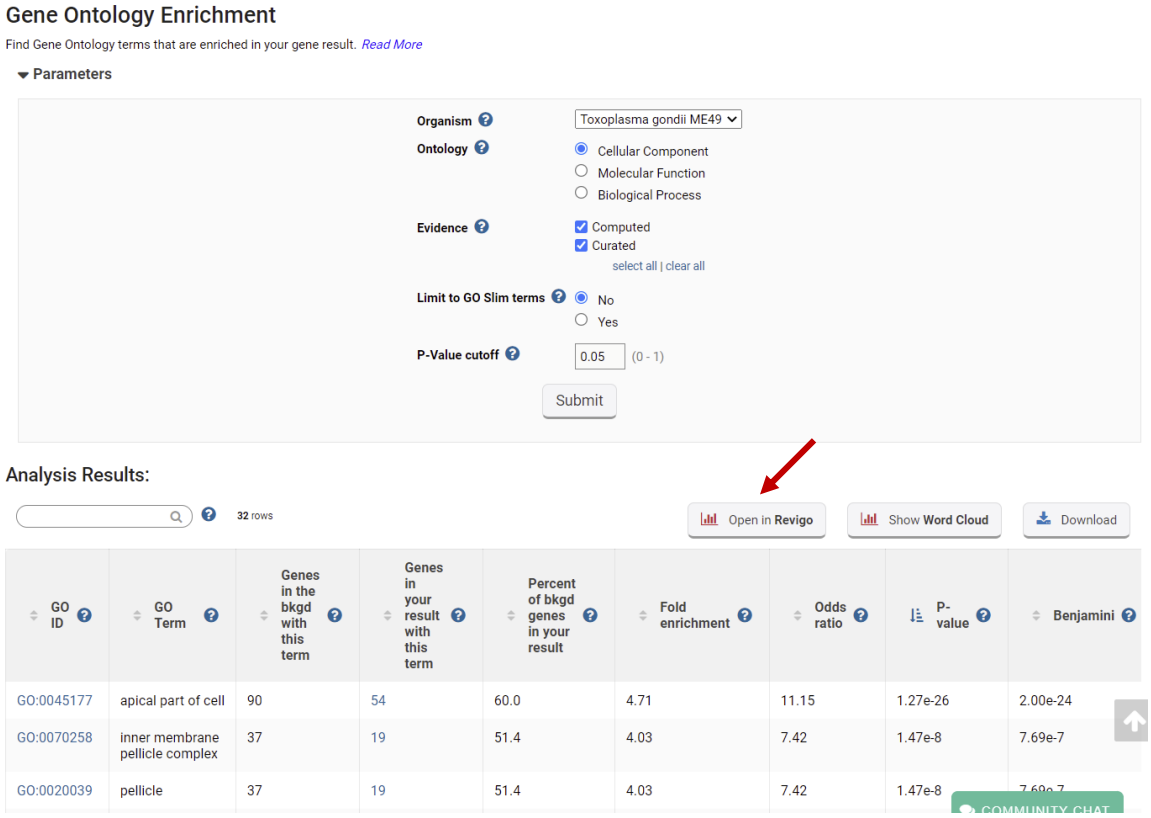

2. To run a GO enrichment analysis on these results, do the following:
   a. Click on the Analyze Results tab just above the list of genes (arrow in image below) to open the enrichment tools. Besides GO enrichment, what other analyses are available?




   b. 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".
   Organism = T. gondii ME49
   Ontology = Cellular Component
   Evidence = Computed and Curated
   Limit to GO Slim terms? = NO



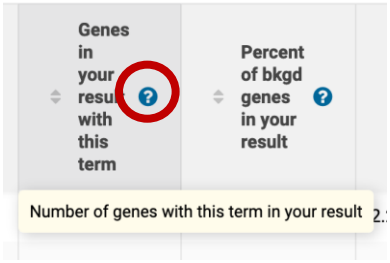

c. What is the top enriched GO term from this analysis? Does this make sense for an enrichment analysis of the cellular component of your Oocyst expressed genes? Notice that the p-value is a rather low, $10^{-24}$.



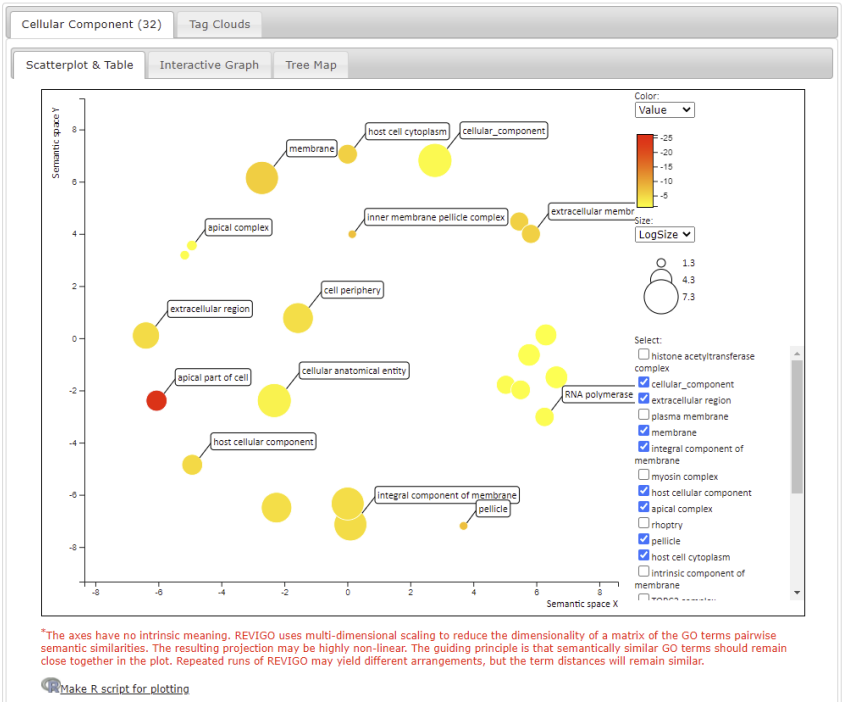

d. What do each of the columns in the analysis table represent? (Hint: move your mouse over the question mark next to each column header)

- 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

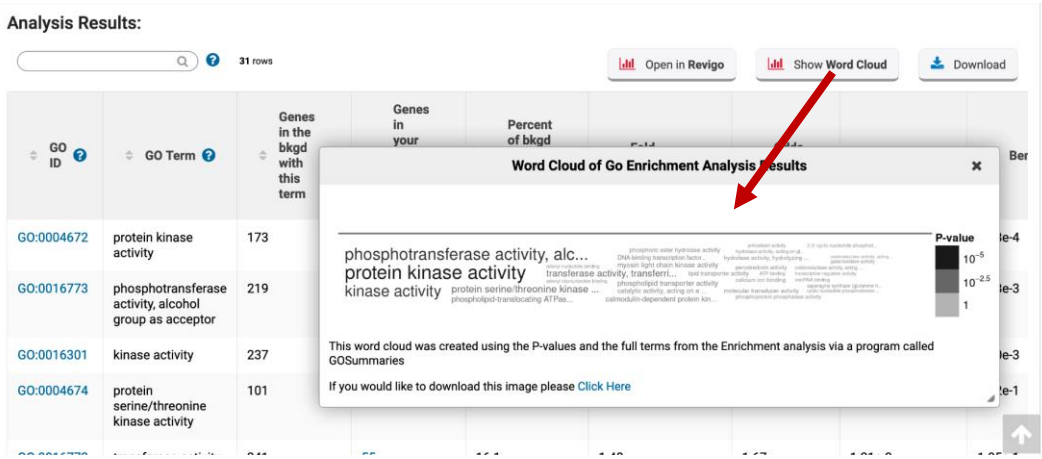



e. Click the Open in Revigo button to port the results to Revigo, the Reduce and Visualize Ontology tool. Once at Revigo, you may need to scroll down to click Start Revigo to run the analysis with default paramters. Revigo provides a scatterplot and table, an integrative map and a tree map to supplement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




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).




**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