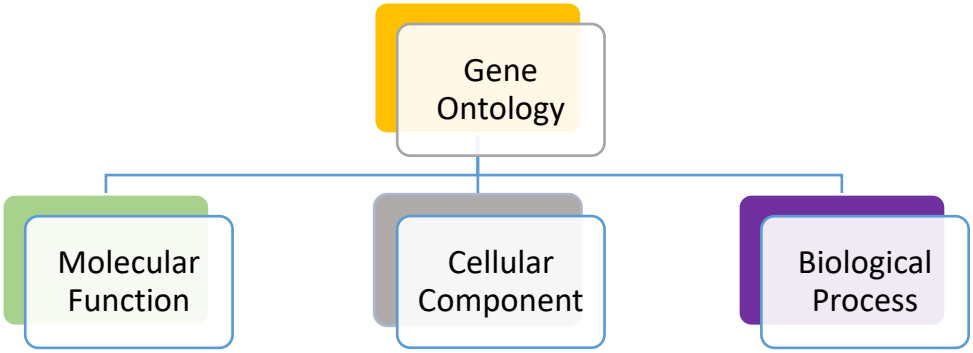
Gene Ontology (GO) Enrichment

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

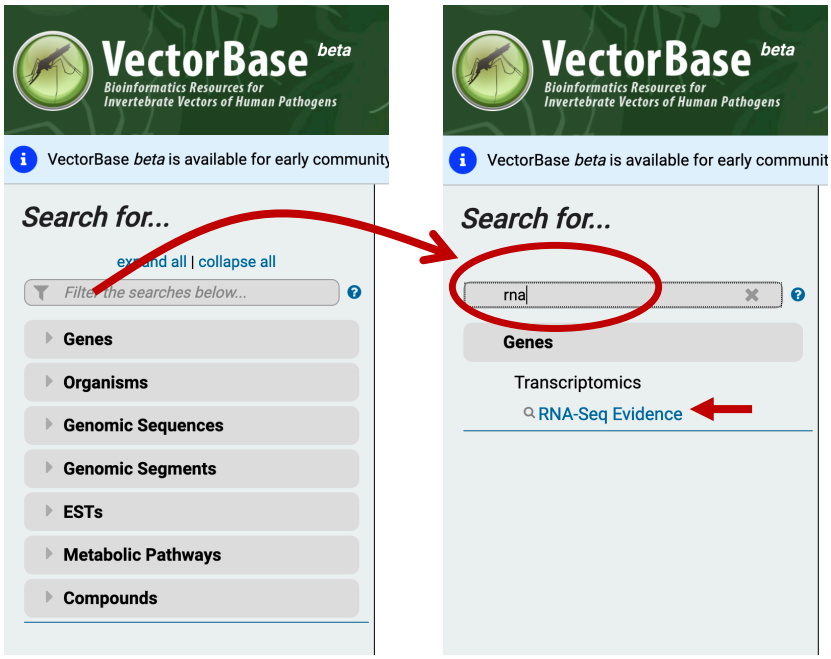
To learn more about Gene ontology please visit: http://geneontology.org/docs/ontology-documentation/

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?

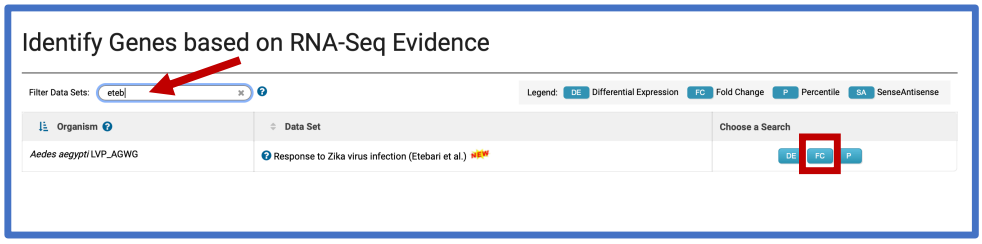
A standard enrichment method is Fisher's exact test which is a statistical test used when analyzing contingency tables. Typically used when you have a small sample size. But when you are doing enrichment analysis on a list of genes with the background being the whole genome, your sample size is not small. 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

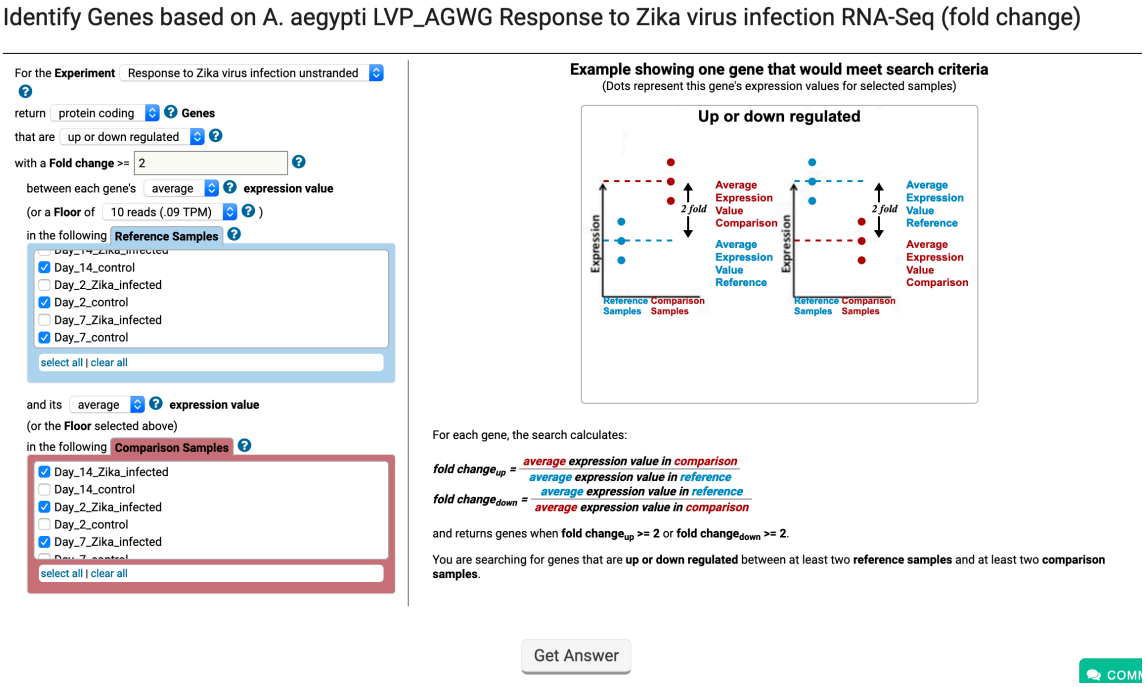
1. In order to run a GO enrichment analysis, we need a list of genes to test. This can be a list of gene IDs from your results that you can upload using the ID search or a gene list resulting from a search you conducted in the database. For this example, in VectorBase, we will identify genes that are differentially regulated when comparing Zika virus infected or uninfected *Aedes aegypti* mosquitos.
    a. Navigate to the RNA-Seq searches and find the data set called "**Response to Zika virus infection**" from Etebari *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 the dataset by Etebari et. al. you can start typing the name 'etebari' in the "Filter Data Sets" box.

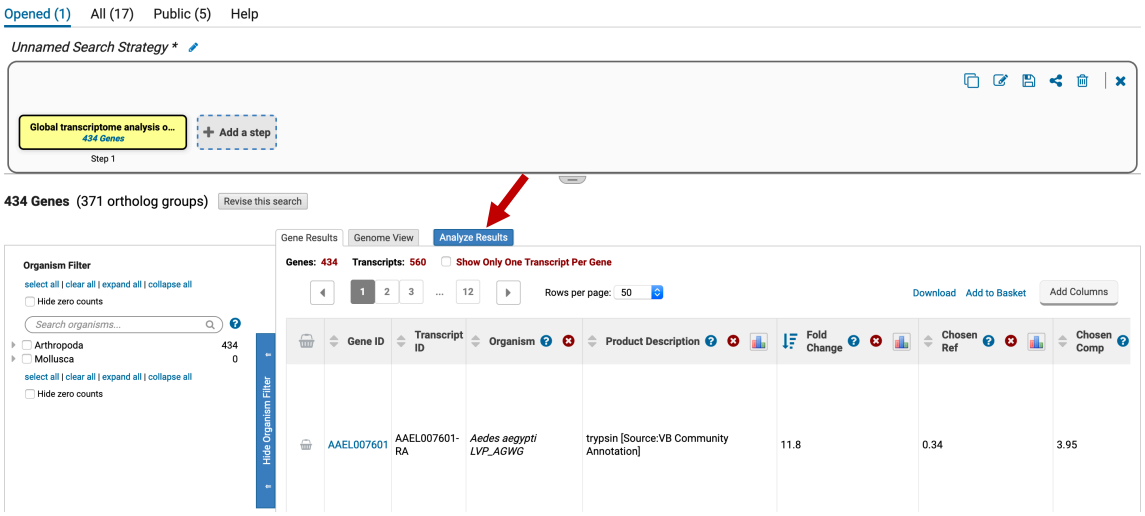


c. Once you find the data set of interest click on the fold change option. This will make available to you all the parameters that you can manipulate to search this data set. For the purpose of this exercise keep all the default parameters and select the control samples as the reference samples and the Zika-infected samples as the comparison samples – see image below.

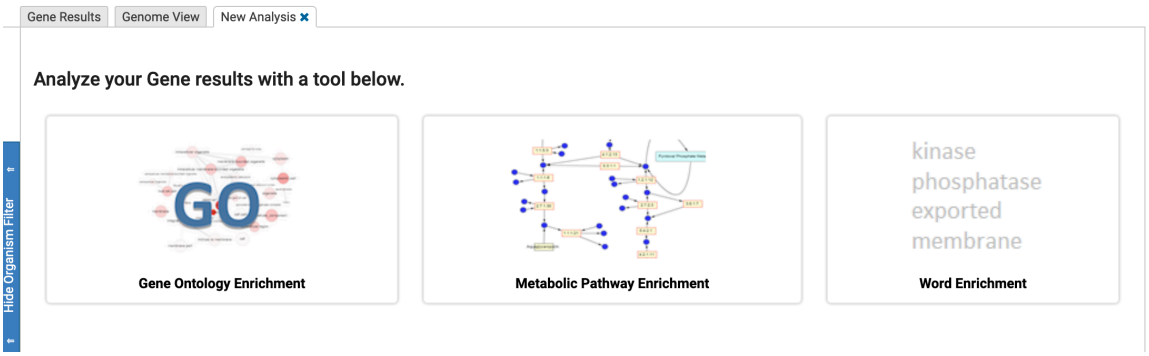


d. Once you have set the parameters you can click on the "Get Answer" button at the bottom of the search. This will return a one-step search strategy. How many genes did you get? (answer: 434)

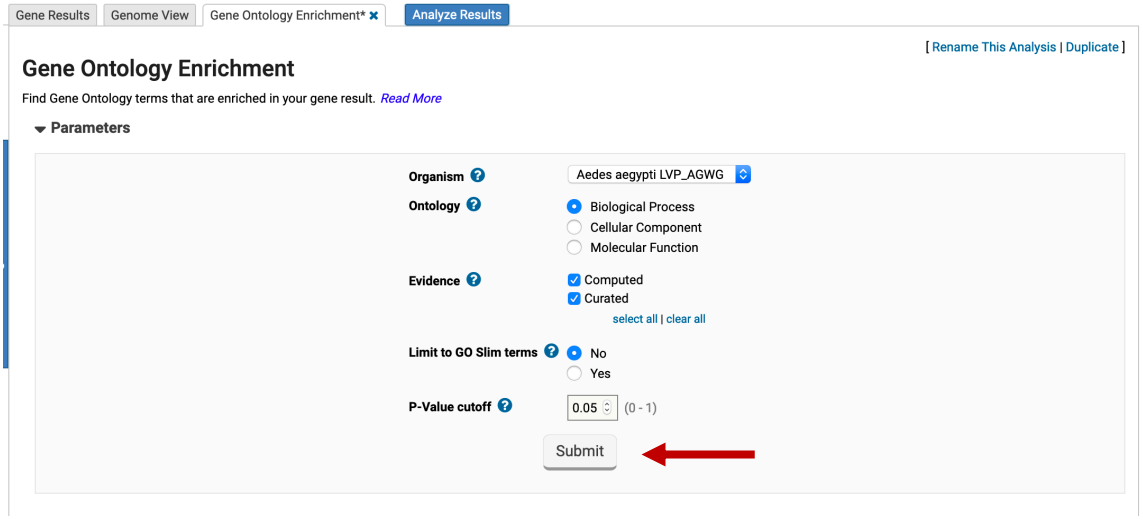
2. To run a GO enrichment analysis on these results, do the following:
   a. Click on the Analyze Results tab right above the list of genes (arrow in image below).



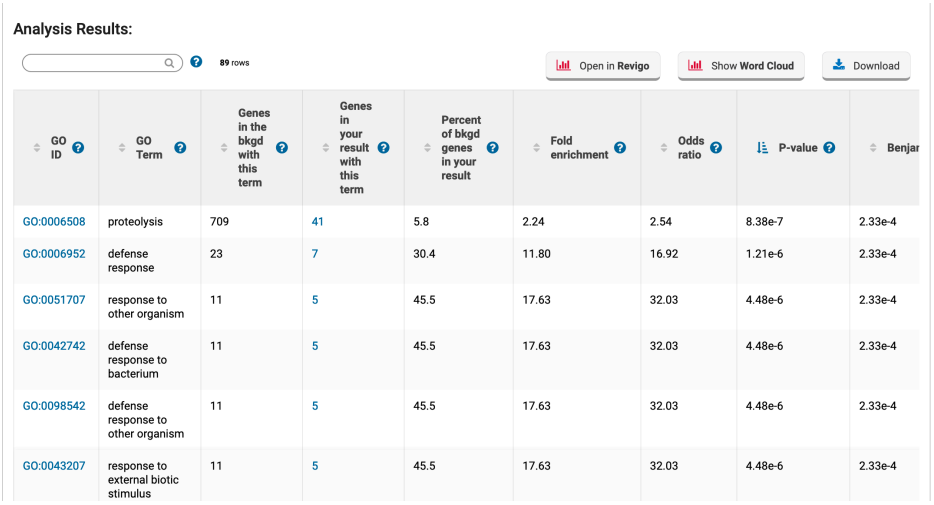
b. Clicking on the “Analyze Results” tab will reveal the different analyses that you can run on your results. Besides GO enrichment what other analyses are available? (Answer: metabolic pathway and word enrichment)



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? (Answer: proteolysis)



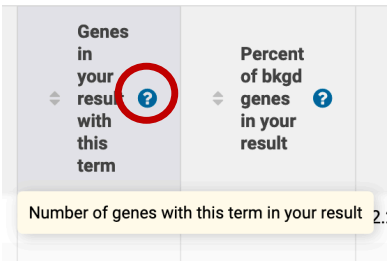
Analysis Results:

89 rows

Open in Revigo | Show Word Cloud | Download

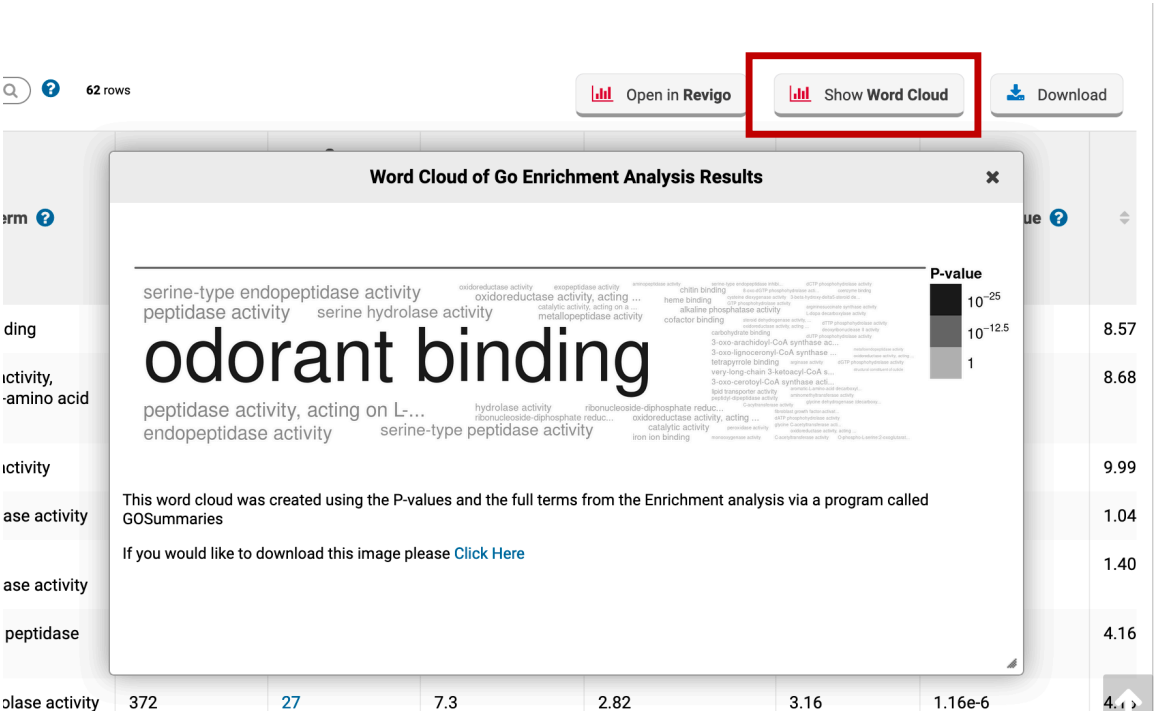
GO ID	GO Term	Genes in the bkgd with this term	Genes in your result with this term	Percent of bkgd genes in your result	Fold enrichment	Odds ratio	P-value	Benjar
GO:0006508	proteolysis	709	41	5.8	2.24	2.54	8.38e-7	2.33e-4
GO:0006952	defense response	23	7	30.4	11.80	16.92	1.21e-6	2.33e-4
GO:0051707	response to other organism	11	5	45.5	17.63	32.03	4.48e-6	2.33e-4
GO:0042742	defense response to bacterium	11	5	45.5	17.63	32.03	4.48e-6	2.33e-4
GO:0098542	defense response to other organism	11	5	45.5	17.63	32.03	4.48e-6	2.33e-4
GO:0043207	response to external biotic stimulus	11	5	45.5	17.63	32.03	4.48e-6	2.33e-4

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)



f. Try rerunning the GO enrichment analysis but this time select the Molecular Function ontology. What is the top enriched GO term? (Answer: odorant binding).

g. Click on the "Word Cloud" button above the analysis results. What does this do? (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