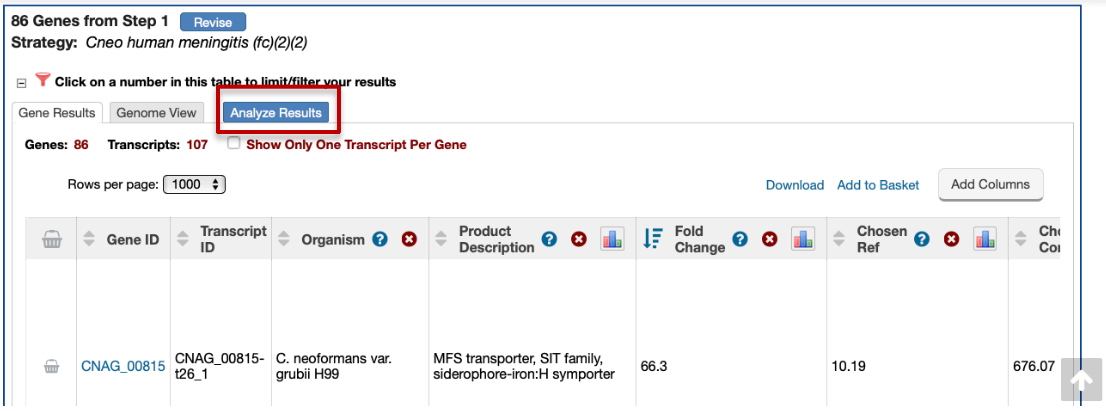
Enrichment Analysis in VEuPathDB

When working with a list of genes such as RNA-Seq results or user-uploaded gene lists one can perform several enrichment analyses to further characterize results into functional categories. Enrichment analysis can be accessed via the blue *Analyze Results* tab and it includes Gene Ontology Enrichment, Metabolic Pathway Enrichment, and Word Enrichment tools.



86 Genes from Step 1

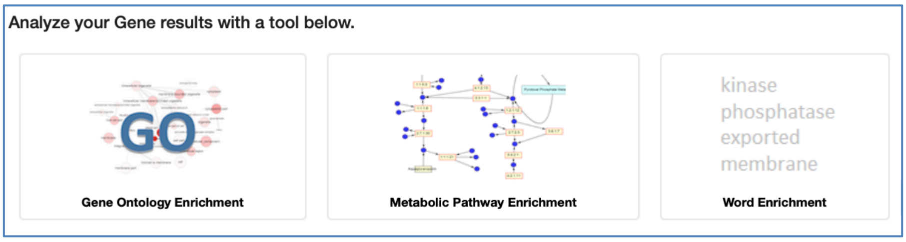
Strategy: *Cneo human meningitis (fc)(2)(2)*

Click on a number in this table to limit/filter your results

Genes: 86 **Transcripts:** 107 ☐ Show Only One Transcript Per Gene

Rows per page: 1000

	Gene ID	Transcript ID	Organism	Product Description	Fold Change	Chosen Ref	Ch Co
	CNAG_00815	CNAG_00815-t26_1	C. neoformans var. grubii H99	MFS transporter, SIT family, siderophore-iron:H symporter	66.3	10.19	676.07



Analyze your Gene results with a tool below.

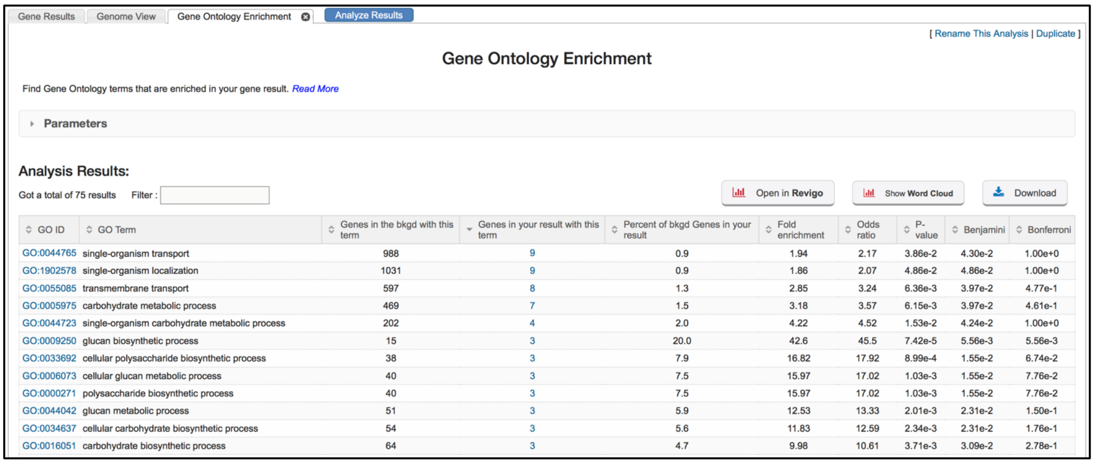
The three types of analysis apply Fisher's Exact test to evaluate ontology terms, over-represented pathways, and product description terms to statistically determine if any annotation terms are enriched. Enrichment is carried out using a Fisher's Exact test with the background defined as all genes from the organism being queried. P-values corrected for multiple testing are provided using both the Benjamini-Hochberg false discovery rate method and the Bonferroni method.

1. GO enrichment

GO enrichment parameters allow users to limit their analysis on either Curated or Computed annotations, or both. Those with a GO evidence code inferred from electronic annotation (IEA) are denoted Computed, while all others have some degree of curation.

Users can also choose to show results for the following functional aspects of the GO ontology: molecular function, cellular component, and biological processes, as well as set a custom P-value cut-off. Gene Ontology terms are standardized phrases that describe the function of gene products. By performing a GO enrichment, you can find gene ontology terms that are enriched in the list of genes generated by your search strategy.

For example, see the Analysis Results table from the GO enrichment focusing on Biological function. The table contains columns with GO IDs and GO terms along with the number of genes in the background and those specific to your analysis results (e.g. RNA-Seq results) are linked in blue.



Gene Ontology Enrichment

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

Analysis Results:

Got a total of 75 results   Filter :

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	Bonferroni
GO:0044765	single-organism transport	988	9	0.9	1.94	2.17	3.86e-2	4.30e-2	1.00e+0
GO:1902578	single-organism localization	1031	9	0.9	1.86	2.07	4.86e-2	4.86e-2	1.00e+0
GO:0055085	transmembrane transport	597	8	1.3	2.85	3.24	6.36e-3	3.97e-2	4.77e-1
GO:0005975	carbohydrate metabolic process	469	7	1.5	3.18	3.57	6.15e-3	3.97e-2	4.61e-1
GO:0044723	single-organism carbohydrate metabolic process	202	4	2.0	4.22	4.52	1.53e-2	4.24e-2	1.00e+0
GO:0009250	glucan biosynthetic process	15	3	20.0	42.6	45.5	7.42e-5	5.56e-3	5.56e-3
GO:0033692	cellular polysaccharide biosynthetic process	38	3	7.9	16.82	17.92	8.99e-4	1.55e-2	6.74e-2
GO:0006073	cellular glucan metabolic process	40	3	7.5	15.97	17.02	1.03e-3	1.55e-2	7.76e-2
GO:0000271	polysaccharide biosynthetic process	40	3	7.5	15.97	17.02	1.03e-3	1.55e-2	7.76e-2
GO:0044042	glucan metabolic process	51	3	5.9	12.53	13.33	2.01e-3	2.31e-2	1.50e-1
GO:0034637	cellular carbohydrate biosynthetic process	54	3	5.6	11.83	12.59	2.34e-3	2.31e-2	1.76e-1
GO:0016051	carbohydrate biosynthetic process	64	3	4.7	9.98	10.61	3.71e-3	3.09e-2	2.78e-1

When the GO Slim option is chosen both the genes of interest and the background are limited to GO terms that are part of the generic GO Slim subset. Users may download a GO enrichment table (with the Gene IDs for each GO term added) as well as view and download a word cloud produced via the GO Summaries R package.

Several additional statistical measurements are also included and are defined below:

- **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** - Determines if the odds of the GO term appearing in the list of interest are the same as that for the background list;
- **P-value** - Assumptions under a null hypothesis, 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;

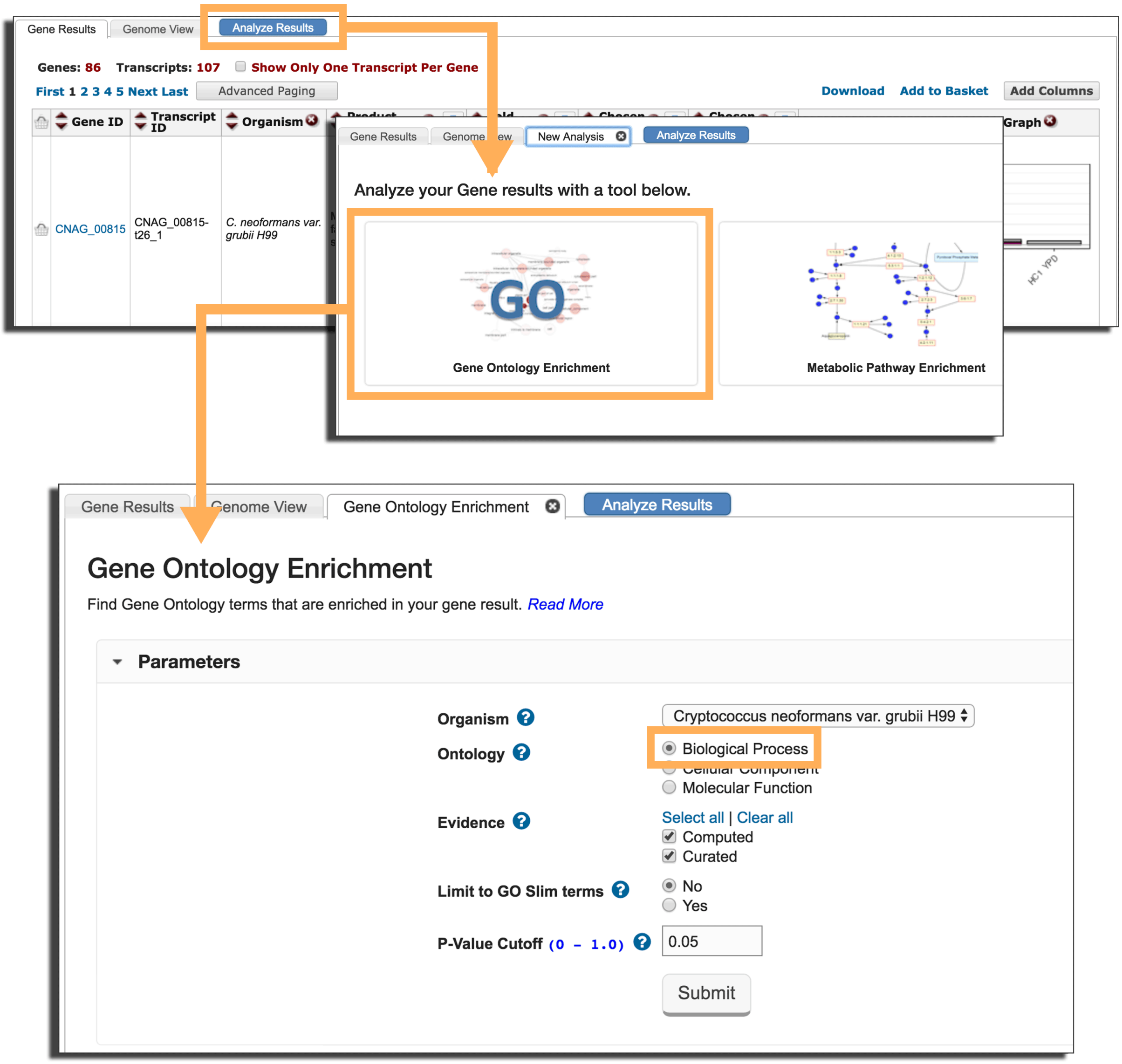
1. Perform GO enrichment analysis

Sample search strategy:
https://fungidb.org/fungidb/im.do?s=bc266b3f01eae513

The strategy linked above returns a list of genes up-regulated in *C. neoformans* when cells are grown in cerebrospinal fluid (single-read data) and, subsequently, looks for any genes that have a predicted signal peptide.

- Perform GO term enrichment on Step 1 by navigating to the results table, selecting *Analyze Results* tab and clicking on *GO Enrichment* button.

- Select *Biological processes* to determine which gene ontology terms are enriched in the list of genes upregulated when *Cryptococcus neoformans* is grown in cerebrospinal fluid and click on the *Submit* button.



Genes: 86 Transcripts: 107 ☐ Show Only One Transcript Per Gene

Gene ID	Transcript ID	Organism
CNAG_00815	CNAG_00815-t26_1	*C. neoformans var. grubii H99*

Analyze your Gene results with a tool below.

Gene Ontology Enrichment

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

Organism: Cryptococcus neoformans var. grubii H99

Ontology: ● Biological Process ○ Cellular Component ○ Molecular Function

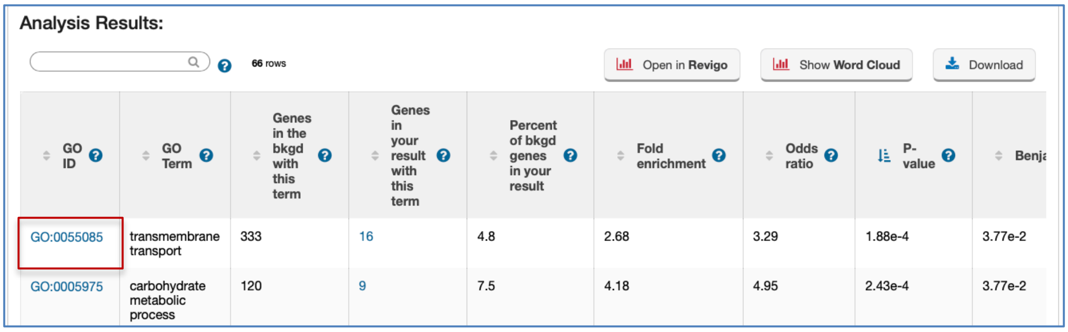
Evidence: ☑ Computed ☑ Curated

Limit to GO Slim terms: ● No ○ Yes

P-Value Cutoff `(0 - 1.0)`: 0.05

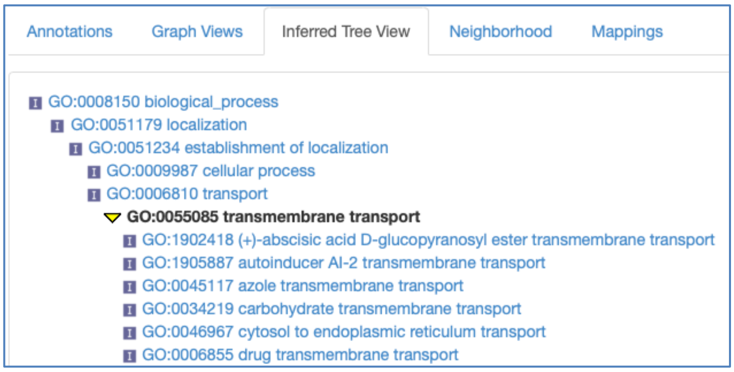
- What kind of biological processes are enriched for the upregulated *C. neoformans* genes? Cerebrospinal fluid is mostly composed of water and carries proteins, glucose, neurotransmitters, and various ions including sodium, magnesium, and chloride. Do the GO enrichment results seem appropriate given the conditions *C. neoformans* would encounter in cerebrospinal fluid?

Note: Clicking on a GO ID opens the associated Inferred Tree View within the AmiGO 2 record page. This offers an image of the ID's position in the hierarchical ontology:



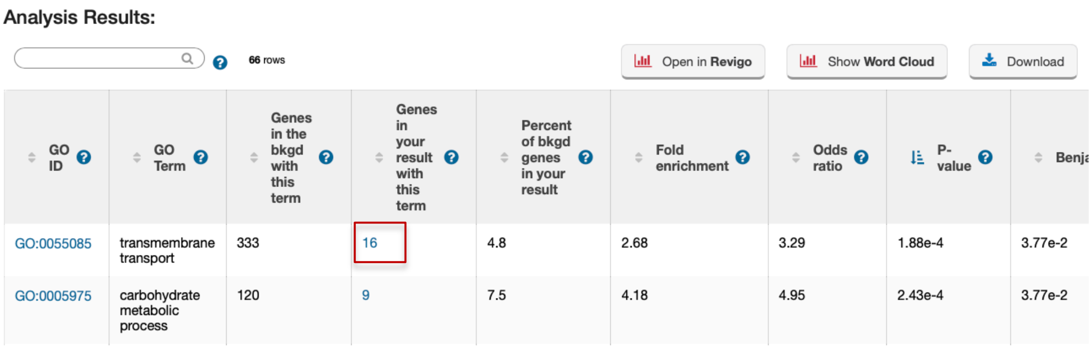
Analysis Results:

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	Benja
GO:0055085	transmembrane transport	333	16	4.8	2.68	3.29	1.88e-4	3.77e-2
GO:0005975	carbohydrate metabolic process	120	9	7.5	4.18	4.95	2.43e-4	3.77e-2



- GO:0008150 biological_process
  - GO:0051179 localization
    - GO:0051234 establishment of localization
      - GO:0009987 cellular process
      - GO:0006810 transport
        - **GO:0055085 transmembrane transport**
          - GO:1902418 (+)-abscisic acid D-glucopyranosyl ester transmembrane transport
          - GO:1905887 autoinducer AI-2 transmembrane transport
          - GO:0045117 azole transmembrane transport
          - GO:0034219 carbohydrate transmembrane transport
          - GO:0046967 cytosol to endoplasmic reticulum transport
          - GO:0006855 drug transmembrane transport

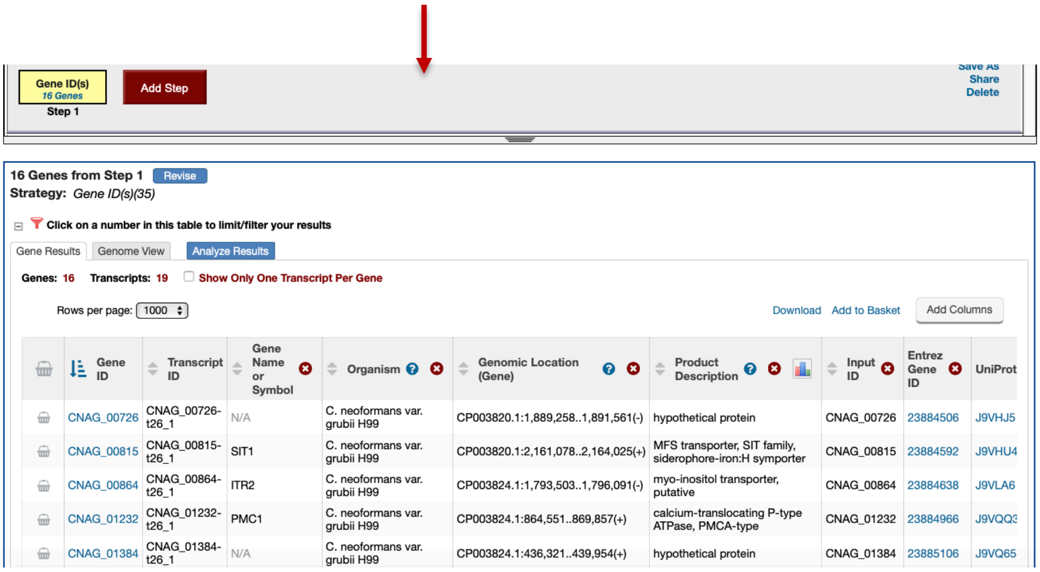
Note: Clicking on the number linked in the "Genes in your result with this term" column will automatically create a new search strategy with the list of GeneIDs in that category:



Analysis Results:

66 rows

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	Benja
GO:0055085	transmembrane transport	333	16	4.8	2.68	3.29	1.88e-4	3.77e-2
GO:0005975	carbohydrate metabolic process	120	9	7.5	4.18	4.95	2.43e-4	3.77e-2



16 Genes from Step 1

Strategy: *Gene ID(s)(35)*

Click on a number in this table to limit/filter your results

Genes: 16 **Transcripts:** 19 ☐ Show Only One Transcript Per Gene

Rows per page: 1000

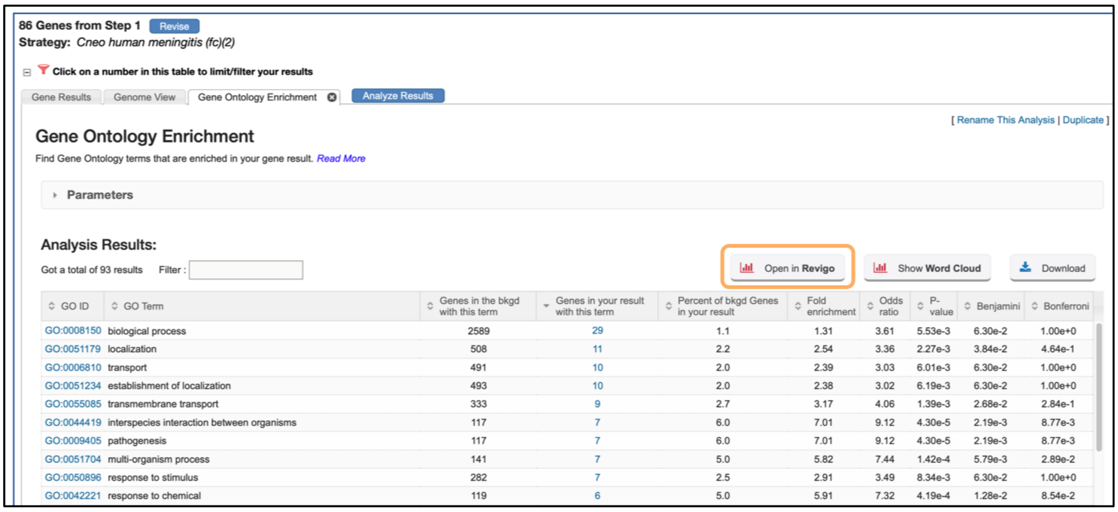
	Gene ID	Transcript ID	Gene Name or Symbol	Organism	Genomic Location (Gene)	Product Description	Input ID	Entrez Gene ID	UniProt
	CNAG_00726	CNAG_00726-t26_1	N/A	C. neoformans var. grubii H99	CP003820.1:1,889,258..1,891,561(-)	hypothetical protein	CNAG_00726	23884506	J9VHJ5
	CNAG_00815	CNAG_00815-t26_1	SIT1	C. neoformans var. grubii H99	CP003820.1:2,161,078..2,164,025(+)	MFS transporter, SIT family, siderophore-iron:H symporter	CNAG_00815	23884592	J9VHU4
	CNAG_00864	CNAG_00864-t26_1	ITR2	C. neoformans var. grubii H99	CP003824.1:1,793,503..1,796,091(-)	myo-inositol transporter, putative	CNAG_00864	23884638	J9VLA6
	CNAG_01232	CNAG_01232-t26_1	PMC1	C. neoformans var. grubii H99	CP003824.1:864,551..869,857(+)	calcium-translocating P-type ATPase, PMCA-type	CNAG_01232	23884966	J9VQQ3
	CNAG_01384	CNAG_01384-t26_1	N/A	C. neoformans var. grubii H99	CP003824.1:436,321..439,954(+)	hypothetical protein	CNAG_01384	23885106	J9VQ65

Note: If you would like to create a new GO enrichment analysis using the same gene list (e.g. perform enrichment for “Cellular Component”), you can duplicate the gene list by clicking on the “Duplicate” link. Each GO enrichment analysis can be given a unique name via “Rename This Analysis” option.

- Visualize enrichment analysis with Revigo

GO enrichment results can be also visualized graphically by transferring the results to an external Reduce + Visualize Gene Ontology (REViGO) website. To do this click on the *Open in Revigo* button.

REViGO visualizes the GO enrichment results while removing redundant GO terms. The resulting similarity-based scatterplots are color-coded and provide interactive graphs and tag clouds as alternative visualizations. Explore your results, learn about color scale and opacity settings, etc. More information is available at Revigo website: http://revigo.irb.hr



86 Genes from Step 1 Revise

Strategy: *Cneo human meningitis (fc)(2)*

Click on a number in this table to limit/filter your results

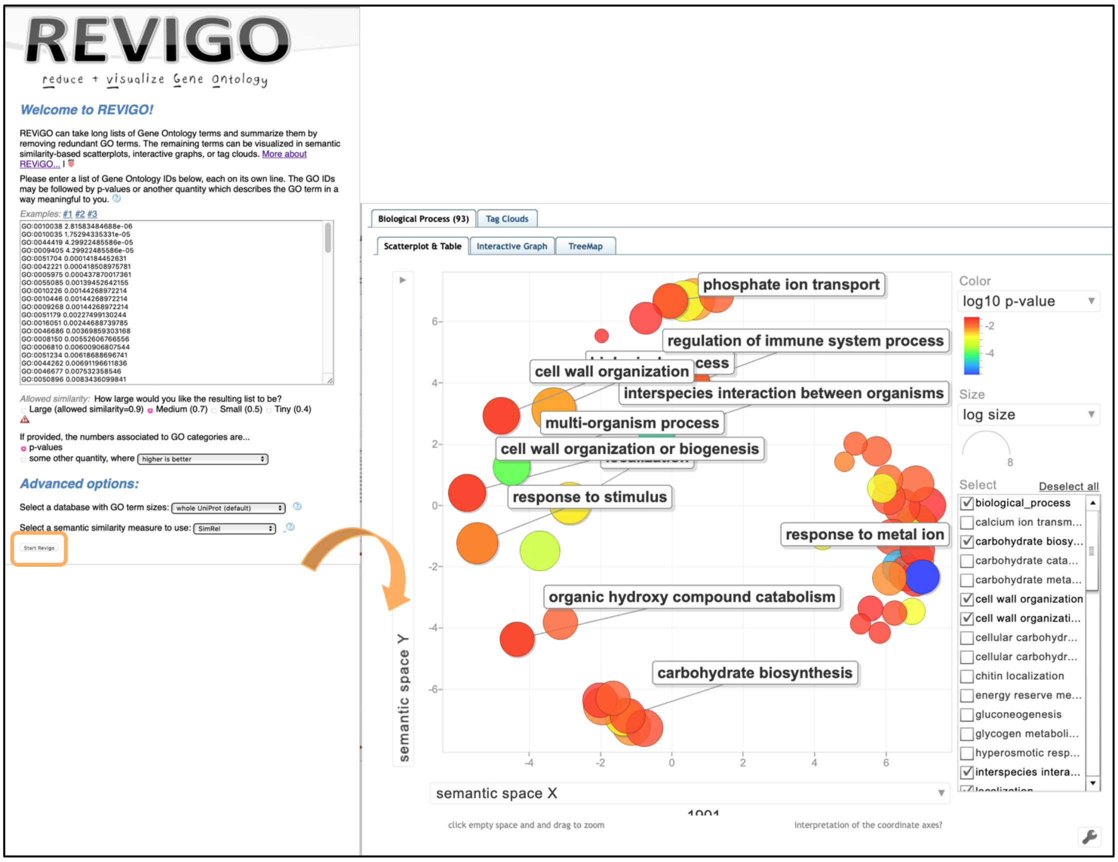
Gene Ontology Enrichment

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

Analysis Results:

Got a total of 93 results Filter :

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	Bonferroni
GO:0008150	biological process	2589	29	1.1	1.31	3.61	5.53e-3	6.30e-2	1.00e+0
GO:0051179	localization	508	11	2.2	2.54	3.36	2.27e-3	3.84e-2	4.64e-1
GO:0006810	transport	491	10	2.0	2.39	3.03	6.01e-3	6.30e-2	1.00e+0
GO:0051234	establishment of localization	493	10	2.0	2.38	3.02	6.19e-3	6.30e-2	1.00e+0
GO:0055085	transmembrane transport	333	9	2.7	3.17	4.06	1.39e-3	2.68e-2	2.84e-1
GO:0044419	interspecies interaction between organisms	117	7	6.0	7.01	9.12	4.30e-5	2.19e-3	8.77e-3
GO:0009405	pathogenesis	117	7	6.0	7.01	9.12	4.30e-5	2.19e-3	8.77e-3
GO:0051704	multi-organism process	141	7	5.0	5.82	7.44	1.42e-4	5.79e-3	2.89e-2
GO:0050896	response to stimulus	282	7	2.5	2.91	3.49	8.34e-3	6.30e-2	1.00e+0
GO:0042221	response to chemical	119	6	5.0	5.91	7.32	4.19e-4	1.28e-2	8.54e-2



Welcome to REVIGO!

REViGO can take long lists of Gene Ontology terms and summarize them by removing redundant GO terms. The remaining terms can be visualized in semantic similarity-based scatterplots, interactive graphs, or tag clouds. More about REViGO... |

Please enter a list of Gene Ontology IDs below, each on its own line. The GO IDs may be followed by p-values or another quantity which describes the GO term in a way meaningful to you.

Examples: #1 #2 #3

```
GO:0010038 2.81583484688e-06
GO:0010035 1.75294335331e-05
GO:0044419 4.29922485586e-05
GO:0009405 4.29922485586e-05
GO:0051704 0.00014184452631
GO:0042221 0.000418508975781
GO:0005975 0.000437870017361
GO:0055085 0.00139452642155
GO:0010226 0.00144268972214
GO:0010446 0.00144268972214
GO:0009268 0.00144268972214
GO:0051179 0.00227499130244
GO:0016051 0.00244688739785
GO:0046686 0.00369859303168
GO:0008150 0.00552606766556
GO:0006810 0.00600906807544
GO:0051234 0.00618688696741
GO:0044262 0.00691196611836
GO:0046677 0.007532358546
GO:0050896 0.0083436099841
```

Allowed similarity: How large would you like the resulting list to be?

○ Large (allowed similarity=0.9) ● Medium (0.7) ○ Small (0.5) ○ Tiny (0.4)

If provided, the numbers associated to GO categories are...

● p-values

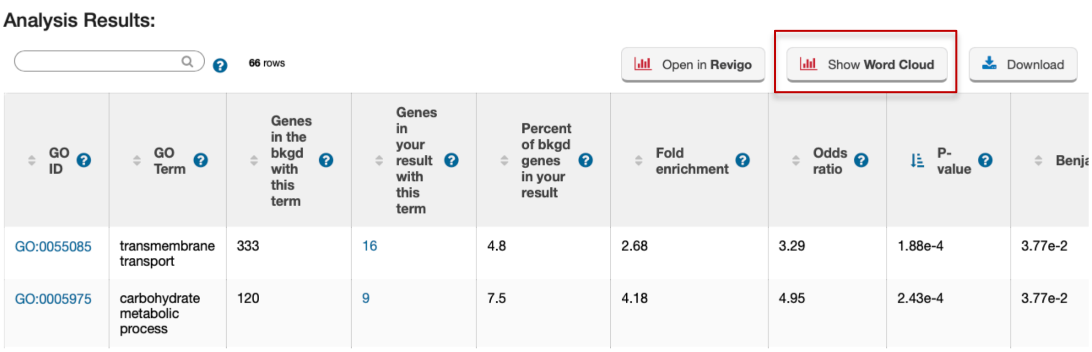
○ some other quantity, where higher is better

Advanced options:

Select a database with GO term sizes: whole UniProt (default)

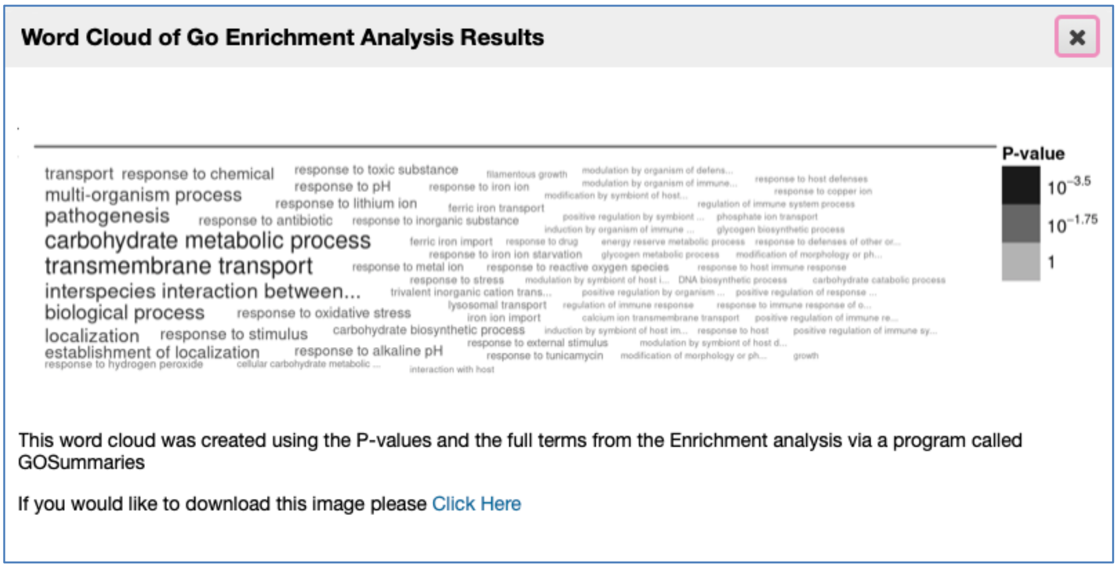
Select a semantic similarity measure to use: SimRel

Note that in addition to REViGO analysis you can also display enrichment results as Word cloud. You can also deploy Word Cloud on GO enrichment analysis from VeEuPathDB analysis pages:



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	Benja
GO:0055085	transmembrane transport	333	16	4.8	2.68	3.29	1.88e-4	3.77e-2
GO:0005975	carbohydrate metabolic process	120	9	7.5	4.18	4.95	2.43e-4	3.77e-2



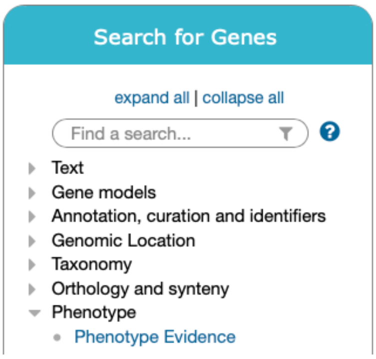
Word Cloud of Go Enrichment Analysis Results

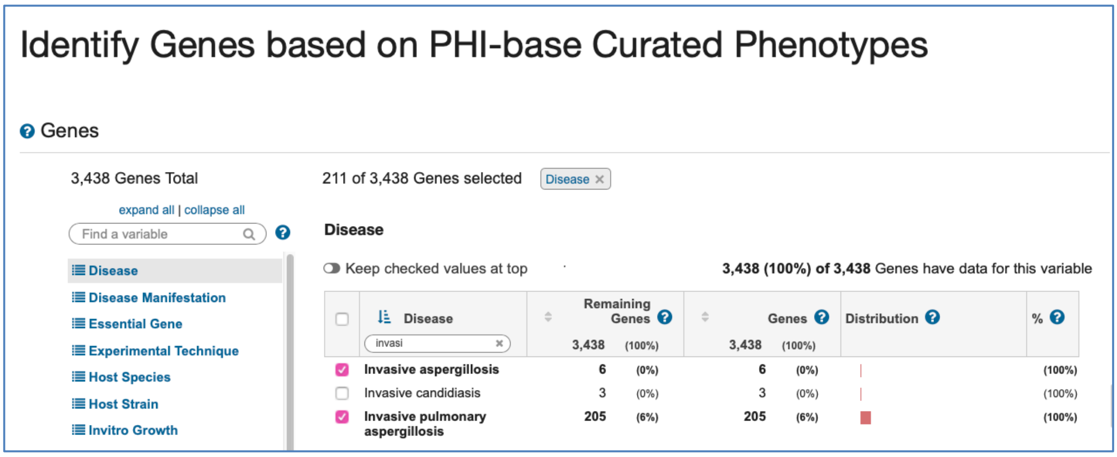
This word cloud was created using the P-values and the full terms from the Enrichment analysis via a program called GOSummaries

If you would like to download this image please Click Here

2. Metabolic Pathway enrichment

Sample strategy: https://fungidb.org/fungidb/im.do?s=c02bd81f37d7e95e
This search uses Phenotype Evidence dataset from PHI-Base to identify genes implicated in invasive aspergillosis:





Identify Genes based on PHI-base Curated Phenotypes

Genes

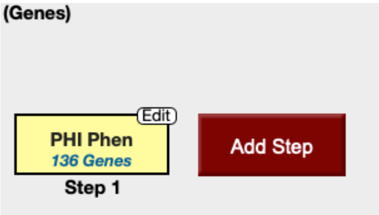
3,438 Genes Total

211 of 3,438 Genes selected Disease ×

Disease

3,438 (100%) of 3,438 Genes have data for this variable

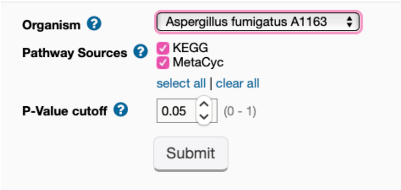
☐	Disease	Remaining Genes		Genes		Distribution	%
	invasi	3,438	(100%)	3,438	(100%)		
☑	Invasive aspergillosis	6	(0%)	6	(0%)		(100%)
☐	Invasive candidiasis	3	(0%)	3	(0%)		(100%)
☑	Invasive pulmonary aspergillosis	205	(6%)	205	(6%)		(100%)



- Navigate to the Analyze Results tab and deploy Metabolic Pathway Enrichment analysis.

  Note that the search returned genes for several *Aspergillus* species (also can be viewed in the Organism parameter drop down menu). For statistical reasons, the analysis can only be run on a set of genes from one organism at a time. Use “Duplicate” and “Rename This Analysis” options to create and track of analysis carried out on multiple species.

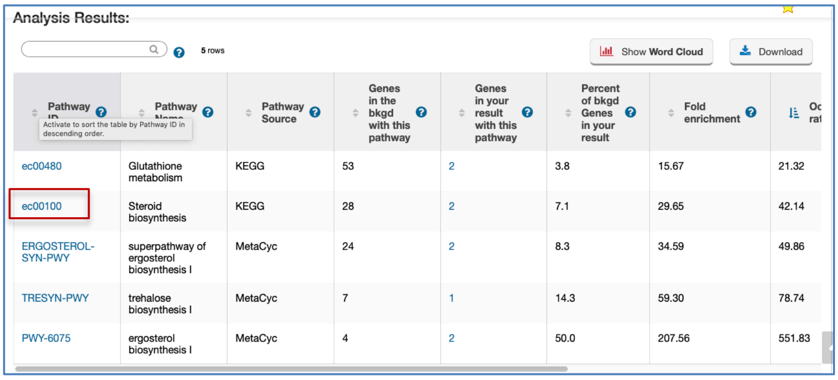
- Deploy Metabolic Pathway Enrichment analysis for *A. fumigatus*.



Organism: Aspergillus fumigatus A1163

Pathway Sources: ☑ KEGG ☑ MetaCyc

P-Value cutoff: 0.05 (0 - 1)

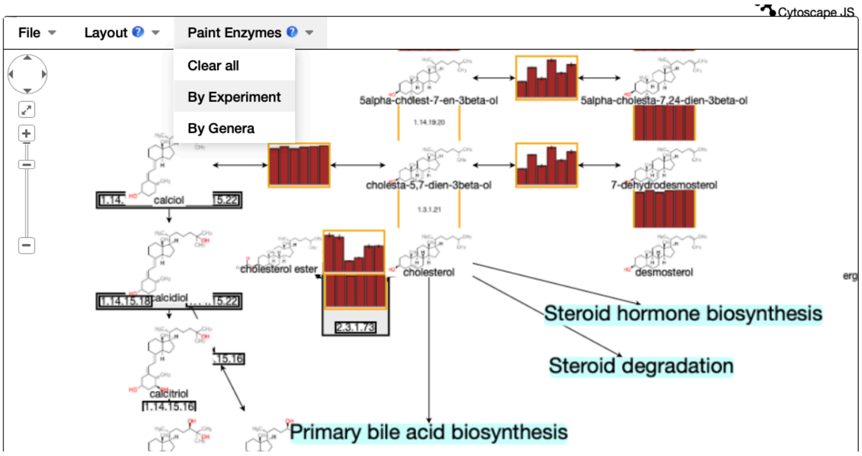


Analysis Results:

5 rows

Pathway ID	Pathway Name	Pathway Source	Genes in the bkgd with this pathway	Genes in your result with this pathway	Percent of bkgd Genes in your result	Fold enrichment	Od rat
ec00480	Glutathione metabolism	KEGG	53	2	3.8	15.67	21.32
ec00100	Steroid biosynthesis	KEGG	28	2	7.1	29.65	42.14
ERGOSTEROL-SYN-PWY	superpathway of ergosterol biosynthesis I	MetaCyc	24	2	8.3	34.59	49.86
TRESYN-PWY	trehalose biosynthesis I	MetaCyc	7	1	14.3	59.30	78.74
PWY-6075	ergosterol biosynthesis I	MetaCyc	4	2	50.0	207.56	551.83

- Pathway IDs are linked to Metabolic pathways record pages that also provide a visual representation of the pathway in Cytoscape JS, which offers interactive display of the metabolic pathways, including an ability to paint experimental data (e.g. gene expression) over each node. For example, you can paint the expression data from the “*A. fumigatus* Af293 Transcriptomes of intraconazole-resistant strains” (Paint > By Experiment).



Hint: Use filter “fumig” in the dataset selection pop-up window to quickly bring up datasets for A*. fumigatus*.

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