FungiDB: GO Enrichment Analysis

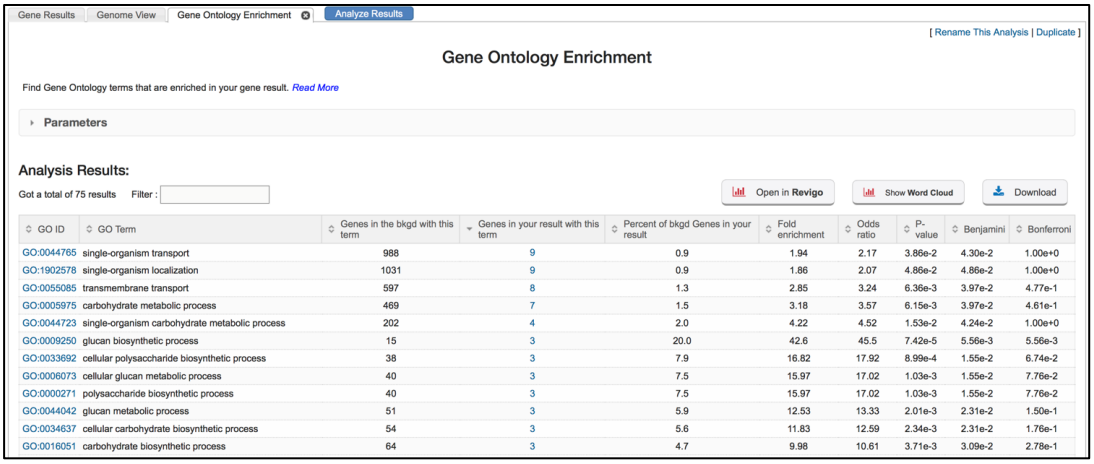
When working with a list of genes such as RNA-Seq results or user-uploaded gene lists one can perform several enrichment analyses to further characterize results into functional categories. Enrichment analysis can be accessed via the blue Analyze Results tab and it includes Gene Ontology, Metabolic Pathway, and Word Enrichment tools. The three types of analysis apply Fisher's Exact test to evaluate ontology terms, over-represented pathways, and product description terms. 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.

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.

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.

For example, the Analysis Results table from the GO enrichment focusing on Biological function (below), contains columns for GO IDs, GO terms, number of genes in the background and those specific to the analyzed gene list.


Gene Results | Genome View | Gene Ontology Enrichment | Analyze Results

[Rename This Analysis | Duplicate]

Gene Ontology Enrichment

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

Parameters

Analysis Results:

Got a total of 75 results Filter : 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	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

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 on previously generated search strategy of *C. neoformans* transcriptomics dataset.

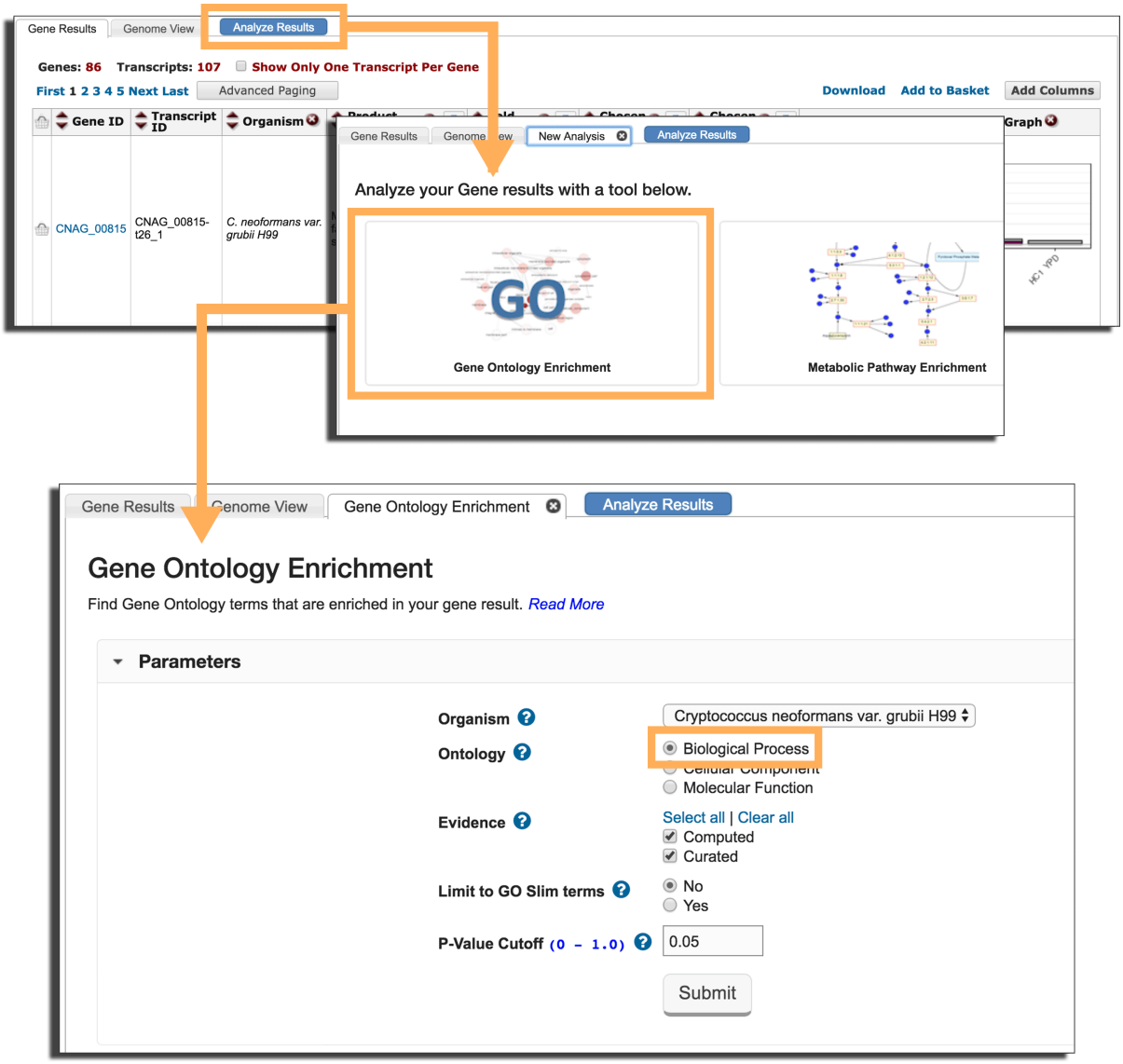
- Perform GO term enrichment on **Step 1** (click on the results in step 1 before proceeding)**.**

➢ Search strategy: https://fungidb.org/fungidb/im.do?s=69a8dd401983653c

- Navigate to the *Analyze Results* tab and click on *GO Enrichment* button.

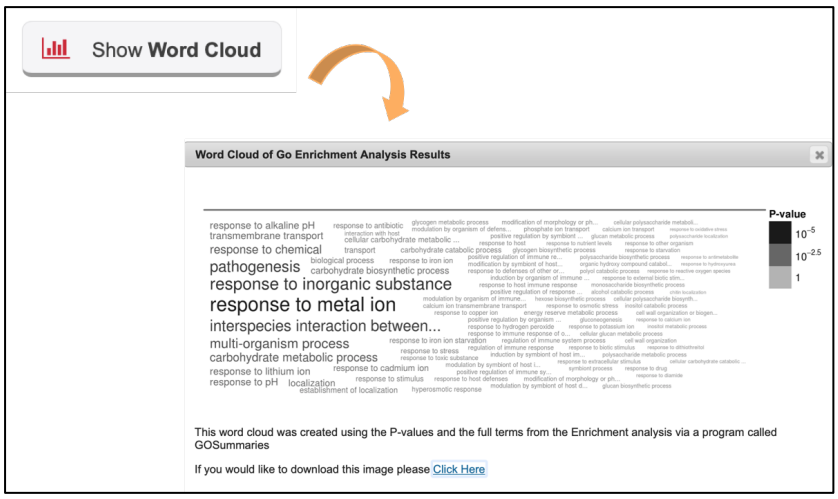
- Perform Gene Ontology (GO) Term enrichment in FungiDB. "Gene Ontology" terms are standardized phrases that describe the function of gene products. By performing a GO enrichment, you can find gene ontology terms that are enriched in the list of genes. This can help you infer if a group of genes participates in certain functions or biological processes in the cell.

- To run a GO enrichment, click on the "Analyze Results" button, and select "Gene Ontology Enrichment". Since we are looking for biological processes that are associated with the upregulated *C. neoformans* genes, select the "Biological Process" ontology and submit the analysis.



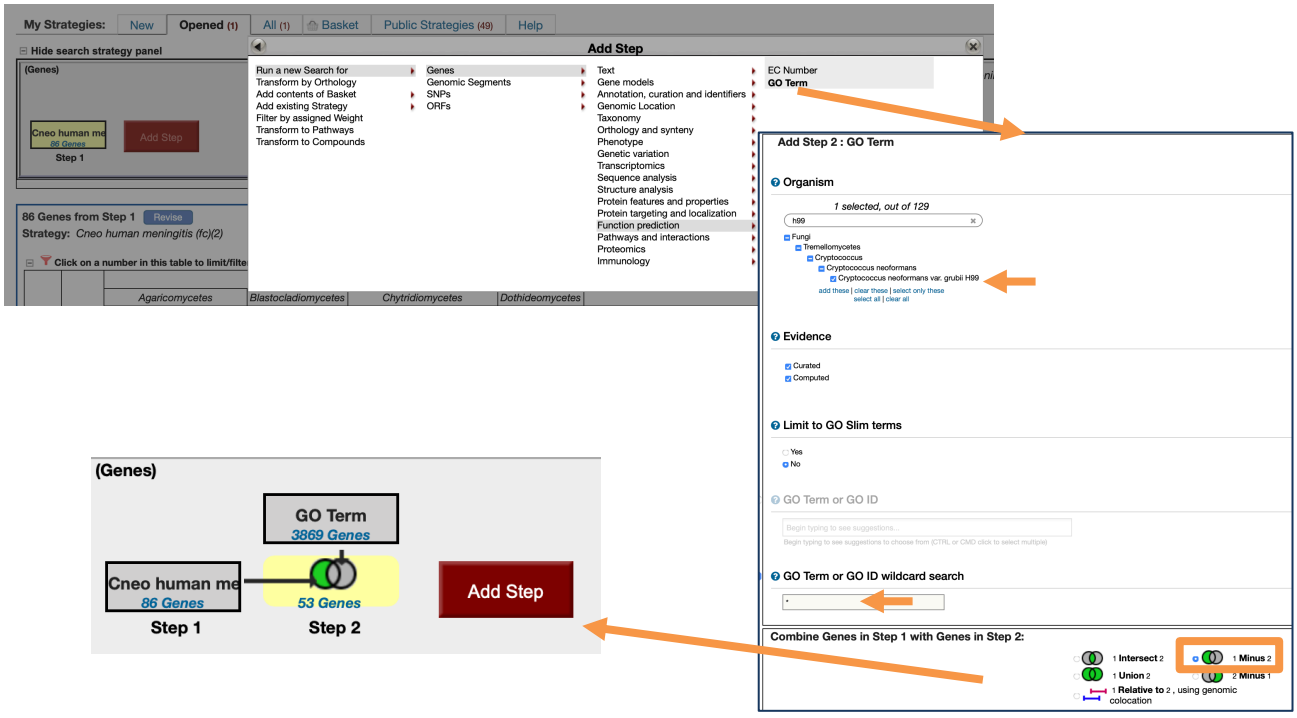
- What kind of biological processes are enriched for the upregulated *C. neoformans* genes? Cerebrospinal fluid is mostly composed of water and carries proteins, glucose, neurotransmitters, and various ions including sodium, magnesium, and chloride. Do the GO enrichment results seem appropriate given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

- Visualize results as Word Cloud that was created using the P-values and the full terms from the Enrichment analysis via a program called GOSummaries

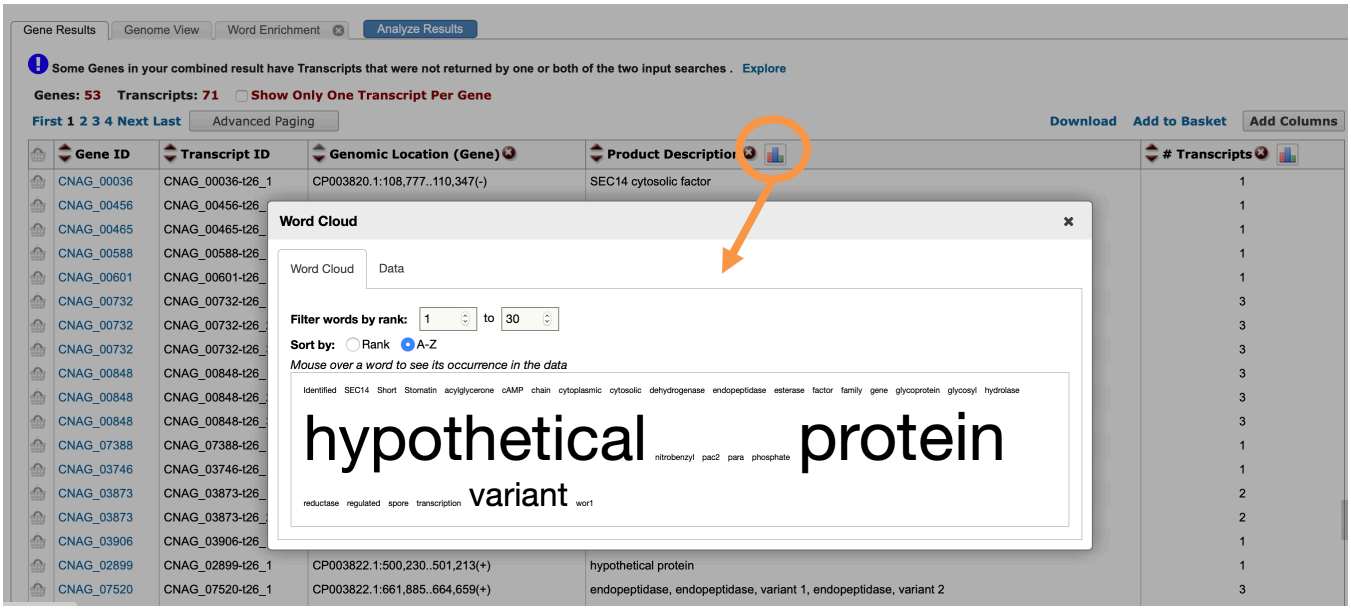


- Do the results make sense?
    - Were all genes included in GO enrichment analysis?
    - What about genes that do not have GO annotation?
    - How can you find determine which genes out of the 86 results lack GO term annotations, and therefore were excluded from your GO enrichment analysis? (See below)

- Add a step and select the GO Term search under the Function prediction category.
- Configure the GO Term search to limit the search on C. neformans var. grubii H99.
- In the GO Term or GO ID wild card search box enter *
- Think carefuly about which boolean operation to use. In this case we are interested in excluding the results for this search from our previous results, ie. To find genes without GO Term annotation.



What kids of genes are in this list? A quick way to look at these results is to generate a word cloud based on the words in the product descriptions of the genes. To do this click on the graph icon in the Product description column.



- Are there any interesting genes in the list?

Other things you might want to try include:

- Look for orthologs in other pathogenic fungi
- Determine which gene products may have a signal peptide or transmembrane domains
- Look at the expression of these genes in other datasets, *etc.*

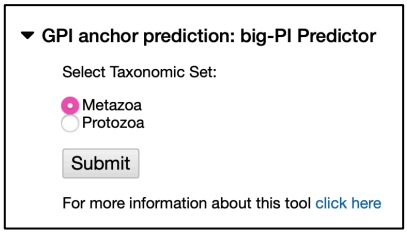
2. Creating queries across FungiDB and SGD

Glycosylphosphatidylinositol (GPI)-anchored proteins are involved in cell wall integrity and cell-cell interactions and perturbations in GPI biosynthesis lead to hypersensitivity to host defenses. A gene in *Lomentospora prolificans*, identified during your genetic screens, piqued your interest - jhhlp_004726. Take advantage of FungiDB and SGD records to learn more about this putative gene. How would you go about confirming that this gene may encode for a GPI protein in *L. prolificans*?

1. Navigate to jhhlp_004726 in FungiDB: http://fungidb.org/gene/jhhlp_004726

Note: To be a GPI-anchored protein, the jhhlp_004726 product must be post-translationally modified by attachment of a GPI anchor. Determine if the jhhlp_004726 protein sequence has any predicted GPI modification sites by running the GPI anchor prediction tool, ***big-PI****. This tool can be found in the* ***Protein features and properties*** *section.*

- What happens if you run the **big-PI Predictor tool** against **Metazoa**? Does the tool predict any potential GPI modification sites? Run the tool again, this time selecting the **Protozoa** taxonomic set, and see if you get a different answer.



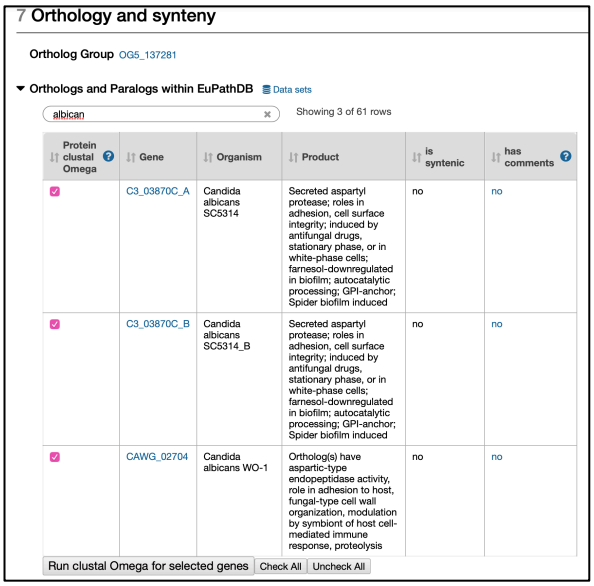
- In eukaryotes, many cell surface proteins are attached to the plasma membrane through GPI anchoring. Is there any information on the FungiDB gene page to suggest that the jhhlp_004726 protein product is secreted?

Hint: see the Proteins Properties and Features section of the gene record page.

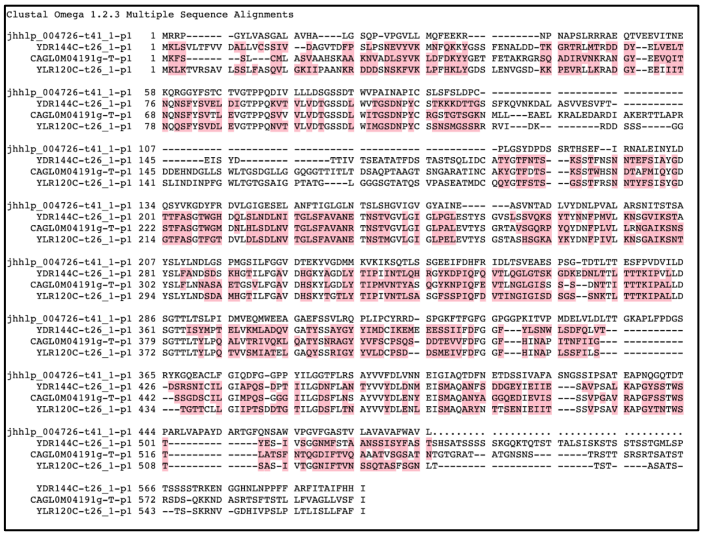
2. One way to predict the function of an uncharacterized gene is by exploring the data available on its more well-understood orthologs. What potential orthologs, if any, does jhhlp_004726 have in ***Saccharomyces cerevisiae*** and ***Candida albicans***?

   - Return to the FungiDB gene page for jhhlp_004726 and visit the **Orthology and synteny** section.

   - Align the jhhlp_004726 protein sequence to orthologous sequences in various *C. albicans* strains and the *S. cerevisiae* reference strain S288C. To do so, use the search box to look for proteins in both species or select directly from the list.



- Click on the **Run clustal Omega on selected genes** at the bottom of the table to launch ClustalOmega alignment.



  - Is there evidence of conservation between *L. prolificans, C. albicans, and S. cerevisiae* with respect to the sequence of the protein product?

From the protein sequence alignment results from ClustalOmega, jhhlp_004726 appears to be a potential *L. prolificans* ortholog of *S. cerevisiae* YDR144C. We can take advantage of other resources, such as the ***Saccharomyces*** **Genome Database** (SGD), to learn more about this protein in model organisms with more extensive

annotation. Let's take a closer look at the functions and interactions of YDR144C in SGD.

3. Navigate to the SGD record page and examine the locus summary page for YDR144C: https://www.yeastgenome.org/locus/YDR144C

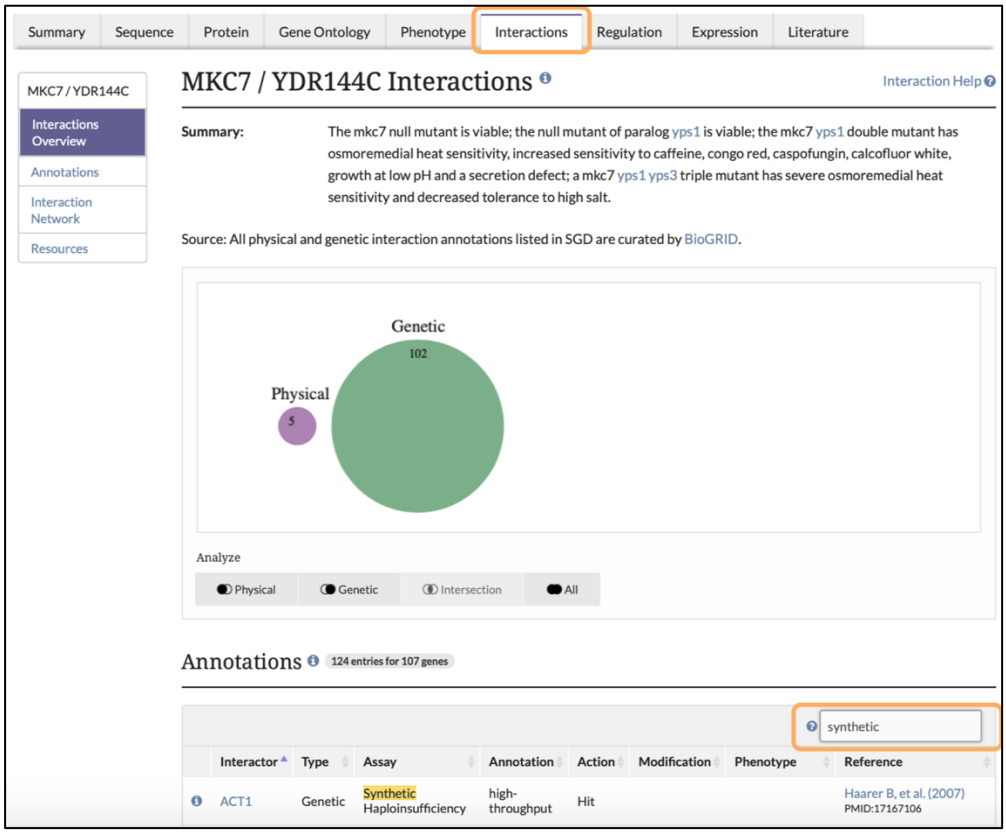
Hint: There are several ways you can navigate to this gene in SGD. You can either search SGD directly or navigate to the SGD record page from the Link outs section of the gene record page in FungiDB.

- Is YDR144C (more commonly known by its standard gene name, MKC7) a GPI anchored protein, as predicted of its *L. prolificans* ortholog?
- What is the function of MKC7 in *S. cerevisiae?*
- Does it encode a protein with enzymatic activity?
- Where in the cell does the protein execute its function? What biological process?

*Hint: see the **Gene Ontology** section on the locus page or click on the Gene Ontology tab at the top of the page.*

Functional relationships between genes and pathways can sometimes be revealed by examining genetic interactions between two or more genes. Genes are described as having a genetic interaction if the simultaneous mutation of both genes produces a phenotype that is unexpected, given the phenotypes of the single mutants.

4. Find known genetic interactions for MKC7.
    - Navigate to the **Interactions** tab at the top of the MKC7 page.
    - The **Annotations** table lists both physical interactions and genetic interactions.
        - Search the table for "genetic" to filter for genetic interactions only.
        - Next, search for "synthetic" in the table. This filters the table to show only the genetic interactions where some sort of synthetic growth defect, haploinsufficiency, or lethality is produced.



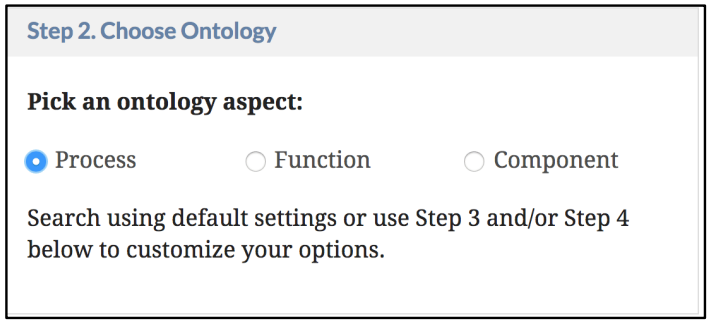
- Click on the **Download** button, which is located under the results table, and save this gene list. *Rename the file to.* ***synthetic.txt***.

Note: Rename the file to. ***synthetic.txt*** *so that we can find it easily later.*

- Click on the **Analyze** button, then on **GO Term Finder.**

- Run a **process** enrichment for the MKC7 genetic interaction genes.

Hint: GO Term Finder finds common Gene Ontology (GO) annotations between genes. To run a Biological Process enrichment, select the Process button as shown below, then submit the form. More ways to customize your GO Term Finder query can be found in the GO Term Finder exercise.



- Scroll down the results page to see the table of enriched biological processes. What kind of processes are associated with the genes we analyzed? What do these results suggest about MKC7's functional relationships in the cell?

- Click on any of the genes shown for a biological process of interest to visit the gene's page on SGD. Use the gene page to uncover how the respective gene is involved in the biological process you were interested in.

Terms from the Process Ontology of gene_association.sgd with p-value <= 0.01

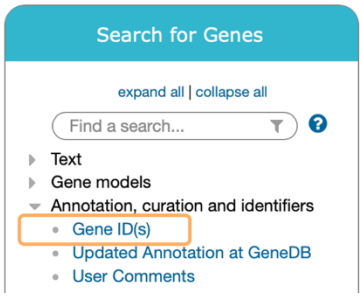
Gene Ontology term	Cluster frequency	Genome frequency	Corrected P-value	FDR	False Positives	Genes annotated to the term
tubulin complex assembly	3 of 9 genes, 33.3%	10 of 7166 genes, 0.1%	2.09e-05	0.00%	0.00	YGR078C, YLR200W, YML094W
protein folding	4 of 9 genes, 44.4%	113 of 7166 genes, 1.6%	0.00088	0.00%	0.00	YLR200W, YGR078C, YKL117W, YML094W
protein complex assembly	5 of 9 genes, 55.6%	295 of 7166 genes, 4.1%	0.00160	0.00%	0.00	YML094W, YLR200W, YGR078C, YKL117W, YHR079C
protein complex biogenesis	5 of 9 genes, 55.6%	295 of 7166 genes, 4.1%	0.00160	0.00%	0.00	YML094W, YLR200W, YGR078C, YKL117W, YHR079C
protein complex subunit organization	5 of 9 genes, 55.6%	355 of 7166 genes, 5.0%	0.00396	0.00%	0.00	YLR200W, YGR078C, YKL117W, YHR079C, YML094W
peptide pheromone maturation	2 of 9 genes, 22.2%	9 of 7166 genes, 0.1%	0.00643	0.33%	0.02	YLR120C, YNL238W
chaperone-mediated protein complex assembly	2 of 9 genes, 22.2%	9 of 7166 genes, 0.1%	0.00643	0.29%	0.02	YLR200W, YKL117W
fungal-type cell wall organization	4 of 9 genes, 44.4%	206 of 7166 genes, 2.9%	0.00955	0.50%	0.04	YLR121C, YFL039C, YLR120C, YHR079C

Now, let's go back to the file of MKC7 "synthetic" genetic interactors we downloaded earlier and find the orthologs of these genes in *Lomentospora prolificans*.

- Open this file in Excel and copy the Gene IDs in the **Interactor Systematic Name** column (not including the header)

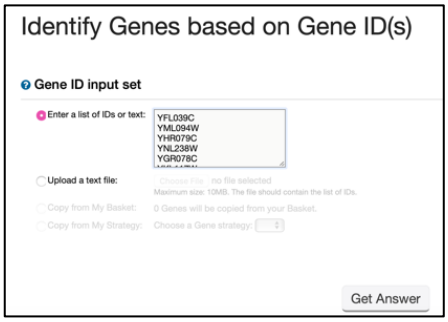
Interactor	Interactor Sy	Interactor	Interactor Systematic Name	Type	Assay	Annotation
MKC7	YDR144C	ACT1	YFL039C	Genetic	Synthetic Ha	high-through
MKC7	YDR144C	GIM5	YML094W	Genetic	Synthetic Gr	high-through
MKC7	YDR144C	IRE1	YHR079C	Genetic	Synthetic Gr	manually cur
MKC7	YDR144C	KEX2	YNL238W	Genetic	Synthetic Let	manually cur
MKC7	YDR144C	PAC10	YGR078C	Genetic	Synthetic Let	high-through
MKC7	YDR144C	SBA1	YKL117W	Genetic	Synthetic Let	high-through
MKC7	YDR144C	YKE2	YLR200W	Genetic	Synthetic Gr	high-through
MKC7	YDR144C	YPS1	YLR120C	Genetic	Synthetic Let	manually cur
MKC7	YDR144C	YPS1	YLR120C	Genetic	Synthetic Let	manually cur
MKC7	YDR144C	YPS1	YLR120C	Genetic	Synthetic Gr	manually cur
MKC7	YDR144C	YPS1	YLR120C	Genetic	Synthetic Let	manually cur
MKC7	YDR144C	YPS3	YLR121C	Genetic	Synthetic Let	manually cur

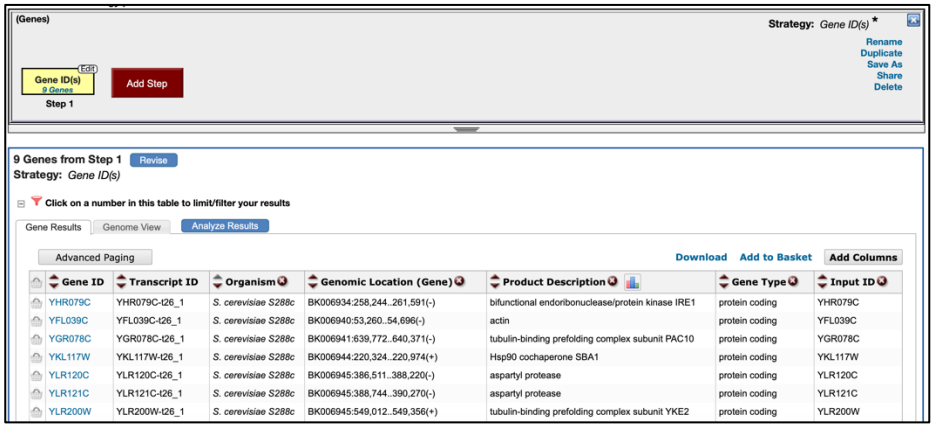
- Visit FungiDB again and initiate the GeneIDs search query



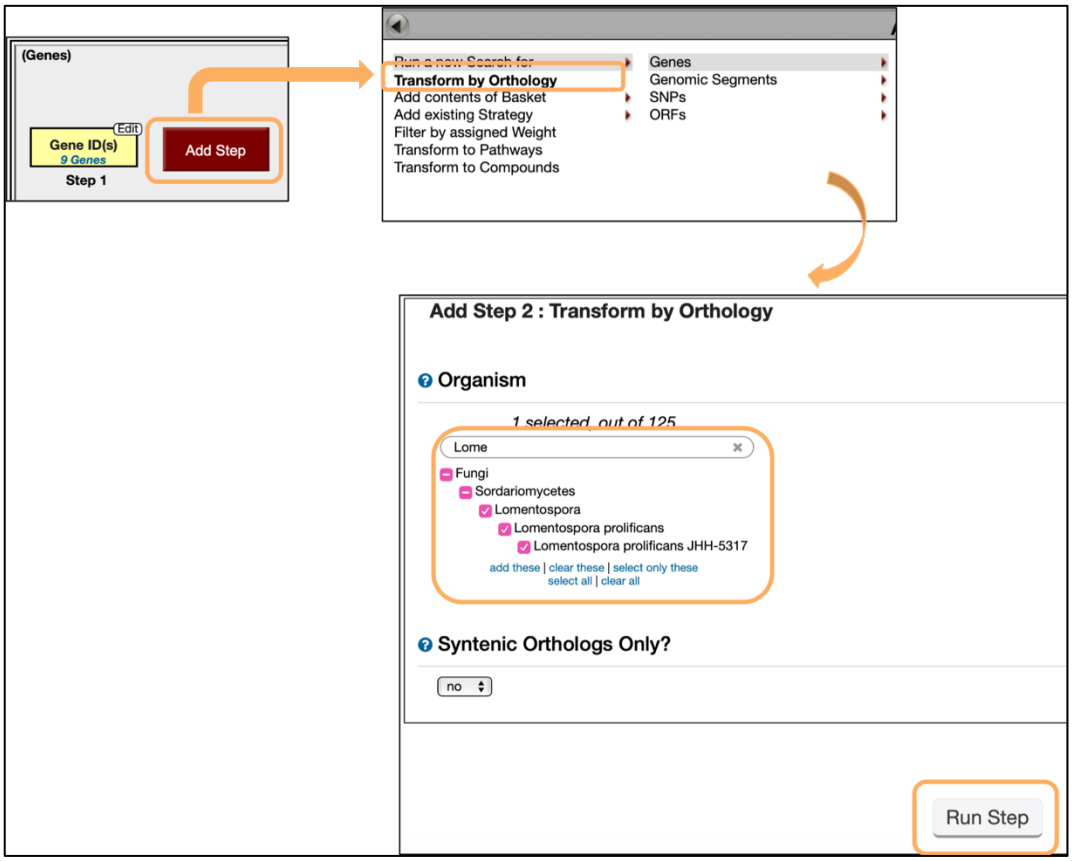
Hint: The query can be deployed from the Search for Genes section on the main page.

- Paste the list of Gene IDs that had the "synthetic" genetic interactions with MKC7 into FungiDB query and click on the **Get Answer** button.

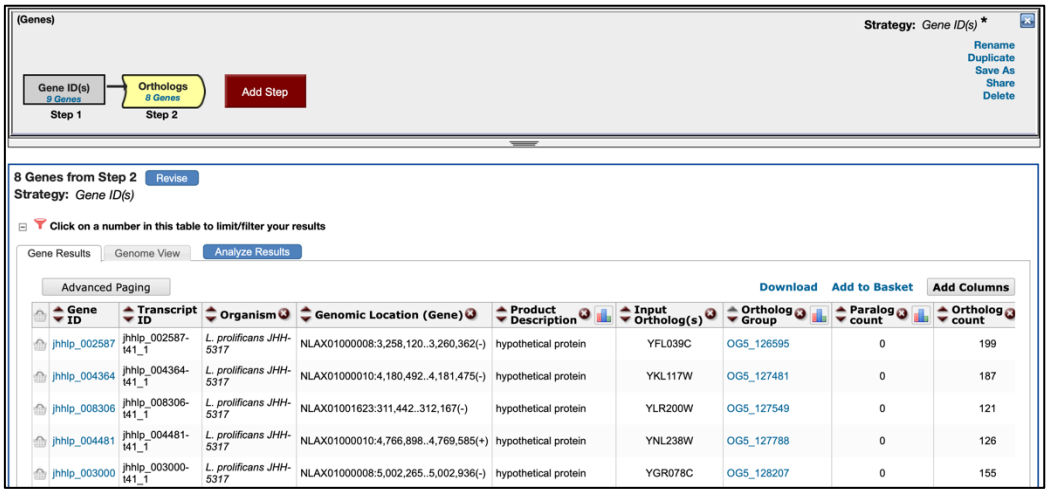




- Add a step to **Transform** the list **by orthology** to *Lomentospora prolificans*



- How many of the interacting *S. cerevisiae* genes have a hypothetical protein ortholog in *Lomentospora prolificans*?



Gene ID	Transcript ID	Organism	Genomic Location (Gene)	Product Description	Input Ortholog(s)	Ortholog Group	Paralog count	Ortholog count
jhhlp_002587	jhhlp_002587-t41_1	*L. prolificans JHH-5317*	NLAX01000008:3,258,120..3,260,362(-)	hypothetical protein	YFL039C	OG5_126595	0	199
jhhlp_004364	jhhlp_004364-t41_1	*L. prolificans JHH-5317*	NLAX01000010:4,180,492..4,181,475(-)	hypothetical protein	YKL117W	OG5_127481	0	187
jhhlp_008306	jhhlp_008306-t41_1	*L. prolificans JHH-5317*	NLAX01001623:311,442..312,167(-)	hypothetical protein	YLR200W	OG5_127549	0	121
jhhlp_004481	jhhlp_004481-t41_1	*L. prolificans JHH-5317*	NLAX01000010:4,766,898..4,769,585(+)	hypothetical protein	YNL238W	OG5_127788	0	126
jhhlp_003000	jhhlp_003000-t41_1	*L. prolificans JHH-5317*	NLAX01000008:5,002,265..5,002,936(-)	hypothetical protein	YGR078C	OG5_128207	0	155

- Given the accumulated biological information we uncovered at SGD and FungiDB, summarize your predictions about the hypothetical *L. prolificans* protein jhhlp_004726.
  - What is jhhlp_004726 ortholog in *S. cerevisiae?*
    - Is this gene a GPI-protein in yeast?
  - Based on the GPI-anchor and ClustalOmega analysis results do you have sufficient information to think that the hypothetical gene in *L. prolificans* may be a putative GPI-anchor protein?
  - How many "synthetic" genetic interactors exist in SGD for MKC7 in yeast?
    - What GO terms were enriched in biological processes associated with MKC7 interactors in *S. cerevisiae*?
    - How many orthologs of these genes are found in *L. prolificans?*
    - Why do you think the number of genes vary between *S. cerevisiae* and *L. prolificans*?

Additional resources:

More info on Fischer's exact test:
http://udel.edu/~mcdonald/statfishers.html

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 enrichment analysis
http://geneontology.org/docs/go-enrichment-analysis/