Performing GO Enrichment analysis

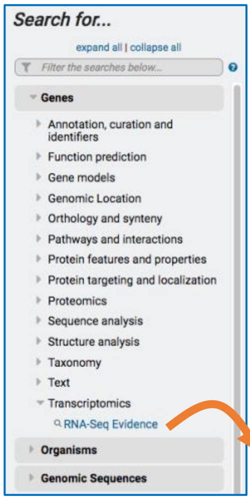
Learning objectives:

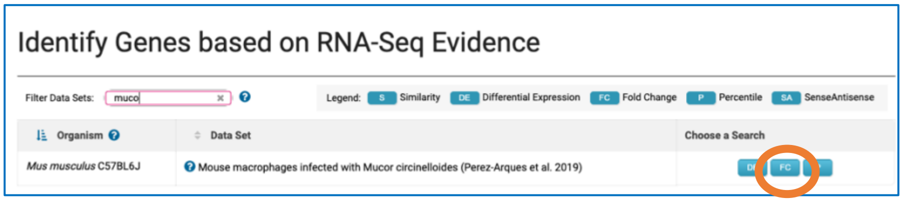
- Explore host responses by running a search strategy in HostDB.org
- Perform a GO enrichment analysis
- Create complex search strategy using both FungiDB and SGD

1. Find host genes that are upregulated in infected mouse cells compared to uninfected ones. For this exercise use http://hostdb.org

HostDB has data from published studies on host-fungal pathogen interactions, where host data I integrated in HostDB.org while fungal data is loaded into FungiDB.org. In this exercises we will the study titled "Mouse macrophages were infected with *Mucor circinelloides"*

- Go to HostDB.org and navigate to the "Transcriptomics" section then select "RNA Seq Evidence". Select the fold change query for the "Mouse macrophages were infected with Mucor circinelloides" experiment.



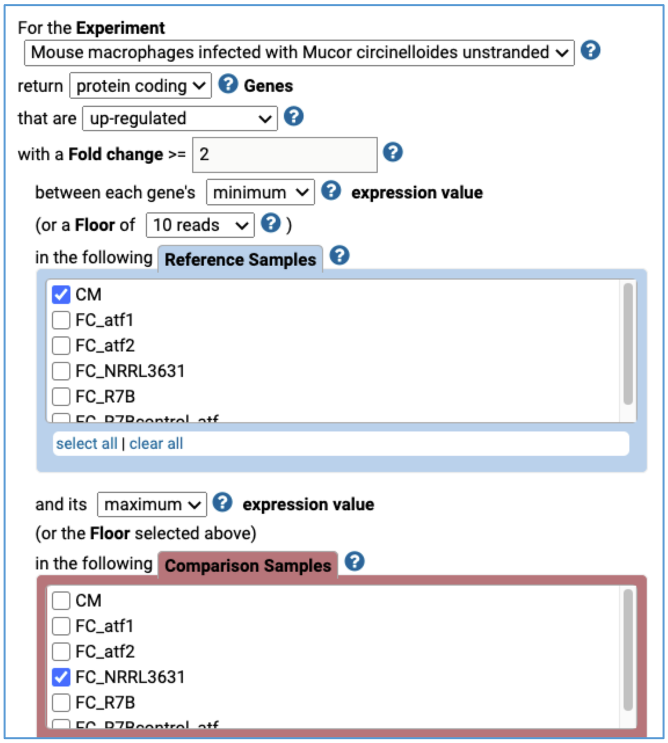


Identify Genes based on RNA-Seq Evidence

Filter Data Sets: muco

Organism	Data Set	Choose a Search
Mus musculus C57BL6J	Mouse macrophages infected with Mucor circinelloides (Perez-Arques et al. 2019)	DE FC P

- Configure the search to return genes that are up-regulated at least 2-fold in the NRRL3661 sample (Fungal spore-macrophage coculture, Mucor NRRL3631 avirulent strain cocultured with mouse macrophages (cell line J774A.1) for 5h) compared to the uninfected control (CM).



For the **Experiment**

Mouse macrophages infected with Mucor circinelloides unstranded

return protein coding **Genes**

that are up-regulated

with a **Fold change** >= 2

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

(or a **Floor** of 10 reads)

in the following **Reference Samples**

☑ CM
□ FC_atf1
□ FC_atf2
□ FC_NRRL3631
□ FC_R7B
□ FC_R7Bcontrol_atf

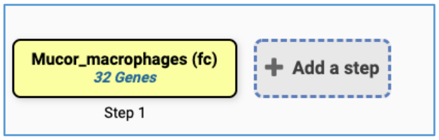
select all | clear all

and its maximum **expression value**

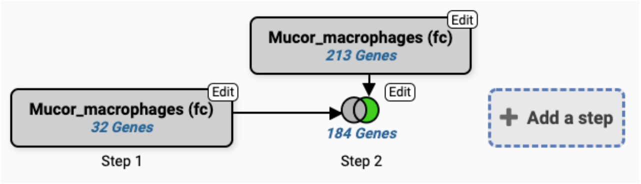
(or the **Floor** selected above)

in the following **Comparison Samples**

□ CM
□ FC_atf1
□ FC_atf2
☑ FC_NRRL3631
□ FC_R7B
□ FC_R7Bcontrol_atf



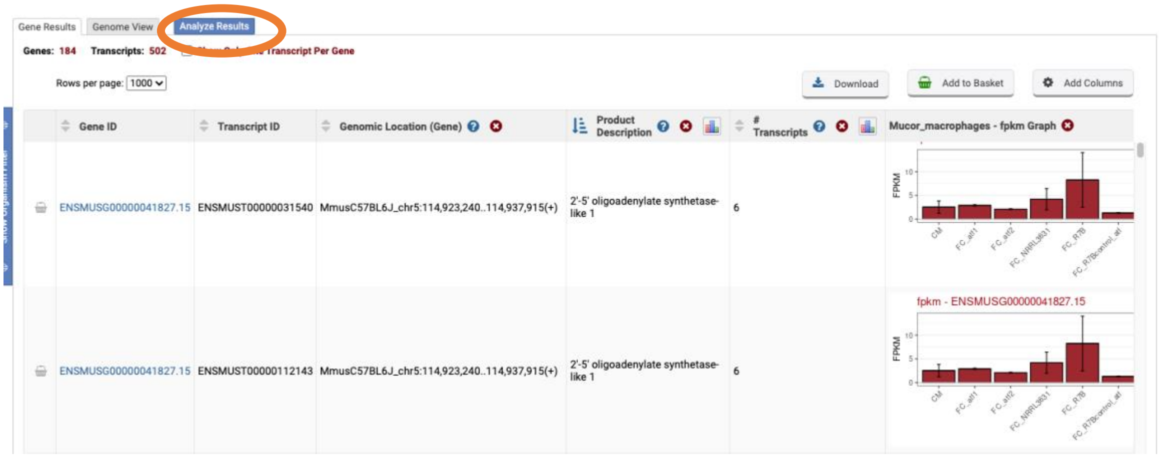
- Add a step and search for genes that are up regulated by at least 2-fold in response to the infection with the R7B virulent strain (use CM as a control).



- Examine the functional characteristics of the genes.
  Hint: click on the “Analyze Results” tab and perform a *Gene Ontology enrichment analysis* for the *Biological process* using the default parameters.

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.

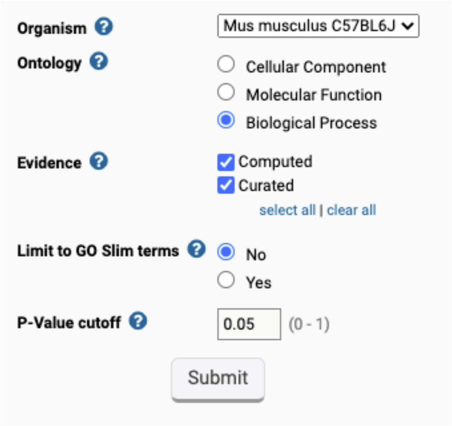


Genes: 184 Transcripts: 502 Transcript Per Gene

Rows per page: 1000

Gene ID	Transcript ID	Genomic Location (Gene)	Product Description	# Transcripts	Mucor_macrophages - fpkm Graph
ENSMUSG00000041827.15	ENSMUST00000031540	MmusC57BL6J_chr5:114,923,240..114,937,915(+)	2'-5' oligoadenylate synthetase-like 1	6	
ENSMUSG00000041827.15	ENSMUST00000112143	MmusC57BL6J_chr5:114,923,240..114,937,915(+)	2'-5' oligoadenylate synthetase-like 1	6	

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.



Organism: Mus musculus C57BL6J

Ontology:
○ Cellular Component
○ Molecular Function
● Biological Process

Evidence:
☑ Computed
☑ Curated
select all | clear all

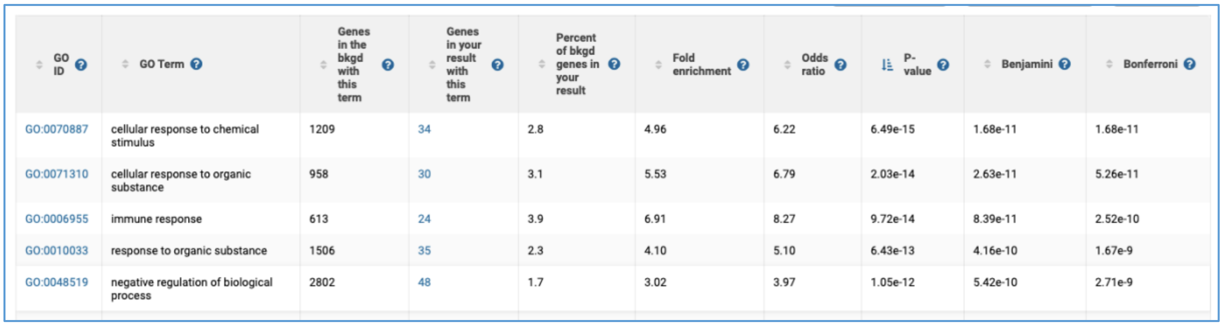
Limit to GO Slim terms:
● No
○ Yes

P-Value cutoff: 0.05 (0 - 1)

Submit

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. The table contains columns with GO IDs and GO terms along with the number of genes in the background and those specific to the RNA-Seq analysis results presented (linked in blue).



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:0070887	cellular response to chemical stimulus	1209	34	2.8	4.96	6.22	6.49e-15	1.68e-11	1.68e-11
GO:0071310	cellular response to organic substance	958	30	3.1	5.53	6.79	2.03e-14	2.63e-11	5.26e-11
GO:0006955	immune response	613	24	3.9	6.91	8.27	9.72e-14	8.39e-11	2.52e-10
GO:0010033	response to organic substance	1506	35	2.3	4.10	5.10	6.43e-13	4.16e-10	1.67e-9
GO:0048519	negative regulation of biological process	2802	48	1.7	3.02	3.97	1.05e-12	5.42e-10	2.71e-9

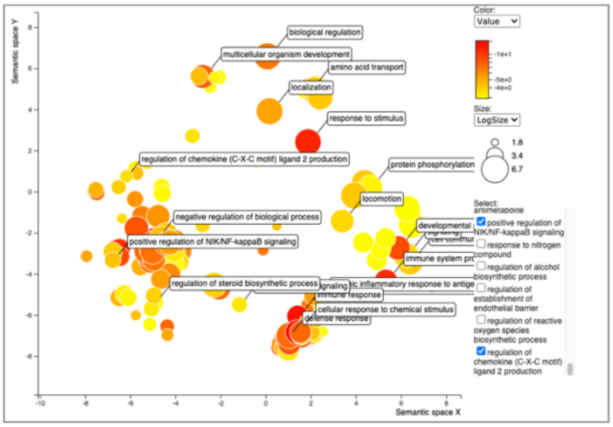
The results table also includes several additional statistical measurements:

- 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

- Examine GO enrichment analysis results.
  What kinds of GO terms are enriched? Does the host immune response appear to be turned on? Is there a particular cellular location that is common in this group of genes?

- Visualize the results in Revigo.
  Hint: Open in REVIGO button and click Start Revigo button on the next screen.

  Color: significant p values
  Proximity reflects semantic similarity



Semantic space Y
Semantic space X
8
6
4
2
0
-2
-4
-6
-8
-10
-8
-6
-4
-2
0
2
4
6
8
biological regulation
multicellular organism development
amino acid transport
localization
response to stimulus
regulation of chemokine (C-X-C motif) ligand 2 production
protein phosphorylation
locomotion
negative regulation of biological process
positive regulation of NIK/NF-kappaB signaling
developmental
immune system pr
regulation of steroid biosynthetic process
ic inflammatory response to antige
immune response
cellular response to chemical stimulus
defense response
Color:
Value
-1e+1
-5e+0
-4e+0
Size:
LogSize
1.8
3.4
6.7
Select:
antimetabolite
☑ positive regulation of NIK/NF-kappaB signaling
□ response to nitrogen compound
□ regulation of alcohol biosynthetic process
□ regulation of establishment of endothelial barrier
□ regulation of reactive oxygen species biosynthetic process
☑ regulation of chemokine (C-X-C motif) ligand 2 production

More about Revigo:
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0021800

This search can be further expanded to determine how the list of genes identified in Step responds to infections with different Candida spp. You can also add a step to convert mouse genes into orthologs in humans and look for enriched GO terms…

Creating queries across FungiDB and SGD

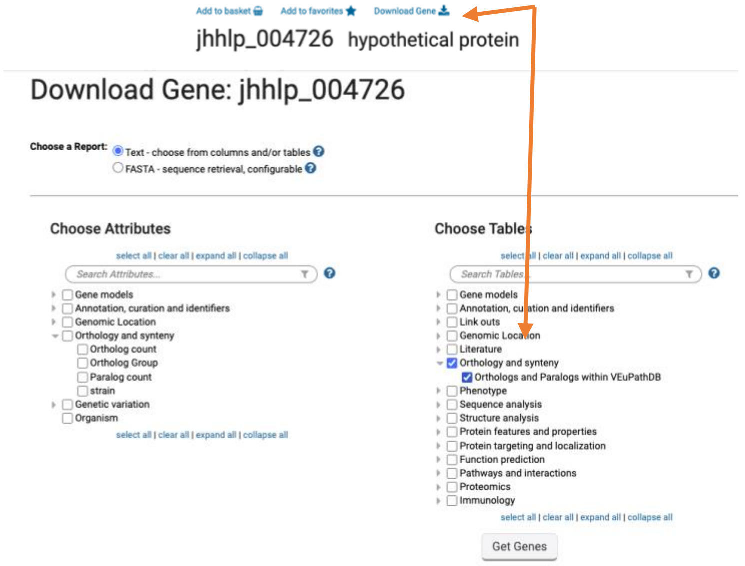
During a genetic screen in *Lomentospora prolificans*, you identified several interesting genes, including jhhlp_004726, which is a hypothetical protein. Take advantage of FungiDB and SGD records to learn more about this gene.

- Navigate to jhhlp_004726 in FungiDB and examine available records https://fungidb.org/fungidb/app/record/gene/jhhlp_004726

- Run an InterPro search and a GPI anchor prediction tool. What did you learn about this protein?

Hint: InterPro and GPI search tools can be found in the Protein features and properties section of the gene record page.

- Export orthologs of this gene.
  Click on the Download gene link and select to export orthologs in VEuPathDB option



Add to basket
Add to favorites
Download Gene

jhhlp_004726 hypothetical protein

Download Gene: jhhlp_004726

Choose a Report:
● Text - choose from columns and/or tables
○ FASTA - sequence retrieval, configurable

Choose Attributes

select all | clear all | expand all | collapse all
Search Attributes...

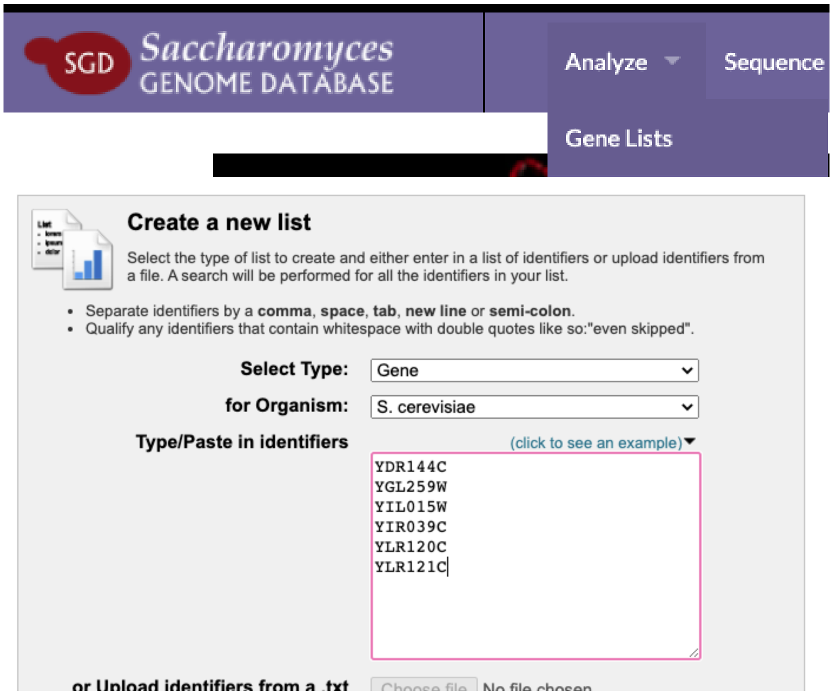
- □ Gene models
- □ Annotation, curation and identifiers
- □ Genomic Location
- □ Orthology and synteny
  - □ Ortholog count
  - □ Ortholog Group
  - □ Paralog count
  - □ strain
- □ Genetic variation
- □ Organism

select all | clear all | expand all | collapse all

Choose Tables

- □ Gene models
- □ Annotation, curation and identifiers
- □ Link outs
- □ Genomic Location
- □ Literature
- ☑ Orthology and synteny
  - ☑ Orthologs and Paralogs within VEuPathDB
- □ Phenotype
- □ Sequence analysis
- □ Structure analysis
- □ Protein features and properties
- □ Protein targeting and localization
- □ Function prediction
- □ Pathways and interactions
- □ Proteomics
- □ Immunology

- Navigate to the SGD gene lists search and copy and paste *S. cerevisiae* orthologs for jhhlp_004726: https://www.yeastgenome.org/locus/YDR144C



Create a new list

Select the type of list to create and either enter in a list of identifiers or upload identifiers from a file. A search will be performed for all the identifiers in your list.

- Separate identifiers by a **comma**, **space**, **tab**, **new line** or **semi-colon**.
- Qualify any identifiers that contain whitespace with double quotes like so:"even skipped".

Select Type: Gene

for Organism: S. cerevisiae

Type/Paste in identifiers (click to see an example)

```
YDR144C
YGL259W
YIL015W
YIR039C
YLR120C
YLR121C
```

or Upload identifiers from a .txt Choose file No file chosen

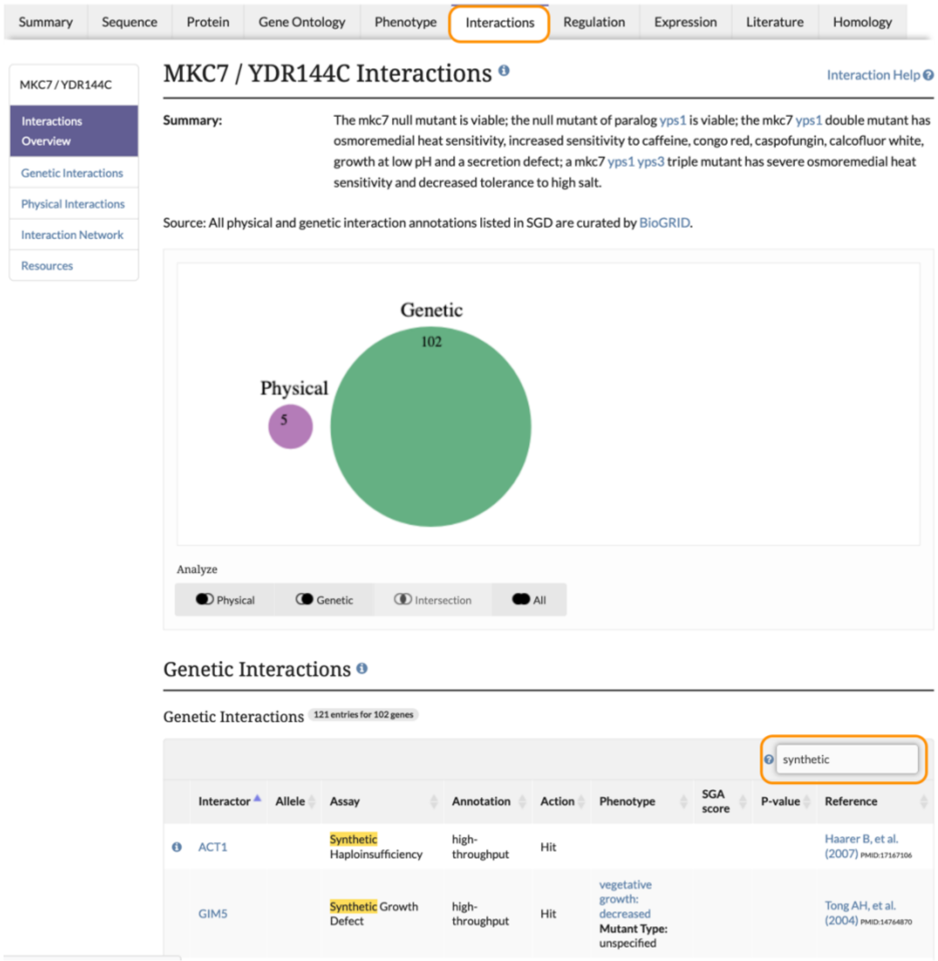
- Click on the GeneIDs to examine *S. cerevisiae* genes. What is the function of MKC7 (YDR144C) in *S. cerevisiae?* Does it encode a protein with enzymatic activity? Where in the cell does the protein execute its function? What biological process?
  Hint: see the **Gene Ontology** section on the locus 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.

- **Find known genetic interactions for MKC7.**

  - In SGD, find the MKC7 locus page and navigate to the **Interactions** tab, which is listed in the Quick Links panel near the top. The interactions are divided into separate physical interactions and genetic interactions tables below the summary.

  - Search for “synthetic” in the **Genetic Interactions** table. The filters the table to show only the genetic interactions where some sort of synthetic growth defect, haploinsufficiency, or lethality is produced.



MKC7 / YDR144C Interactions

Summary: The mkc7 null mutant is viable; the null mutant of paralog yps1 is viable; the mkc7 yps1 double mutant has osmoremedial heat sensitivity, increased sensitivity to caffeine, congo red, caspofungin, calcofluor white, growth at low pH and a secretion defect; a mkc7 yps1 yps3 triple mutant has severe osmoremedial heat sensitivity and decreased tolerance to high salt.

Source: All physical and genetic interaction annotations listed in SGD are curated by BioGRID.

Genetic Interactions

Genetic Interactions 121 entries for 102 genes

synthetic

	Interactor	Allele	Assay	Annotation	Action	Phenotype	SGA score	P-value	Reference
	ACT1		Synthetic Haploinsufficiency	high-throughput	Hit				Haarer B, et al. (2007) PMID:17167106
	GIM5		Synthetic Growth Defect	high-throughput	Hit	vegetative growth: decreased **Mutant Type:** unspecified			Tong AH, et al. (2004) PMID:14764870

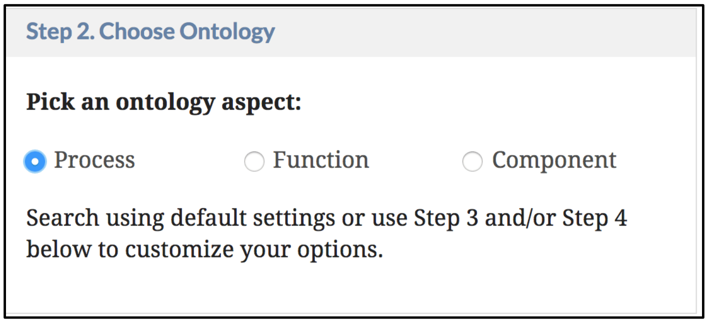
- Click on the **Download** button, which is located under the results table, and save this gene list. *Rename the file to* ***synthetic.txt***.

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.



Step 2. Choose Ontology

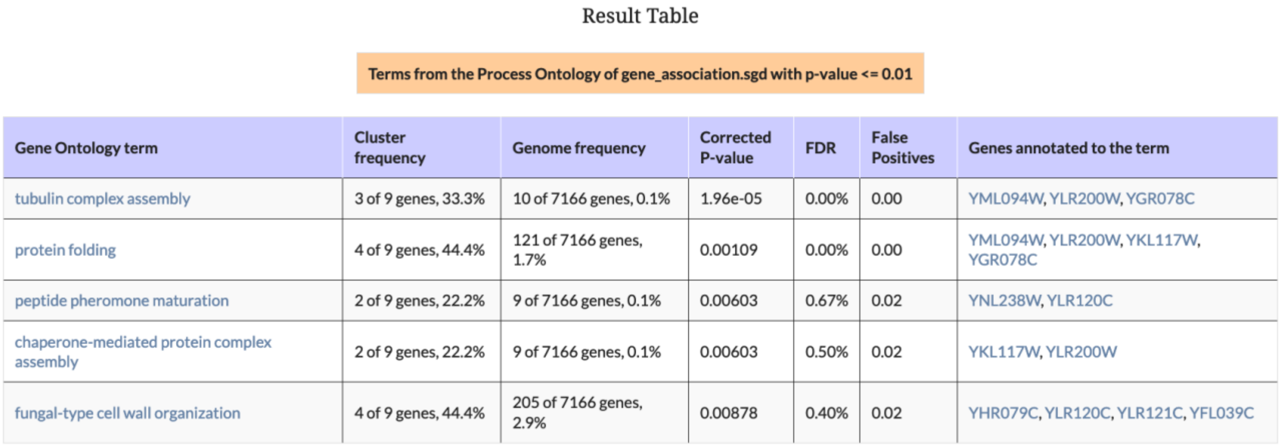
Pick an ontology aspect:

● Process ○ Function ○ Component

Search using default settings or use Step 3 and/or Step 4 below to customize your options.

- 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.



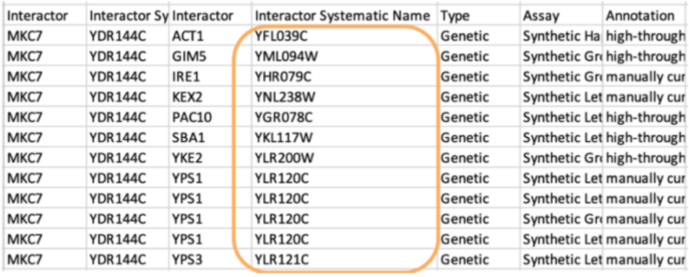
Result Table

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%	1.96e-05	0.00%	0.00	YML094W, YLR200W, YGR078C
protein folding	4 of 9 genes, 44.4%	121 of 7166 genes, 1.7%	0.00109	0.00%	0.00	YML094W, YLR200W, YKL117W, YGR078C
peptide pheromone maturation	2 of 9 genes, 22.2%	9 of 7166 genes, 0.1%	0.00603	0.67%	0.02	YNL238W, YLR120C
chaperone-mediated protein complex assembly	2 of 9 genes, 22.2%	9 of 7166 genes, 0.1%	0.00603	0.50%	0.02	YKL117W, YLR200W
fungal-type cell wall organization	4 of 9 genes, 44.4%	205 of 7166 genes, 2.9%	0.00878	0.40%	0.02	YHR079C, YLR120C, YLR121C, YFL039C

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*.

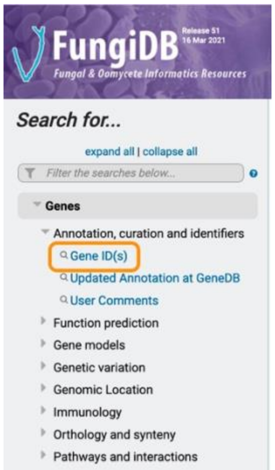
- Open this file in Excel and copy the Gene IDs in the **Interactor Systematic Name** column (not including the header)



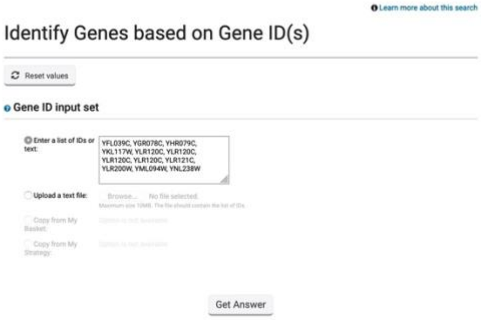
Interactor	Interactor Sy	Interactor	Interactor Systematic Name	Type	Assay	Annotation
MKC7	YDR144C	ACT1	YFL039C	Genetic	Synthetic Ha	high-through
MKC7	YDR144C	GIM5	YML094W	Genetic	Synthetic Gr	high-through
MKC7	YDR144C	IRE1	YHR079C	Genetic	Synthetic Gr	manually cur
MKC7	YDR144C	KEX2	YNL238W	Genetic	Synthetic 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.


FungiDB
Release 51
16 Mar 2021
Fungal & Oomycete Informatics Resources
Search for...
expand all | collapse all
Filter the searches below...
Genes
Annotation, curation and identifiers
Gene ID(s)
Updated Annotation at GeneDB
User Comments
Function prediction
Gene models
Genetic variation
Genomic Location
Immunology
Orthology and synteny
Pathways and interactions

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

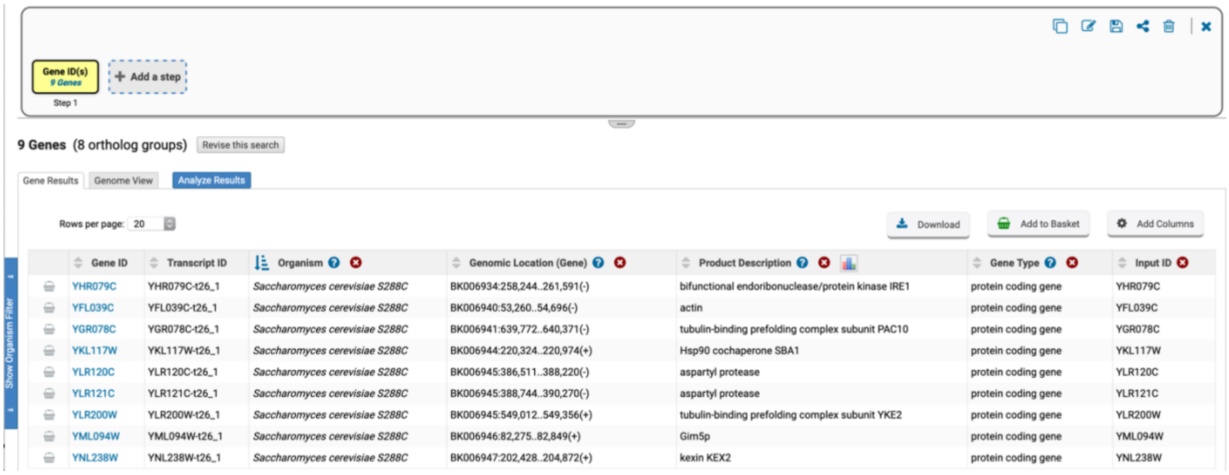

Learn more about this search
Identify Genes based on Gene ID(s)
Reset values
Gene ID input set

● Enter a list of IDs or text: YFL039C, YGR078C, YHR079C, YKL117W, YLR120C, YLR120C, YLR120C, YLR120C, YLR121C, YLR200W, YML094W, YNL238W

○ Upload a text file: Browse... No file selected.
Maximum size 10MB. The file should contain the list of IDs.

○ Copy from My Basket: Option is not available

○ Copy from My Strategy: Option is not available



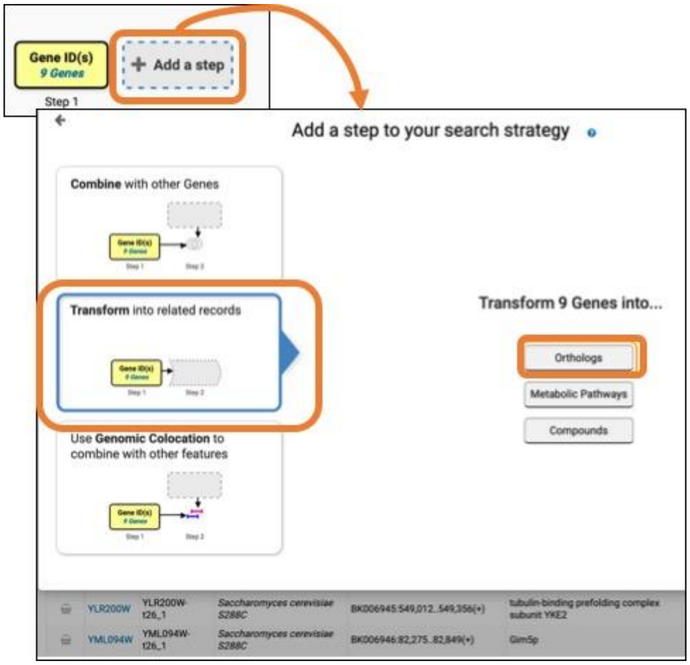
9 Genes (8 ortholog groups) Revise this search

Rows per page: 20

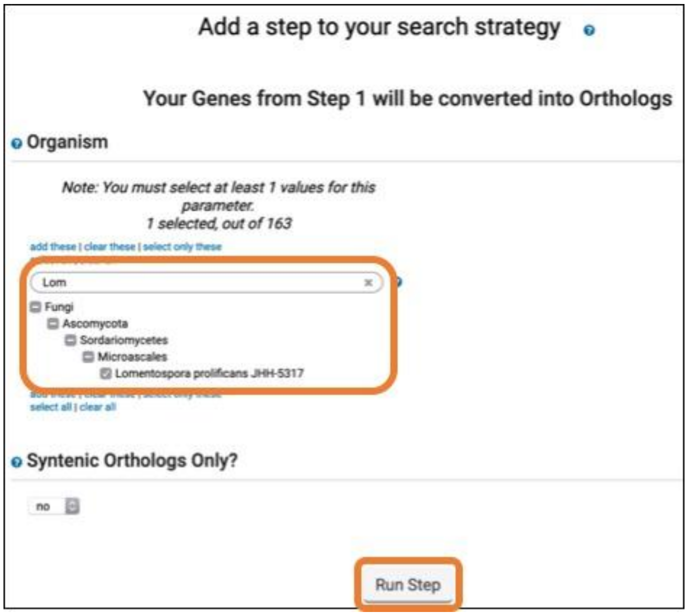
Gene ID	Transcript ID	Organism	Genomic Location (Gene)	Product Description	Gene Type	Input ID
YHR079C	YHR079C-t26_1	*Saccharomyces cerevisiae S288C*	BK006934:258,244..261,591(-)	bifunctional endoribonuclease/protein kinase IRE1	protein coding gene	YHR079C
YFL039C	YFL039C-t26_1	*Saccharomyces cerevisiae S288C*	BK006940:53,260..54,696(-)	actin	protein coding gene	YFL039C
YGR078C	YGR078C-t26_1	*Saccharomyces cerevisiae S288C*	BK006941:639,772..640,371(-)	tubulin-binding prefolding complex subunit PAC10	protein coding gene	YGR078C
YKL117W	YKL117W-t26_1	*Saccharomyces cerevisiae S288C*	BK006944:220,324..220,974(+)	Hsp90 cochaperone SBA1	protein coding gene	YKL117W
YLR120C	YLR120C-t26_1	*Saccharomyces cerevisiae S288C*	BK006945:386,511..388,220(-)	aspartyl protease	protein coding gene	YLR120C
YLR121C	YLR121C-t26_1	*Saccharomyces cerevisiae S288C*	BK006945:388,744..390,270(-)	aspartyl protease	protein coding gene	YLR121C
YLR200W	YLR200W-t26_1	*Saccharomyces cerevisiae S288C*	BK006945:549,012..549,356(+)	tubulin-binding prefolding complex subunit YKE2	protein coding gene	YLR200W
YML094W	YML094W-t26_1	*Saccharomyces cerevisiae S288C*	BK006946:82,275..82,849(+)	Gim5p	protein coding gene	YML094W
YNL238W	YNL238W-t26_1	*Saccharomyces cerevisiae S288C*	BK006947:202,428..204,872(+)	kexin KEX2	protein coding gene	YNL238W

- Find orthologs in *Lomentospora prolificans*.

Hint: Click Add a step to **Transform** the list **into related records.** Select the option to transform into **orthologs**, then use the search bar to filter on *Lomentospora prolificans* and **Run Step**.


Gene ID(s)
9 Genes
+ Add a step
Step 1
Add a step to your search strategy
Combine with other Genes
Transform into related records
Use Genomic Colocation to combine with other features
Transform 9 Genes into...
Orthologs
Metabolic Pathways
Compounds

YLR200W	YLR200W-t26_1	*Saccharomyces cerevisiae S288C*	BK006945:549,012..549,356(+)	tubulin-binding prefolding complex subunit YKE2
YML094W	YML094W-t26_1	*Saccharomyces cerevisiae S288C*	BK006946:82,275..82,849(+)	Gim5p



Add a step to your search strategy

Your Genes from Step 1 will be converted into Orthologs

Organism

Note: You must select at least 1 values for this parameter.
1 selected, out of 163

add these | clear these | select only these

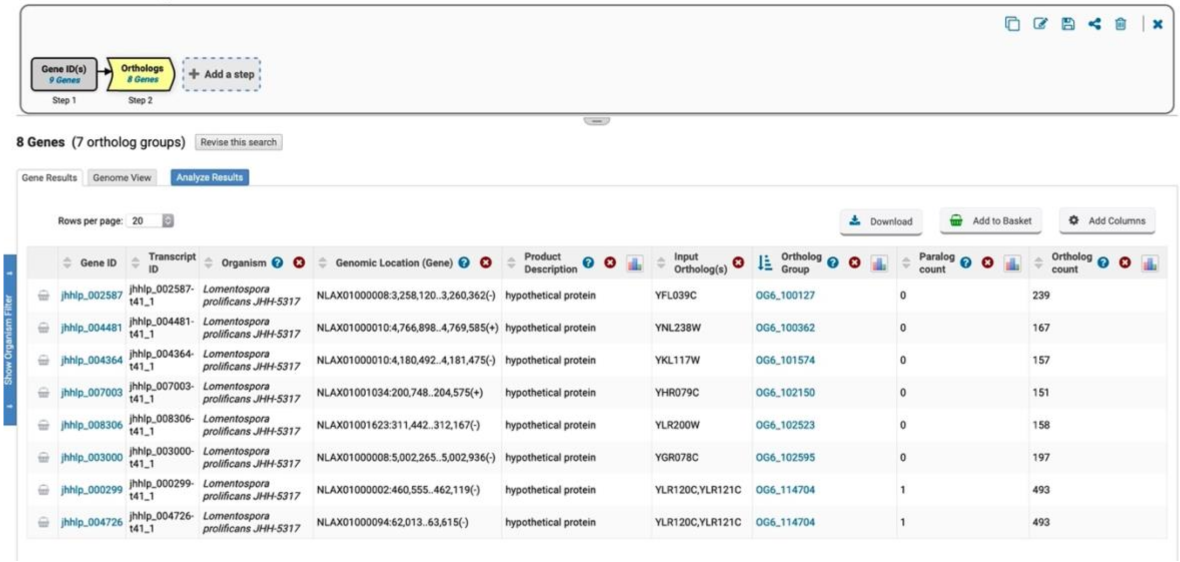
Lom

⊟ Fungi
⊟ Ascomycota
⊟ Sordariomycetes
⊟ Microascales
☑ Lomentospora prolificans JHH-5317

select all | clear all

Syntenic Orthologs Only?

no



8 Genes (7 ortholog groups)

Rows per page: 20

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	*Lomentospora prolificans JHH-5317*	NLAX01000008:3,258,120..3,260,362(-)	hypothetical protein	YFL039C	OG6_100127	0	239
jhhlp_004481	jhhlp_004481-t41_1	*Lomentospora prolificans JHH-5317*	NLAX01000010:4,766,898..4,769,585(+)	hypothetical protein	YNL238W	OG6_100362	0	167
jhhlp_004364	jhhlp_004364-t41_1	*Lomentospora prolificans JHH-5317*	NLAX01000010:4,180,492..4,181,475(-)	hypothetical protein	YKL117W	OG6_101574	0	157
jhhlp_007003	jhhlp_007003-t41_1	*Lomentospora prolificans JHH-5317*	NLAX01001034:200,748..204,575(+)	hypothetical protein	YHR079C	OG6_102150	0	151
jhhlp_008306	jhhlp_008306-t41_1	*Lomentospora prolificans JHH-5317*	NLAX01001623:311,442..312,167(-)	hypothetical protein	YLR200W	OG6_102523	0	158
jhhlp_003000	jhhlp_003000-t41_1	*Lomentospora prolificans JHH-5317*	NLAX01000008:5,002,265..5,002,936(-)	hypothetical protein	YGR078C	OG6_102595	0	197
jhhlp_000299	jhhlp_000299-t41_1	*Lomentospora prolificans JHH-5317*	NLAX01000002:460,555..462,119(-)	hypothetical protein	YLR120C,YLR121C	OG6_114704	1	493
jhhlp_004726	jhhlp_004726-t41_1	*Lomentospora prolificans JHH-5317*	NLAX01000094:62,013..63,615(-)	hypothetical protein	YLR120C,YLR121C	OG6_114704	1	493

How many of the interacting *S. cerevisiae* genes have a hypothetical protein ortholog in *Lomentospora prolificans*? Can you find jhhlp_004726 amongst these genes?

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. 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?
- 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/