Phenotypic data

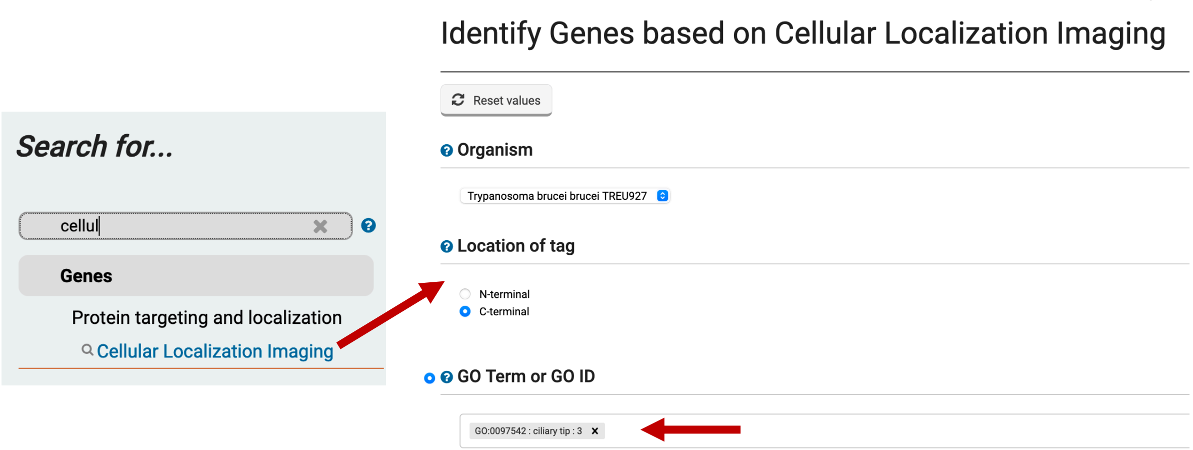
Learning objectives:

- Explore how to combine different phenotypic data
- Explore high throughput mutagenesis data
- Explore curated phenotypic data
- Explore high throughput subcellular localization data

1. **Identify genes that are targeted to the ciliary tip of *Trypanosoma brucei* that are also essential for parasite fitness.**
   Note for this exercise use http://tritrypdb.org

   a. TriTrypDB integrates data from the TrypTag project (http://tryptag.org). Genes from *T. brucei* were N- and C-terminally tagged with a fluorescent protein and subcellular localization determined by microscopy. The description of the localization was done using gene ontology terms.

   - Start by finding the "Cellular Localization Imaging" search.



Search for...

cellul

Genes

Protein targeting and localization

Cellular Localization Imaging

Identify Genes based on Cellular Localization Imaging

Organism: Trypanosoma brucei brucei TREU927

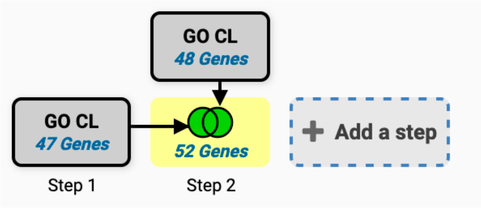
Location of tag:

○ N-terminal

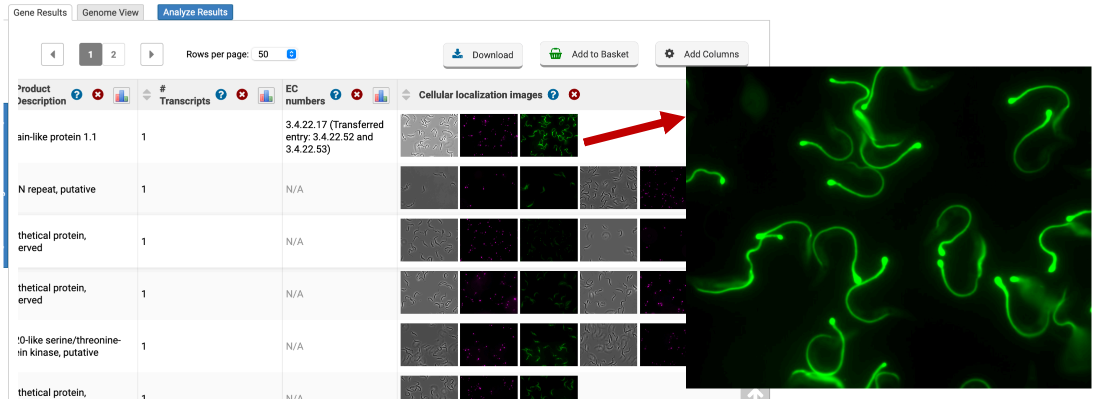
● C-terminal

GO Term or GO ID: GO:0097542 : ciliary tip : 3

   - Configure the search to identify the GO term "Ciliary Tip" – notice that when you start typing the autocomplete function offers you selectable options.
   - Since the experiment examined both N and C terimnal fusions proteins, you will have to run the search twice and combine the results from both searches. Did you use a union or an intersect to combine the results?



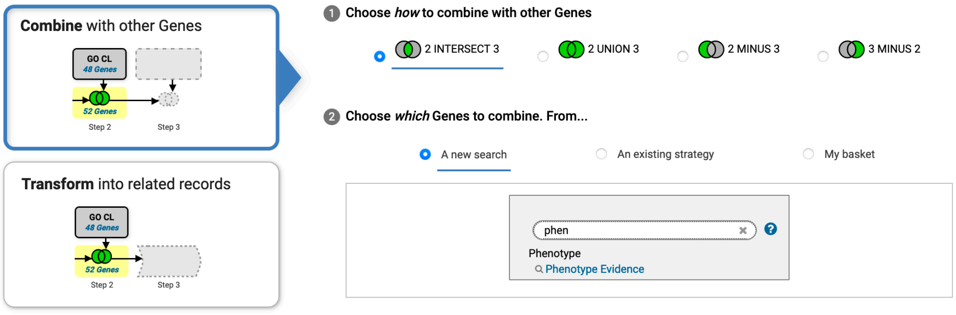
- Explore the results you got. Scroll down to the results section, then scroll to the right of the results window to reveal the subcellular localization images. These are very small, but you can right click on them to open a larger image in a new window.



Rows per page: 50

›roduct ›escription	# Transcripts	EC numbers	Cellular localization images
ain-like protein 1.1	1	3.4.22.17 (Transferred entry: 3.4.22.52 and 3.4.22.53)	
N repeat, putative	1	N/A	
thetical protein, erved	1	N/A	
thetical protein, erved	1	N/A	
·0-like serine/threonine- ·in kinase, putative	1	N/A	
thetical protein,	1	N/A	

b. Add a step to identify how many genes are essential for the fitness of the parasite. Click on Add step, then search for the phenotype searches. Click on the Phenotype Evidence option.



1 Choose *how* to combine with other Genes

● 2 INTERSECT 3 ○ 2 UNION 3 ○ 2 MINUS 3 ○ 3 MINUS 2

2 Choose *which* Genes to combine. From...

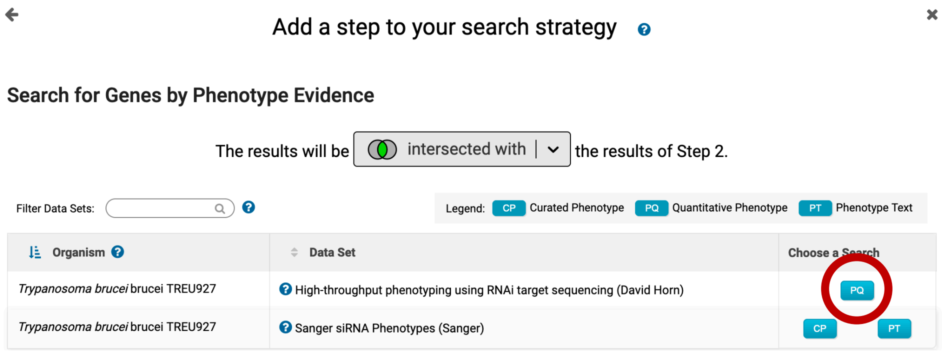
● A new search ○ An existing strategy ○ My basket

phen

Phenotype

Phenotype Evidence

- Select the “High-throughput phenotyping using RNAi target sequencing (David Horn)”.



Add a step to your search strategy

Search for Genes by Phenotype Evidence

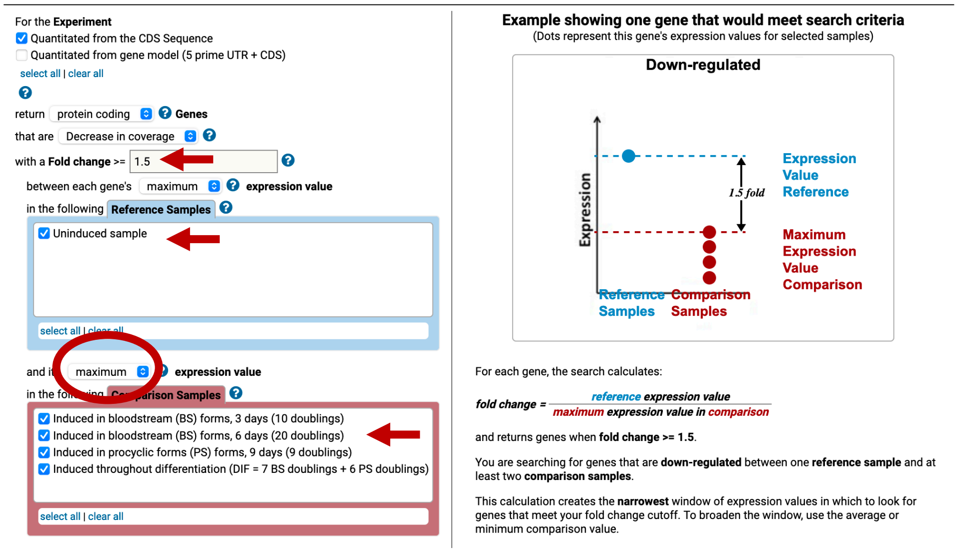
The results will be intersected with the results of Step 2.

Filter Data Sets:

Legend: CP Curated Phenotype PQ Quantitative Phenotype PT Phenotype Text

Organism	Data Set	Choose a Search
Trypanosoma brucei brucei TREU927	High-throughput phenotyping using RNAi target sequencing (David Horn)	PQ
Trypanosoma brucei brucei TREU927	Sanger siRNA Phenotypes (Sanger)	CP PT

- Configure the search to return genes that are decreased in coverage by 1.5 fold when comparing the maximum expression value of all induced samples to the uninduced sample.



For the **Experiment**

☑ Quantitated from the CDS Sequence

☐ Quantitated from gene model (5 prime UTR + CDS)

select all | clear all

return protein coding **Genes**

that are Decrease in coverage

with a **Fold change** >= 1.5

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

in the following **Reference Samples**

☑ Uninduced sample

select all | clear all

and its maximum **expression value**

in the following **Comparison Samples**

☑ Induced in bloodstream (BS) forms, 3 days (10 doublings)

☑ Induced in bloodstream (BS) forms, 6 days (20 doublings)

☑ Induced in procyclic forms (PS) forms, 9 days (9 doublings)

☑ Induced throughout differentiation (DIF = 7 BS doublings + 6 PS doublings)

select all | clear all

Example showing one gene that would meet search criteria

(Dots represent this gene's expression values for selected samples)

For each gene, the search calculates:

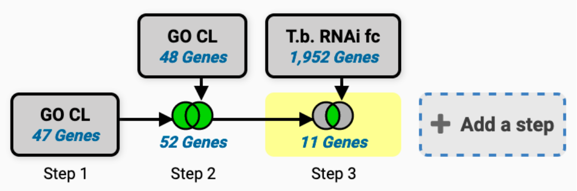
$$\textit{fold change} = \frac{\textit{reference expression value}}{\textit{maximum expression value in comparison}}$$

and returns genes when **fold change >= 1.5**.

You are searching for genes that are **down-regulated** between one **reference sample** and at least two **comparison samples**.

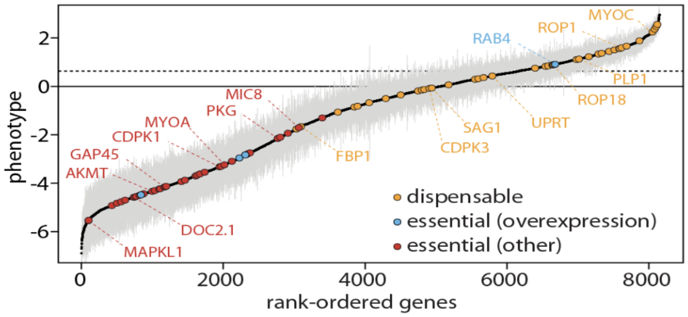
This calculation creates the **narrowest** window of expression values in which to look for genes that meet your fold change cutoff. To broaden the window, use the average or minimum comparison value.

- How many genes did you get?

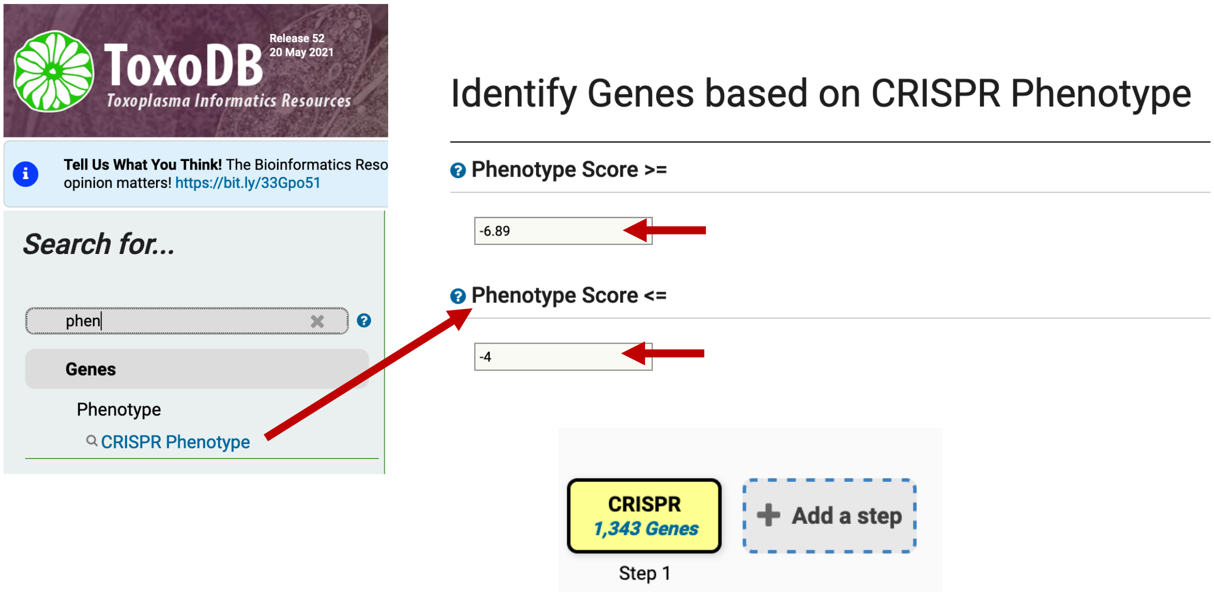


2. Finding genes based on high throughput mutagenesis and fitness analysis.
Note for this exercise use http://toxodb.org

- Navigate to the CRISPR phenotype search. Note that this search form is quite simple just requiring a range of fitness values. The defaults return all genes not limiting the search at all. This is only useful in as much as it tells you which genes were assayed which is nearly the entire genome. The tricky bit is deciding where to make the cutoffs. Again, the description on the search form is very helpful in this regard (as is the link to the paper … remember these phenotypes were assayed under specific conditions so just because a particular gene doesn't show a phenotype doesn't mean it wouldn't in other conditions (or infecting an



actual host). The plot showing the phenotype score (fitness) is particularly useful. Red points along the plot are genes known to be essential under these conditions while yellow are known to be expendable. This will help you determine where to set the values. The scores range from 2.96 (least "essential) to -6.89 (most "essential). Try it running this search by limiting the range from -6.89 to -4. Do you get the expected results based on the above graph and the number of genes returned in your search results?



Tell Us What You Think! The Bioinformatics Reso
opinion matters! https://bit.ly/33Gpo51

Search for...

phen

Genes

Phenotype

CRISPR Phenotype

Identify Genes based on CRISPR Phenotype

Phenotype Score >=

-6.89

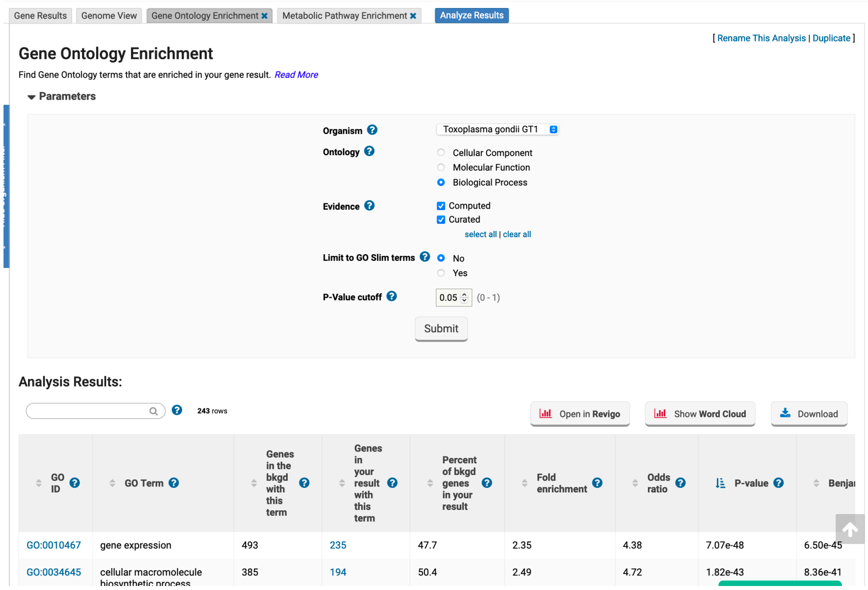
Phenotype Score <=

-4

CRISPR
1,343 Genes

Step 1

- What kinds of genes are in your results? What kinds of genes would you expect to be essential? One way to explore the data is to run a GO enrichment analysis to determine if any biological processes are enriched in your results. Give this a try. What do you results look like and do they make sense?


Gene Results
Genome View
Gene Ontology Enrichment
Metabolic Pathway Enrichment
Analyze Results
[Rename This Analysis | Duplicate]

Gene Ontology Enrichment

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

Parameters

Organism Toxoplasma gondii GT1

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

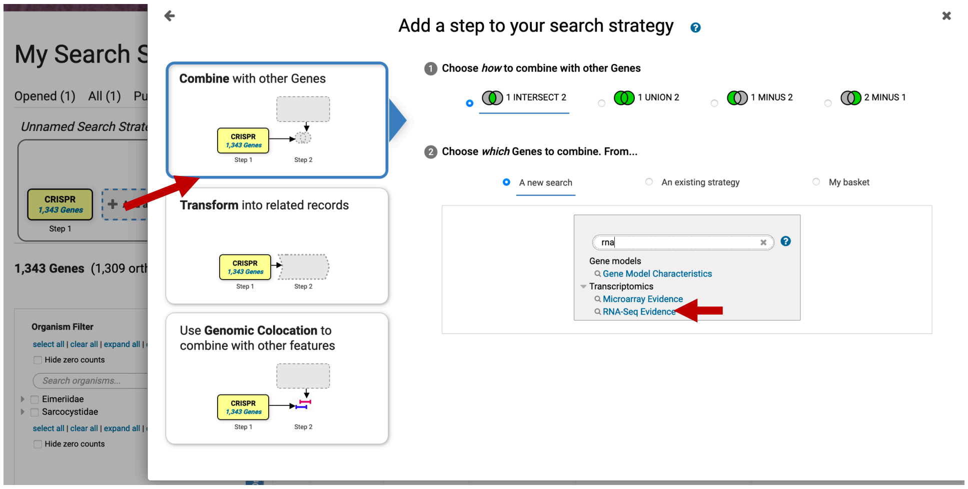
Analysis Results:

243 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:0010467	gene expression	493	235	47.7	2.35	4.38	7.07e-48	6.50e-45
GO:0034645	cellular macromolecule biosynthetic process	385	194	50.4	2.49	4.72	1.82e-43	8.36e-41

- How many of these genes are upregulated in *in vivo* chronic stages of *T. gondii*?

- Click on add step and elect the RNAseq searches under the Transcriptomics category



Add a step to your search strategy

1 Choose *how* to combine with other Genes

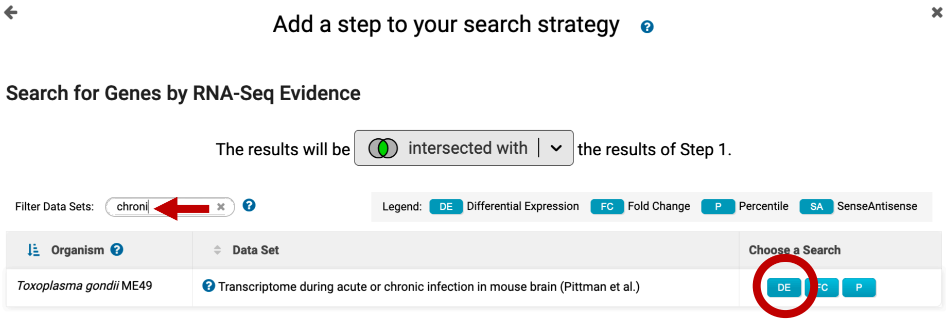
● 1 INTERSECT 2 ○ 1 UNION 2 ○ 1 MINUS 2 ○ 2 MINUS 1

2 Choose *which* Genes to combine. From...

● A new search ○ An existing strategy ○ My basket

rna

- Find the experiment with chronic stages and run a search based on differentially expressed genes (DE).



Add a step to your search strategy

Search for Genes by RNA-Seq Evidence

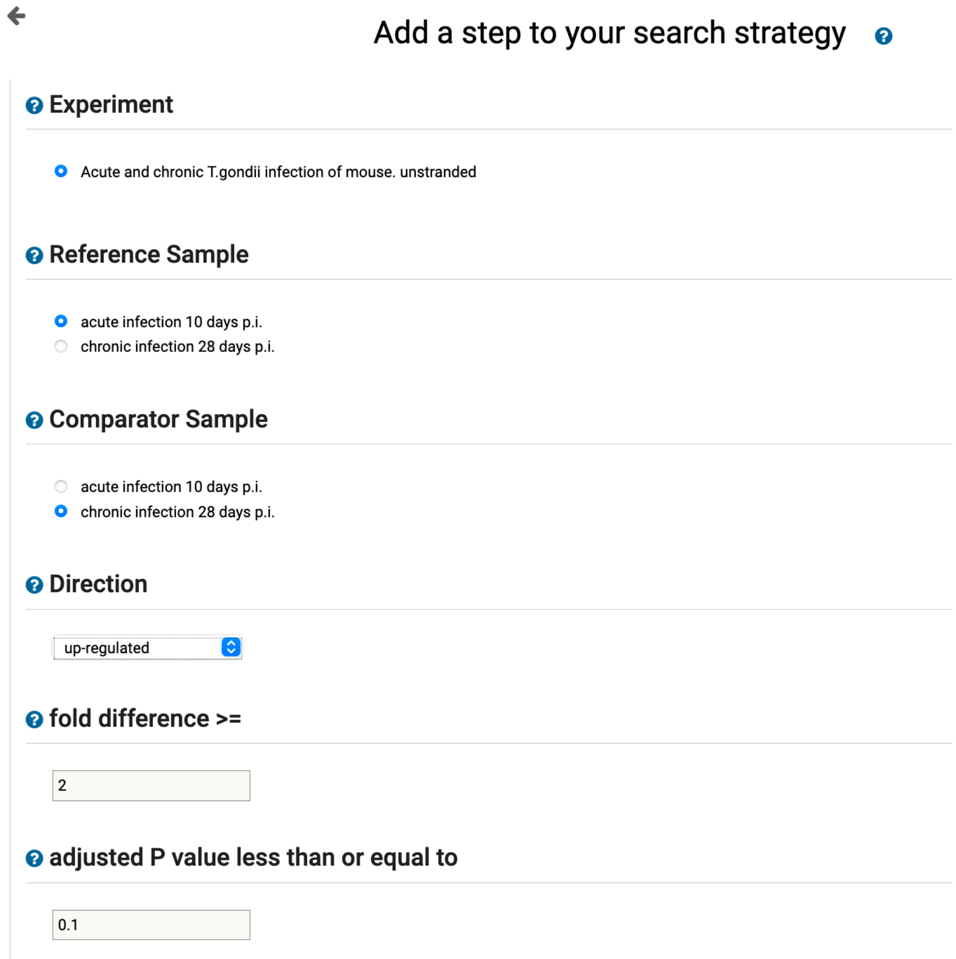
The results will be intersected with the results of Step 1.

Filter Data Sets: chroni

Legend: DE Differential Expression FC Fold Change P Percentile SA SenseAntisense

Organism	Data Set	Choose a Search
Toxoplasma gondii ME49	Transcriptome during acute or chronic infection in mouse brain (Pittman et al.)	DE FC P

- Intersect genes that are 2-fold upregulated in chronic stages compared to acute stages.



Add a step to your search strategy

Experiment

● Acute and chronic T.gondii infection of mouse. unstranded

Reference Sample

● acute infection 10 days p.i.
○ chronic infection 28 days p.i.

Comparator Sample

○ acute infection 10 days p.i.
● chronic infection 28 days p.i.

Direction

up-regulated

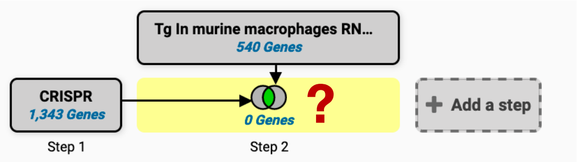
fold difference >=

2

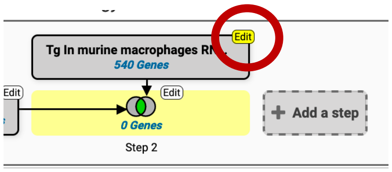
adjusted P value less than or equal to

0.1

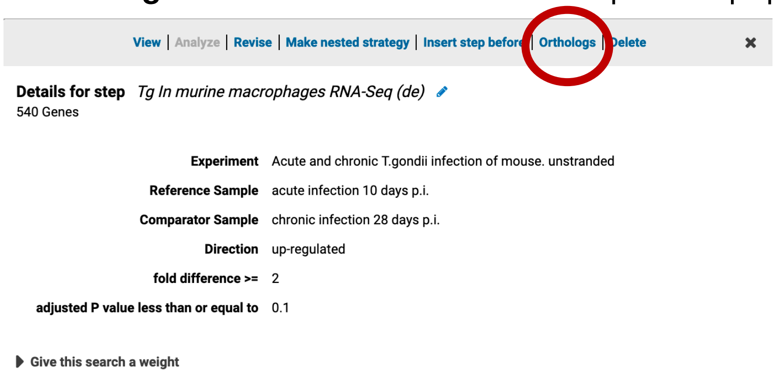
- Did you get zero results? This is to be expected since the CRISPR data was analyzed using the GT1 strain of *Toxoplasma* and the RNA-Seq data is from the ME49 strain. How can you fix this?



- Hint: transform the results in step 2 from *T. gondii* ME49 to *T. gondii GT1*. Click on the step edit button (move your mouse over the step and select edit).



- Select **orthologs** from the menu items at the top of the pop window.

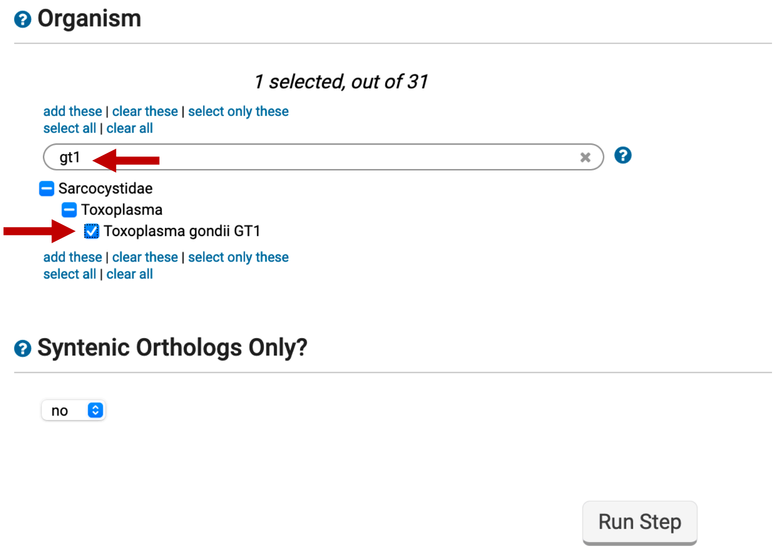

View | Analyze | Revise | Make nested strategy | Insert step before | Orthologs | Delete
✖

Details for step *Tg In murine macrophages RNA-Seq (de)*
540 Genes

Experiment	Acute and chronic T.gondii infection of mouse. unstranded
Reference Sample	acute infection 10 days p.i.
Comparator Sample	chronic infection 28 days p.i.
Direction	up-regulated
fold difference >=	2
adjusted P value less than or equal to	0.1

▸ Give this search a weight

- Select *T. gondii* GT1 from the list of organisms and click on Run Step.


add these | clear these | select only these
select all | clear all
add these | clear these | select only these
select all | clear all
Run Step

Organism

1 selected, out of 31

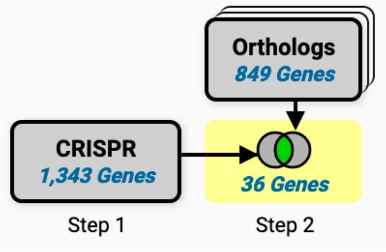
gt1

- Sarcocystidae
  - Toxoplasma
    - ☑ Toxoplasma gondii GT1

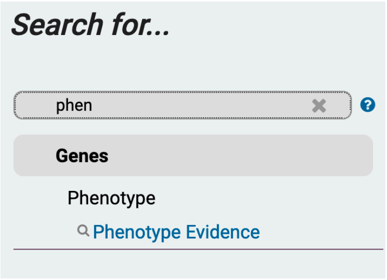
Syntenic Orthologs Only?

no

- Now what do your results look like?



3. **Identify essential *Plasmodium falciparum* genes that are highly expressed in schizont stages of the parasite.**
   Note for this exercise use https://plasmodb.org
   - You can start by exploring the phenotype data in PlasmoDB.
   - Select and run the search associated with the dataset: piggyBac insertion mutagenesis (John Adams).



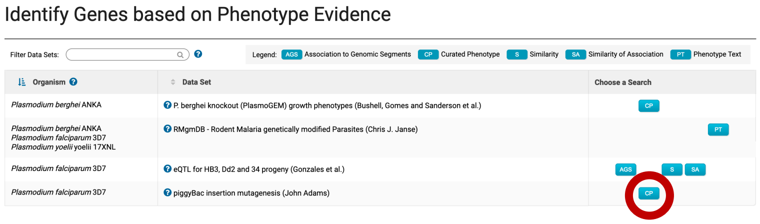
Search for...

phen

Genes

Phenotype

Phenotype Evidence



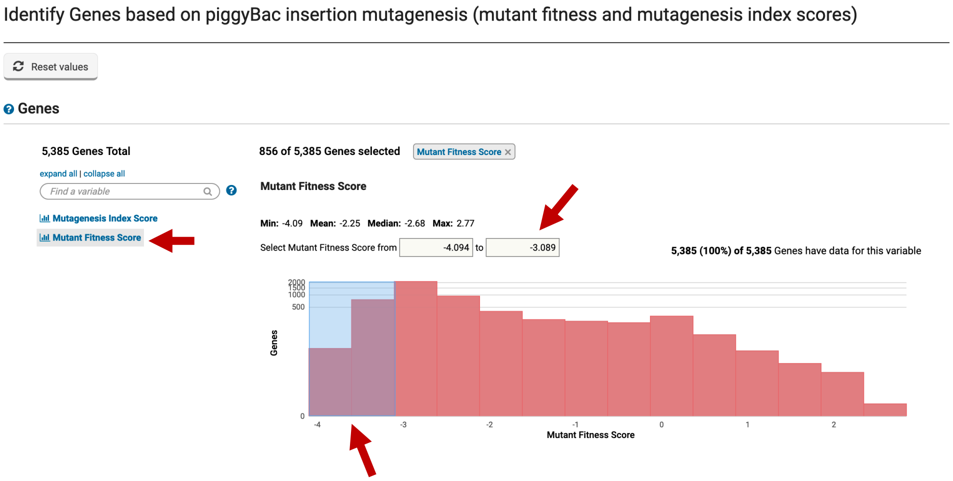
Identify Genes based on Phenotype Evidence

Filter Data Sets:

Legend: AGS Association to Genomic Segments CP Curated Phenotype S Similarity SA Similarity of Association PT Phenotype Text

Organism	Data Set	Choose a Search
Plasmodium berghei ANKA	P. berghei knockout (PlasmoGEM) growth phenotypes (Bushell, Gomes and Sanderson et al.)	CP
Plasmodium berghei ANKA *Plasmodium falciparum* 3D7 *Plasmodium yoelii* yoelii 17XNL	RMgmDB - Rodent Malaria genetically modified Parasites (Chris J. Janse)	PT
Plasmodium falciparum 3D7	eQTL for HB3, Dd2 and 34 progeny (Gonzales et al.)	AGS S SA
Plasmodium falciparum 3D7	piggyBac insertion mutagenesis (John Adams)	CP

   - Configure the search to identify genes with a *mutant fitness score* of less that -3. Note that you can select the range by either clicking and dragging you mouse over the histogram or by typing the values in the selection boxes.



Identify Genes based on piggyBac insertion mutagenesis (mutant fitness and mutagenesis index scores)

Reset values

Genes

5,385 Genes Total

expand all | collapse all

Find a variable

Mutagenesis Index Score

Mutant Fitness Score

856 of 5,385 Genes selected   Mutant Fitness Score ×

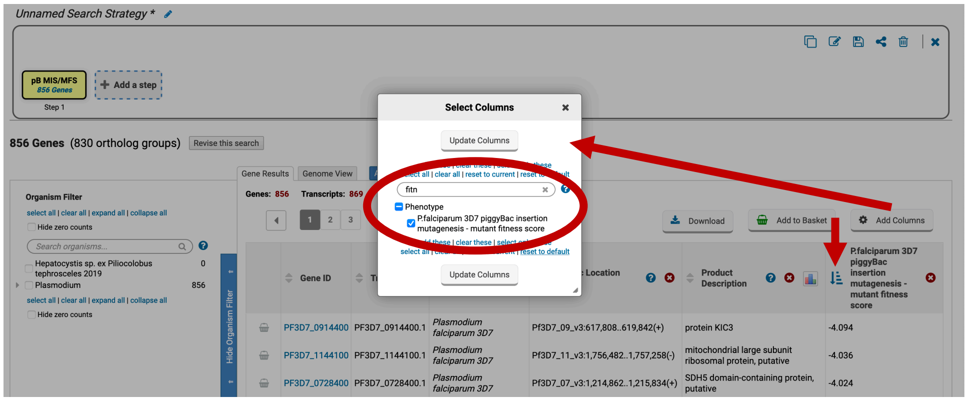
Mutant Fitness Score

Min: -4.09 **Mean:** -2.25 **Median:** -2.68 **Max:** 2.77

Select Mutant Fitness Score from -4.094 to -3.089

5,385 (100%) of 5,385 Genes have data for this variable

- How many genes did you identify? Which gene has the lowest fitness score? Note that you might need to add the fitness score column, by clicking on add columns then filtering the options with the word “fitness”.



Unnamed Search Strategy *

pB MIS/MFS 856 Genes — Step 1

856 Genes (830 ortholog groups)

Genes: 856 Transcripts: 869

Select Columns

fitn

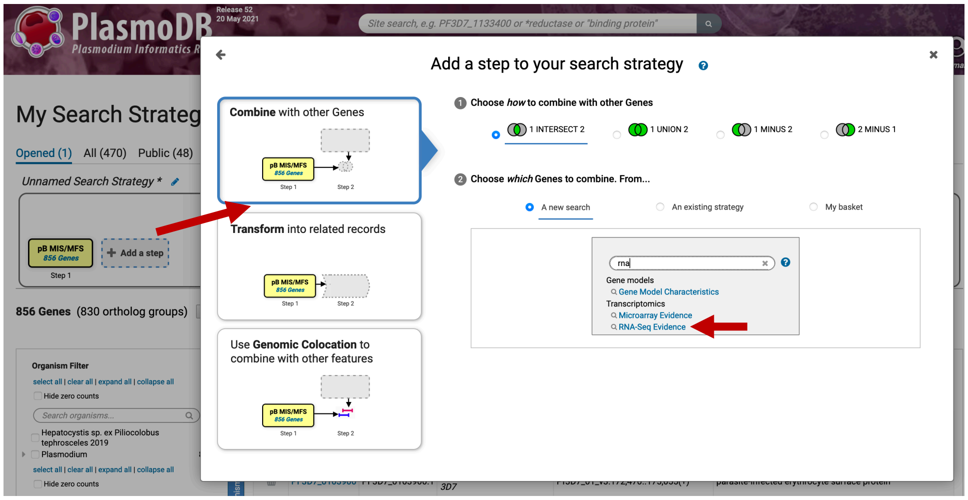
☐ Phenotype

☑ P.falciparum 3D7 piggyBac insertion mutagenesis - mutant fitness score

Gene ID	Transcript ID	Organism	Genomic Location	Product Description	P.falciparum 3D7 piggyBac insertion mutagenesis - mutant fitness score
PF3D7_0914400	PF3D7_0914400.1	*Plasmodium falciparum 3D7*	Pf3D7_09_v3:617,808..619,842(+)	protein KIC3	-4.094
PF3D7_1144100	PF3D7_1144100.1	*Plasmodium falciparum 3D7*	Pf3D7_11_v3:1,756,482..1,757,258(-)	mitochondrial large subunit ribosomal protein, putative	-4.036
PF3D7_0728400	PF3D7_0728400.1	*Plasmodium falciparum 3D7*	Pf3D7_07_v3:1,214,862..1,215,834(+)	SDH5 domain-containing protein, putative	-4.024

Add a step
Revise this search
Gene Results
Genome View
Organism Filter
select all | clear all | expand all | collapse all
Hide zero counts
Search organisms...
Hepatocystis sp. ex Piliocolobus tephrosceles 2019 0
Plasmodium 856
Hide Organism Filter
Update Columns
select all | clear all | reset to current | reset to default
add these | clear these | select only these
1 2 3
Download
Add to Basket
Add Columns

- Click on Add Step and find the RNA-Seq searches.



Add a step to your search strategy

1 Choose *how* to combine with other Genes

● 1 INTERSECT 2 ○ 1 UNION 2 ○ 1 MINUS 2 ○ 2 MINUS 1

2 Choose *which* Genes to combine. From...

● A new search ○ An existing strategy ○ My basket

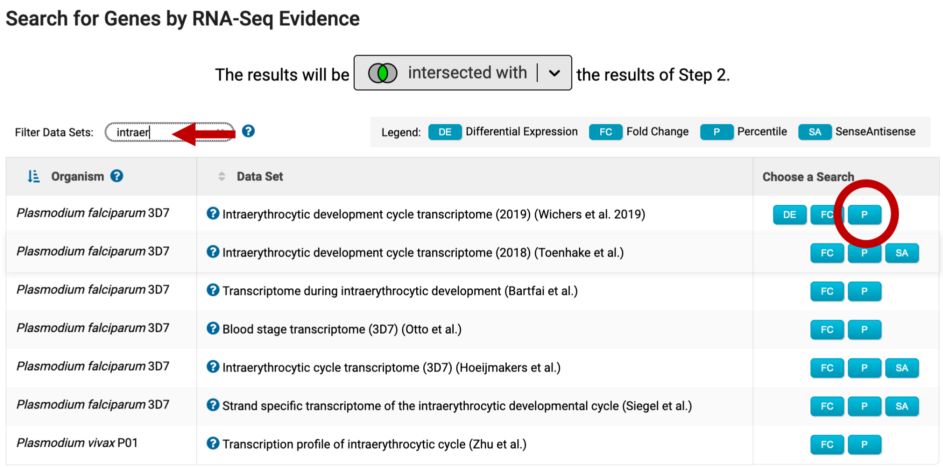
rna

Gene models
- Gene Model Characteristics

Transcriptomics
- Microarray Evidence
- RNA-Seq Evidence

PlasmoDB
Plasmodium Informatics R
Release 52
20 May 2021
Site search, e.g. PF3D7_1133400 or *reductase or "binding protein"
My Search Strateg
Opened (1) All (470) Public (48)
Unnamed Search Strategy *
pB MIS/MFS
856 Genes
Add a step
Step 1
Step 2
856 Genes (830 ortholog groups)
Combine with other Genes
Transform into related records
Use Genomic Colocation to combine with other features
Organism Filter
select all | clear all | expand all | collapse all
Hide zero counts
Search organisms...
Hepatocystis sp. ex Piliocolobus tephrosceles 2019
Plasmodium

- Find the search called “Intraerythrocytic development cycle transcriptome (2019)” and select the percentile search.



Search for Genes by RNA-Seq Evidence

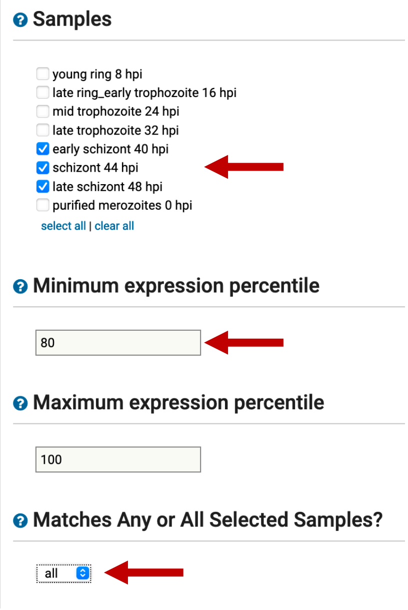
The results will be intersected with the results of Step 2.

Filter Data Sets: intraer

Legend: DE Differential Expression FC Fold Change P Percentile SA SenseAntisense

Organism	Data Set	Choose a Search
Plasmodium falciparum 3D7	Intraerythrocytic development cycle transcriptome (2019) (Wichers et al. 2019)	DE FC P
Plasmodium falciparum 3D7	Intraerythrocytic development cycle transcriptome (2018) (Toenhake et al.)	FC P SA
Plasmodium falciparum 3D7	Transcriptome during intraerythrocytic development (Bartfai et al.)	FC P
Plasmodium falciparum 3D7	Blood stage transcriptome (3D7) (Otto et al.)	FC P
Plasmodium falciparum 3D7	Intraerythrocytic cycle transcriptome (3D7) (Hoeijmakers et al.)	FC P SA
Plasmodium falciparum 3D7	Strand specific transcriptome of the intraerythrocytic developmental cycle (Siegel et al.)	FC P SA
Plasmodium vivax P01	Transcription profile of intraerythrocytic cycle (Zhu et al.)	FC P

- Configure the search to identify all genes that are in the 80-100 percentile in all three available schizont samples. Remember to change the parameter to require matching all samples.
- How many genes did you get? Are any of these genes interesting? How many are predicted to be secreted?



Samples

☐ young ring 8 hpi
☐ late ring_early trophozoite 16 hpi
☐ mid trophozoite 24 hpi
☐ late trophozoite 32 hpi
☑ early schizont 40 hpi
☑ schizont 44 hpi
☑ late schizont 48 hpi
☐ purified merozoites 0 hpi

select all | clear all

Minimum expression percentile

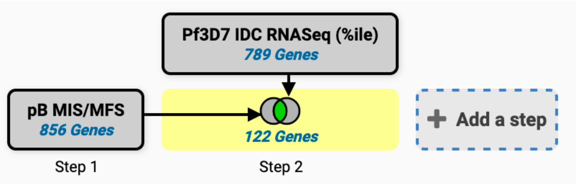
80

Maximum expression percentile

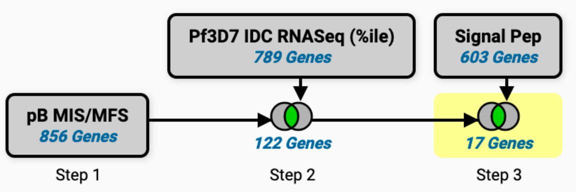
100

Matches Any or All Selected Samples?

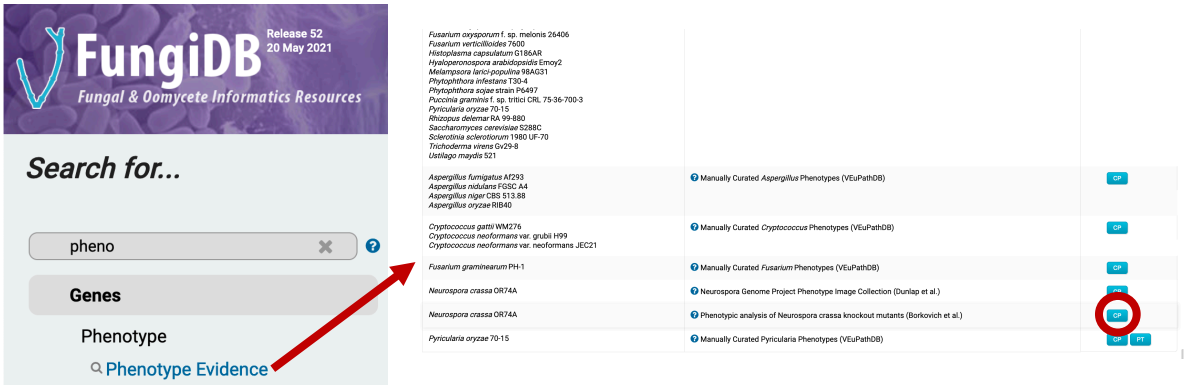
all



- How did you identify the secreted genes? Hint, add a step and search for genes that have a predicted secretory signal peptide.



4. **Identify Neurospora crassa genes that affect conidia formation. Note for the exercise use https://fungidb.org**
    - Start by locating the phenotype searches.



Search for...

pheno

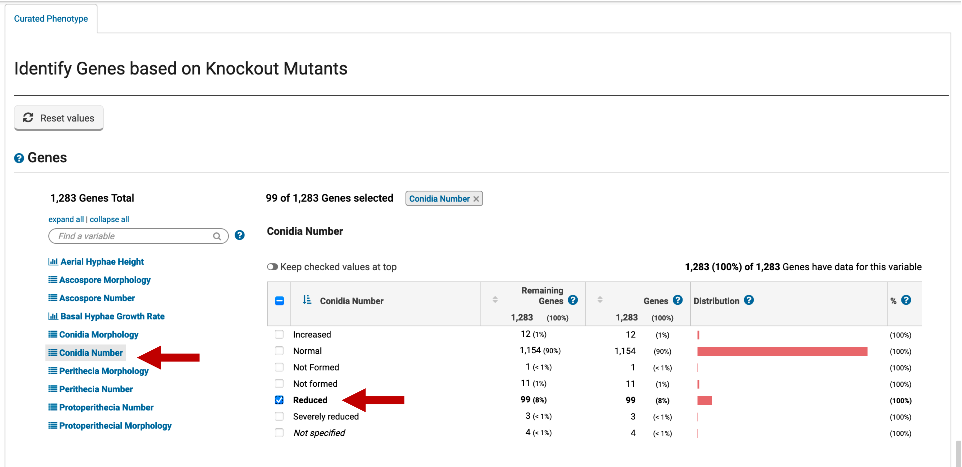
Genes

Phenotype

Phenotype Evidence

Fusarium oxysporum f. sp. melonis 26406 *Fusarium verticillioides* 7600 *Histoplasma capsulatum* G186AR *Hyaloperonospora arabidopsidis* Emoy2 *Melampsora larici-populina* 98AG31 *Phytophthora infestans* T30-4 *Phytophthora sojae* strain P6497 *Puccinia graminis* f. sp. tritici CRL 75-36-700-3 *Pyricularia oryzae* 70-15 *Rhizopus delemar* RA 99-880 *Saccharomyces cerevisiae* S288C *Sclerotinia sclerotiorum* 1980 UF-70 *Trichoderma virens* Gv29-8 *Ustilago maydis* 521		
Aspergillus fumigatus Af293 *Aspergillus nidulans* FGSC A4 *Aspergillus niger* CBS 513.88 *Aspergillus oryzae* RIB40	Manually Curated *Aspergillus* Phenotypes (VEuPathDB)	CP
Cryptococcus gattii WM276 *Cryptococcus neoformans* var. grubii H99 *Cryptococcus neoformans* var. neoformans JEC21	Manually Curated *Cryptococcus* Phenotypes (VEuPathDB)	CP
Fusarium graminearum PH-1	Manually Curated *Fusarium* Phenotypes (VEuPathDB)	CP
Neurospora crassa OR74A	Neurospora Genome Project Phenotype Image Collection (Dunlap et al.)	CP
Neurospora crassa OR74A	Phenotypic analysis of Neurospora crassa knockout mutants (Borkovich et al.)	CP
Pyricularia oryzae 70-15	Manually Curated Pyricularia Phenotypes (VEuPathDB)	CP PT

- This search provides you the option to filter based on categories on the left. Notice how when you select a different category on the left the filtering options in the middle change. Select the **Conidia number** category. Next select the "Reduced" value.



Identify Genes based on Knockout Mutants

Genes

1,283 Genes Total

99 of 1,283 Genes selected Conidia Number ×

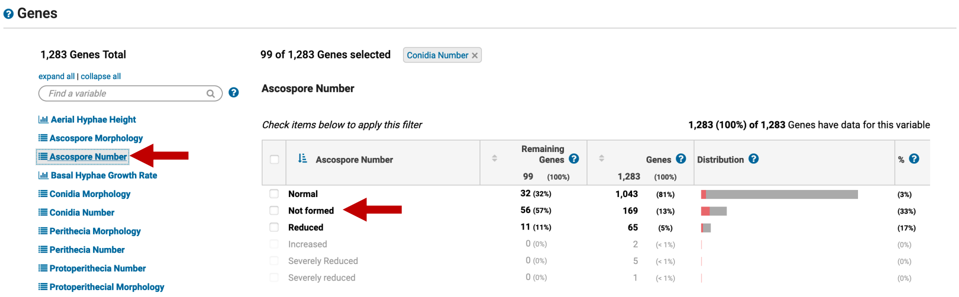
Conidia Number

Keep checked values at top

1,283 (100%) of 1,283 Genes have data for this variable

▬	Conidia Number	Remaining Genes 1,283 (100%)	Genes 1,283 (100%)	Distribution	%
□	Increased	12 (1%)	12 (1%)		(100%)
□	Normal	1,154 (90%)	1,154 (90%)		(100%)
□	Not Formed	1 (< 1%)	1 (< 1%)		(100%)
□	Not formed	11 (1%)	11 (1%)		(100%)
☑	**Reduced**	**99 (8%)**	**99 (8%)**		**(100%)**
□	Severely reduced	3 (< 1%)	3 (< 1%)		(100%)
□	*Not specified*	4 (< 1%)	4 (< 1%)		(100%)

- Notice that this search allows you to explore your results even before you click on the "Get Answer" button! Click around on the other categories on the left and see if the genes that are involved in a reduced number of conidia may also be involved in other phenotypes. For example, click on the **Ascospore Number** category, how maybe of your genes also have a phenotype with no ascospore formation?



Genes

1,283 Genes Total

99 of 1,283 Genes selected Conidia Number ×

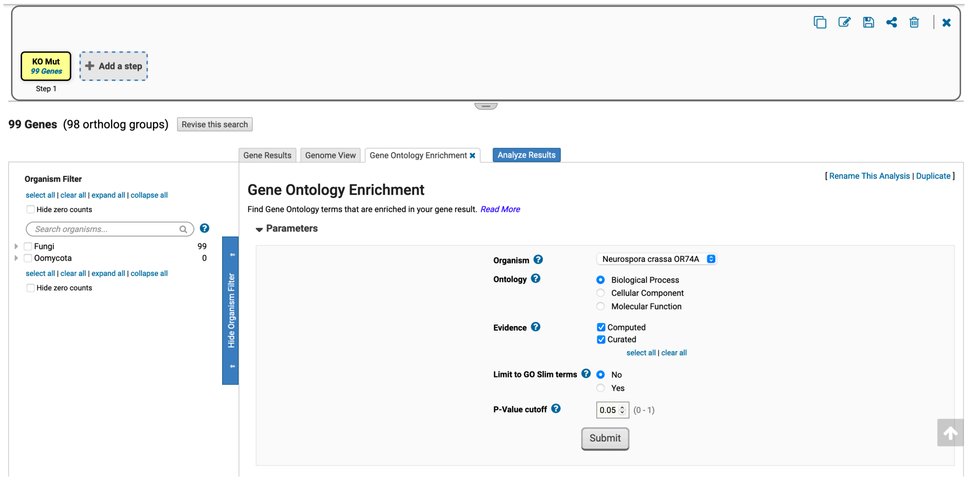
Ascospore Number

Check items below to apply this filter

1,283 (100%) of 1,283 Genes have data for this variable

□	Ascospore Number	Remaining Genes 99 (100%)	Genes 1,283 (100%)	Distribution	%
□	Normal	32 (32%)	1,043 (81%)		(3%)
□	Not formed	56 (57%)	169 (13%)		(33%)
□	Reduced	11 (11%)	65 (5%)		(17%)
□	Increased	0 (0%)	2 (< 1%)		(0%)
□	Severely Reduced	0 (0%)	5 (< 1%)		(0%)
□	Severely reduced	0 (0%)	1 (< 1%)		(0%)

- Click on get answer. What kinds of genes are in your results? Try analysing the results to see if there are any biological processes enriched in your results.



99 Genes (98 ortholog groups) Revise this search

Organism	Count
Fungi	99
Oomycota	0

Gene Ontology Enrichment

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

Parameters

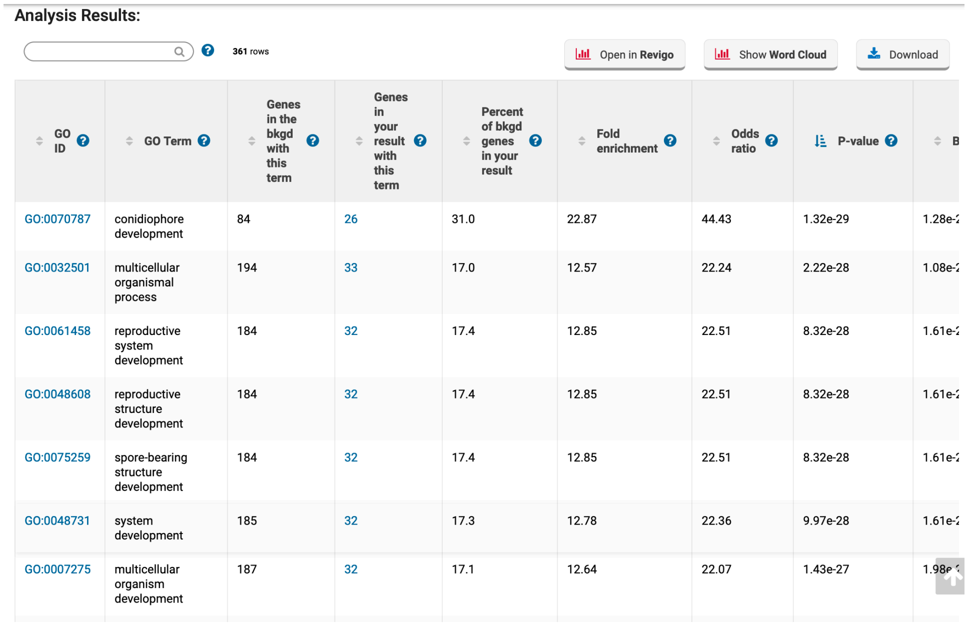
Organism: Neurospora crassa OR74A

Ontology: ● Biological Process ○ Cellular Component ○ Molecular Function

Evidence: ☑ Computed ☑ Curated

Limit to GO Slim terms: ● No ○ Yes

P-Value cutoff: 0.05 (0 - 1)



Analysis Results:

361 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	B
GO:0070787	conidiophore development	84	26	31.0	22.87	44.43	1.32e-29	1.28e-2
GO:0032501	multicellular organismal process	194	33	17.0	12.57	22.24	2.22e-28	1.08e-2
GO:0061458	reproductive system development	184	32	17.4	12.85	22.51	8.32e-28	1.61e-2
GO:0048608	reproductive structure development	184	32	17.4	12.85	22.51	8.32e-28	1.61e-2
GO:0075259	spore-bearing structure development	184	32	17.4	12.85	22.51	8.32e-28	1.61e-2
GO:0048731	system development	185	32	17.3	12.78	22.36	9.97e-28	1.61e-2
GO:0007275	multicellular organism development	187	32	17.1	12.64	22.07	1.43e-27	1.98e-