

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>

- 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.

The screenshot shows the TriTrypDB search interface. On the left, a sidebar titled "Search for..." contains a search bar with "cellu|" and a "Genes" section with "Protein targeting and localization" and "Cellular Localization Imaging" (highlighted with a red arrow). The main panel is titled "Identify Genes based on Cellular Localization Imaging" and includes a "Reset values" button. It has three sections: "Organism" with "Trypanosoma brucei brucei TREU927", "Location of tag" with "N-terminal" and "C-terminal" (the latter is selected with a red arrow), and "GO Term or GO ID" with "GO:0097542 : ciliary tip : 3" (highlighted with a red arrow).

- 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 terminal 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?

The screenshot shows the 'GO CL 48 Genes' interface. It includes a table with columns for 'Product Description', 'Transcripts', 'EC numbers', and 'Cellular localization images'. The 'Cellular localization images' column contains a grid of small images. A red arrow points from one of these images to a larger, expanded view of the images on the right side of the screen.

- 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.
- 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.

The screenshot shows the 'Combine with other Genes' and 'Transform into related records' sections. The 'Combine with other Genes' section has two sub-sections: 'Choose how to combine with other Genes' and 'Choose which Genes to combine. From...'. The 'Choose how to combine with other Genes' section has four options: '2 INTERSECT 3', '2 UNION 3', '2 MINUS 3', and '3 MINUS 2'. The 'Choose which Genes to combine. From...' section has three options: 'A new search', 'An existing strategy', and 'My basket'. The 'A new search' option is selected, and a search bar is shown with the text 'phen' and a dropdown menu with 'Phenotype' and 'Phenotype Evidence' options.

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

The screenshot shows the 'Add a step to your search strategy' dialog box. It has a title 'Search for Genes by Phenotype Evidence' and a subtitle 'The results will be [GO CL 48 Genes] intersected with [] the results of Step 2.' Below this is a 'Filter Data Sets' section with a search bar and a legend. The legend includes 'CP Curated Phenotype', 'PQ Quantitative Phenotype', and 'PT Phenotype Text'. Below the legend is a table with two columns: 'Organism' and 'Data Set'. The table has two rows: 'Trypanosoma brucei/brucei TREU927' and 'Trypanosoma brucei/brucei TREU927'. The first row has a data set 'High-throughput phenotyping using RNAi target sequencing (David Horn)' and the second row has a data set 'Sanger siRNA Phenotypes (Sanger)'. To the right of the table is a 'Choose a Search' section with three buttons: 'CP', 'PQ', and 'PT'. The 'PQ' button is circled in red.

- 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 ☐ Increase in coverage ☒ **Decrease in coverage**

with a **Fold change** \geq

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

in the following **Reference Samples**

☒ Uninduced sample

[select all](#) | [clear all](#)

and ☐ 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:

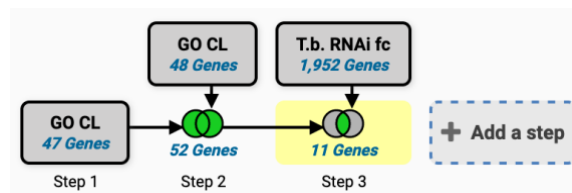
$$\text{fold change} = \frac{\text{reference expression value}}{\text{maximum expression value in comparison}}$$

and returns genes when **fold change** \geq 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.

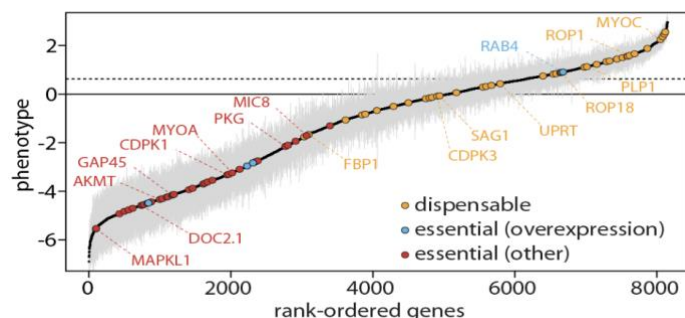
- How many genes did you get?



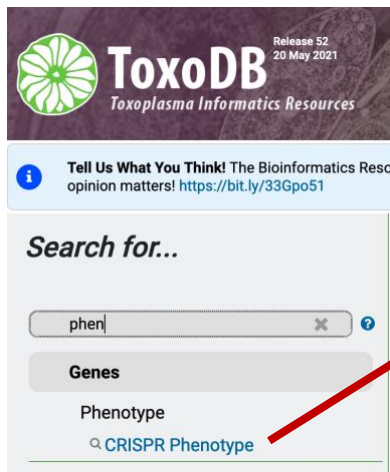
2. Finding genes based on high throughput mutagenesis and fitness analysis.

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



Identify Genes based on CRISPR Phenotype

Phenotype Score >=

-6.89

Phenotype Score <=

-4

CRISPR
1,343 Genes

+ Add a step

Step 1

graph and the number of genes returned in your search results?

- 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 your 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 [Open in Revigo](#) [Show Word Cloud](#) [Download](#)

GO ID	GO Term	Genes in the bgkd with this term	Genes in your result with this term	Percent of bgkd 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

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

- 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.

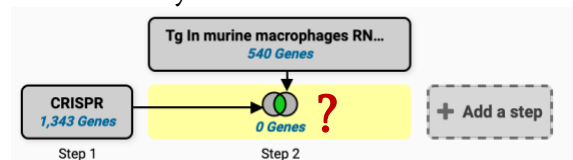
Direction

up-regulated

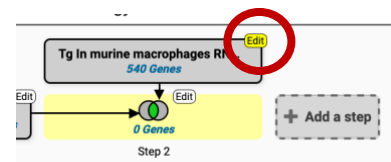
fold difference >=

adjusted P value less than or equal to

- 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).



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 **orthologs** from the menu items at the top of the pop window.
- Select *T. gondii* GT1 from the list of organisms and click on Run Step.
- Now what do your results look like?

Organism

1 selected, out of 31

add these | clear these | select only these
select all | clear all

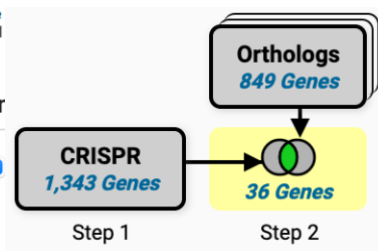
gt1

☐ Sarcocystidae
☐ Toxoplasma
☒ Toxoplasma gondii GT1

add these
select all |

Synter

no



Run Step

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

Search for...

phen

Genes

Phenotype

Phenotype Evidence

Identify Genes based on Phenotype Evidence

Filter Data Sets: Legend: Assoc. Association to Genomic Segments Curated Phenotype S Similarity SA Similarity of Association PT Phenotype Text

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

mutagenesis (John Adams).

- Configure the search to identify genes with a *mutant fitness score* of less than -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 to

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”.



The screenshot shows the PlasmoDB search results for 'pB MIS/MFS 856 Genes'. A 'Select Columns' dialog box is open, allowing the user to choose which columns to display. The search bar in the dialog contains the text 'fit', and the results list includes 'P.falciparum 3D7 piggyBac insertion mutagenesis - mutant fitness score'. A red arrow points from the 'Add Columns' button in the top right of the search results to the dialog box.


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

The screenshot shows the 'Add a step to your search strategy' dialog box in PlasmoDB. The dialog box has three main sections: 'Combine with other Genes', 'Transform into related records', and 'Use Genomic Colocation to combine with other features'. The 'Combine with other Genes' section has three options: '1 INTERSECT 2', '1 UNION 2', and '1 MINUS 2'. The 'Choose which Genes to combine. From...' section has three options: 'A new search', 'An existing strategy', and 'My basket'. A red arrow points to the 'RNA-Seq Evidence' option in the 'Choose which Genes to combine. From...' section.

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

Search for Genes by RNA-Seq Evidence

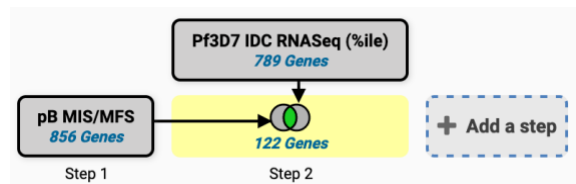
The results will be  intersected with  the results of Step 2.

Filter Data Sets: 

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

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

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

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

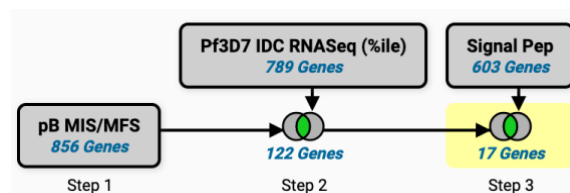
80

Maximum expression percentile

100

Matches Any or All Selected Samples?

all 



4. Identify *Neurospora crassa* genes that affect conidia formation.

Note for the exercise use <https://fungidb.org>

- Start by locating the phenotype searches.

FungiDB Release 52, 20 May 2021
Fungal & Oomycete Informatics Resources

Search for...

pheno

Genes

Phenotype

Phenotype Evidence

Species	Phenotype	Source
<i>Fusarium oxysporum</i> f. sp. melonis 26406		
<i>Fusarium verticillioides</i> 7603		
<i>Histioplasma capsulatum</i> G1 B6A3		
<i>Hyaloperonospora arabidopsidis</i> Emoy2		
<i>Metarhizium larici populina</i> 98A031		
<i>Phytophthora infestans</i> T304		
<i>Phytophthora sojae</i> strain P6497		
<i>Puccinia graminis</i> f. sp. tritici CRL 75-36-700-3		
<i>Pyricularia oryzae</i> 70-15		
<i>Rhizopus delemar</i> RA 99-880		
<i>Saccharomyces cerevisiae</i> S288C		
<i>Sclerotinia sclerotiorum</i> 1980 UF-70		
<i>Trichoderma reesei</i> QM949		
<i>Ustilago maydis</i> S21		
<i>Aspergillus fumigatus</i> AF293	Manually Curated Aspergillus Phenotypes (VFuPathDB)	CP
<i>Aspergillus nidulans</i> FGSC A4		
<i>Aspergillus niger</i> CBS 513.05		
<i>Aspergillus oryzae</i> RB80		
<i>Cryptococcus gatti</i> WM276	Manually Curated Cryptococcus Phenotypes (VFuPathDB)	CP
<i>Cryptococcus neoformans</i> var. grubii H99		
<i>Cryptococcus neoformans</i> var. neoformans JEC21		
<i>Fusarium graminearum</i> PH-1	Manually Curated Fusarium Phenotypes (VFuPathDB)	CP
<i>Neurospora crassa</i> OR74A	Neurospora Genome Project Phenotype Image Collection (Dunlap et al.)	CP
<i>Neurospora crassa</i> OR74A	Phenotypic analysis of <i>Neurospora crassa</i> knockout mutants (Borkovich et al.)	CP
<i>Pyricularia oryzae</i> 70-15	Manually Curated Pyricularia Phenotypes (VFuPathDB)	CP

- 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.

Curated Phenotype

Identify Genes based on Knockout Mutants

Reset values

Genes

1,283 Genes Total

99 of 1,283 Genes selected

Conidia Number

Keep checked values at top

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

Conidia Number	Remaining Genes	Genes	Distribution	%
<input type="checkbox"/> Increased	12 (1%)	12 (1%)		(100%)
<input type="checkbox"/> Normal	1,154 (90%)	1,154 (90%)		(100%)
<input type="checkbox"/> Not Formed	1 (< 1%)	1 (< 1%)		(100%)
<input type="checkbox"/> Not formed	11 (1%)	11 (1%)		(100%)
<input checked="" type="checkbox"/> Reduced	99 (8%)	99 (8%)		(100%)
<input type="checkbox"/> Severely reduced	3 (< 1%)	3 (< 1%)		(100%)
<input type="checkbox"/> 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
 expand all | collapse all
 Find a variable

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

Ascospore Number

Check items below to apply this filter

	Remaining Genes	Genes	Distribution	%
<input type="checkbox"/> Normal	32 (32%)	1,043 (81%)		(3%)
<input type="checkbox"/> Not formed	56 (57%)	169 (13%)		(33%)
<input type="checkbox"/> Reduced	11 (11%)	65 (5%)		(17%)
<input type="checkbox"/> Increased	0 (0%)	2 (< 1%)		(0%)
<input type="checkbox"/> Severely Reduced	0 (0%)	5 (< 1%)		(0%)
<input type="checkbox"/> 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.

Step 1

KO Mut 99 Genes

Add a step

99 Genes (98 ortholog groups) [Revise this search](#)

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

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 - 1)

[Submit](#)

Analysis Results:

361 rows

[Open in Revigo](#) [Show Word Cloud](#) [Download](#)

GO ID	GO Term	Genes in the bkgd with this term	Genes in your result with this term	Percent of bkgd genes in your result	Fold enrichment	Odds ratio	P-value	B
GO:0070787	conidiophore development	84	26	31.0	22.87	44.43	1.32e-29	1.28e-29
GO:0032501	multicellular organismal process	194	33	17.0	12.57	22.24	2.22e-28	1.08e-28
GO:0061458	reproductive system development	184	32	17.4	12.85	22.51	8.32e-28	1.61e-28
GO:0048608	reproductive structure development	184	32	17.4	12.85	22.51	8.32e-28	1.61e-28
GO:0075259	spore-bearing structure development	184	32	17.4	12.85	22.51	8.32e-28	1.61e-28
GO:0048731	system development	185	32	17.3	12.78	22.36	9.97e-28	1.61e-28
GO:0007275	multicellular organism development	187	32	17.1	12.64	22.07	1.43e-27	1.98e-27