

## Exploring Transcriptomic data

### 1. Exploring RNA sequence data in *Plasmodium falciparum*.

Note: For this exercise use <http://www.plasmodb.org>

- a. Find all genes in *P. falciparum* that are up-regulated during the later stages of the intraerythrocytic cycle.
  - Hint: Use the fold change search for the data set “**Transcriptome during intraerythrocytic development (Bartfai et al.)**”. For this data set, synchronized Pf3D7 parasites were assayed by RNA-seq at 8 time-points during the iRBC cycle. We want to find genes that are up-regulated in the later time points (30, 35, 40 hours) using the early time points (5, 10, 15, 20, 25 hours) as reference.

The image shows two screenshots from the Plasmodb.org website. The top screenshot, titled "Identify Genes based on RNA Seq Evidence", shows a search results table with columns for Organism and Data Set. The "Filter Data Sets" field is set to "development". The table lists four data sets for *P. falciparum* 3D7, with the third one, "Transcriptome during intraerythrocytic development (Bartfai et al.)", highlighted. A red circle highlights the "FC" button in the "Choose a search" column for this data set. The bottom screenshot shows the search configuration page for the selected data set. It includes a "Fold Change" field set to "12" and a "Direction" dropdown set to "Up or down regulated". A graph titled "Up or down regulated" shows a line graph with a rising slope, indicating up-regulation. The "Get Answer" button is visible at the bottom.

- There are a number of parameters to manipulate in this search. As you modify parameters on the left side note the dynamic help on the right side. See screenshots.
- **Direction:** the direction of change in expression. **Choose up-regulated.**
- **Fold Change** $\geq$  the intensity of difference in expression needed before a gene is returned by the search. **Choose 12** but feel free to modify this.

- **Reference Sample:** the samples that will serve as the reference when comparing expression between samples. **choose 5, 10, 15, 20, 25**
- **Between each gene's AVERAGE expression value:** This parameter appears once you have chosen two Reference Samples and defines the operation applied to reference samples. Fold change is calculated as the ratio of two values (upregulated ratio = expression in comparison)/(expression in reference). When you choose multiple samples to serve as reference, we generate one number for the fold change calculation by using the minimum, maximum, or average. **Choose average**
- **(or a Floor of 10 reads):** This parameter defines a lower limit of aligned reads for a gene to avoid unreliable fold change calculations. (Low numbers of aligned reads means low expression but the low values may be may be technically inaccurate. Dividing by small numbers creates large numbers.  $2000\text{FPKM}/10 = 200$ ;  $2000/0.1 = 20,000$ ) If a gene has fewer than 10 aligned reads, it is assigned 10 reads before the fold change calculation is made. **Leave this as default at 10 reads.**
- **Comparison Sample:** the sample that you are comparing to the reference. In this case you are interested in genes that are up-regulated in later time points **choose 30, 35, 40**
- **And its AVERAGE expression value:** This parameter appears once you have chosen two Comparison Samples and defines the operation applied to comparison samples. See explanation above. **Choose average**

## Identify Genes based on P. falciparum 3D7 Transcriptome during intraerythrocytic development RNASeq (fold change) Tutorial

For the Experiment: Transcriptome during intraerythrocytic development scaled unstranded

return: protein coding Genes

that are: up-regulated

with a Fold change >= 12

between each gene's: average expression value  
(or a Floor of 10 reads (1 FPKM))

in the following: Reference Samples

- Hour 5
- Hour 10
- Hour 15
- Hour 20
- Hour 25

select all | clear all

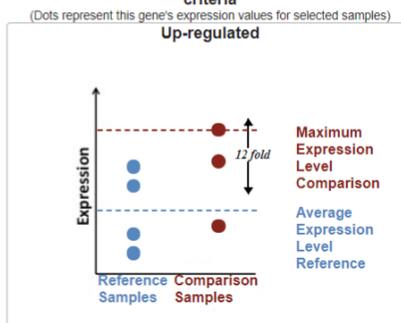
and its: maximum expression value  
(or the Floor selected above)

in the following: Comparison Samples

- Hour 20
- Hour 25
- Hour 30
- Hour 35
- Hour 40

select all | clear all

### Example showing one gene that would meet search criteria



*A maximum of four samples are shown when more than four are selected.*  
You are searching for genes that are **up-regulated** between at least two **reference samples** and at least two **comparison samples**.

For each gene, the search calculates:

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

and returns genes when **fold change** >= 12

To narrow the window, use the maximum reference value, or average or minimum comparison value. To broaden the window, use the minimum reference value.

See the [detailed help](#) for this search.

\* or FPKM Floor, whichever is greater

Get Answer

**3D7 iRBC RNASeq**  
 969 Genes  
**Step 1**

Add Step

b. For the genes returned by the search, how does the RNA-sequence data compare to microarray data?

- Hint: PlasmoDB contains data from a similar experiment that was analyzed by microarray instead of RNA sequencing. This experiment is called: **Erythrocytic expression time series (3D7, DD2, HB3) (Bozdech et al. and Linas et al.)**. IDC 48 hr Marray – Expr Graph shows normalized expression values. To directly compare the data for genes returned by the RNA-seq search that you just ran, add the column called “Pf-iRBC 48hr - Graph”.

The screenshot displays the PlasmoDB interface. At the top, a strategy named "3D7 iRBC RNASeq (fc)" is shown with a yellow box around the search results and an "Add Step" button. Below this, a table lists 978 genes from Step 1. A "Select Columns" dialog box is open, showing a list of available columns. The "IDC 48 hr Marray - Expr Graph" option is selected and highlighted with a red box. Below the dialog, three graphs are displayed: two for "3D7 iRBC RNASeq - fpkm Graph" and one for "IDC 48 hr Marray - Expr Graph". The graphs show expression levels over time for a specific gene (PF3D7\_0207600).

OPTIONAL: You can also run a fold change search using this experiment to compare results on a genome scale. Add a step to your strategy and intersect your current results (genes upregulated 12 fold in later IDC time periods) with a fold change search using the “Erythrocytic expression time series (3D7, Dd2, HB3) (Bozdech et al. and Linas et al.)” experiment (under microarray evidence). Configure it similarly to the RNA-seq experiment although you will probably need to make the fold change smaller (try 2 or 3) due to the decreased dynamic range of microarrays compared to RNA-seq.

**Add Step**

**Add Step 2 : P.falciparum Erythrocytic expression time series (3D7, DD2, HB3) Microarray (fold change)**

For the **Experiment** | IRBC HB3 (48 Hour scaled) | **Genes**  
 return | protein coding | **Genes**  
 that are | up-regulated | **Genes**  
 with a **Fold change** >= 2 | **Genes**  
 between each gene's | average | expression value | **Genes**  
 in the following | **Reference Samples** | **Genes**

**28 selected, out of 46**

Filter list below...

- 1-16 Hours
- 17-30 Hours
- 31-48 Hours

select all | clear all | expand all | collapse all

and its | average | expression value | **Genes**  
 in the following | **Comparison Samples** | **Genes**

**18 selected, out of 46**

Filter list below...

- 1-16 Hours
- 17-30 Hours
- 31-48 Hours
  - 31-39 Hours
  - 40-48 Hours

select all | clear all | expand all | collapse all

**Up-regulated**

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

A maximum of four samples are shown when more than four are selected.

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

For each gene, the search calculates:

$$\text{fold change} = \frac{\text{average expression level in comparison}}{\text{average expression level in reference}}$$

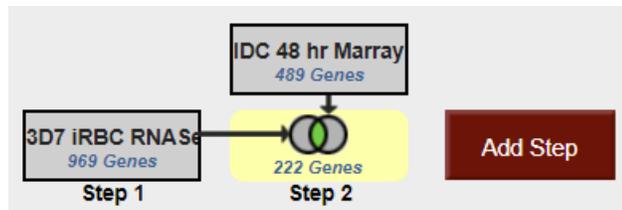
and returns genes when **fold change >= 2**.

To narrow the window, use the maximum reference value, or minimum comparison value. To broaden the window, use the minimum reference value, or maximum comparison value.

See the detailed help for this search.

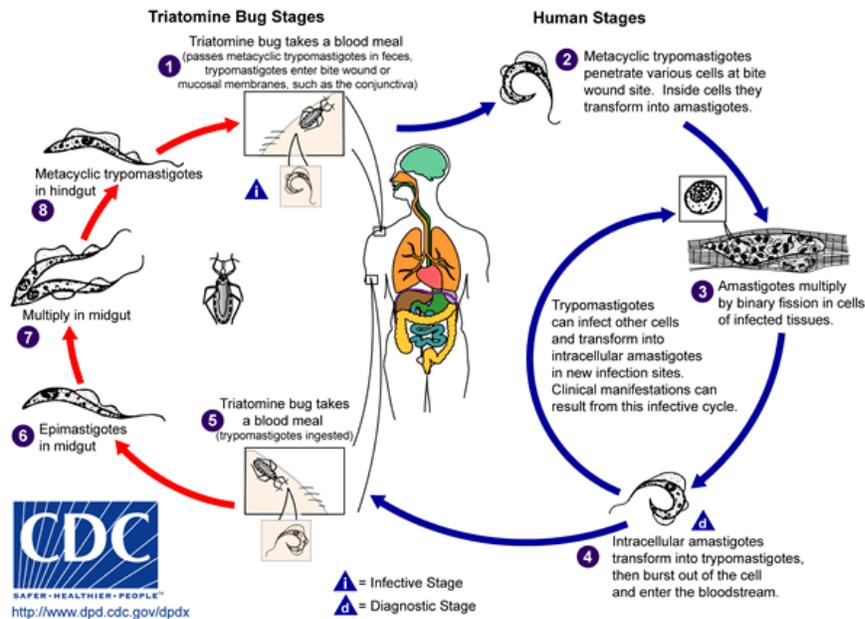
Combine Genes in Step 1 with Genes in Step 2:

- 1 Intersect 2
- 1 Minus 2
- 1 Union 2
- 2 Minus 1
- 1 Relative to 2, using genomic colocation



## 2. Exploring microarray data in TriTrypDB.

Note: For this exercise use <http://www.tritrypdb.org>



- a. Find *T. cruzi* protein coding genes that are upregulated in amastigotes compared to trypomastigotes. Go to the transcript expression section then select **microarray**. Choose the fold change (FC) search for the data set called: **Transcriptomes of Four Life-Cycle Stages (Minning et al.)**.

Fold Change | Percentile

### Identify Genes based on *T. cruzi* CL Brener Esmeraldo-like Transcriptomes of Four Life-Cycle Stages Microarray (fold change)

Tutorial

For the Experiment  
 Transcriptomes of Four Life-Cycle Stages trcuCLBrenerEsmeraldo-lik

return **protein coding** Genes  
 that are **up-regulated**  
 with a **Fold change >= 2.0**

between each gene's expression value  
 in the following **Reference Samples**

- amastigotes
- trypomastigotes
- epimastigotes
- metacyclics

select all | clear all

and its expression value  
 in the following **Comparison Samples**

- amastigotes
- trypomastigotes
- epimastigotes
- metacyclics

select all | clear all

#### Example showing one gene that would meet search criteria

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

**Up-regulated**

Comparison  
Reference  
Reference Samples Comparison Samples

You are searching for genes that are up-regulated between one reference sample and one comparison sample.

For each gene, the search calculates:

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

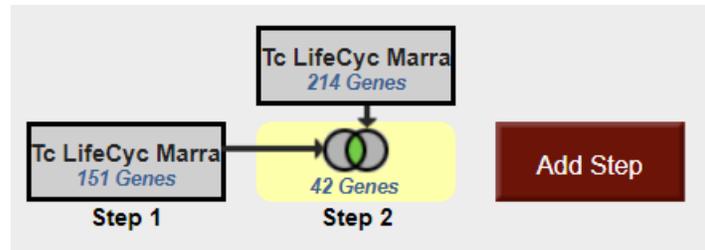
and returns genes when **fold change >= 2.0**.

See the [detailed help](#) for this search.

Advanced Parameters

Get Answer

- Select the direction of regulation, your reference sample and your comparison sample. For the fold change keep the default value 2.
- How many genes did you find? Do the results seem plausible?
- Are any of these genes also up-regulated in the replicative insect stage compared to the transmissive insect stage? How can you find this out? (*Hint*: add a step and run a microarray search comparing expression of epimastigotes to metacyclics).



- Do these genes have orthologs in other kinetoplastids? (*Hint*: add a step and transform your results into orthologs in all other organisms in TriTrypDB (select all for the ortholog transform).

How many orthologs exist in *L. braziliensis* MHOM/BR/75/M2903? (*Hint*: look at the filter table between the strategy panel and your result list. Click on the number in the table under a species to view results from a specific species). Explore your results. Scan the product descriptions for this list of genes. Did you find anything interesting? Perhaps a GO enrichment analysis would support your ideas.

48 Genes from Step 3 [Revise](#)

Strategy: *Tc LifeCyc Marra (fc)*

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

All Results	Ortholog Groups	<i>B. ayaxii</i>	<i>B. bodoi</i>	<i>C. trichidia</i>	<i>E. montenegroi</i>	<i>Laethiopicca</i>	<i>L. amazonensis</i>	<i>L. arabica</i>	<i>L. braziliensis</i> (99)	<i>L. donovani</i> (149)	<i>L....</i>
2261	36	36	60	66	47	48	38	46	48	51	46

Gene Results | Genome View | Gene Ontology Enrichment\* | [Analyze Results](#)

Gene Ontology Enrichment [ Rename This Analysis | Duplicate ]

Find Gene Ontology terms that are enriched in your gene result. [Read More](#)

Parameters

Organism: *Leishmania braziliensis* MHOM/BR/75/M2903

Ontology:  Cellular Component  Molecular Function  Biological Process

Evidence:  Computed  Curated

Limit to GO Slim terms:  No  Yes

P-Value cutoff: 0.05 (0 - 1)

[Submit](#)

3. Finding genes based on RNAseq evidence and inferring function of hypothetical genes.  
 Note: Use <http://plasmodb.org> for this exercise.

- a. Find all genes in *P. falciparum* that are up-regulated at least 50-fold in ookinetes compared to other stages: “Transcriptomes of 7 sexual and asexual life stages (Lopez-Barragan et al.)”. For this search select “average” for the operation applied on the reference samples.

### Identify Genes based on P. falciparum 3D7 Transcriptomes of 7 sexual and asexual life stages RNASeq (fold change) Tutorial

For the Experiment: Transcriptomes of 7 sexual and asexual life stages unstranded

return: protein coding Genes

that are: up-regulated

with a Fold change >= 50

between each gene's: average expression value

(or a Floor of 10 reads (.88 FPKM) NEW)

in the following: Reference Samples

- Ring
- Early Trophozoite
- Late Trophozoite
- Schizont
- Gametocyte II

select all | clear all

and its expression value

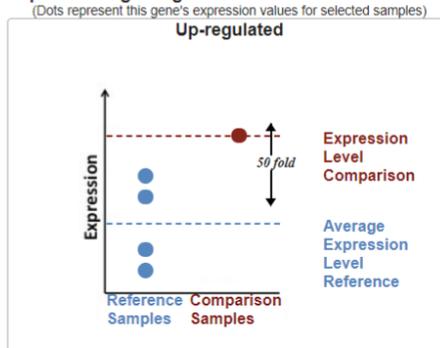
(or the Floor selected above)

in the following: Comparison Samples

- Late Trophozoite
- Schizont
- Gametocyte II
- Gametocyte V
- Ookinete

select all | clear all

**Example showing one gene that would meet search criteria**



*A maximum of four samples are shown when more than four are selected.*  
 You are searching for genes that are **up-regulated** between at least two reference samples and one comparison sample.

For each gene, the search calculates:

$$\text{fold change} = \frac{\text{comparison expression level}}{\text{average expression level in reference}^*}$$

and returns genes when **fold change >= 50**.

To narrow the window, use the maximum reference value. To broaden the window, use the minimum reference value.

See the detailed help for this search.

\* or FPKM Floor, whichever is greater

Get Answer

- b. The above search will give you all genes that are up-regulated by 50 fold in ookinetes compared to the average expression level of other stages. Despite the high fold change, some genes in the list may be highly expressed in the other stages. How can you remove genes from the list that are highly expressed in the other stages?

**3D7 7 Stages RNA**

94 Genes

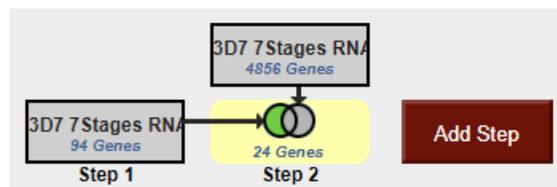
**Step 1**

Add Step

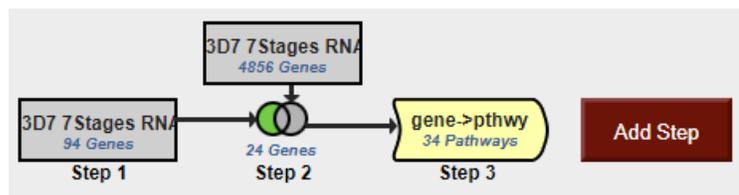
- Hint: Add a search for genes based on RNA Seq evidence from the same experiment, but this time select the percentile search: *P.f.* seven stages - RNA Seq (percentile). What

minimal percentile values should you choose? 40 – 100%? How does setting the any / all samples impact the result .... Which would be better in this case?

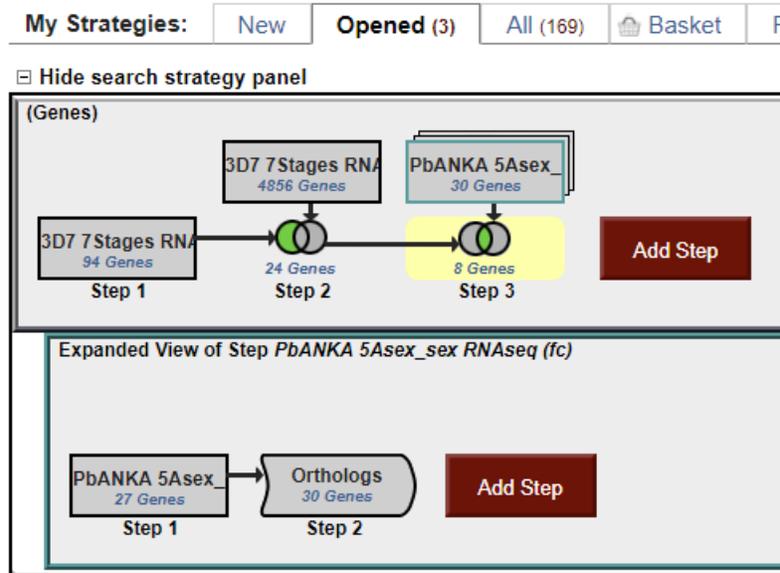
- Hint II: Try changing the operator from average to maximum for the set of non-ookinete stages in your initial fold change search. What does this do? How do the resulting genes compare with the two step strategy you generated in the first hint? Which hint do you think works better?



- c. Which metabolic pathways are represented in this gene list? Hint: add a step and transform results to pathways. How does this result compare to running a pathways enrichment on step 2?



- d. What happens if you revise the first step and modify the fold difference to a lower value - 10 for example? Compare results when you also modify the “between each genes” parameter. What happens if you set this to maximum? Which value do you think is most stringent for ensuring at 10 fold up regulation compared to the other samples?
- e. PlasmoDB also has an experiment examining gene expression during sexual development in *Plasmodium berghei* (rodent malaria). Can you determine if there are genes that are up-regulated in both human and rodent ookinetes (compared to all other stages)? *Hint*: start by deleting the last step you added in this exercise (transform to pathways). To do this click on edit then delete in the popup. Next, add steps for the *P. berghei* experiments “P berghei ANKA 5 asexual and sexual stage transcriptomes RNASeq”. Note that you will have to use a nested strategy or by running a separate strategy then combining both strategies.



4. Find genes that are essential in procyclics but not in blood form *T. brucei*.  
 Note: for this exercise use <http://TriTrypDB.org>.
- Find the query for High Throughput Phenotyping. Think about how to set up this query (*Hint*: you will have to set up a two-step strategy). Remember you can play around with the parameters but there is no one correct way of setting them up –

Quantitative Phenotype Learn more about this search

## Identify Genes based on High-Throughput Phenotyping Tutorial

For the **Experiment** Quantitated from the CDS Sequence return protein coding **Genes** that are Decrease in coverage with a **Fold change**  $\geq 1.5$

between each gene's **expression value** in the following **Reference Samples**

- Uninduced sample

and its **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)

**Down-regulated**

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

For each gene, the search calculates:

$$\text{fold change} = \frac{\text{reference expression level}}{\text{comparison expression level}}$$

and returns genes when **fold change**  $\geq 1.5$ .  
See the detailed help for this search.

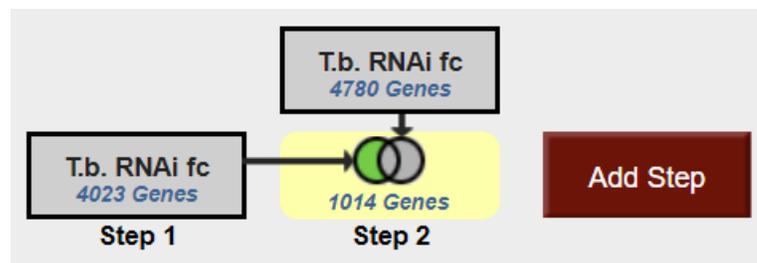
[Get Answer](#)

**T.b. RNAi fc**  
4023 Genes

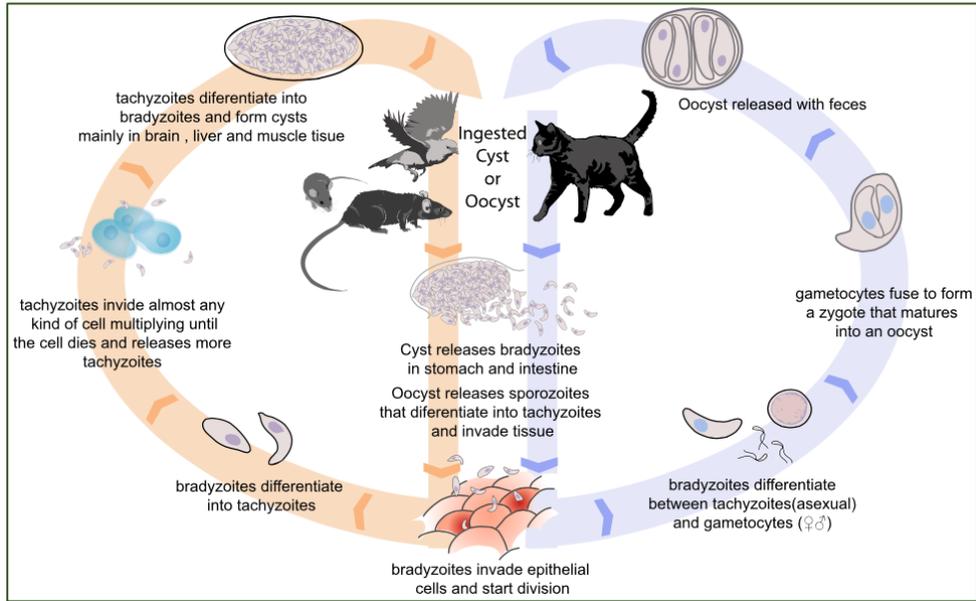
**Step 1**

Add Step

- Next add a step and run the same search except this time select the “induced bloodstream form” samples.
- How did you combine the results? Remember you want to find genes that are essential in procyclics and not in blood form.



5. Finding oocyst expressed genes in *T. gondii* based on microarray evidence.  
 Note: For this exercise use <http://toxodb.org>



- a. Find genes that are expressed at 10 fold higher levels in one of the oocyst stages than in any other stage in the “Oocyst, tachyzoite, and bradyzoite developmental expression profiles (M4) (John Boothroyd)” microarray experiment.

**Search for Genes**

expand all | collapse all

Find a search...

- Text
- Gene models
- Annotation, curation and identifiers
- Genomic Location
- Taxonomy
- Orthology and synteny
- Phenotype
- Genetic variation
- Epigenomics
- Transcriptomics
  - EST Evidence
  - Microarray Evidence
  - RNA Seq Evidence

Filter Data Sets | oocyst

Legend: 2 Similarity FC Fold Change P Percentile

Organism	Data Set	Choose a search
<i>T. gondii</i> ME49 (filtered from 11 total entries)	Oocyst, tachyzoite, and bradyzoite developmental expression profiles (M4) (Fritz and Buchholz et al.)	FC P

Show All Data Sets

Fold Change Percentile

### Identify Genes based on *T. gondii* ME49 Oocyst, tachyzoite, and bradyzoite developmental expression profiles (M4) Microarray (fold change)

Learn more about this search

**For the Experiment**  
 Oocyst, tachyzoite, and bradyzoite developmental expression profiles (M4)

return | protein coding | Genes  
 that are | up or down regulated |  
 with a Fold change >= 10

between each gene's | average | expression value  
 in the following | Reference Samples |

- 10 days sporulated
- 2 days in vitro
- 4 days in vitro
- 6 days in vitro
- 21 days in vivo

select all | clear all

and its | average | expression value  
 in the following | Comparison Samples |

- unsporulated
- 4 days sporulated
- 10 days sporulated
- 2 days in vitro
- 4 days in vitro

select all | clear all

**Example showing one gene that would meet search criteria**  
 (Dots represent this gene's expression values for selected samples)

**Up or down regulated**

You are searching for genes that are up or down regulated between at least two reference samples and at least two comparison samples

For each gene, the search calculates:

$$\text{fold change}_{\text{up}} = \frac{\text{average expression level in comparison}}{\text{average expression level in reference}}$$

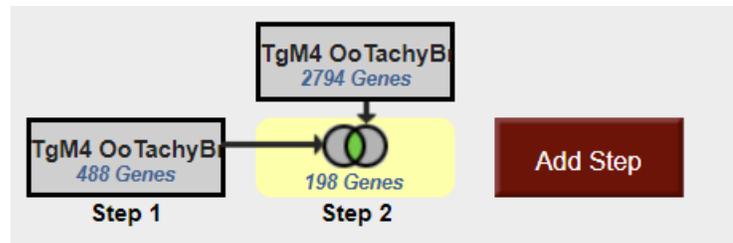
$$\text{fold change}_{\text{down}} = \frac{\text{average expression level in reference}}{\text{average expression level in comparison}}$$

and returns genes when  $\text{fold change}_{\text{up}} \geq 10$  or  $\text{fold change}_{\text{down}} \geq 10$ .  
 See the detailed help for this search.

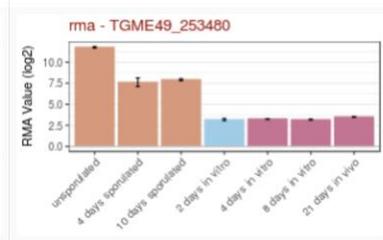
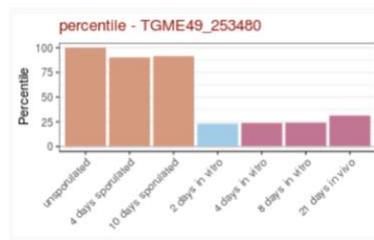
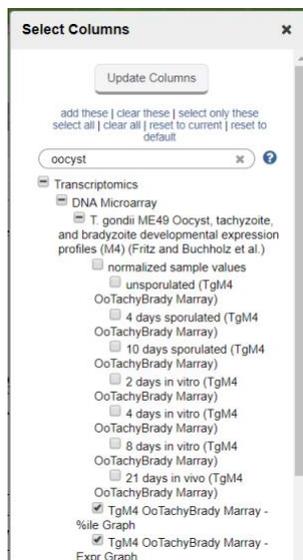
Get Answer

b. Add a step to limit this set of genes to only those for which all the non-oocyst stages are expressed below 50<sup>th</sup> percentile ... ie likely not expressed at those stages. (*Hint: after you click on add step find the same experiment under microarray expression and chose the percentile search*).

- Select the 4 **non-oocyst** samples.
- We want all to have less than 50<sup>th</sup> percentile so set **minimum percentile to 0** and **maximum percentile to 50**.
- Since we want all of them to be in this range, choose **ALL** in the “*Matches Any or All Selected Samples*”.



- To view the graphs in the final result table, turn on the columns called “TgM4 OoTachyBrady Marray - Expr Graph” and “TgM4 OoTachyBrady Marray - %ile Graph” (inside the “T. gondii ME49 Oocyst, tachyzoite, and bradyzoite developmental expression profiles (M4) (Fritz and Buchholz et al.)” Microarray).



1. Comparing RNA abundance and Protein abundance data.  
Note: for this exercise use <http://TriTrypDB.org>.

In this exercise we will compare genes that show differential RNA abundance levels between procyclic and blood form stages in *T. brucei* with genes that show differential protein abundance in these same stages.

- a. Find genes that are down-regulated 2-fold in procyclic form cells. Go to the search page for Genes by Microarray Evidence and select the fold change search for the “Expression profiling of five life cycle stages (Marilyn Parsons)” experiment and configure the search to return protein-coding genes that are down-regulated 2 fold in procyclic form (PCF) relative to the Blood Form reference sample. Since there are two PCF samples, it is reasonable to choose both and average them.

**Search for Genes**

expand all | collapse all

Find a search...

- Text
- Gene models
- Annotation, curation and identifiers
- Genomic Location
- Taxonomy
- Orthology and synteny
- Phenotype
- Genetic variation
- Transcriptomics
  - EST Evidence
  - Microarray Evidence
  - RNA Seq Evidence
- Sequence analysis
- Structure analysis
- Protein properties
- Protein targeting and localization
- Function prediction
- Pathways and interactions
- Proteomics
- Immunology

expand all | collapse all

**Identify Genes based on Microarray Evidence**

Filter Data Sets: Type keyword(s) to filter

Legend:  DC Direct Co...  FC Fold Chan...  P Percentile

Organism	Data Set	Choose a search
<i>L. infantum</i> JPCM5	Promastigote-to-amastigote differentiation (L.d. Samples) (Lahav et al.)	<input type="checkbox"/> FC <input type="checkbox"/> P
<i>L. infantum</i> JPCM5	Axenic and intracellular amastigote profiles (Rochette et al.)	<input type="checkbox"/> DC <input type="checkbox"/> P
<i>L. major</i> strain Friedlin	Three Developmental Stages (Stephen M. Beverley)	<input type="checkbox"/> DC <input type="checkbox"/> P
<i>T. brucei</i> brucei TREU927	Expression profiling of in vitro differentiation (Queiroz et al.)	<input type="checkbox"/> FC <input type="checkbox"/> P
<i>T. brucei</i> brucei TREU927	Expression profiling of five life cycle stages (Marilyn Parsons)	<input checked="" type="checkbox"/> FC <input type="checkbox"/> P
<i>T. brucei</i> brucei TREU927	Procyclic trypanosomes: heat shock vs untreated control (Kramer et al.)	<input type="checkbox"/> DC <input type="checkbox"/> P
<i>T. brucei</i> brucei TREU	Identify Genes based on <i>T. brucei</i> Expression profiling of five life cycle stages Microarray (fold change)	<input type="checkbox"/> DC <input type="checkbox"/> P
<i>T. brucei</i> brucei TREU		<input type="checkbox"/> FC <input type="checkbox"/> P
<i>T. brucei</i> brucei TREU		<input type="checkbox"/> DC <input type="checkbox"/> P
<i>T. cruzi</i> CL Brener E5n		<input type="checkbox"/> FC <input type="checkbox"/> P

For the Experiment: Expression profiling of five life cycle stages

return:  protein coding  Genes

that are:  down-regulated

with a Fold change  $\geq$  2.0

between each gene's average expression value in the following Reference Samples

- Blood Form
- Slender
- Stumpy
- PCF Log
- PCF Stat

select all | clear all

and its average expression value in the following Comparison Samples

- Blood Form
- Slender
- Stumpy
- PCF Log
- PCF Stat

select all | clear all

**Example showing one gene that would meet search criteria**

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

**Down-regulated**

Expression

Average Reference

Average Comparison

Reference Samples

Comparison Samples

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

For each gene, the search calculates:

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

and returns genes when fold change  $\geq$  2.0. To narrow the window, use the minimum reference value, or maximum comparison value. To broaden the window, use the maximum reference value, or minimum comparison value.

See the detailed help for this search.

Get Answer

**Tb LifeCyc Marra**  
378 Genes  
Step 1

Add Step

- b. Add a step to compare with quantitative protein expression. Select protein expression then “Quantitative Mass Spec Evidence” and the "Quantitative phosphoproteomes of bloodstream and procyclic forms (Tb427) (Urbanik et al.)" experiment. Configure this search to return genes that are down-regulated in procyclic form relative to blood form.

**Add Step**

**Add Step 2 : T. brucei brucei TREU927 Quantitative phosphoproteomes of bloodstream and procyclic forms (Tb427) Proteomics (direct comparison)**

**Experiment**  
 Quantitative phosphoproteomes of bloodstream and procyclic forms (Tb427)

**Direction**  
 down-regulated

**Comparison**  
 Pcf-Bsf ratio

**Fold difference >=**  
 2

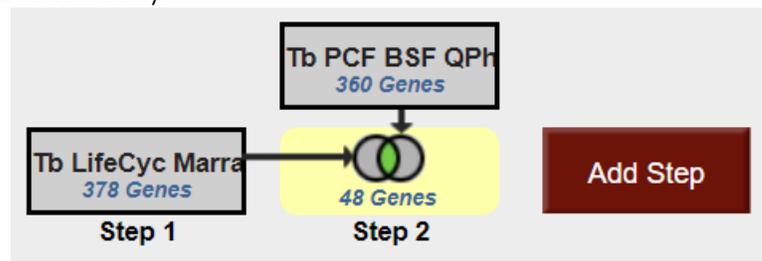
---

**Combine Genes in Step 1 with Genes in Step 2:**

1 Intersect 2      1 Minus 2  
 1 Union 2      2 Minus 1  
 1 Relative to 2, using genomic colocation

**Run Step**

- c. How many genes are in the intersection? Does this make sense? Make certain that you set the directions correctly.



- d. Try changing directions and compare up-regulated genes/proteins. (*Hint*: revise the existing strategy ... you might want to duplicate it so you can keep both). When you change one of the steps but not the other do you have any genes in the intersection? Why might this be?
- e. Can you think of ways to provide more confidence (or cast a broader net) in the microarray step? (*Hint*: you could insert steps to restrict based on percentile or add a RNA Sequencing step that has the same samples).

2. Find genes with evidence of protein phosphorylation in intracellular *Toxoplasma* tachyzoites.

For this exercise use <http://www.toxodb.org>

Phosphorylated peptides can be identified by searching the appropriate experiments in the [Mass Spec Evidence](#) search page.

7a. Find all genes with evidence of protein phosphorylation in intracellular tachyzoites. Navigate to the Post-Translational Modification search. Select the “Infected host cell, phosphopeptide-enriched (peptide discovery against TgME49)” sample under the experiment called “Tachyzoite phosphoproteome from purified parasite or infected host cell (RH) (Treeck et al.)”

The screenshot shows the Toxodb.org search interface. On the left, a sidebar titled "Search for Genes" has a search bar and a list of categories. "Post-Translational Modification" is selected, indicated by a blue arrow. The main content area is titled "Identify Genes based on Post-Translational Modification". It features several filter sections:
 

- Type of Post-Translational Modification:** A dropdown menu set to "phosphorylation site".
- Experiments and Samples:** A dropdown menu showing "1 selected, out of 9". The selected item is "Tachyzoite phosphoproteome from purified parasite or infected host cell (RH) (Treeck et al.)", which is expanded to show sub-experiments:
  - Toxoplasma gondii**
    - Toxoplasma gondii GT1**
      - Tachyzoite phosphoproteome from purified parasite or infected host cell (RH) (Treeck et al.)
        - Infected host cell, phosphopeptide-depleted (peptide discovery against TgGT1)
        - Infected host cell, phosphopeptide-enriched (peptide discovery against TgGT1)
        - Purified tachyzoites phosphopeptide-depleted (peptide discovery against TgGT1)
        - Purified tachyzoites phosphopeptide-enriched (peptide discovery against TgGT1)
      - Toxoplasma gondii ME49**
        - Tachyzoite phosphoproteome - Calcium dependent (RH) (Nebi et al.)
          - phosphopeptide-enriched (via Mascot)
          - phosphopeptide-depleted (via Sequest)
          - phosphopeptide-enriched (via Sequest)
        - Tachyzoite phosphoproteome from purified parasite or infected host cell (RH) (Treeck et al.)
          - Infected host cell, phosphopeptide-enriched (peptide discovery against TgME49)
          - Purified tachyzoites phosphopeptide-enriched (peptide discovery against TgME49)
- Number of modifications is:** A dropdown menu set to "Greater than or equal to".
- Number of Modifications:** A text input field containing the number "1".

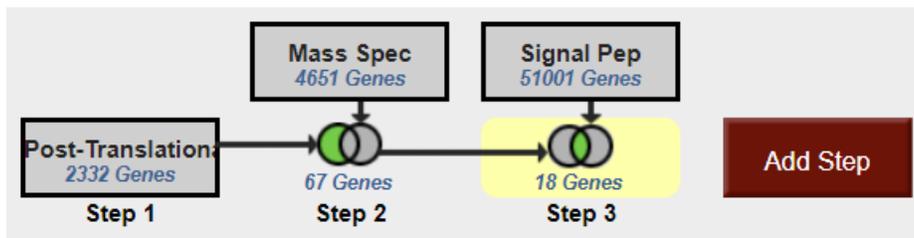
 At the bottom right, there is a "Get Answer" button.

7b. Remove all genes with phosphorylation evidence from purified tachyzoites and the phosphopeptide depleted fractions.

Hint: Use the Mass Spec Evidence search to access the tachyzoite and depleted fractions. Subtract (1 minus 2) these results from your first search.

7d. Explore your results. What kinds of genes did you find? *Hint: use the Product description word column or perform a GO enrichment analysis of your results.*

7e. Are any of these genes likely to be secreted? *Hint: add a step searching for genes with secretory signal peptides.*



7f. Pick one or two of the hypothetical genes in your results and visit their gene pages. Can you infer anything about their function? *Hint: explore the protein and expression sections.*

7g. What about polymorphism data? Go back to your strategy and add columns for SNP data found under the population biology section. Explore the gene page for the gene that has the most number of non-synonymous SNPs. *Hint: you can sort the columns by clicking on the up/down arrows next to the column names.*



- Navigate to the RNAseq searches and identify the experiment of cat enterocyte stages. Configure the search to identify call *T. gondii* genes that are upregulated by at least 2-fold in late and very late enteroepithelial stages (EES4 and EES5) compared to all other stages available from this experiment.

**Search for Genes**

expand all | collapse all

Find a search...

- Text
- Gene models
- Annotation, curation and identifiers
- Genomic Location
- Taxonomy
- Orthology and synteny
- Phenotype
- Genetic variation
- Epigenomics
- Transcriptomics
  - EST Evidence
  - Microarray Evidence
  - RNA Seq Evidence
- Sequence analysis
- Structure analysis
- Protein features and properties
- Protein targeting and localization
- Function prediction
- Pathways and interactions
- Proteomics
- Immunology

expand all | collapse all

**Identify Genes based on RNA Seq Evidence**

Filter Data Sets: enter

Legend: FC Differential Expression FC Fold Change % Percentile AS Sense/Antisense

Organism	Data Set	Change a search
T. gondii ME49	Feline enterocyte, tachyzoite, bradyzoite stage transcriptome (Hehl, Ramakrishnan et al.)	FC % AS
T. gondii ME49	Tachyzoite and merozoite transcriptomes (Hehl et al.) <span style="color: red;">NEW</span>	FC % AS

(Filtered from 25 total entries)

**Identify Genes based on T. gondii ME49 Feline enterocyte, tachyzoite, bradyzoite stage transcriptome RNASeq (fold change)**

[Tutorial](#)

For the Experiment: Feline enterocyte, tachyzoite, bradyzoite stage transcriptome - Sense

where: protein-coding  Gene

that are: up-regulated

with a Fold change  $\geq 2$

(or a Floor of: 10 reads (24 FPKM)

in the following Reference Samples:

- EES1
- EES2
- EES3
- EES4
- EES5
- Tachyzoites
- Tissue cysts

select all | clear all

and the maximum expression value

(or the Floor selected above)

in the following Comparison Samples:

- EES1
- EES2
- EES3
- EES4
- EES5
- Tachyzoites
- Tissue cysts

select all | clear all

**Toxo Cat RNaseq**

255 Genes

Step 1

Add Step

**Example showing one gene that would meet search criteria**

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

Up-regulated

Minimum Expression Level Comparison

Maximum Expression Level Reference

Reference Comparison

Samples Samples

4 maximum of four samples are shown when more than four are selected.

You are searching for genes that are up-regulated relative to at least two reference samples and at least two comparison samples.

For each gene, the search calculates:

Fold change =  $\frac{\text{minimum expression level in comparison}}{\text{maximum expression level in reference}}$

and returns genes when fold change  $\geq 2$ .

This calculation creates the narrowed 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 reference value, or average or maximum comparison value.

See the related help for this search.

\* or FPKM Floor, whichever is greater

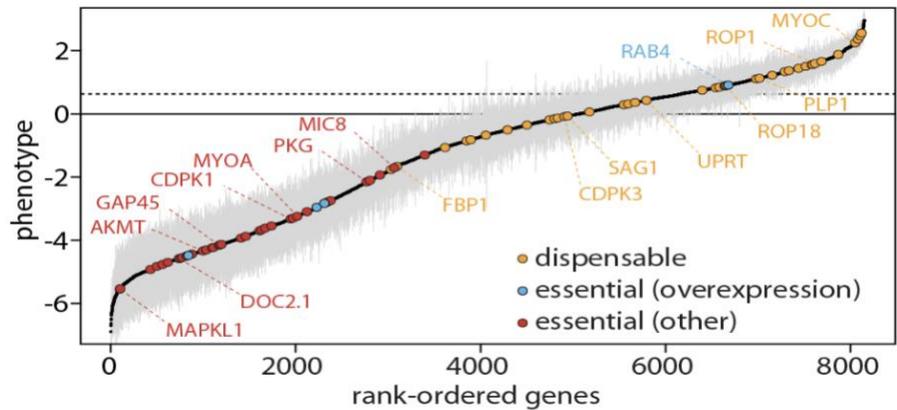


## 8. Finding genes based on high throughput mutagenesis and fitness analysis.

In EuPathDB we have a variety of studies where genome scale phenotypic analyses were carried out. In this exercise we'll use [ToxoDB.org](http://ToxoDB.org) and look at fitness following CRISPR mutagenesis. You could also explore phenotyping studies in PlasmoDB or FungiDB if you prefer, the principles are the same.

- 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 last essential gene has a fitness score just  $\geq$  than -4 so setting the phenotype score  $\leq$  -2 would provide a pretty stringent search but still return more than 1000 genes. Try it. Do you get the expected results based on the number of genes returned?

**Search for Genes**

expand all | collapse all

Find a search...

- ▶ Text
- ▶ Gene models
- ▶ Annotation, curation and identifiers
- ▶ Genomic Location
- ▶ Taxonomy
- ▶ Orthology and synteny
- ▶ Phenotype
  - CRISPR Phenotype
- ▶ Genetic variation
- ▶ Epigenomics
- ▶ Transcriptomics
- ▶ Sequence analysis
- ▶ Structure analysis
- ▶ Protein features and properties
- ▶ Protein targeting and localization
- ▶ Function prediction
- ▶ Pathways and interactions
- ▶ Proteomics
- ▶ Immunology

expand all | collapse all

### Identify Genes based on CRISPR Phenotype

Phenotype Score  $\geq$

Phenotype Score  $\leq$

**CRISPR**

1542 Genes

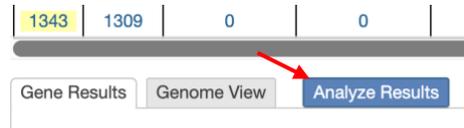
**Step 1**

Edit

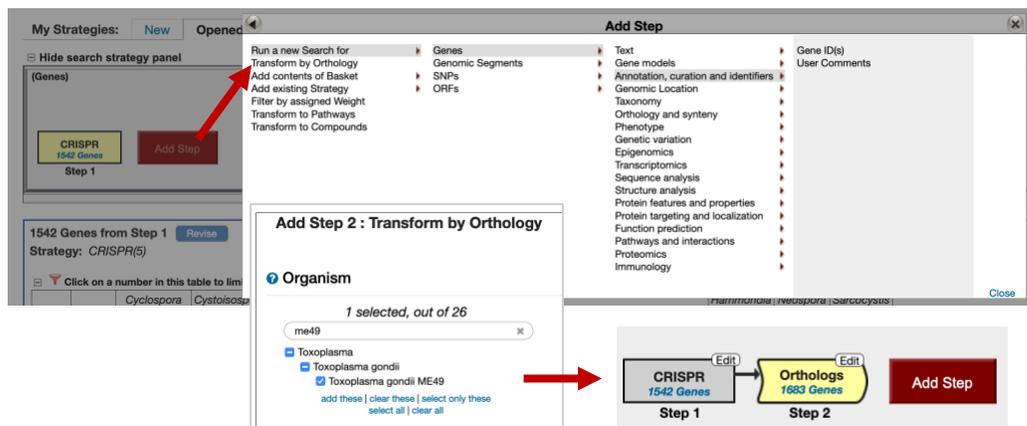
Add Step

Get Answer

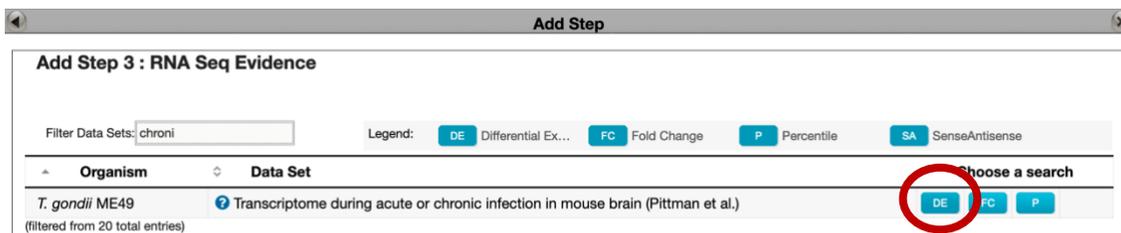
- Can you find additional evidence that these genes are essential? One way is to use the analysis tools to assess biological process and go function. Are the results what you would expect?



- Try adding columns to show additional data or intersecting these results with other queries, perhaps expression queries, to further assess this list. NOTE: this experiment was done in GT1 while all *T. gondii* functional data in ToxoDB is mapped to ME49 so an ortholog transform to ME49 is required before adding any additional functional studies.
- To do this, click on add step and select the Transform to orthologs option and select *T. gondii* ME49 to transform to.



- How many of these genes are upregulated in *in vivo* chronic stages of *T. gondii*?
  - Click on add step
  - Select 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 Step 3 : T. gondii ME49 Transcriptome during acute or chronic infection in mouse brain RNASeq (Differential Expression)

**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

**Combine Genes in Step 2 with Genes in Step 3:**

2 Intersect 3     2 Minus 3  
 2 Union 3     3 Minus 2  
 2 Relative to 3, using genomic colocation

- What do these results look like? Do you find any interesting genes?

## 9. Find genes with evidence of protein phosphorylation in intracellular *Toxoplasma* tachyzoites. For this exercise use <http://www.toxodb.org>

- Phosphorylated peptides can be identified by searching the appropriate experiments in the Mass Spec Evidence search page.

**Search for Genes**

expand all | collapse all

Find a search...

- ▶ Text
- ▶ Gene models
- ▶ Annotation, curation and identifiers
- ▶ Genomic Location
- ▶ Taxonomy
- ▶ Orthology and synteny
- ▶ Phenotype
- ▶ Genetic variation
- ▶ Epigenomics
- ▶ Transcriptomics
- ▶ Sequence analysis
- ▶ Structure analysis
- ▶ Protein features and properties
- ▶ Protein targeting and localization
- ▶ Function prediction
- ▶ Pathways and interactions
- ▼ Proteomics
  - Mass Spec. Evidence
  - Post-Translational Modification
  - Quantitative Mass Spec. Evidence
- ▶ Immunology

**Identify Genes based on Post-Translational Modification**

**Type of Post-Translational Modification**  
phosphorylation site

**Experiments and Samples**  
1 selected, out of 9

Filter list below:

- ▼ **Toxoplasma gondii**
  - ▼ **Toxoplasma gondii GT1**
    - Tachyzoite phosphoproteome from purified parasite or infected host cell (RH) (Treck et al.)
      - infected host cell, phosphopeptide-depleted (peptide discovery against TgGT1)
      - infected host cell, phosphopeptide-enriched (peptide discovery against TgGT1)
      - Purified tachyzoites phosphopeptide-depleted (peptide discovery against TgGT1)
      - Purified tachyzoites phosphopeptide-enriched (peptide discovery against TgGT1)
  - ▼ **Toxoplasma gondii ME49**
    - Tachyzoite phosphoproteome - Calcium dependent (RH) (Nebi et al.)
      - phosphopeptide-enriched (via Mascot)
      - phosphopeptide-depleted (via Sequest)
      - phosphopeptide-enriched (via Sequest)
    - Tachyzoite phosphoproteome from purified parasite or infected host cell (RH) (Treck et al.)
      - infected host cell, phosphopeptide-enriched (peptide discovery against TgME49)
      - Purified tachyzoites phosphopeptide-enriched (peptide discovery against TgME49)

select all | clear all | expand all | collapse all

**Number of modifications is**  
Greater than or equal to

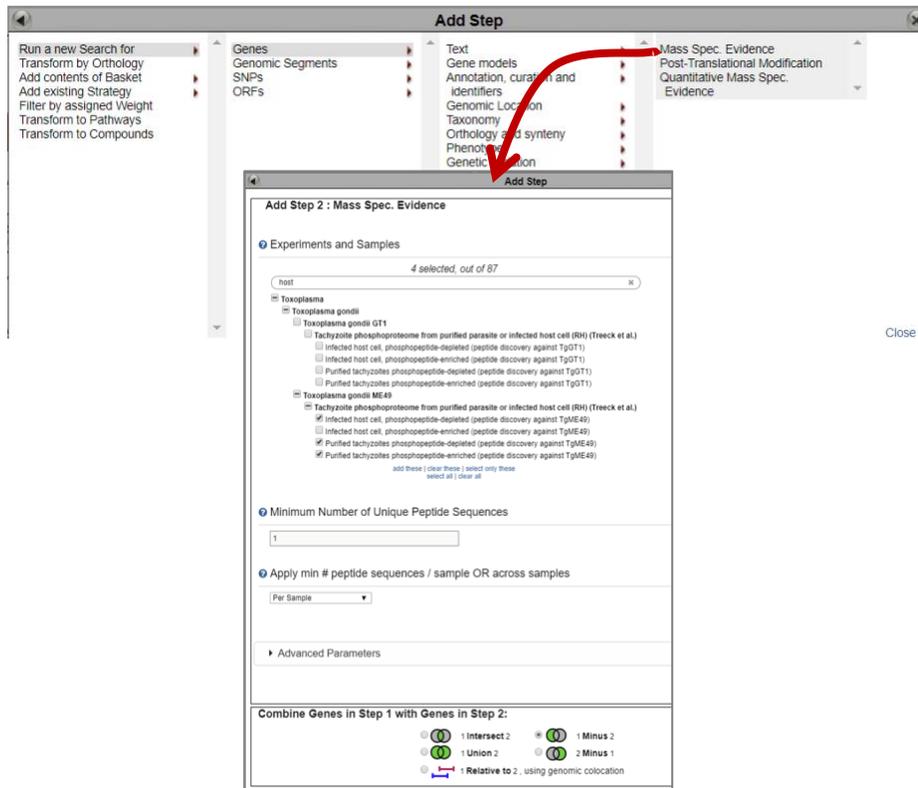
**Number of Modifications**  
1

Get Answer

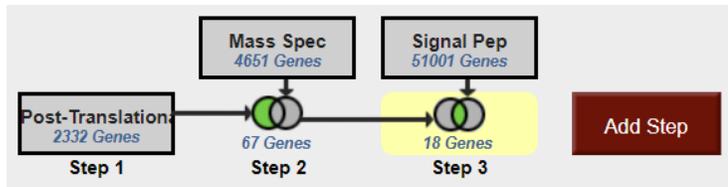
Find all genes with evidence of protein phosphorylation in intracellular tachyzoites. Navigate to the Post-Translational Modification search. Select the **“Infected host cell, phosphopeptide-enriched (peptide discovery against TgME49)”** sample under the experiment called **“Tachyzoite phosphoproteome from purified parasite or infected host cell (RH) (Trecek et al.)”**

- Remove all genes with phosphorylation evidence from purified tachyzoites and the phosphopeptide depleted fractions.

Hint: Use the Mass Spec Evidence search to access the tachyzoite and depleted fractions. Subtract (1 minus 2) these results from your first search.



- Explore your results. What kinds of genes did you find?  
*Hint: use the Product description word column or perform a GO enrichment analysis of your results.*



- Are any of these genes likely to be secreted? Hint: add a step searching for genes with secretory signal peptides.
- Pick one or two of the hypothetical genes in your results and visit their gene pages. Can

My Strategies: New Opened (2) All (2) Basket Public Strategies (14) Help

Hide search strategy panel

Strategy: Post-Translational Mod

18 Genes from Step 5

Strategy: Post-Translational Mod

Click on a number in this table to filter your results

Ortholog Group	C. layuana	C. zeylanoides	C. casei	E. aspergillus	E. brunellii	E. faeciformis	E. maxima	E. mitis	E. nidulans	E. parvulus	E. terrestris	H. hammondi	N. crassa	S. cerevisiae	T. reesei
18	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Gene Results

Gene ID	Transcript ID	Product Description	# Transcripts	Non-Coding SNPs All Strains	Non-Syn/Syn SNP Ratio All Strains	Non-Synonymous SNPs All Strains	SNPs with Stop Codons All Strains	Synonymous SNPs All Strains	Total SNPs All Strains
TGME49_288370	TGME49_288370-28_1	hypothetical protein	1	83	2.34	75	0	32	190
TGME49_288800	TGME49_288800-28_1	hypothetical protein	1	158	3.29	56	0	17	231
TGME49_243200	TGME49_243200-28_1	hypothetical protein	1	216	1.08	43	0	40	299
TGME49_205625	TGME49_205625-28_1	hypothetical protein	1	207	1.02	55	0	34	296
TGME49_259830	TGME49_259830-28_1	disaccharidase kinase catalytic domain-containing protein	1	139	0.61	14	0	23	176
TGME49_257595	TGME49_257595-28_1	hypothetical protein	1	131	2.32	130	0	56	317
TGME49_229600	TGME49_229600-28_1	hypothetical protein	1	28	0	0	0	5	33

you infer anything about their function? Hint: explore the protein and expression sections.

- What about polymorphism data? Go back to your strategy and add columns for SNP data found under the population biology section. Explore the gene page for the gene that has the highest number of non-synonymous SNPs. Hint: you can sort the columns by clicking on the up/down arrows next to the column names.