

## 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.

**Identify Genes based on RNA Seq Evidence**

Filter Data Sets: Type keyword(s) to filter Legend: DE Differential E... FC Fold Change P Percentile

Organism	Data Set	Choose a search
<i>P. berghei</i> ANKA	5 asexual and sexual stage transcriptomes (Hoeijmakers et al.)	FC P
<i>P. chabaudi</i> chabaudi	Trophozoite transcriptomes after mosquito transmission or direct injection into mice (Spence et al.)	DE FC P
<i>P. falciparum</i> 3D7	NSR-seq Transcript Profiling of malaria-infected pregnant women and children (Vignali et al.)	FC P
<i>P. falciparum</i> 3D7	Polysomal and steady-state asexual stage transcriptomes (Bunnik et al.)	FC P
<i>P. falciparum</i> 3D7	Blood stage transcriptome (3D7) (Otto et al.)	FC P
<i>P. falciparum</i> 3D7	Ribosome and steady state mRNA sequencing of asexual cell cycle stages (Caro et al.)	FC P
<i>P. falciparum</i> 3D7	Transcriptomes of 7 sexual and asexual life stages (Lopez-Barragan et al.)	FC P
<i>P. falciparum</i> 3D7	Intraerythrocytic cycle transcriptome (3D7) (Hoeijmakers et al.)	FC P
<i>P. falciparum</i> 3D7	Strand specific transcriptomes of 4 life cycle stages (Lopez-Barragan et al.)	FC P
<i>P. falciparum</i> 3D7	Transcriptome during intraerythrocytic development (Bartfai et al.)	FC P
<i>P. yoelii</i> yoelii 17X		FC P

**Search for Genes**

expand all | collapse all

Find a search...

- Text
- Gene models
- Annotation, curation and identifiers
- Genomic Location
- Taxonomy
- Orthology and synten
- Phenotype
- Genetic variation
- Epigenomics
- 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 P.falciparum Transcriptome during intraerythrocytic development RNASeq (fold change)**

For the Experiment: Transcriptome during intraerythrocytic development scaled HTSeq union - Ser

return protein coding Genes that are up or down regulated with a Fold change >= 2

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

Hour 5  
Hour 10  
Hour 15  
Hour 20  
Hour 25  
select all | clear all

and its expression value in the following Comparison Samples

Hour 5  
Hour 10  
Hour 15  
Hour 20  
Hour 25  
select all | clear all

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

Up or down regulated

This graphic will help you visualize the parameter choices you make at the left. It will begin to display when you choose a Reference Sample or a Comparison Sample. See the detailed help for this search.

Get Answer

- 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 >=** the intensity of difference in expression needed before a gene is returned by the search. **Choose 12** but feel free to modify this.
- **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**

- **Reference Sample:** the samples that will serve as the reference when comparing expression between samples. **choose 5, 10, 15, 20, 25**
- **And its AVERAGE expression value:** This is the operation applied to comparison samples. see explanation above. **Choose average**
- **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**

Fold Change
Percentile

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

Tutorial
YouTube

For the **Experiment**

Transcriptome during intraerythrocytic development scaled HTSeq union - Se

return protein coding Genes

that are up-regulated

with a **Fold change** >= 12

between each gene's average expression value

in the following Reference Samples

☒ Hour 20

☒ Hour 25

☐ Hour 30

☐ Hour 35

☐ Hour 40

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

and its average expression value

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

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

**Up-regulated**

Expression

Reference Samples Comparison Samples

Average Comparison

Average Reference

12 fold

*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:

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

and returns genes when **fold change** >= 12. 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](#).

Get Answer

- 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.)** or **Pf-iRBC 48hr** for shorter column headings. 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".

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 the results of 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.

The screenshot shows a bioinformatics software interface with a strategy named "P.f. RBC" containing 79 genes. A "Select Columns" dialog is open, displaying a tree of data types. A red arrow points from the "Add Columns" button in the main interface to the "Select Columns" dialog. Another red arrow points from the "Add Columns" button in the "Select Columns" dialog to a graph titled "Pf-RBC 48hr - Graph".

**My Strategies:** New Opened (1) All (1) Basket Examples Help

**Strategy: P.f. RBC\***

**79 Genes from Step 1**  
Strategy: P.f. RBC

**Filter results by species** (results removed by the filter)

**Gene Results** Genome View

**First 1 2 3 4 Next Last** Advanced Pa

**Gene ID** **Organism** **Product**

**PF3D7\_0207600** *P. falciparum* 3D7 serine re

**Select Columns**

Update Columns

clear all | expand all | collapse all  
reset to current | reset to default

- ☒ Search-Specific
  - ☒ Text, IDs, Species
  - ☐ Genomic Position
  - ☐ Gene Attributes
  - ☐ Protein Attributes
  - ☐ Protein Features
  - ☐ Transcript Expression
    - ☐ Pf-Gametocytogenesis
    - ☒ Pf-IRBC 48hr
      - ☒ Pf-IRBC 48hr - Graph
      - ☐ Pf-IRBC 48hr %ile - Graph
      - ☐ Pf-IRBC 48hr Max Exp Timing
      - ☐ Pf-IRBC 48hr Min Exp Timing
      - ☐ Pf-IRBC 48hr Max %ile
      - ☐ Pf-IRBC 48hr Max FC
    - ☐ Pf-IRBC+Spz+Gam
    - ☐ Pf-Trans Var Time Series
    - ☐ Pf-Invasion KO
    - ☐ Pf-Si2 KO
    - ☐ Pf-CQ Treatment
    - ☐ Pf-DD2\_X\_HB3
    - ☐ Pb-Dozi KO
    - ☐ Pb-IRBC HP/HPE
    - ☐ Py-Liver Stages
  - ☒ RNASeq
  - ☐ Putative Function

**Add 79 Genes to Basket | Download 79 Genes**

**Add Columns**

**Chosen Comp (log2)**

**Pf-RBC Infected RNASeq - Graph**

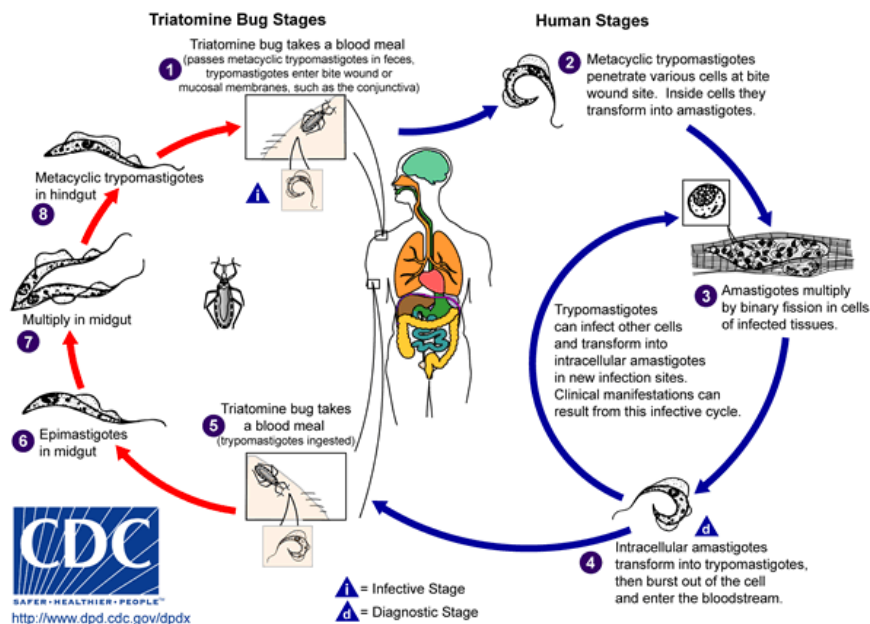
**Pf-RBC 48hr - Graph**

**Pf-RBC Infected RNASeq - Graph**

**Pf-IRBC 48hr - Graph**

## 2. Exploring microarray data in TriTrypDB.

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



- 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](#) [YouTube](#)

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

return **protein coding** **Genes**  
 that are **up-regulated**  
 with a Fold change  $\geq 2.0$

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:

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

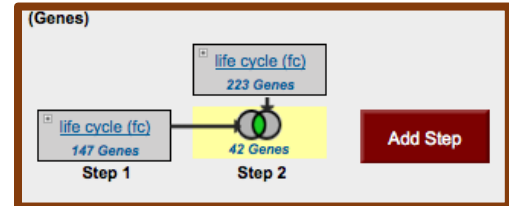
and returns genes when fold change  $\geq 2.0$ .

[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 (epimastigotes)? 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 run an ortholog transform on your results).
- How many orthologs exist in *L. braziliensis*? (*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.



My Strategies: [New](#) [Opened \(1\)](#) [All \(212\)](#) [Basket](#) [Public Strategies \(9\)](#) [Help](#)

(Genes) Strategy: *Tc LifeCyc Marray (fc)* [Rename](#) [Duplicate](#) [Save As](#) [Share](#) [Delete](#)

Step 1: *Tc LifeCyc Marray* 147 Genes → Step 2: *Tc LifeCyc Marray* 42 Genes → Step 3: *Orthologs* 57 Genes [Add Step](#)

57 Genes from Step 3  
Strategy: *Tc LifeCyc Marray (fc)* [Add 57 Genes to Basket](#) | [Download 57 Genes](#)

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

All Results	Ortholog Groups	Leishmania										T.brucei (nr Genes: 39)		T.congolense		Total
		C.fasciculata		L.braziliensis (nr Genes: 58)		L.donovani	L.infantum	L.major	L.mexicana	L.tarentolae	TREU927	gambiense DAL972	IL3000	CL Brer Esmeraldc		
		strain Cf-CI	MHOMBR /75/M2903	MHOMBR /75/M2904	BPK282A1	JPCM5	strain Friedlin	MHOMGT /2001/U1103	Parrot-Tartil							
1760	37	85	46	57	52	57	59	57	59	36	39	36	34	330		

Gene Results [Genome View](#) [Analyze Results](#) **BETA**

First 1 2 3 Next Last [Advanced Paging](#) [Add Columns](#)

Gene ID	Organism	Genomic Location	Product Description	Input Ortholog(s)	Ortholog Group	Paralog count	Ortholog count
LbrM.02.0350	<i>L. braziliensis</i> MHOMBR /75/M2904	LbrM.02: 147,781 - 154,645 (-)	ABC1 transporter, putative	TcCLB.510149.80	OG5_126568	8	112
LbrM.11.0960	<i>L. braziliensis</i> MHOMBR /75/M2904	LbrM.11: 439,107 - 444,425 (+)	ABC transporter, putative	TcCLB.510149.80	OG5_126568	8	112

### 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.

**Revise Step 1 : P falciparum 3D7 Transcriptomes of 7 sexual and asexual life stages RNASeq (fold change)**

For the Experiment  
 Transcriptomes of 7 sexual and asexual life stagesP. falciparum Su Seven Sta

return  Genes  
 that are   
 with a Fold change  $\geq$  50  
 between each gene's  expression value  
 in the following **Reference Samples**

☒ Ring  
☒ Early Trophozoite  
☒ Late Trophozoite  
☒ Schizont  
☒ Gametocyte II  
[select all](#) | [clear all](#)

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

☐ Late Trophozoite  
☐ Schizont  
☐ Gametocyte II  
☐ Gametocyte V  
☒ Ookinete  
[select all](#) | [clear all](#)

Global min / max in selected time points

Advanced Parameters

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

**Up-regulated**

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:

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

and returns genes when fold change  $\geq$  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](#).

- b. The above search will give you all genes that are up-regulated by 50 fold in ookinetes compared to the 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?
  - Hint: Run 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%





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.



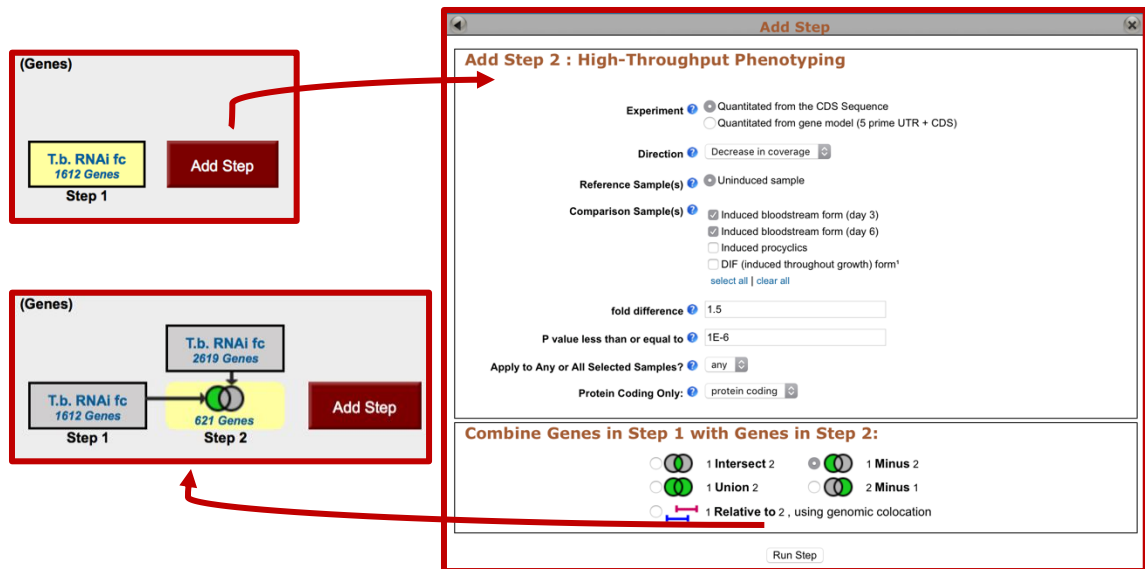
4. Find genes that are essential in procyclics but not in blood form *T. brucei*.  
Note: for this exercise use <http://TriTrypDB.org>.

- 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 –

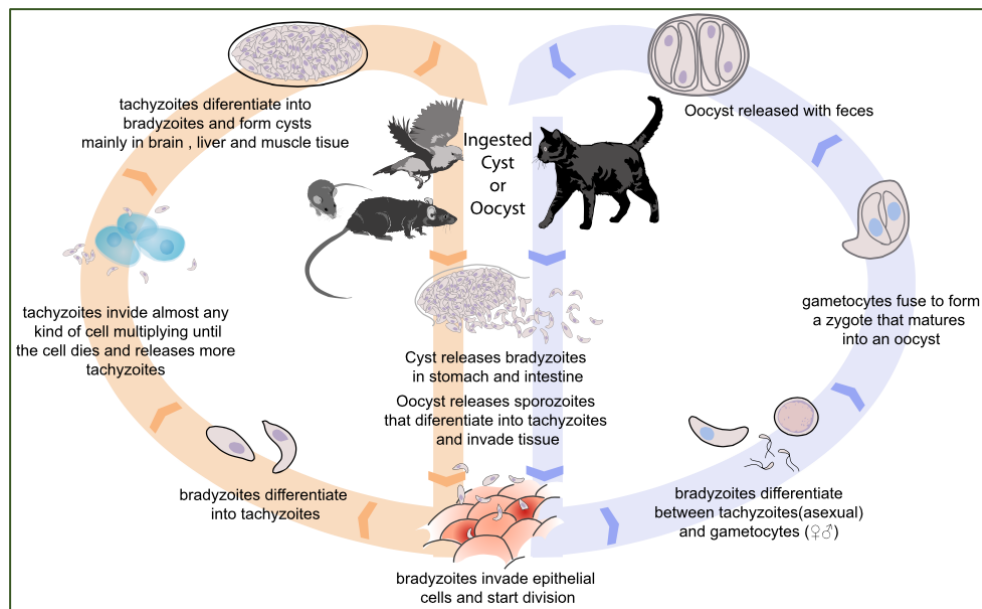
The screenshot shows two interfaces. The left panel is 'Search for Genes' with a sidebar containing categories like Text, Gene models, Annotation, Genomic Location, Taxonomy, Orthology and synteny, Phenotype, and High-Throughput Phenotyping. The right panel is 'Identify Genes based on High-Throughput Phenotyping' with a form for setting parameters. The form includes fields for Experiment (Quantitated from the CDS Sequence), Direction (Decrease in coverage), Reference Sample(s) (Uninduced sample), Comparison Sample(s) (Induced bloodstream form (day 3), Induced bloodstream form (day 6), Induced procyclics, DIF (induced throughout growth) form), fold difference (1.5), P value less than or equal to (1E-6), Apply to Any or All Selected Samples? (any), and Protein Coding Only (protein coding). A 'Get Answer' button is at the bottom. A red arrow points from the 'Get Answer' button to a 'Step 1' box in a separate window, which contains 'T.b. RNAi fc' (1612 Genes) and an 'Add Step' button.

- 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. In this example, the maximum expression value between genes in the reference and comparison groups was used to determine the fold difference.

**Search for Genes**

Find a search...

- Text
- Gene models
- Annotation, curation and identifiers
- Genomic Location
- Taxonomy
- Ontology and keywords
- Phenotype
- Genetic variation
- Epigenomics
  - 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: S Similarity FC Fold Change P Percentile

Organism	Data Set	Search
T. gondii ME49	Oocyst, tachyzoite, and bradyzoite developmental expression profiles (M4) (John Boothroyd)	<span>FC</span> <span>P</span>
T. gondii ME49	Bradyzoite Differentiation Expression Profiles (ME49, GT1, CTGara) (Michael W White)	<span>FC</span> <span>P</span>
T. gondii ME49	Mutants and wild-type expression profiles during bradyzoite differentiation (Mariana Matrajt)	<span>FC</span> <span>P</span>
T. gondii ME49	Bradyzoite Differentiation (3-day time series)(Pru) (John Boothroyd)	<span>FC</span> <span>P</span>
T. gondii ME49	Expression profiling of 3 archetypal lineages (David S. Roos)	<span>FC</span> <span>P</span>
T. gondii ME49	Bradyzoite Differentiation (Multiple 6-hr time points and Extended time series) (Paul H. Davis)	<span>FC</span> <span>P</span>
T. gondii ME49	Cell Cycle Expression Profiles (RH) (Michael W White)	<span>S</span> <span>FC</span> <span>P</span>
T. gondii ME49	Tachyzoite transcriptome during invasion (RH) (Vern B. Carruthers)	<span>FC</span> <span>P</span>
T. gondii ME49	Differential Expression Profiling GCN5-A mutant (William Sullivan)	<span>FC</span> <span>P</span>

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

**Tutorial**

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

return protein coding **Genes**

that are up-regulated

with a Fold change  $\geq 10$

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

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

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)

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:

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

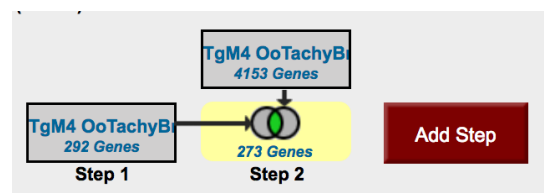
and returns genes when fold change  $\geq 10$ . To narrow the window, use the maximum reference value, or minimum comparison value. To broaden the window, use the minimum reference value, or maximum comparison value.

See the detailed help for this search.

Get Answer

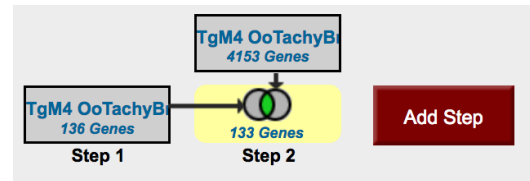
b. Add a step to limit this set of genes to only those for which all the non-oocyst stages are expressed below 50<sup>th</sup> percentile ... ie likely not expressed at those stages. (Hint: after you click on add step find the same experiment under microarray expression and chose the percentile search).

- 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 "Tg-M4 Life Cycle Stages – graph" and "Tg-M4 Life Cycle Stage %ile- graph" (inside the "Tg-Life Cycle" Microarray).



- c. Revise the first step of this strategy and compare the maximum expression of the reference samples to the minimum of the comparison samples.

- Does this result look cleaner/more convincing? Why?
- Would you consider these genes to be oocyst specific?



## 6. Comparing RNA abundance and Protein abundance data.

Note: for this exercise use <http://TriTrypDB.org>.

In this exercise we will compare the list of genes that show differential RNA abundance levels between procyclic and blood form stages in *T. brucei* with the list of 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 T.brucei 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"/> P <input type="checkbox"/> P
<i>T. cruzi</i> CL Brener Esr		<input type="checkbox"/> C <input type="checkbox"/> P

For the Experiment: Expression profiling of five life cycle stages

return: ☒ protein coding ☐ Genes

that are: ☒ down-regulated ☐ up-regulated

with a Fold change >= 2.0

between each gene's average expression value

in the following Reference Samples

☒ Blood Form ☐ Slender ☐ Stumpy ☐ PCF Log ☐ PCF Stat

and its average expression value

in the following Comparison Samples

☐ Blood Form ☐ Slender ☐ Stumpy ☒ PCF Log ☒ PCF Stat

**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 at least two reference samples and at least two comparison samples.

For each gene, the search calculates:

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

and returns genes when fold change >= 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](#)

- 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) (Urbaniak et al.)" experiment. Configure this search to return genes that are down-regulated in procyclic form relative to blood form.

The screenshot illustrates the workflow for adding a new step to a search. On the left, a box labeled 'Step 1' contains 'Tb LifeCyc Marra 360 Genes'. A red arrow points from this box to the 'Add Step' dialog. The 'Add Step' dialog has a list of categories on the left, including 'Genes', 'Text', and 'Mass Spec. Evidence'. The 'Mass Spec. Evidence' category is selected, and a sub-dialog 'Add Step 2 : Quantitative Mass Spec. Evidence' is open. This sub-dialog shows a list of data sets with 'T. brucei TREU927' selected. The 'Combine Genes in Step 1 with Genes in Step 2' section shows 'Intersect 2' selected. A red arrow points from the 'Intersect 2' option to the 'Run Step' button.

- 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).

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

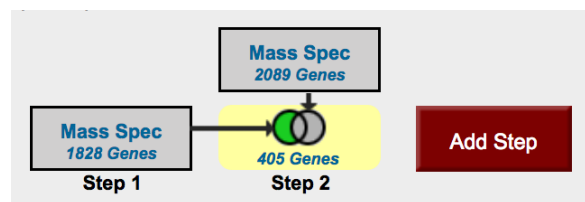
For this exercise use <http://www.toxodb.org>

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

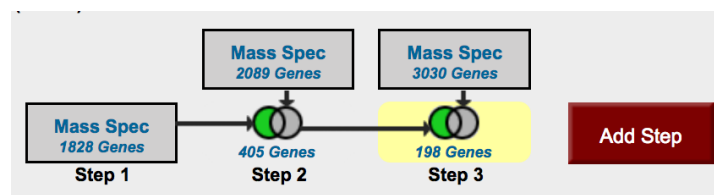
7a. Find all genes with evidence of phosphorylation in intracellular tachyzoites. Select the “Infected host cell, phosphopeptide-enriched (peptide discovery against TgME49)” sample under the experiment called “Tachyzoite phosphoproteome from purified parasite or infected host cell (RH) (Treeck et al.)”

The screenshot shows the ToxoDB search interface. On the left, under 'Search for Genes', the 'Proteomics' section is expanded, and 'Mass Spec. Evidence' is selected. On the right, the 'Identify Genes based on Mass Spec. Evidence' page shows a tree of experiments. The experiment 'Tachyzoite phosphoproteome from purified parasite or infected host cell (RH) (Treeck et al.)' is selected, and within it, the sample 'Infected host cell, phosphopeptide-enriched (peptide discovery against TgME49)' is checked. Below the tree, filters for 'Minimum Number of Spectra / Sample' and 'Minimum Number of Unique Peptide Sequences' are set to 1. A 'Get Answer' button is at the bottom right.

7b. Remove all genes with phosphorylation evidence from purified tachyzoites.



7c. Remove all genes that are also present in the phosphopeptide-depleted fractions (select both intracellular and extracellular).



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.* Could you achieve this same 105 genes with a two step strategy? *Hint: remove depleted and tachyzoite proteins in one step rather than two.*

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

My Strategies: [New](#) [Opened \(1\)](#) [All \(1\)](#) [Basket](#) [Public Strategies \(14\)](#) [Help](#)

(Genes) Strategy: [Mass Spec](#)

Mass Spec 1828 Genes Step 1 → Mass Spec 2089 Genes Step 2 → Mass Spec 3030 Genes Step 3 → Signal Pep 1945 Genes Step 4

33 Genes from Step 4  
Strategy: Mass Spec

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

All Results	Ortholog Groups	Eimeria							Hammondia	Neospora	Sarcocystis		Toxoplasma				
		E.acervulina	E.brunetti	E.falciformis	E.maxima	E.mitis	E.necatrix	E.praecox	E.tenella	H.hammondi	N.caninum	S.neurona	(nr Genes: 0)	T.gondii	(nr Genes: 33)		
		Houghton	Houghton	Bayer Haberkorn 1970	Weybridge	Houghton	Houghton	Houghton	strain Houghton	strain H.H.34	Liverpool	SN3	SO SN1	GT1	ME49	RH	VEG
33	33	0	0	0	0	0	0	0	0	0	0	0	0	0	33	0	0

Filter by strains (advanced)

Gene Results Genome View [Analyze Results](#)

First 1 2 Next Last Advanced Paging Download Add to Basket Add Columns

Gene ID	Transcript ID	Gene Group (representative gene)	Genomic Location (Gene)	Product Description	# Transcripts
TGME49_208830	TGME49_208830-i26_1	TGME49_208830-i26_1	TGME49_chrib:888,008..891,283(-)	hypothetical protein	1
TGME49_321640	TGME49_321640-i26_1	TGME49_321640-i26_1	TGME49_chrib:1,665,247..1,675,489(-)	cell division protein CDC48AP	1
TGME49_223140	TGME49_223140-i26_1	TGME49_223140-i26_1	TGME49_chrib:1,469,476..1,475,491(+)	tRNA binding domain-containing protein	1
TGME49_252360	TGME49_252360-i26_1	TGME49_252360-i26_1	TGME49_chrib:512,377..515,416(+)	roptry kinase family protein ROP24 (incomplete catalytic triad)	1
TGME49_288370	TGME49_288370-i26_1	TGME49_288370-i26_1	TGME49_chrib:2,478,472..2,482,708(-)	hypothetical protein	1

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.*

Gene Results Genome View [Analyze Results](#)

First 1 2 Next Last Advanced Paging Download Add to Basket Add Columns

Gene ID	Transcript ID	Product Description	Total SNPs All Strains	Non-Coding SNPs All Strains	NonSyn/Syn SNP Ratio All Strains	NonSynonymous SNPs All Strains	SNPs with Stop Codons All Strains	Synonymous SNPs All Strains
TGME49_224280	TGME49_224280-i26_1	CPSF A subunit region protein	922	635	0.98	142	0	145
TGME49_202490	TGME49_202490-i26_1	AP2 domain transcription factor AP2Vila-7	593	328	0.95	129	0	136
TGME49_311080	TGME49_311080-i26_1	transporter, cation channel family protein	551	360	1.51	115	0	76
TGME49_321640	TGME49_321640-i26_1	cell division protein CDC48AP	548	446	1.04	52	0	50
TGME49_205120	TGME49_205120-i26_1	hypothetical protein	447	185	2.2	180	0	82
TGME49_313280	TGME49_313280-i26_1	WD domain, G-beta repeat-containing protein	443	367	1.81	49	0	27
TGME49_286120	TGME49_286120-i26_1	prolyl endopeptidase	427	366	0.79	27	0	34
TGME49_219640	TGME49_219640-i26_1	hypothetical protein	382	263	2.5	85	0	34
TGME49_315700	TGME49_315700-i26_1	hypothetical protein	339	266	1.15	39	0	34
TGME49_239440	TGME49_239440-i26_1	protein kinase (incomplete catalytic triad)	333	215	1.13	62	1	55
TGME49_220350	TGME49_220350-i26_1	tRNA ligases class II (D, K and N) domain-containing protein	317	200	1.79	75	0	42
TGME49_257595	TGME49_257595-i26_1	hypothetical protein	317	131	2.32	130	0	56
TGME49_205625	TGME49_205625-i26_1	hypothetical protein	294	206	1.59	54	0	34
TGME49_288880	TGME49_288880-i26_1	hypothetical protein	231	158	3.29	56	0	17
TGME49_223140	TGME49_223140-i26_1	tRNA binding domain-containing protein	197	172	1.08	13	0	12
TGME49_216840	TGME49_216840-i26_1	hypothetical protein	189	89	1.17	54	0	46
TGME49_288370	TGME49_288370-i26_1	hypothetical protein	189	83	2.42	75	0	31
TGME49_308070	TGME49_308070-i26_1	hypothetical protein	188	123	1.95	43	0	22
TGME49_214080	TGME49_214080-i26_1	toxofillin	177	63	3.67	88	2	24
TGME49_314280	TGME49_314280-i26_1	AAR2 protein	163	113	1.78	32	0	18