

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: Legend: Fold Change Fold Change... Percentile

| Organism | Data Set | Choose a search |
|--------------------------|---|--|
| <i>P. falciparum</i> 3D7 | Transcriptome during intraerythrocytic development (Bartfai et al.) | <input type="button" value="FC"/> <input type="button" value="FCpV"/> <input type="button" value="P"/> |
| <i>P. falciparum</i> 3D7 | Transcriptomes of 7 sexual and asexual life stages (Lopez-Barragan et al.) | <input type="button" value="FC"/> <input type="button" value="FCpV"/> <input type="button" value="P"/> |
| <i>P. falciparum</i> 3D7 | Strand specific transcriptomes of 4 life cycle stages (Lopez-Barragan et al.) | <input type="button" value="FC"/> <input type="button" value="P"/> |
| <i>P. falciparum</i> 3D7 | NSR-seq Transcript Profiling of malaria-infected pregnant women and children (Vignali et al.) | <input type="button" value="FC"/> <input type="button" value="FCpV"/> <input type="button" value="P"/> |

Identify Genes based on P.f. post infection (RBC) RNA-seq time series (fold change)

For the Experiment

return Genes that are with a Fold change >=

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

- Hour 5
- Hour 10
- Hour 15
- Hour 20
- Hour 25
- Hour 30
- ...

select all | clear all

and its expression value in the following Comparison Samples

- Hour 5
- Hour 10
- Hour 15
- Hour 20
- Hour 25
- Hour 30
- ...

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

Expression

Expression

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.

Advanced Parameters

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 (expression in reference)/(expression in comparison). 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 it's 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
Fold Change with pValue
Percentile

Identify Genes based on P.f. post infection (RBC) RNA-seq time series (fold change)

Tutorial

For the Experiment Post-Infection (RBC) RNA-Seq time Series

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 5
 Hour 10
 Hour 15
 Hour 20
 Hour 25
 Hour 30
 Hour 35
 Hour 40
 Hour 45
 Hour 50

and its average expression value

in the following **Comparison Samples**

Hour 15
 Hour 20
 Hour 25
 Hour 30
 Hour 35
 Hour 40

Advanced Parameters

[Get Answer](#)

Example showing one gene that would meet search criteria

(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:

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

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.](#)

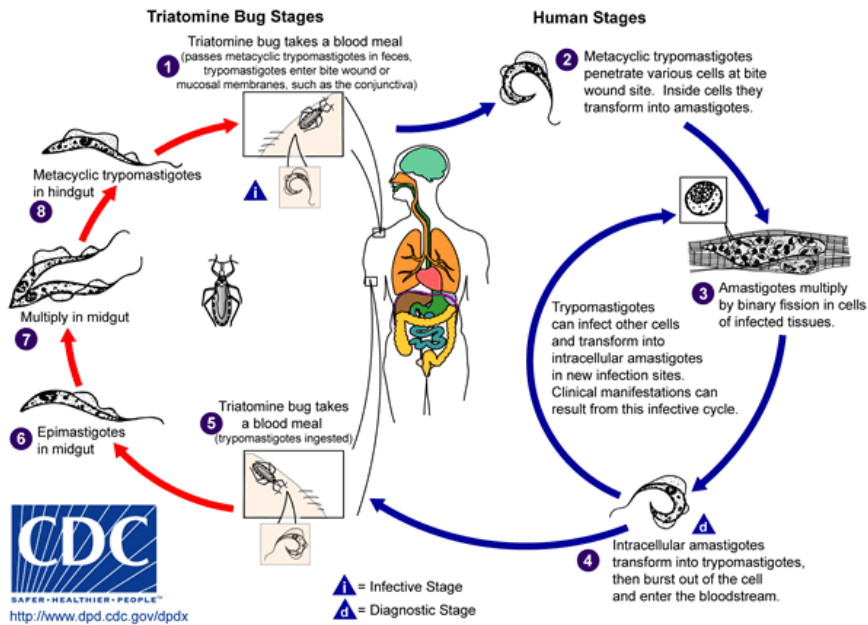
- b. For the genes returned by the search, how does the RNA-sequence data compare to microarray data?
- Hint: PlasmoDB contains data from a similar experiment that was analyzed by microarray instead of RNA sequencing. This experiment is called: **Erythrocytic expression time series (3D7, DD2, HB3) (Bozdech 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”.

The screenshot displays the PlasmoDB interface for a strategy named "P.f. RBC". A "Select Columns" dialog box is open, showing a list of data sources. The "Microarray" section is expanded, and "Pf-iRBC 48hr - Graph" is selected. A red arrow points from this selection to the "Add Columns" button in the main interface. Below the dialog, a grid of four graphs compares "Pf-RBC Infected RNASeq - Graph" and "Pf-iRBC 48hr - Graph" for gene PF3D7_0207600. The graphs show RPKM (log2) and Log2(Ratio) expression values over time, with the RNA-seq data generally showing higher expression levels than the microarray data.

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.

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 tcrucLBrenerEsmeraldo-lik

return protein coding Genes
 that are up-regulated down-regulated no change

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

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

For each gene, the search calculates:

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

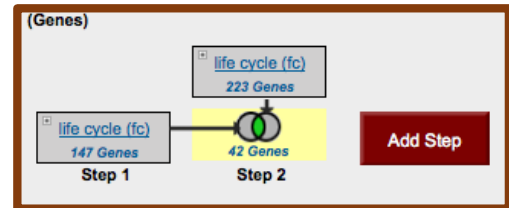
and returns genes when fold change \geq 2.0.

See the detailed help for this search.

Advanced Parameters

Get Answer

- Select the direction of regulation, your reference sample and your Life 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 of gene 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) *

Tc LifeCyc Marray 147 Genes Step 1 → Tc LifeCyc Marray 223 Genes Step 2 → Orthologs 57 genes Step 3

[Add Step](#) [Rename](#) [Duplicate](#) [Save As](#) [Share](#) [Delete](#)

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 | Crithidia | | Leishmania | | | | | | | | | | |
|-------------|-----------------|---------------|------------------|-------------------------------|----------|------------|-----------------|---------------------|---------------|-------------------|-------------------------|------------------|--------------|-------------------|
| | | C.fasciculata | | L.braziliensis (nr Genes: 58) | | L.donovani | L.infantum | L.major | L.mexicana | L.tarentolae | T.brucei (nr Genes: 39) | | T.congolense | |
| | | strain Cf-CI | MHOMBR /75/M2903 | MHOMBR /75/M2904 | BPK282A1 | JPCM5 | strain Friedlin | MHOMGT /2001/U11103 | Parrot-Tarill | Lister strain 427 | TREU927 | gambiense DAL972 | IL3000 | CL.Brer Esmeraldc |
| 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 | L. braziliensis MHOMBR /75/M2904 | LbrM.02: 147,781 - 154,645 (-) | ABC1 transporter, putative | TcCLB.510149.80 | OG5_126568 | 8 | 112 |
| LbrM.11.0960 | L. braziliensis 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 stages P. falciparum Su Seven Sta

return Genes
 that are
 with a Fold change >= 50
 between each gene's expression value
 in the following Reference Samples

Ring
 Early Trophozoite
 Late Trophozoite
 Schizont
 Gametocyte II
 Gametocyte V
 Ookinete
 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 Don't care

Advanced Parameters

Example showing one gene that would meet search criteria
 (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 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 >= 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.

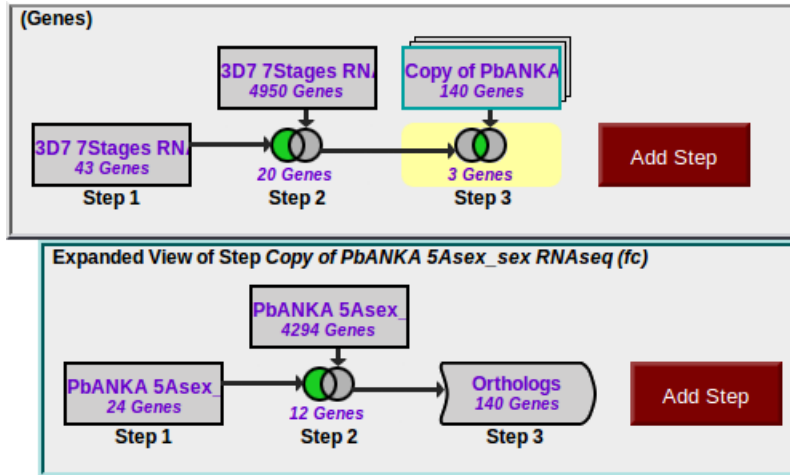
- b. The above search will give you all genes that are up-regulated by 50 fold in ookinetes compared to the other stages. Despite the high fold change, some genes in the list may be highly expressed in the other stages. How can you remove genes from the list that are highly expressed in the other stages?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

- 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?

| Pathway Id | Pathway | Source | No. of Enzymes | Total Pathway Enzymes | Total Pathway Compounds | Map - Painted With Transformed Genes (new window) |
|------------|---|---------|----------------|-----------------------|-------------------------|---|
| ec00230 | Purine metabolism | ec00230 | 1 | 177 | 100 | Pathway Map |
| ec00231 | Puromycin biosynthesis | ec00231 | 1 | 7 | 10 | Pathway Map |
| ec00240 | Pyrimidine metabolism | ec00240 | 1 | 114 | 73 | Pathway Map |
| ec00563 | Glycosylphosphatidylinositol(GPI)-anchor biosynthesis | ec00563 | 1 | 9 | 15 | Pathway Map |
| ec00983 | Drug metabolism - other enzymes | ec00983 | 1 | 31 | 32 | Pathway Map |

- d. What happens if you revise the first step and modify the fold difference to a lower value - 10 for example?
- 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 – try the default parameters first and select the “induced procyclics” as the comparison sample.

- 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.

(Genes)
T.b. RNAi fc
1612 Genes
Step 1

(Genes)
T.b. RNAi fc
1612 Genes
Step 1

T.b. RNAi fc
2619 Genes
Step 2

Add Step 2 : High-Throughput Phenotyping

Experiment Quantitated from the CDS Sequence
 Quantitated from gene model (5 prime UTR + CDS)

Direction

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

fold difference

P value less than or equal to

Apply to Any or All Selected Samples?

Protein Coding Only:

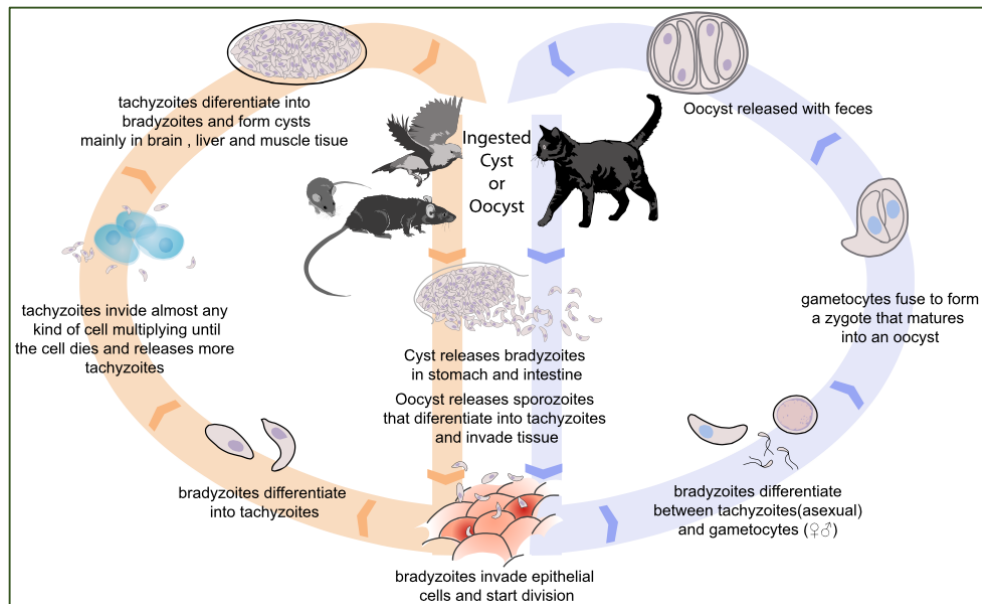
Combine Genes in Step 1 with Genes in Step 2:

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

Run Step

5. Finding oocyst expressed genes in *T. gondii* based on microarray evidence.

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



- 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.

Identify Genes based on Microarray Evidence

Filter Data Sets: Type keyword(s) to filter Legend: FC Fold Chan... FCC Fold Chan... P Percentile S Similarity

| Organism | Data Set | FC | FCC | P | S |
|----------------|--|----|-----|---|---|
| T. gondii ME49 | Differential Expression Profiling GCN5-A mutant (William Sullivan) | FC | FCC | P | |
| T. gondii ME49 | Bradyzoite Differentiation (Multiple 6-hr time points and Extended time series) (Paul H. Davis) | FC | | P | |
| T. gondii ME49 | Expression profiling of the 3 archetypal lineages (David S. Roos) | | FCC | P | |
| T. gondii ME49 | Transcript Profiling Infection (Vern B. Carruthers) | FC | FCC | P | |
| T. gondii ME49 | Mutants and wild-type during bradyzoite differentiation in vitro (Mariana Matrajt) | FC | FCC | P | |
| T. gondii ME49 | Bradyzoite Differentiation (Single Time-Point) (Michael W White) | | | P | |
| T. gondii ME49 | Cell Cycle Expression Profiles (Michael W White) | FC | | P | S |
| T. gondii ME49 | Expression Profiling of oocyst, tachyzoite, and bradyzoite development in strain M4 (John Boothroyd) | FC | | P | |

Identify Genes based on T.g. Life Cycle Stages (fold change) Tutorial

For the Experiment: Oocyst, Tachyzoite and Bradyzoite Development

return: protein coding Genes

that are: up-regulated

with a Fold change >= 10

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

- unsporulated
- 4 days sporulated
- 10 days sporulated
- 2 days in vitro
- 4 days in vitro
- 8 days in vitro
- 21 days in vivo

and its maximum expression value in the following Comparison Samples

- unsporulated
- 4 days sporulated
- 10 days sporulated
- 2 days in vitro
- 4 days in vitro
- 8 days in vitro
- 21 days in vivo

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

Maximum Comparison

Maximum Reference

10 fold

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:

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

and returns genes when fold change >= 10. To narrow the window, use the average or minimum comparison value. To broaden the window, use the average or minimum reference value.

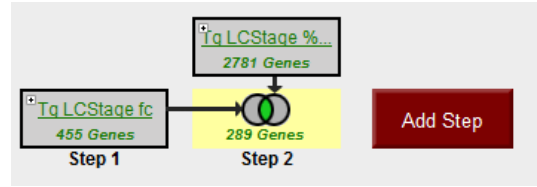
See the detailed help for this search.

Advanced Parameters

Get Answer

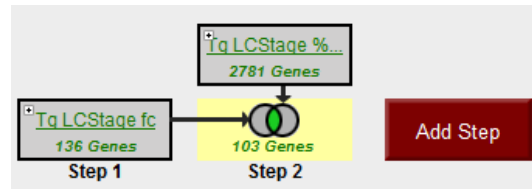
- b. Add a step to limit this set of genes to only those for which all the non-oocyst stages are expressed below 50th percentile ... ie likely not expressed at those stages. (Hint: after you click on add step find the same experiment under microarray expression and chose the percentile search).
- Select the 4 **non-oocyst** samples.
 - We want all to have less than 50th 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?



Save this strategy so that you can use it for an exercise we are doing later during the course.

d. Revise the first step of this strategy to find genes that are 3 fold higher in day 4 oocysts than any other life cycle stage in this experiment.

- Do all these genes have day 4 oocysts as the global maximum time point?
- Note that we still have the step to limit the percentile of non-oocyst samples to $\leq 50^{\text{th}}$ percentile. What happens if you revise this step to also include the unsporulated and day 10 oocyst samples in this percentile range? Do you get more of fewer results back? Why?

My Strategies: [Now](#) [Opened \(1\)](#) [All \(1\)](#) [Basket](#) [Examples](#) [Help](#)

Strategy: **Tg LCStage fc***

Step 1: Tg LCStage fc (67 Genes) → Step 2: Tg LCStage %ile (4 Genes)

4 Genes from Step 2
Strategy: **Tg LCStage fc**

Filter by organism or strain (results removed by the filter will not be combined into the next step.)
Filter by strains (advanced) (results removed by the filter will not be combined into the next step.)

| Gene ID | Gene Group (representative gene) | Genomic Location | Product Description | Tg-M4 Life Cycle Stages - graph | Tg-M4 Life Cycle Stage %ile- graph |
|---------------|----------------------------------|---|---|---------------------------------|------------------------------------|
| TGME49_258800 | TGGT1_258800 | TGME49_chrVIIb: 3,177,133 - 3,178,728 (+) | rhoptyr kinase family protein ROP31 (ROP31) | | |
| TGME49_233300 | TGGT1_233300 | TGME49_chrVIIb: 2,569,523 - 2,577,098 (-) | RhoGAP domain-containing protein | | |

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.

The screenshot displays the TriTrypDB search interface. On the left, a sidebar titled "Identify Genes by:" lists various evidence types, with "Microarray Evidence" highlighted in a red box. An arrow points from this box to the main search results area. The main area is titled "Identify Genes based on Microarray Evidence" and shows a list of search results for *T. brucei* TREU927, including "Expression profiling of five life cycle stages (Marilyn Parsons)". Below this, a configuration window titled "Identify Genes based on T.b. Expression profiling of five life cycle stages Microarray (fold change)" is shown. This window allows users to set the fold change to 2, select "Down-regulated", and choose "Blood Form" as the reference sample and "PCF Log" and "PCF Stat" as comparison samples. A graph titled "Down-regulated" illustrates the search criteria, showing a 2-fold decrease in expression relative to the average reference sample. The graph plots "Expression" on the y-axis and "Reference Comparison Samples" on the x-axis. Below the graph, text explains the search criteria: "You are searching for genes that are down-regulated between at least two reference samples and at least two comparison samples." and "For each gene, the search calculates: fold change = average expression value in reference samples / average expression value in comparison samples" and returns genes when fold change >= 2. The interface also includes a "Protein Coding Only" checkbox and an "Advanced Parameters" link.

- 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 shows a bioinformatics software interface with several panels:

- My Strategies:** A panel on the left showing a strategy named "Tb LifeCyc Marra" with 553 Genes and an "Add Step" button.
- Add Step:** A central panel with a menu "Run a new Search for" containing options like "Transform by Orthology", "Add contents of Basket", "Add existing Strategy", "Filter by assigned Weight", "Transform to Pathways", and "Transform to Compounds".
- Search Criteria:** A list of criteria including Genes, Genomic Segments, SNPs, ORFs, Text, IDs, Organism, Genomic Position, Gene Attributes, Protein Attributes, Protein Features, Similarity/Pattern, Mass Spec. Evidence, and Quantitative Mass Spec. Evidence.
- Add Step 2 : Quantitative Mass Spec. Evidence:** A panel with a "Filler Data Sets" field, a "Legend" (DC, Dir..., FC, Fol..., FCF, Fol...), and a table of data sets. The table has columns for "Organism" and "Data Set". The "Data Set" column contains entries like "Long slender vs short stumpy blood stage quantitative proteomes (Gunasekera et al)", "Quantitative phosphoproteomes of bloodstream and procyclic forms (Tb427) (Urbaniak et al.)", and "Proteome of Procyclic vs Bloodstream forms by SILAC (Butter et al.)". The "DC" button for the second entry is circled in red.
- Add Step 2 : T.bru. Quantitative phosphoproteomes of bloodstream and procyclic forms (Tb427) Proteomics (direct comparison):** A panel with a "Direction" dropdown set to "down-regulated", "Samples" set to "PcF-Bsf ratio", and a "Fold difference >=" field set to "2".
- Combine Genes in Step 1 with Genes in Step 2:** A panel with radio buttons for "Intersect 2", "Union 2", "Relative to 2, using genomic collocation", "1 Minus 2", and "2 Minus 1".

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

Identify Genes based on Mass Spec. Evidence

Experiment/Samples ? select all | clear all | expand all | collapse all | reset to default

- Eimeria
- Toxoplasma
 - Toxoplasma gondii*
 - Oocyst Partially Sporulated Proteome (VEG) (Possenti, et al.)
 - Oocyst proteome (M4 Typell) (Wastling)
 - Oocyst proteome - Fractionated (M4 type II) (Fritz et al.)
 - Proteome During Infection in H. sapiens (Wastling)
 - Tachyzoite Intra- and Extracellular Lysine-Acetylomes (RH) (Jeffers and Xue)
 - Tachyzoite Rhoptyr proteome (RH) (Bradley et al.)
 - Tachyzoite conoid proteome (RH) (Hu et al.)
 - Tachyzoite membrane and cytosolic fractions (RH) (Dybas et al.)
 - Tachyzoite phosphoproteome - Calcium dependent (RH) (Nebl et al.)
 - Tachyzoite phosphoproteome from purified parasite or infected host cell (RH) (Treeck et al.)
 - Infected host cell, phosphopeptide-depleted (peptide discovery against TgME49)
 - Infected host cell, phosphopeptide-depleted (peptide discovery against TgGT1)
 - Infected host cell, phosphopeptide-enriched (peptide discovery against TgME49)
 - Infected host cell, phosphopeptide-enriched (peptide discovery against TgGT1)
 - Purified tachyzoites phosphopeptide-depleted (peptide discovery against TgGT1)
 - Purified tachyzoites phosphopeptide-depleted (peptide discovery against TgME49)
 - Purified tachyzoites phosphopeptide-enriched (peptide discovery against TgGT1)
 - Purified tachyzoites phosphopeptide-enriched (peptide discovery against TgME49)
 - Tachyzoite secretome (RH) (Zhou et al.)
 - Tachyzoite subcellular fractions (Moreno)
 - Tachyzoite total proteome (RH) (Wastling)

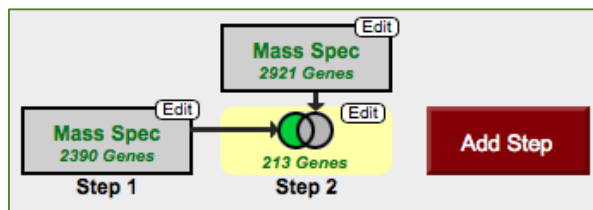
select all | clear all | expand all | collapse all | reset to default

Minimum Number of Unique Peptide Sequences ?

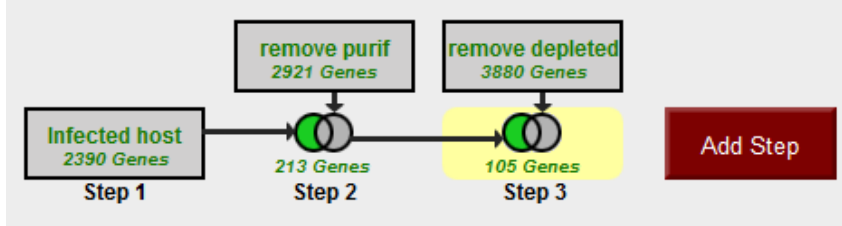
Minimum Number of Spectra ?

Advanced Parameters

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

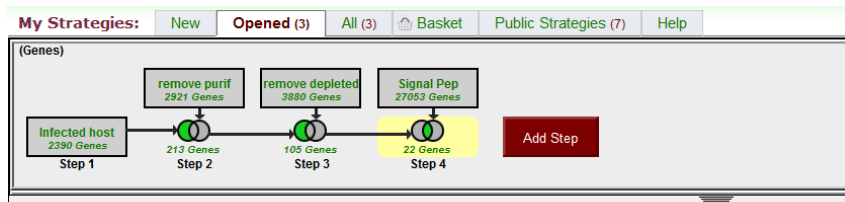


7c. Remove all genes that are also present in the phosphopeptide-depleted fractions (select both intracellular and extracellular).



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 tachozoite 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.*



22 Genes from Step 4
Strategy: *Infected host*

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

| All Results | Ortholog Groups | <i>Eimeria</i> | | | | | | | | | | <i>Hammondia</i> | N |
|-------------|-----------------|---------------------------------|-------------------------------|--|------------------------------|----------------------------|-------------------------------|------------------------------|-------------------------------------|------------------------------------|---|------------------|---|
| | | <i>E.acervulina</i> Houghton | <i>E.brunetti</i> Houghton | <i>E.falciformis</i> Bayer Haberkorn 1970 | <i>E.maxima</i> Weybridge | <i>E.mitis</i> Houghton | <i>E.necatrix</i> Houghton | <i>E.praecox</i> Houghton | <i>E.tenella</i> strain Houghton | <i>H.hammondi</i> strain H.H.34 | L | | |
| 22 | 22 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |

Filter by strains (advanced)

Gene Results | Genome View | Analyze Results **BETA**

| Gene ID | Gene Group (representative gene) | Genomic Location | Product Description |
|---------------|----------------------------------|---|--|
| TGME49_294940 | TGGT1_294940 | TGME49_chria: 1,282,608 - 1,287,925 (-) | hypothetical protein |
| TGME49_222870 | TGGT1_222870 | TGME49_chrl1: 1,271,864 - 1,275,140 (+) | hypothetical protein |
| TGME49_320150 | TGGT1_320150 | TGME49_chrlV: 464,394 - 473,129 (-) | elongation factor Tu GTP binding domain-containing protein |

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 BETA

First 1 2 Next Last Advanced Paging Add Columns

| Gene ID | Product Description | Total SNPs All Strains | NonSynonymous SNPs All Strains | Synonymous SNPs All Strains | Non-Coding SNPs All Strains | SNPs with Stop Codons All Strains | NonSyn/Syn SNP Ratio All Strains |
|---------------|--|------------------------|--------------------------------|-----------------------------|-----------------------------|-----------------------------------|----------------------------------|
| TGME49_271110 | hypothetical protein | 890 | 157 | 44 | 679 | 10 | 3.57 |
| TGME49_257595 | hypothetical protein | 317 | 123 | 51 | 131 | 12 | 2.41 |
| TGME49_219640 | hypothetical protein | 382 | 85 | 34 | 263 | 0 | 2.5 |
| TGME49_288370 | hypothetical protein | 224 | 82 | 35 | 105 | 2 | 2.34 |
| TGME49_216840 | hypothetical protein | 189 | 75 | 23 | 89 | 2 | 3.26 |
| TGME49_257640 | hypothetical protein | 110 | 66 | 12 | 31 | 1 | 5.5 |
| TGME49_320150 | elongation factor Tu GTP binding domain-containing protein | 378 | 65 | 22 | 286 | 5 | 2.95 |
| TGME49_235960 | hypothetical protein | 155 | 58 | 14 | 77 | 6 | 4.14 |
| TGME49_288880 | hypothetical protein | 220 | 56 | 17 | 147 | 0 | 3.29 |
| TGME49_269750 | CrcB family protein | 95 | 54 | 20 | 18 | 3 | 2.7 |
| TGME49_315700 | hypothetical protein | 338 | 54 | 14 | 265 | 5 | 3.86 |
| TGME49_308070 | hypothetical protein | 188 | 43 | 22 | 123 | 0 | 1.95 |
| TGME49_269420 | hypothetical protein | 45 | 37 | 8 | 0 | 0 | 4.63 |
| TGME49_200440 | hypothetical protein | 72 | 35 | 11 | 24 | 2 | 3.18 |
| TGME49_259830 | diacylglycerol kinase catalytic domain-containing protein | 176 | 32 | 3 | 139 | 2 | 10.67 |
| TGME49_236220 | PCI domain-containing protein | 383 | 28 | 18 | 332 | 5 | 1.56 |
| TGME49_231180 | hypothetical protein | 54 | 25 | 9 | 18 | 2 | 2.78 |
| TGME49_294940 | hypothetical protein | 137 | 16 | 7 | 111 | 3 | 2.29 |