

# Functional Genomics I

## Transcriptomics

### 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 the Plasmodb.org website interface for identifying genes based on RNA-seq evidence. A red box highlights the 'Identify Genes based on RNA Seq Evidence' section. A red arrow points from the 'Identify Genes by:' sidebar to the 'RNA Seq Evidence' option. Another red arrow points from the 'Choose a search' dropdown menu to the 'FC' (Fold Change) button. The 'Identify Genes based on P.f. post infection (RBC) RNA-seq time series (fold change)' section is also highlighted with a blue box. This section includes a 'For the Experiment' dropdown set to 'Post-Infection (RBC) RNA-Seq time series', a 'return' dropdown set to 'protein coding', and a 'that are' dropdown set to 'up or down regulated'. The 'with a Fold change >= 2' option is selected. Below this, there are two lists of time points: 'Reference Samples' and 'Comparison Samples', both with checkboxes for Hour 5, Hour 10, Hour 15, Hour 20, Hour 25, and Hour 30. To the right, an 'Example showing one gene that would meet search criteria' is displayed, featuring a graph titled 'Up or down regulated' showing expression levels over time. The graph has two y-axes labeled 'Expression' and two x-axes labeled 'Time'. The graph shows a single data point at Hour 5 and a single data point at Hour 30, with a line connecting them, indicating a change in expression. Below the graph, text explains that the graphic will help visualize parameter choices and that it will begin to display when a Reference Sample or a Comparison Sample is chosen. At the bottom, there is an 'Advanced Parameters' section and a 'Get Answer' button.


- Hint: 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.
- **Between each gene's AVERAGE expression value:** This parameter sets 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  $\geq$  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  
[select all](#) | [clear all](#)

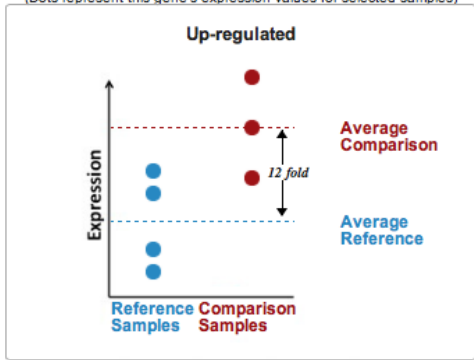
and its average expression value

in the following Comparison Samples

☐ Hour 15  
☐ 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**

**Reference Comparison Samples 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  $\geq$  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.

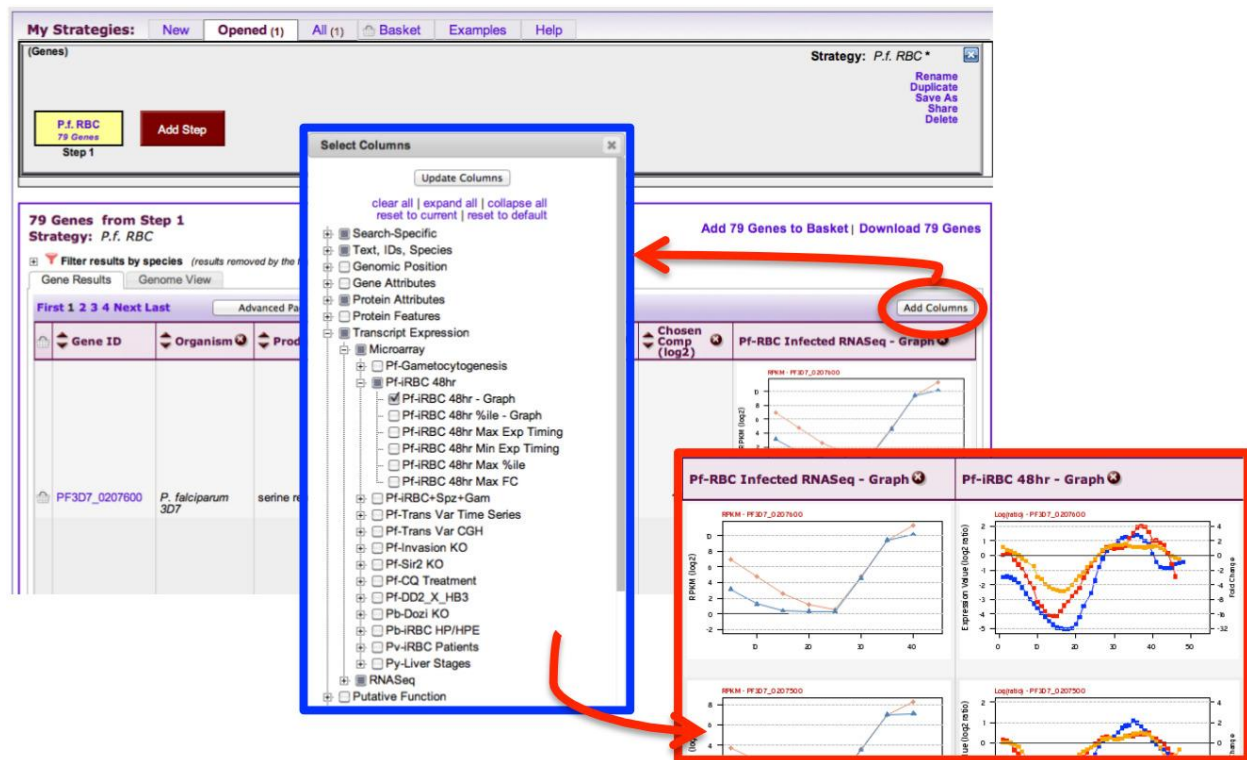
[See the detailed help for this search.](#)

Advanced Parameters

Get Answer

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

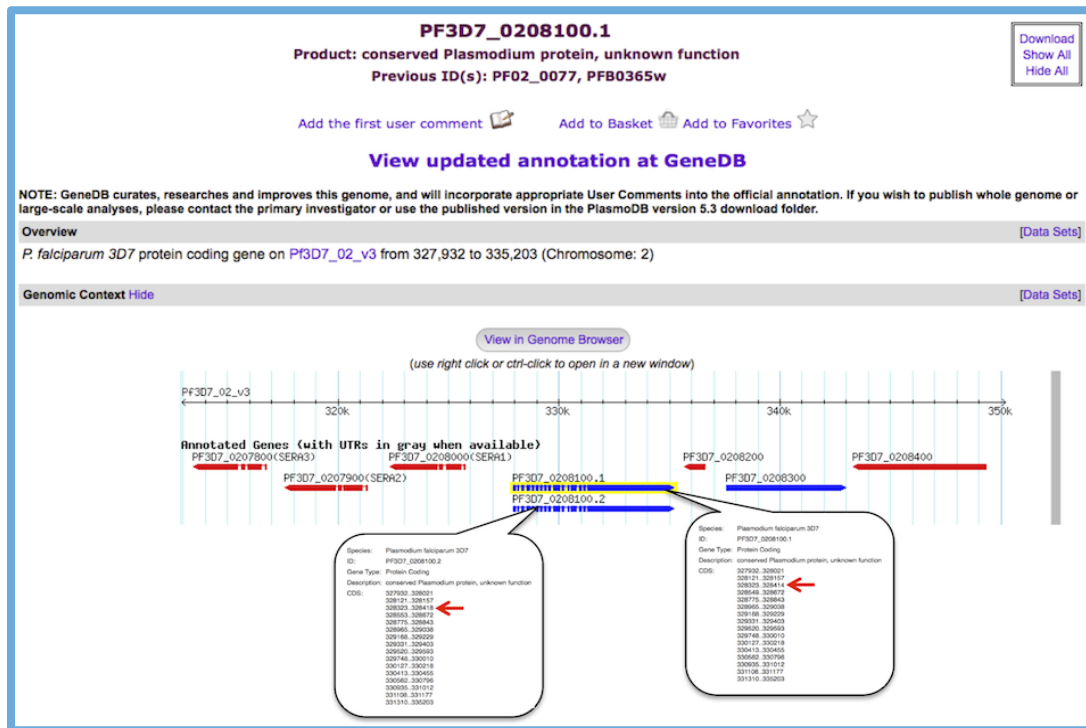
- Hint: PlasmoDB contains data from a similar experiment that was analyzed by microarray instead of RNA sequencing. This experiment is called: Erythrocytic expression time series (3D7, DD2, HB3) (Bozdech et al. and Linas et al.). To directly compare the data for genes returned by the RNA seq search that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



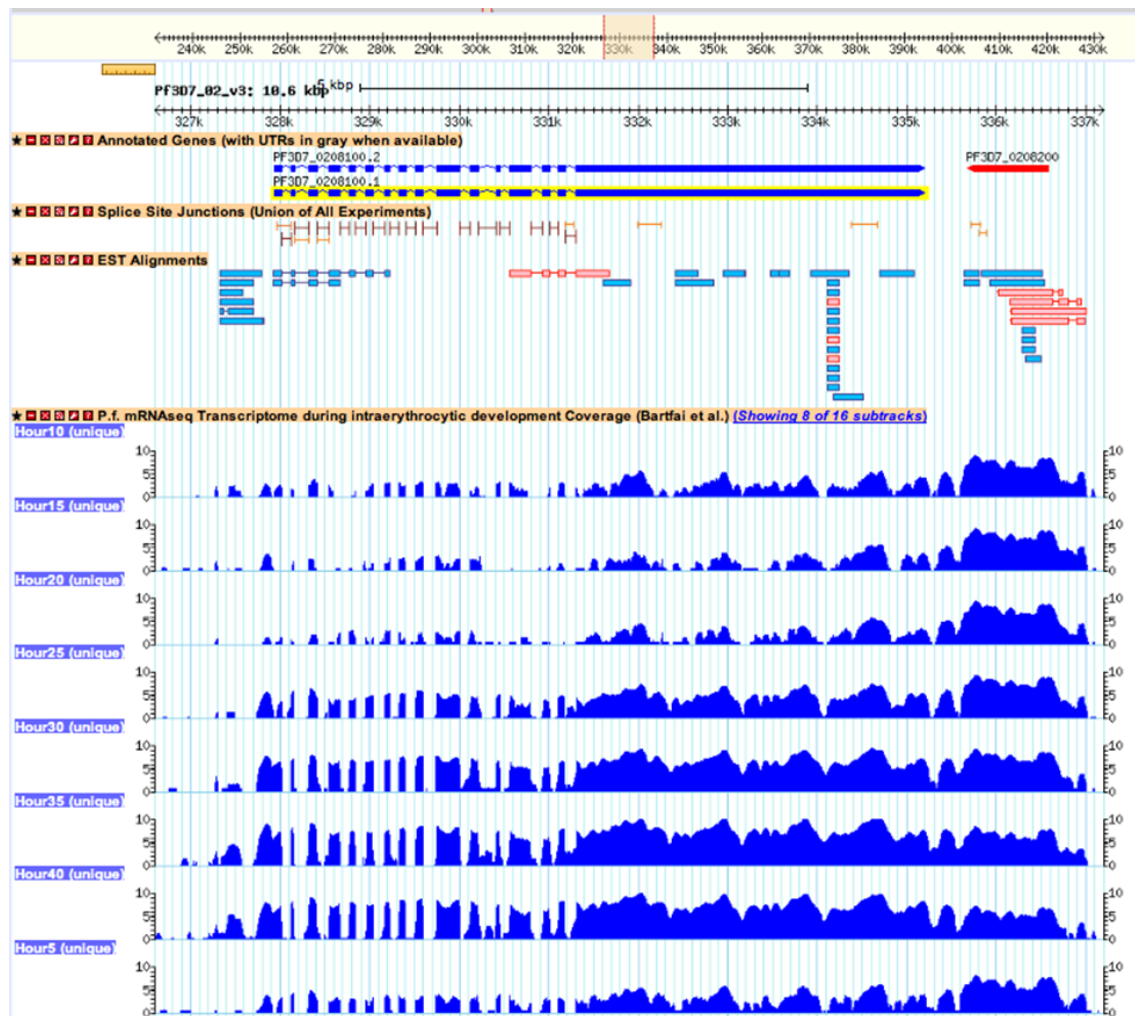
c. How many genes in this result have 16 exons?

- Hint: add a column for number of exons. To help you find the genes with 16 exons, you can sort the columns using the arrows that precede the column heading. Also, clicking the histogram icon in the column heading will provides options for viewing the column data as a table or histogram.
- There are three gene IDs with 16 exons each. Two have similar gene IDs. What does this mean?

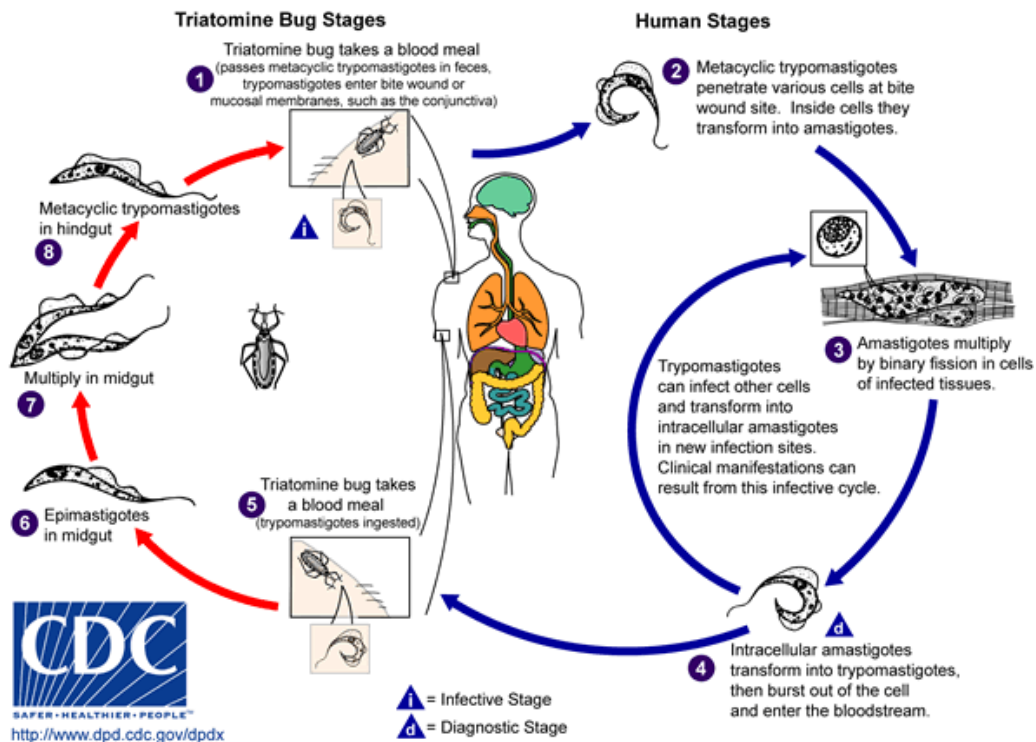
- d. Click on one of the two similar gene IDs from above. Look at the gene page. Take note of the Gene ID. Mouse over the gene models in the genomic context view and explore the popup. What information does it contain? Note that the CDS section includes exon coordinates. Compare the coordinates for the two alternative splice variants of this gene - can you identify the difference (it is very subtle)?



- e. View this gene in the genome browser and load the RNA-seq tracks for this experiment. The track is named: "Transcriptome during intraerythrocytic development mRNAseq Coverage aligned to P falciparum 3D7 (Bartfai et al.) (log plot).
- Do these tracks match the differential expression results you got above? Is this gene differentially regulated between the early time points and the late ones?
  - Do you agree with the alternative splice call? Are there other possible splice variants? (Hint: turn on the track called "Splice Site Junctions (Union of All Experiments)").
- f. What other data type can you load to help in looking at gene structure? (Hint: Look in the transcript expression section of the gbrowse tracks... how about ESTs?).



## 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. The experiment is called: Transcriptomes of Four Life-Cycle Stages (Minning et al.)

**Identify Genes by:**

Expand All | Collapse All

- ☒ Text, IDs, Organism
- ☐ Genomic Position
- ☐ Gene Attributes
- ☐ Protein Attributes
- ☐ Protein Features
- ☐ Similarity/Pattern
- ☒ Transcript Expression
  - ☒ EST Evidence
  - ☒ SAGE Tag Evidence
  - ☒ Microarray Evidence
  - ☒ RNA Seq Evidence
- ☐ Protein Expression
- ☐ Cellular Location
- ☐ Putative Function
- ☐ Evolution
- ☐ Population Biology

**Identify Genes based on Microarray Evidence**

Filter Data Sets:  Legend: DC Direct Com... FC Fold Change FCC Fold Chan... P Percentile

Organism	Data Set	Choose a search	
<i>L. infantum</i> JPCM5	1 Promastigote-to-amastigote differentiation (L.d. Samples) (Lahav et al.)	<input type="checkbox"/> FC	<input type="checkbox"/> P
<i>L. infantum</i> JPCM5	2 Axenic and intracellular amastigote profiles (Rochette et al.)	<input type="checkbox"/> FCC	<input type="checkbox"/> P
<i>L. major</i> strain Friedlin	3 Three Developmental Stages (Stephen M. Beverley)	<input type="checkbox"/> DC	<input type="checkbox"/> P
<i>T. brucei</i> TREU927	4 Life cycle stages and differentiation time course (Kabani et al.)	<input type="checkbox"/> FC	<input type="checkbox"/> P
<i>T. brucei</i> TREU927	5 Expression profiling of five life cycle stages (Marilyn Parsons)	<input type="checkbox"/> FC	<input type="checkbox"/> P
<i>T. brucei</i> TREU927	6 TbDRBD3 Depleted Procyclic Gene Expression (Estevez AM)	<input type="checkbox"/> DC	<input type="checkbox"/>
<i>T. brucei</i> TREU927	7 Expression profiling of in vitro differentiation (Queiroz et al.)	<input type="checkbox"/> FC	<input type="checkbox"/>
<i>T. brucei</i> TREU927	8 mRNA profiles of induced DHH1 vs DEAD:DQAD mutant (Kramer et al.)	<input type="checkbox"/> FCC	<input type="checkbox"/> P
<i>T. brucei</i> TREU927	9 Procyclic trypanosomes: heat shock vs untreated control (Kramer et al.)	<input type="checkbox"/> DC	<input type="checkbox"/> P
<i>T. cruzi</i> CL Brener Esmeraldo-like	10 Transcriptomes of Four Life-Cycle Stages (Minning et al.)	<input checked="" type="checkbox"/> FC	<input type="checkbox"/> P



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 ?

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)

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

For each gene, the search calculates:

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

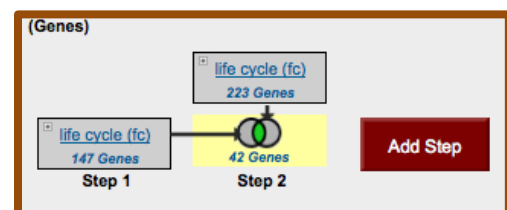
and returns genes when **fold change**  $\geq 2.0$ .

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 of gene to view results from a specific species).

My Strategies: **New** **Opened (1)** All (212) Basket Public Strategies (9) Help

(Genes) Strategy: Tc LifeCyc Marray (fc) \*

Step 1: Tc LifeCyc Marray (147 Genes) → Step 2: 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											
		C.fasciculata	L.braziliensis (nr Genes: 58)	L.donovani	L.infantum	L.major	L.mexicana	L.tarentolae	T.brucei (nr Genes: 39)	T.congolense			
1760	37	85	46	57	52	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 MHOM/BR/75/M2904	LbrM.02: 147,781 - 154,645 (-)	ABC1 transporter, putative	TcCLB.510149.80	OG5_126568	8	112
LbrM.11.0960	L. braziliensis MHOM/BR/75/M2904	LbrM.11: 439,107 - 444,425 (+)	ABC transporter, putative	TcCLB.510149.80	OG5_126568	8	112

- Explore your results. Did you find anything interesting?

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

- 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. 1  
 return protein coding Genes  
 that are up-regulated  
 with a Fold change >= 50

between each gene's average 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 Don't care

Advanced Parameters

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

**Up-regulated**

Expression

Comparison

50 fold

Average Reference

Reference Samples Comparison 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.



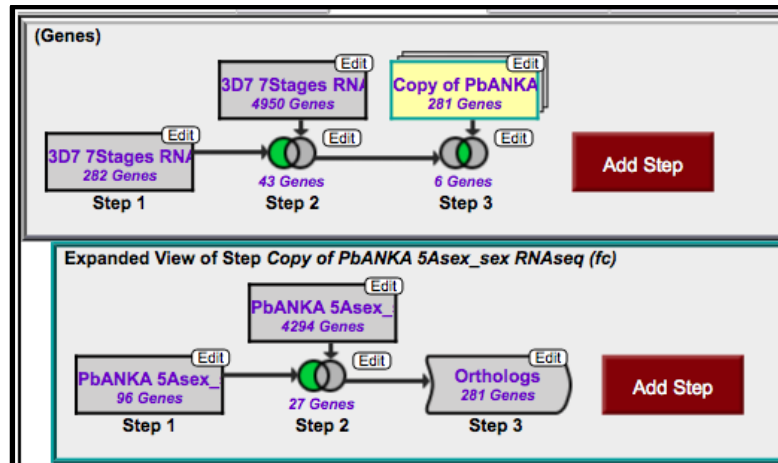
- b. The above search will give you all genes that are up-regulated by 50 fold in ookinetes compared to the other stages. However, this does not mean that these genes are not expressed well in the other stages. How can you remove genes from the list that are likely not expressed in the other stages?
- Hint: run a search for genes based on RNAseq 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? Try different values - for example, 40 (minimum) and 100(maximum).

- c. Which metabolic pathways are represented in this gene list? (Hint: add a step and transform results to metabolic pathways).

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 metabolic 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.

Identify Genes by:

Expand All | Collapse All

- Text IDs, Organism
- Genomic Position
- Gene Attributes
- Protein Attributes
- Protein Features
- Similarity/Pattern
- Transcript Expression
- Protein Expression
- Cellular Location
- Putative Function
- GO Term
- EC Number
- Metabolic Pathway BETA
- Phenotype
- High-Throughput Phenotyping
- Evolution
- Population Biology

Identify Genes based on High-Throughput Phenotyping

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

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

fold difference 1.5

P value less than or equal to 1E-6

Apply to Any or All Selected Samples? any

Protein Coding Only: protein coding

[Advanced Parameters](#)

[Get Answer](#)

### Identify Genes based on High-Throughput Phenotyping

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

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

**fold difference** 1.5

**P value less than or equal to** 1E-6

**Apply to Any or All Selected Samples?** any ▾

**Protein Coding Only:** protein coding ▾

**My Strategies:** New Opened  
 (Genes)  

T.b. RNAi fc  
1612 Genes  
Step 1

- 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.

**My Strategies:** New Opened  
 (Genes)  

T.b. RNAi fc  
1612 Genes  
Step 1

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

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

**fold difference** 1.5

**P value less than or equal to** 1E-6

**Apply to Any or All Selected Samples?** any ▾

**Protein Coding Only:** protein coding ▾

**My Strategies:** New Opened

(Genes)

T.b. RNAi fc  
1612 Genes  
Step 1

→

T.b. RNAi fc  
2619 Genes  
Step 2

→

621 Genes  
Step 2

→

Add Step

## 5. Exploring Expression Quantitative Trait Locus (eQTL) data in PlasmoDB.

Genetic crosses were instrumental in implicating the PfCRT gene in chloroquine resistance. PlasmoDB contains expression quantitative trait locus data from Gonzales *et. al.* PLoS Biol 6(9): e238. The trait that was examined in this study was gene expression using microarray experiments.

- a. Go to the gene page for the gene with the ID PF3D7\_0630200. Can you identify the genomic region (haplotype block) that is “most” associated with this gene, ie. has the highest LOD score? (Hint: examine the table called “Regions/Spans associated by eQTL experiment on HB3 x DD2 progeny” on the gene page.

SNPs Alignment <a href="#">Show</a> <a href="#">[Data Sets]</a>					
Gene Location <a href="#">Show</a> <a href="#">[Data Sets]</a>					
Regions/Spans associated by eQTL experiment on HB3 x DD2 progeny (LOD cut off = 1.5) <a href="#">Hide</a> <a href="#">[Data Sets]</a>					
Haplotype Block	Genomic Segment (Liberal)	Genomic Segment (Conservative)	LOD Score (opens a haplotype plot)	Search for Genes (Liberal by Default)	Search for Genes (Liberal by Default)
PF3D7_05_v3_68.8	<a href="#">PF3D7_05_v3:959929-1010786</a>	<a href="#">PF3D7_05_v3:1007897-1008018</a>	4.94	<a href="#">Genes Contained in this Region</a>	<a href="#">Genes Associated to this Region</a>
PF3D7_05_v3_68.8	<a href="#">PF3D7_05_v3:1010972-1040241</a>	<a href="#">PF3D7_05_v3:1018620-1018825</a>	4.94	<a href="#">Genes Contained in this Region</a>	<a href="#">Genes Associated to this Region</a>
PF3D7_05_v3_65.9	<a href="#">PF3D7_05_v3:870388-1007896</a>	<a href="#">PF3D7_05_v3:918503-959928</a>	4.9	<a href="#">Genes Contained in this Region</a>	<a href="#">Genes Associated to this Region</a>

- b. What kinds of genes do you find in this region? Click on the first link in the column “Genomic Segment (liberal)”. Now examine the gene table on the genomic segment record page.

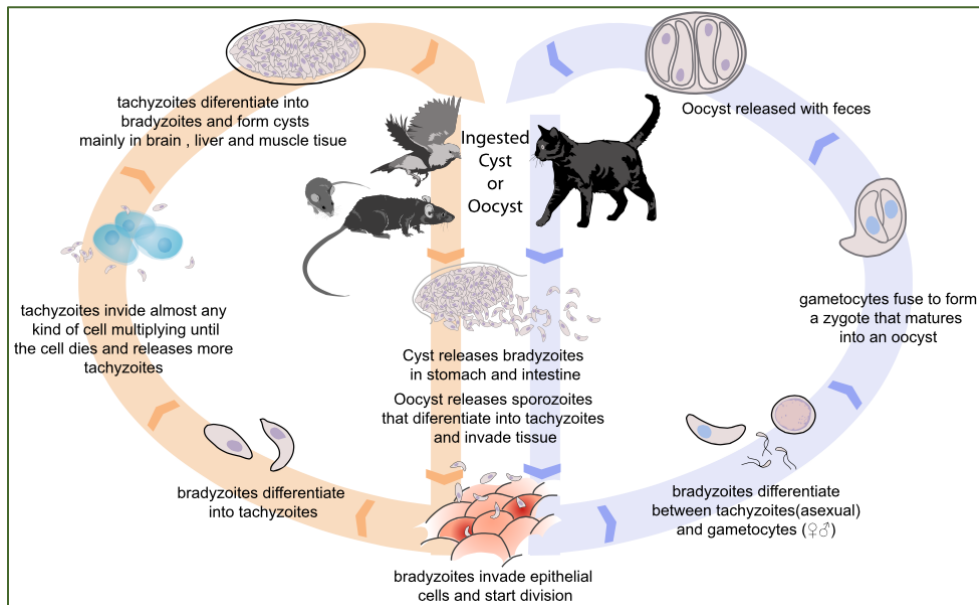
Genes <a href="#">Hide</a>				
Gene ID	Start	End	Strand	Product Description
<a href="#">PF3D7_0523000</a>	957890	962149	forward	multidrug resistance protein (MDR1)
<a href="#">PF3D7_0523100</a>	963227	965044	reverse	mitochondrial processing peptidase alpha subunit, putative
<a href="#">PF3D7_0523200</a>	966123	969737	forward	conserved Plasmodium protein, unknown function
<a href="#">PF3D7_0523300</a>	970266	970962	reverse	conserved Plasmodium protein, unknown function
<a href="#">PF3D7_0523400</a>	973518	975876	forward	DnaJ protein, putative
<a href="#">PF3D7_0523500</a>	976690	977815	reverse	outer arm dynein lc3, putative
<a href="#">PF3D7_0523600</a>	978665	979870	forward	conserved Plasmodium protein, unknown function
<a href="#">PF3D7_0523700</a>	980754	985354	reverse	conserved Plasmodium membrane protein, unknown function
<a href="#">PF3D7_0523800</a>	990005	992059	forward	transporter, putative
<a href="#">PF3D7_0523900</a>	993433	994607	reverse	conserved Plasmodium membrane protein, unknown function
<a href="#">PF3D7_0524000</a>	998753	1002124	forward	karyopherin beta (KASbeta)
<a href="#">PF3D7_0524100</a>	1004237	1008108	forward	conserved Plasmodium protein, unknown function
<a href="#">PF3D7_0524200</a>	1008636	1009404	reverse	conserved Plasmodium membrane protein, unknown function

- c. What other genes are associated with this block?
- Hint: go back to the eQTL table on the gene page, and click the “genes associated with this region” link. Run the search on the next page and examine the list of genes. It might be useful to sort this list based on the LOD scores.

6 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 “Expression Profiling of oocyst, tachyzoite, and bradyzoite development in strain M4 (John Boothroyd)” microarray experiment. In this experiment,



Identify Genes by:

Expand All | Collapse All

- ☒ Text, IDs, Organism
- ☒ Genomic Position
- ☒ Gene Attributes
- ☒ Protein Attributes
- ☒ Protein Features
- ☒ Similarity/Pattern
- ☐ Transcript Expression
- ☐ EST Evidence
- ☐ SAGE Tag Evidence
- ☒ Microarray Evidence
- ☐ RNA Seq Evidence
- ☐ ChIP on Chip Evidence
- ☐ Protein Expression
- ☐ Cellular Location
- ☐ Putative Function

# Identify Genes based on Microarray Evidence

Filter Data Sets:

Legend:

FC
Fold Chan...

FCC
Fold Chan...

P
Percentile

S
Similarity

Organism	Data Set	Choose a search			
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	

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 T.g. Life Cycle Stages (fold change)

For the Experiment Oocyst, Tachyzoite and Bradyzoite Development

return protein coding Genes

that are up-regulated

with a Fold change  $\geq 10$

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

select all | clear all

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

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:

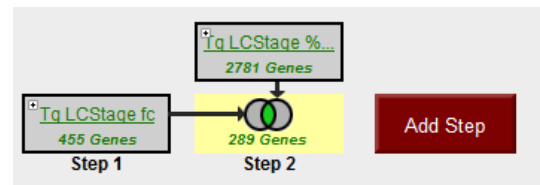
$$\text{fold change} = \frac{\text{maximum expression value in comparison samples}}{\text{maximum expression value in reference samples}}$$

and returns genes when fold change  $\geq 10$ . 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.

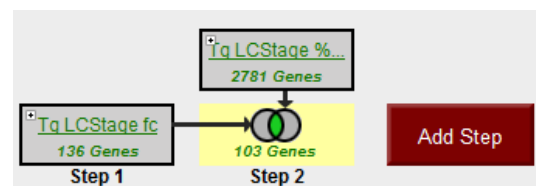
b. Add a step to limit this set of genes to only those for which all the non-oocyst stages are expressed below 50<sup>th</sup> percentile ... ie likely not expressed at those stages. (*Hint: after you click on add step find the same experiment under microarray expression and chose the percentile search*).

- 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"**.
- Note: you can 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) to view the graphs in the final result table.



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?

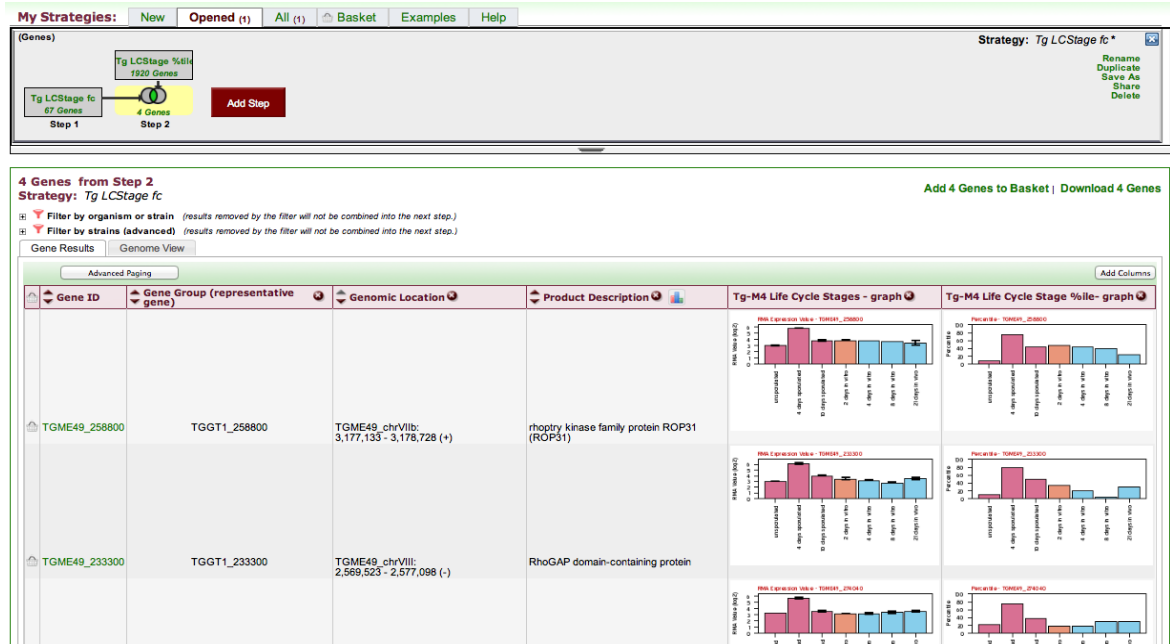


Save this strategy so that you can use it for an exercise we are doing later during the course.



- 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 or fewer results back? Why?



## 7 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 Expression 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.

### Identify Genes by:

- ☐ Expand All | Collapse All
- ☐ Text, IDs, Organism
- ☐ Genomic Position
- ☐ Gene Attributes
- ☐ Protein Attributes
- ☐ Protein Features
- ☐ Similarity/Pattern
- ☐ Transcript Expression
- ☐ EST Evidence
- ☒ **Microarray Evidence**
- ☐ Protein Expression
- ☐ Cellular Location
- ☐ Putative Function
- ☐ Evolution
- ☐ Population Biology

### Identify Genes based on Microarray Evidence

Filter Data Sets:  Legend: DC Direct Comparison FC Fold Change P Percentile

Organism	Data Set	Choose a search
<i>L. infantum</i> JPCM5	Expression profiling of the promastigote time-course (L.d. Samples) (Peter Myler)	<span>FC</span> <span>P</span>
<i>L. infantum</i> JPCM5	axenic and intracellular amastigote profiles (Barbara Papadopoulos)	<span>P</span>
<i>L. major</i> strain Friedlin	Three Developmental Stages (Stephen M. Beverley)	<span>DC</span> <span>P</span>
<i>T. brucei</i> TREU927	Dynamic mRNA Expression analysis of cells undergoing synchronous life-cycle differentiation (Keith R. Matthews)	<span>FC</span> <span>P</span>
<i>T. brucei</i> TREU927	Expression profiling of five life cycle stages (Marilyn Parsons)	<span>FC</span> <span>P</span>
<i>T. brucei</i> TREU927	Procytic TbDRBD3 Depletion (Antonio Estevez)	<span>DC</span>
<i>T. brucei</i> TREU927	Expression profiling of in vitro differentiation time series (Christine Clayton)	<span>FC</span>
<i>T. brucei</i> TREU927	Induced DHH1 in wild type and DEAD:DQAD mutant (Mark Carrington)	<span>P</span>
<i>T. brucei</i> TREU927	Procytic trypanosomes treated with heat shock (Mark Carrington)	<span>DC</span> <span>P</span>
<i>T. cruzi</i> CL Brener Esmeraldo-like	Life-Cycle Stages (Rick Tarleton)	<span>FC</span> <span>P</span>

Fold Change
Percentile

### Identify Genes based on T.b. Expression profiling of five life cycle stages Microarray (fold change)

[Tutorial](#)

For the Experiment  return genes that are  with a Fold change  $\geq$

between each gene's  expression value in the following

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

and to  expression value in the following

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

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

(Dots represent the gene's expression values for selected 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:

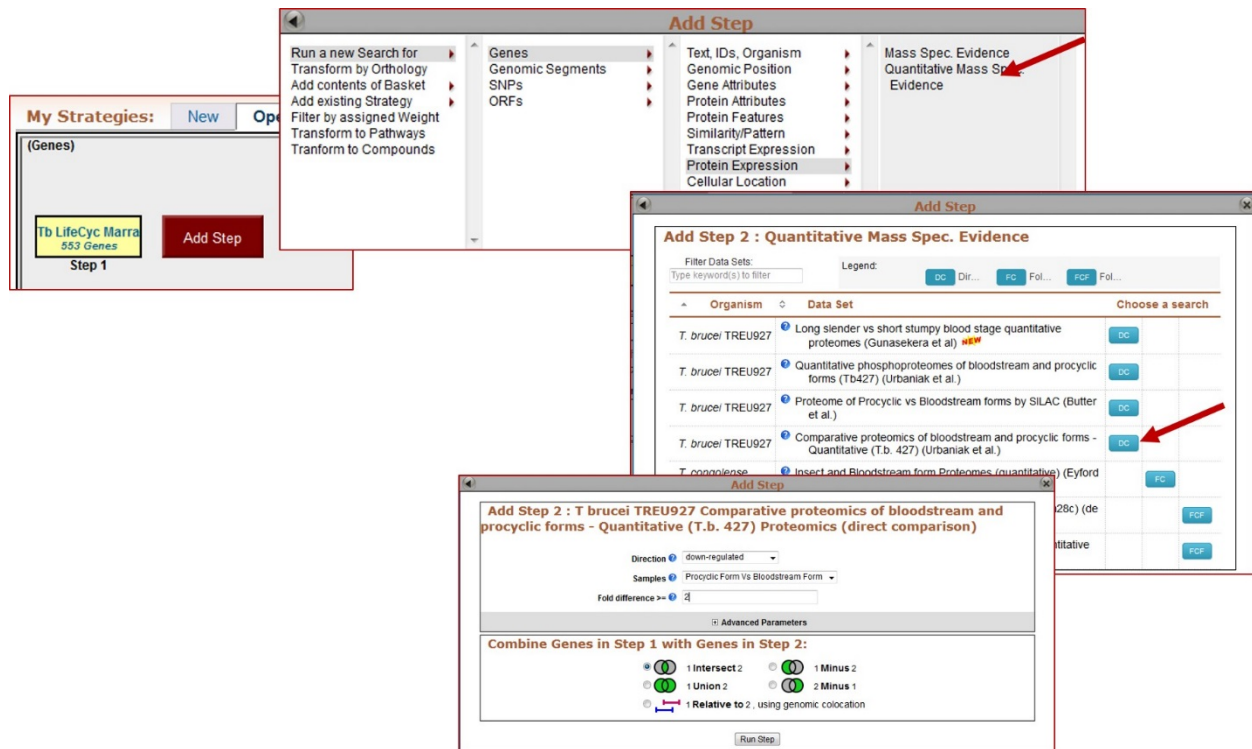
$$\text{fold change} = \frac{\text{average expression value in reference samples}}{\text{average expression value in comparison samples}}$$

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

Protein Coding Only:

- b. Add a step to compare with quantitative protein expression. Select protein expression then “Quantitative Mass Spec Evidence”. Configure this search to return genes that are down-regulated in procyclic form relative to blood form.



- 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).