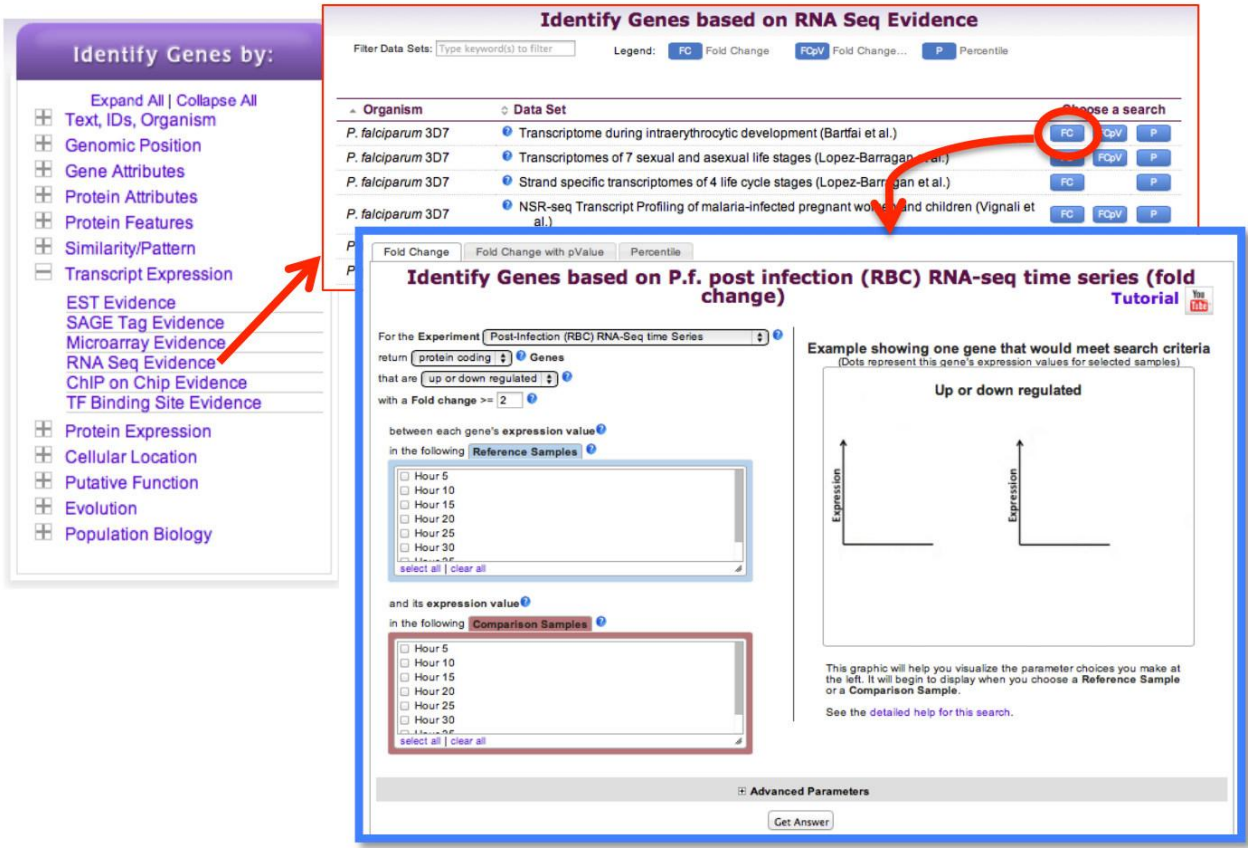
Exploring Transcriptomic data

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.



- 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.

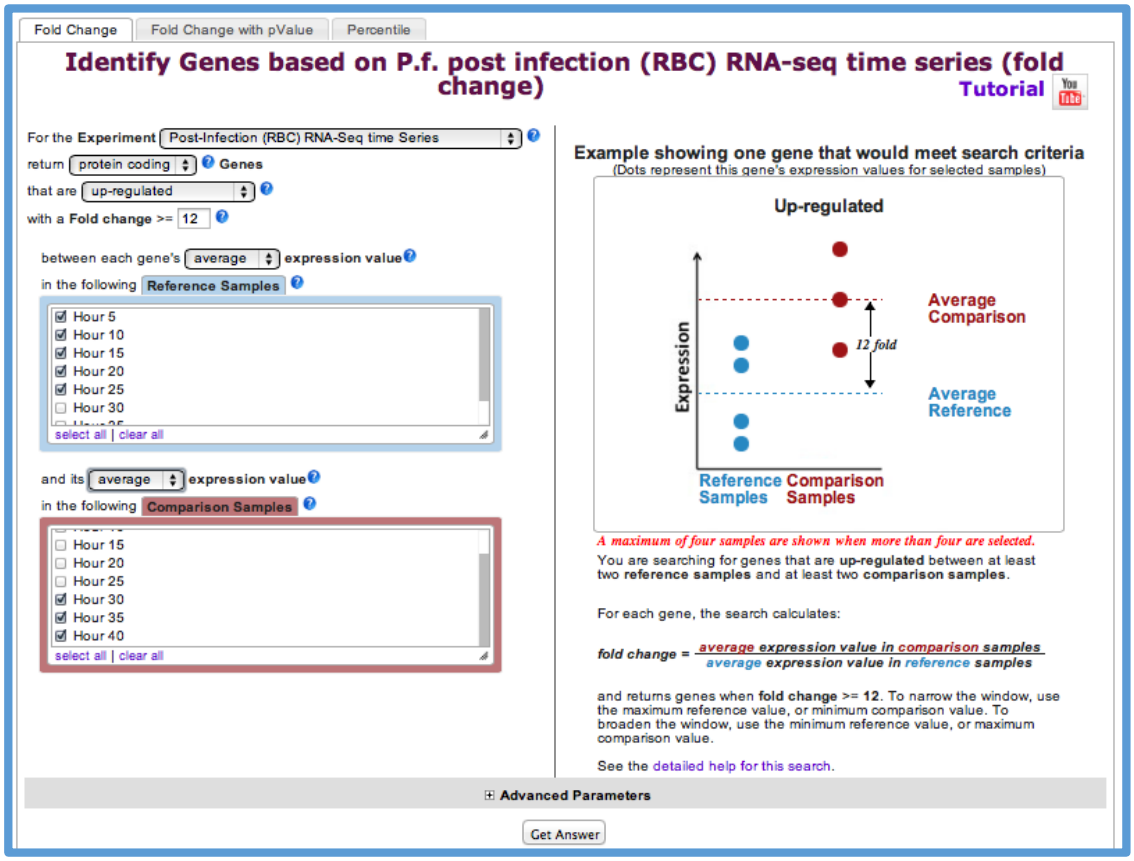
- Fold Change>= the intensity of difference in expression needed before a gene is returned by the search. Choose 12 but feel free to modify this.

- 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

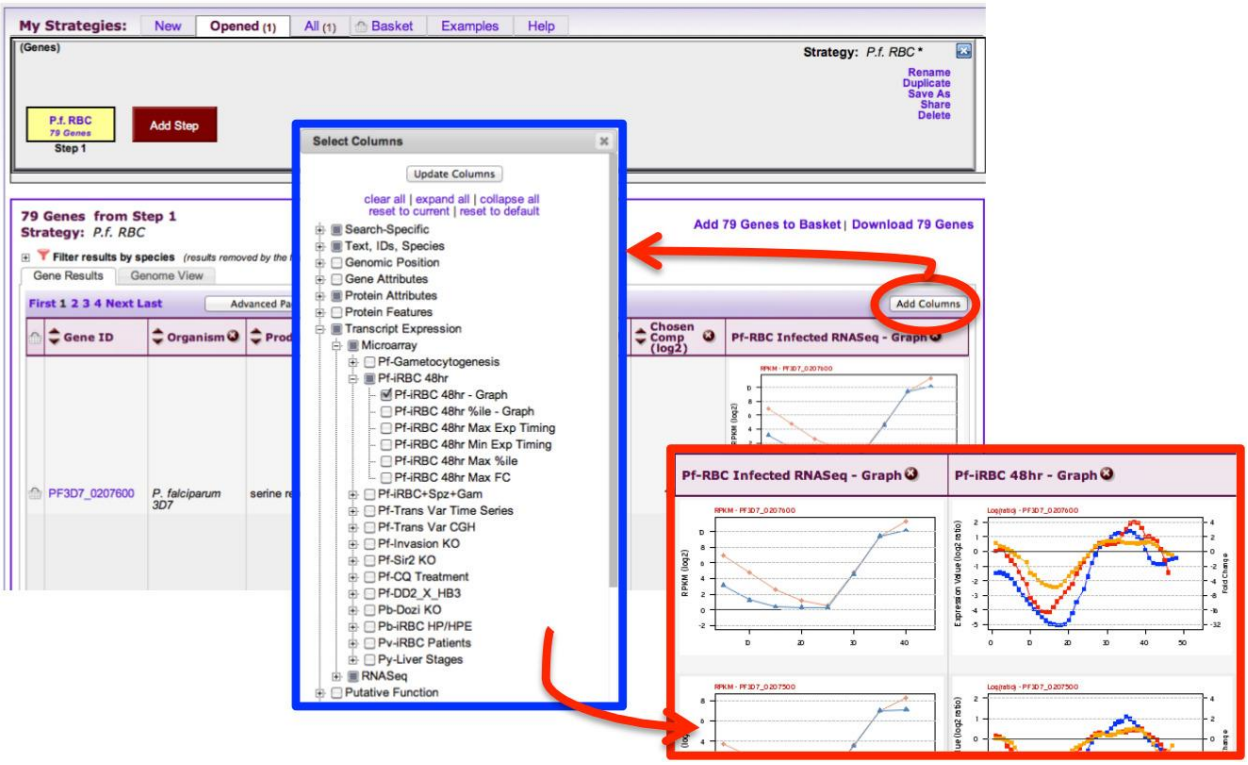
- And it's AVERAGE expression value: This is the operation applied to comparison samples. see explanation above. Choose average

- 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



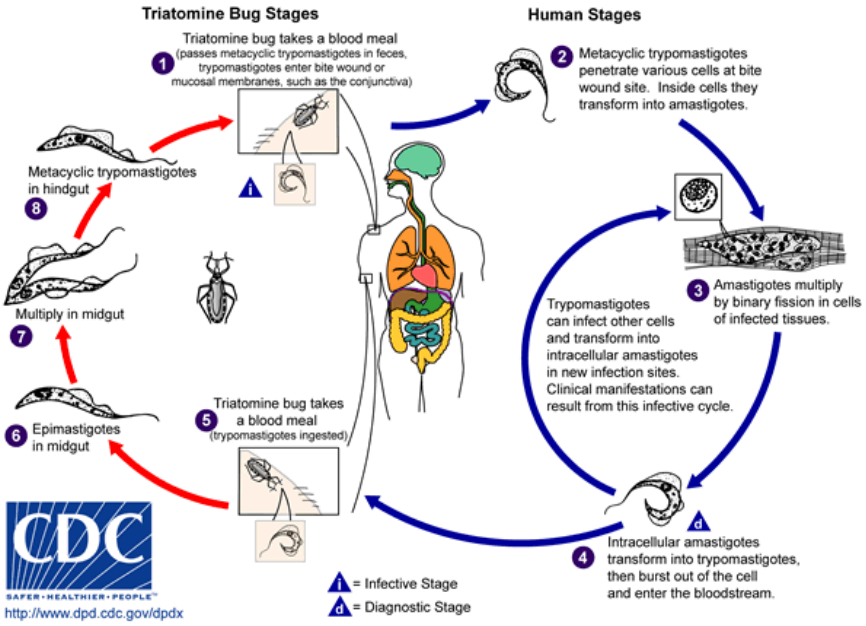
b. For the genes returned by the search, how does the RNA-sequence data compare to microarray data?

- 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".

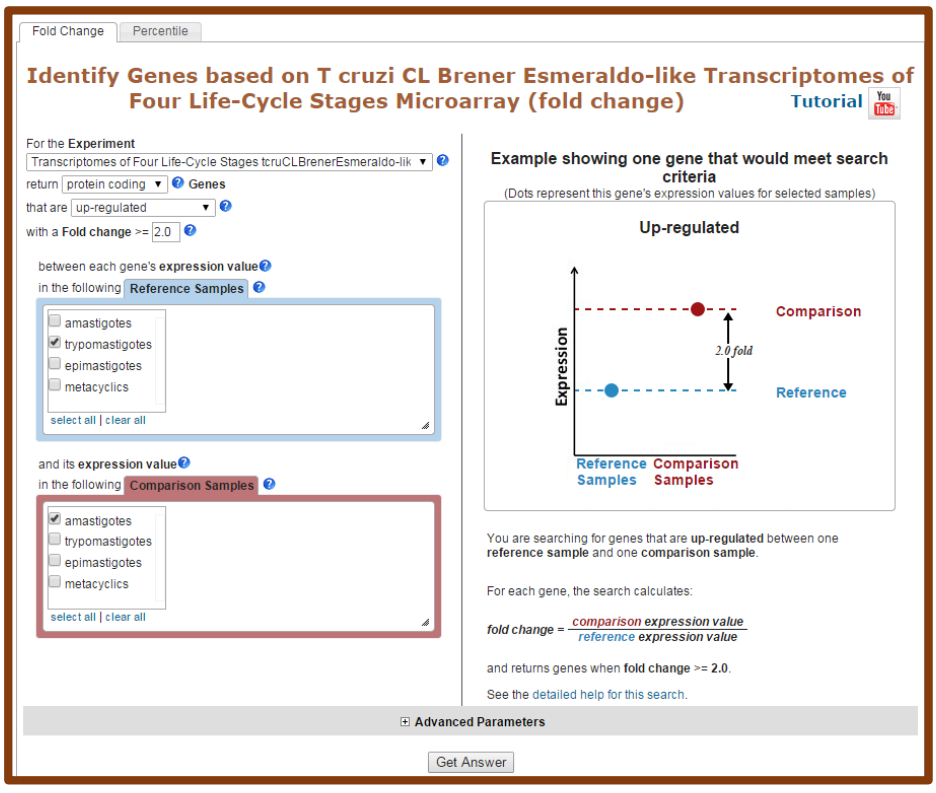


OPTIONAL: You can also run a fold change search using this experiment to compare results on a genome scale. Add a step to your strategy and intersect the results of a fold change search using the "Erythrocytic expression time series (3D7, Dd2, HB3) (Bozdech et al. and Linas et al.)" experiment (under 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.

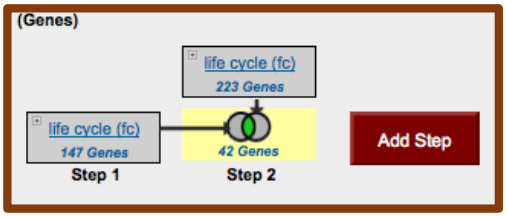
2. Exploring microarray data in TriTrypDB.
   Note: For this exercise use http://www.tritrypdb.org



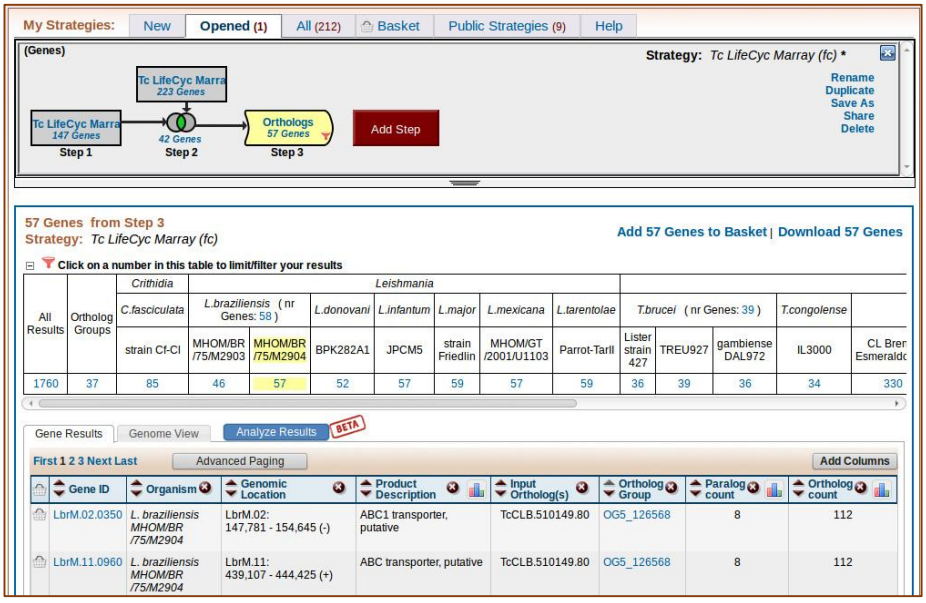
a. Find *T. cruzi* protein coding genes that are upregulated in amastigotes compared to trypomastigotes. Go to the transcript expression section then select microarray. Choose the fold change (FC) search for the data set called: **Transcriptomes of Four Life-Cycle Stages (Minning et al.)**.



- 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).

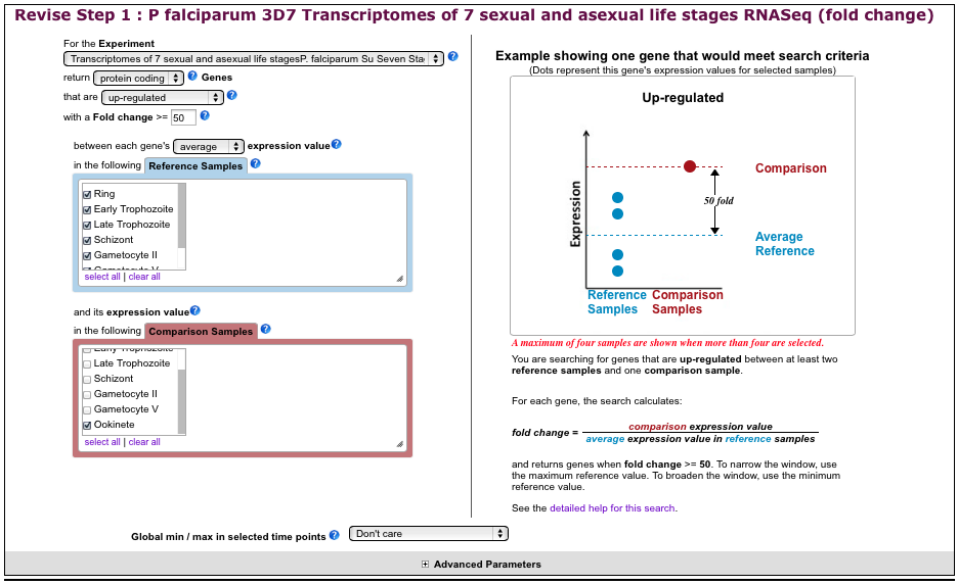


- Do these genes have orthologs in other kinetoplastids? (*Hint*: add a step and run an ortholog transform on your results).
- How many orthologs exist in *L. braziliensis*? (*Hint*: look at the filter table between the strategy panel and your result list. Click on the number in of gene to view results from a specific species). 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.



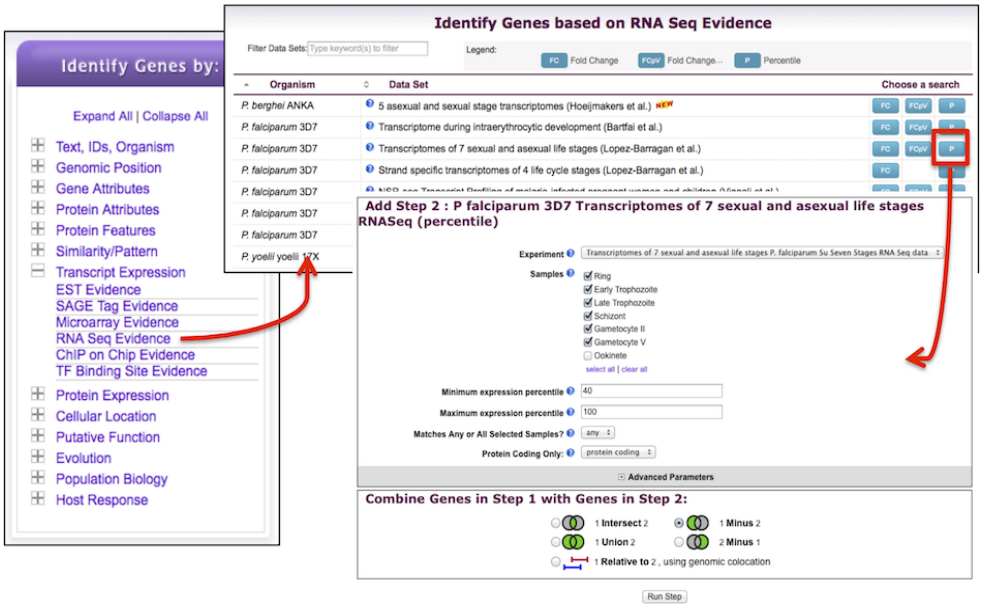
3. Finding genes based on RNAseq evidence and inferring function of hypothetical genes.
Note: Use http://plasmodb.org for this exercise.

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.

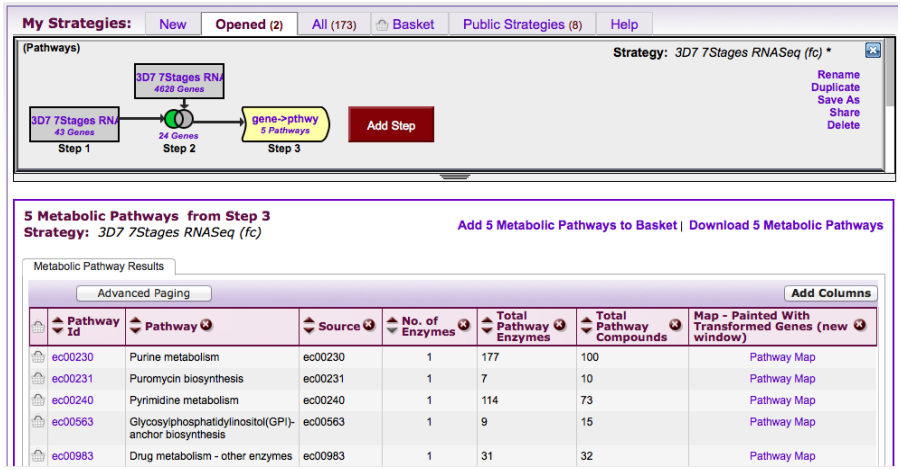


b. The above search will give you all genes that are up-regulated by 50 fold in ookinetes compared to the other stages. Despite the high fold change, some genes in the list may be highly expressed in the other stages. How can you remove genes from the list that are highly expressed in the other stages?

- *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%?*



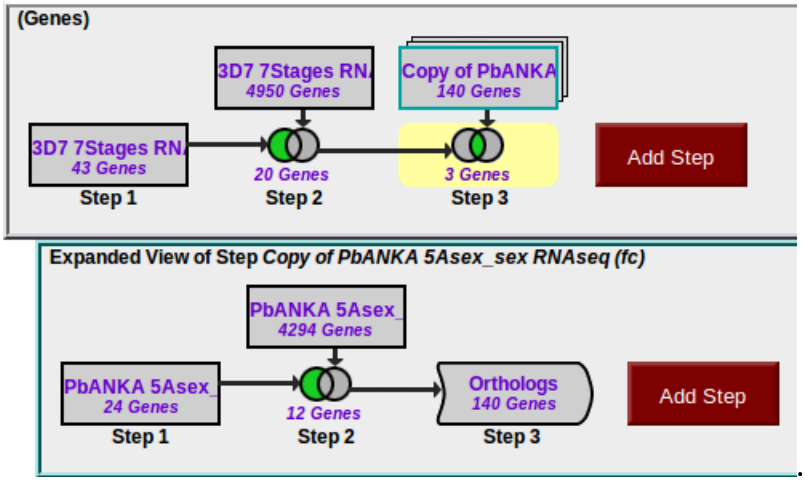
c. Which metabolic pathways are represented in this gene list? *Hint:* add a step and transform results to pathways. How does this result compare to running a pathways enrichment on step 2?



d. What happens if you revise the first step and modify the fold difference to a lower value - 10 for example?

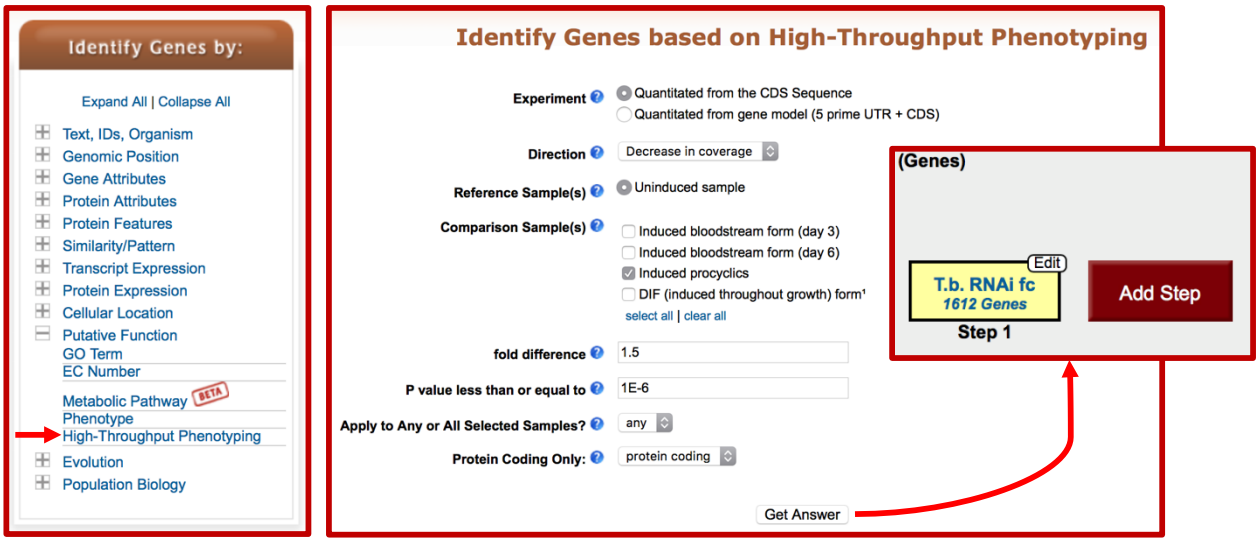
e. PlasmoDB also has an experiment examining gene expression during sexual development in *Plasmodium berghei* (rodent malaria). Can you determine if there are genes that are up-regulated in both human and rodent ookinetes (compared to all other stages)? *Hint:* start by deleting the last step you added 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.



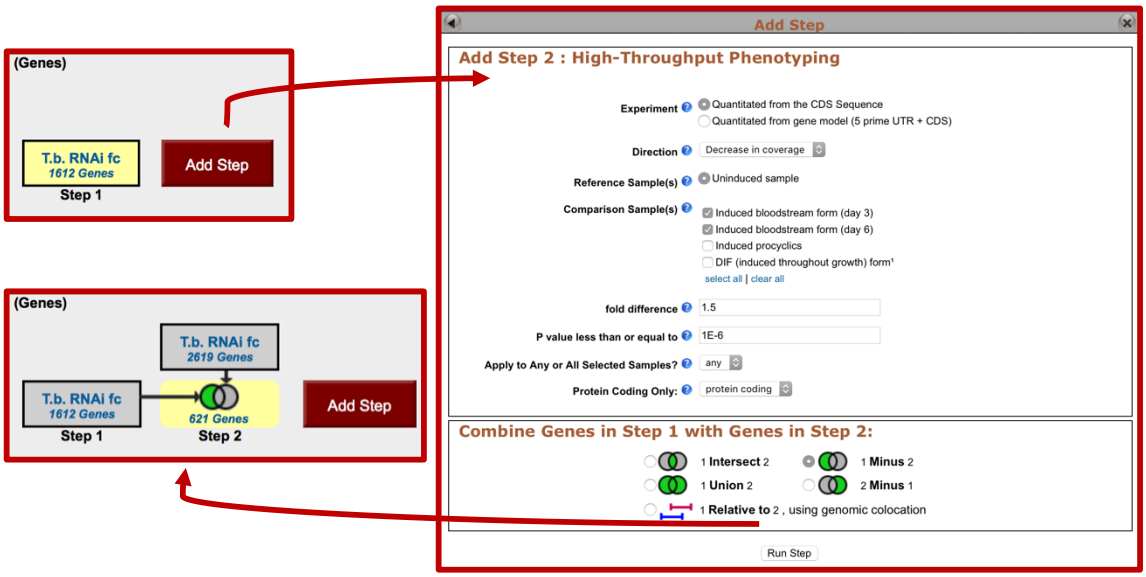
4. Find genes that are essential in procyclics but not in blood form *T. brucei*.
Note: for this exercise use http://TriTrypDB.org.

- 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.

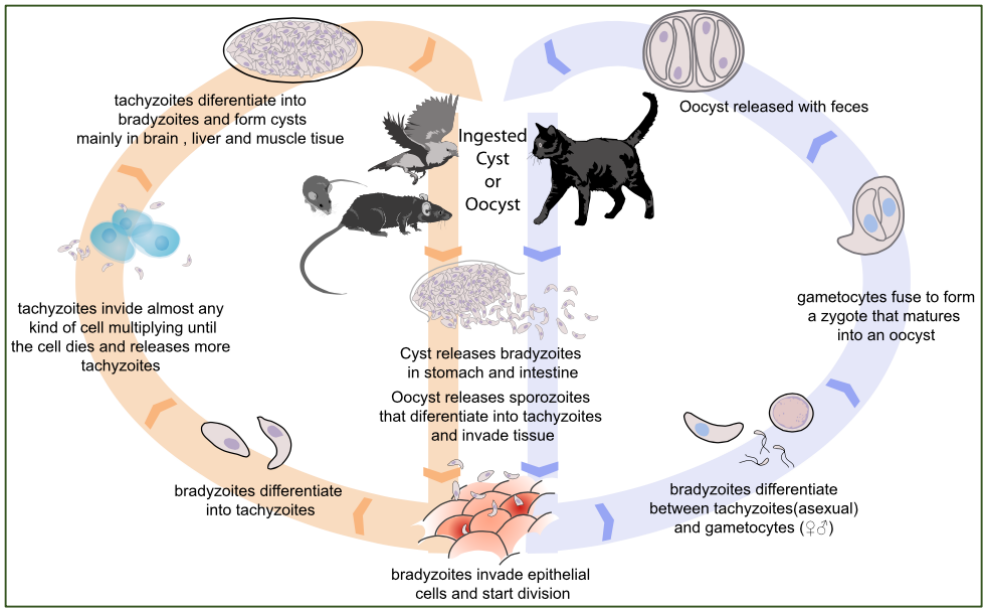


- Next add a step and run the same search except this time select the "induced bloodstream form" samples.

- How did you combine the results? Remember you want to find genes that are essential in procyclics and not in blood form.

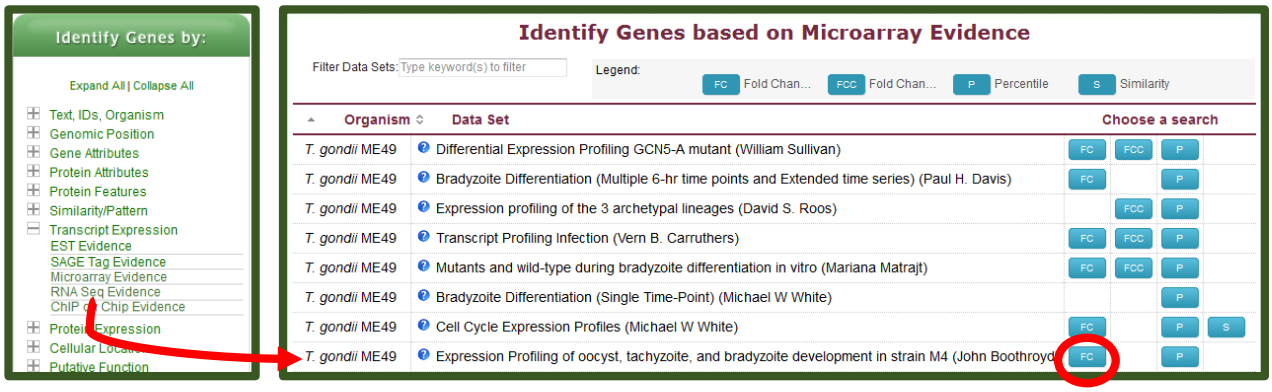


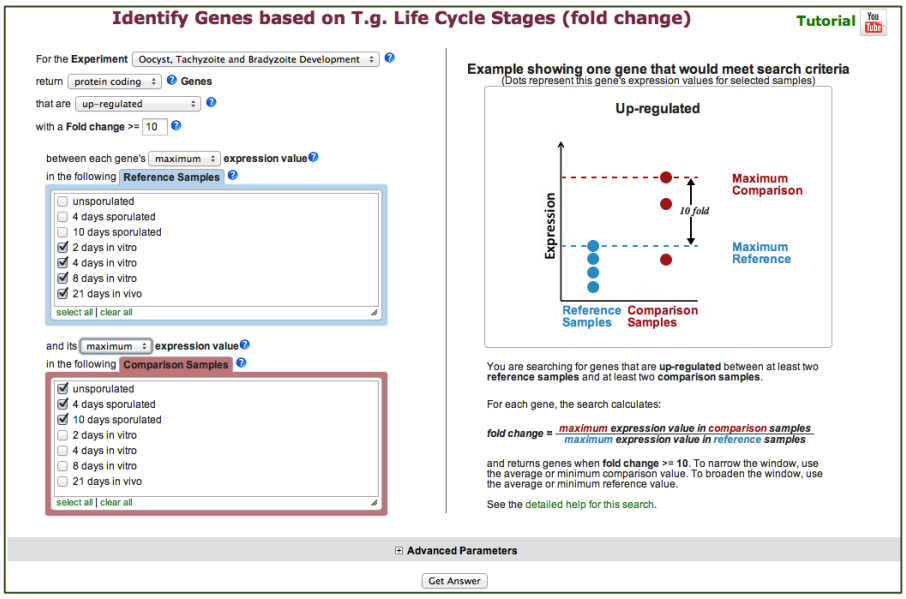
5. Finding oocyst expressed genes in *T. gondii* based on microarray evidence.
Note: For this exercise use http://toxodb.org



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.

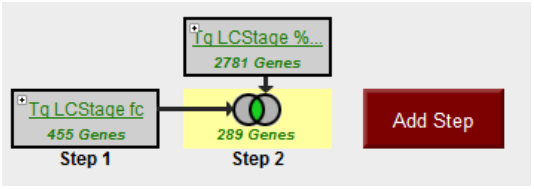




b. Add a step to limit this set of genes to only those for which all the non-oocyst stages are expressed below 50th percentile … ie likely not expressed at those stages. *(Hint*: after you click on add step find the same experiment under microarray expression and chose the percentile search).

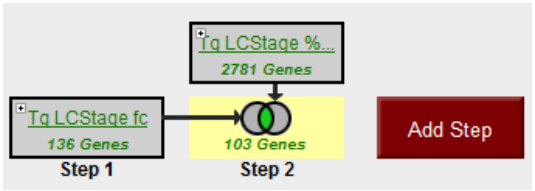
- Select the 4 **non-oocyst** samples.
- We want all to have less than 50th percentile so set ***minimum percentile*** to 0 and ***maximum percentile*** to 50.

- 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.

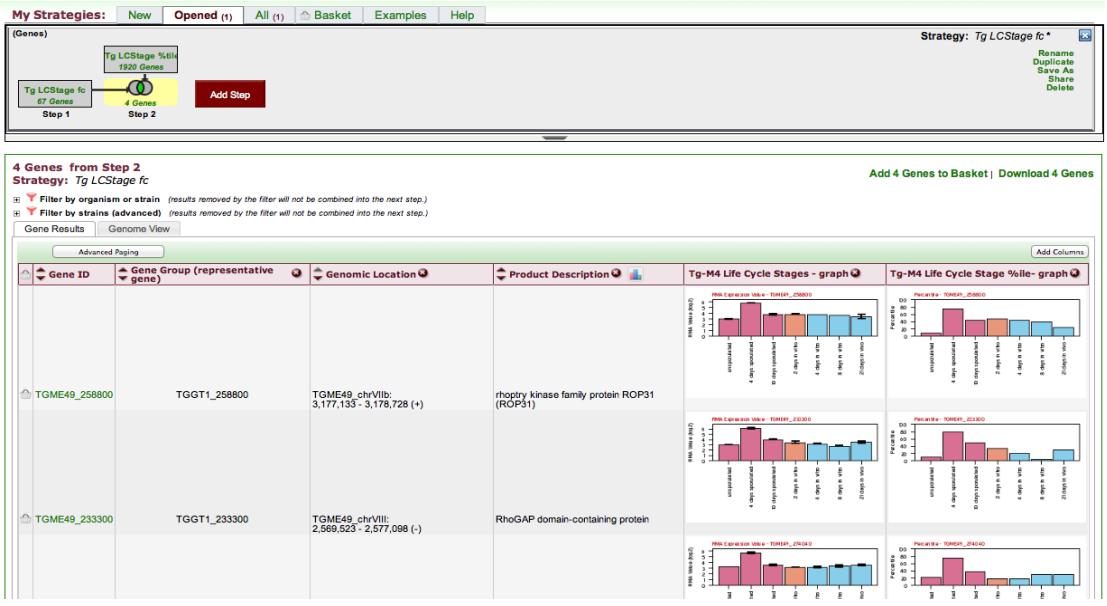
- Does this result look cleaner/more convincing? Why?
- Would you consider these genes to be oocyst specific?



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.

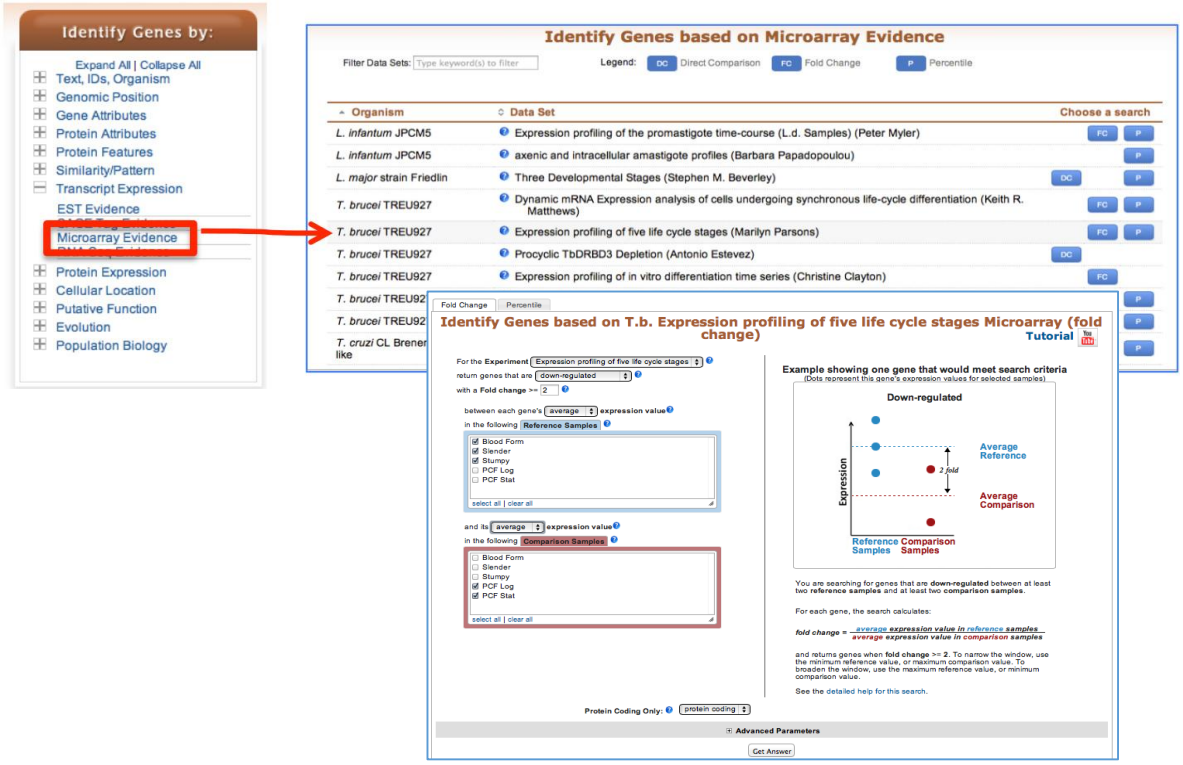
- Do all these genes have day 4 oocysts as the global maximum time point?
- Note that we still have the step to limit the percentile of non-oocyst samples to <= 50th percentile. What happens if you revise this step to also include the unsporulated and day 10 oocyst samples in this percentile range? Do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



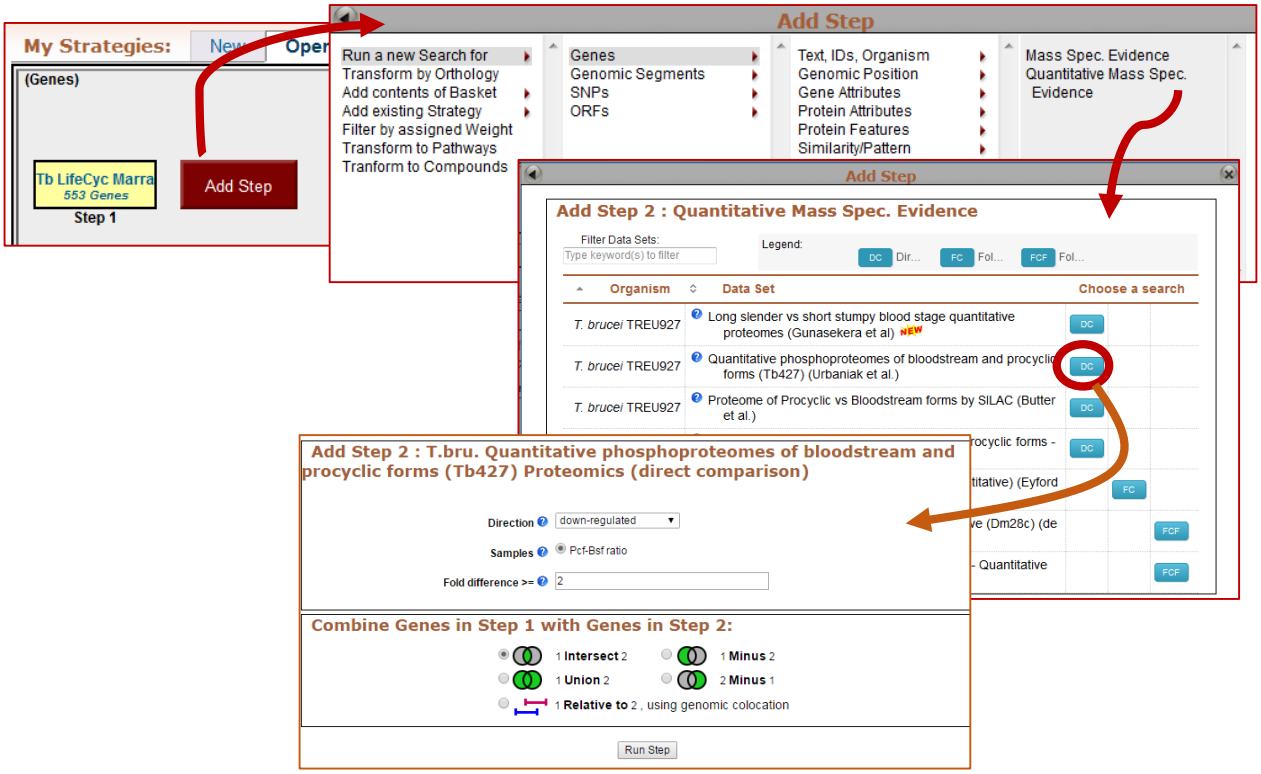
6. Comparing RNA abundance and Protein abundance data.
Note: for this exercise use http://TriTrypDB.org.

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.



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.



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?

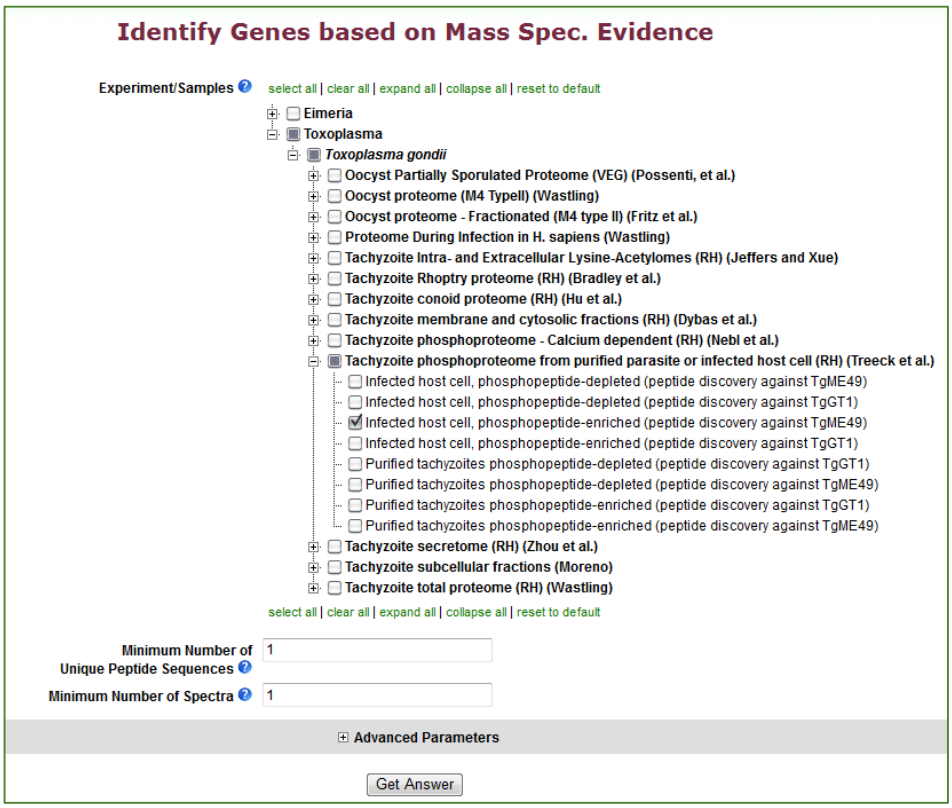
e. Can you think of ways to provide more confidence (or cast a broader net) in the microarray step? (*Hint:* you could insert steps to restrict based on percentile or add a RNA Sequencing step that has the same samples).

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

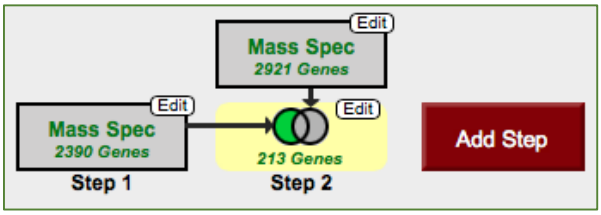
For this exercise use http://www.toxodb.org

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

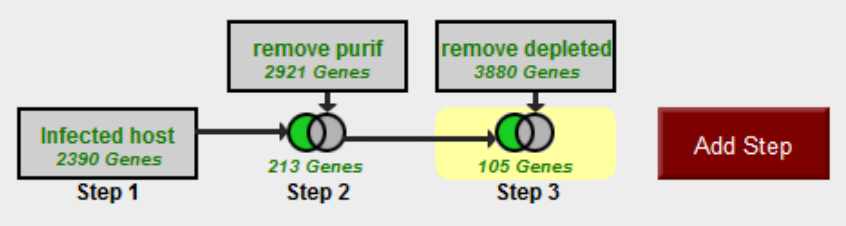
7a. Find all genes with evidence of phosphorylation in intracellular tachyzoites. Select the "**Infected host cell, phosphopeptide-enriched** (peptide discovery against **TgME49**)" sample under the experiment called "**Tachyzoite phosphoproteome from purified parasite or infected host cell (RH) (Treeck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



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

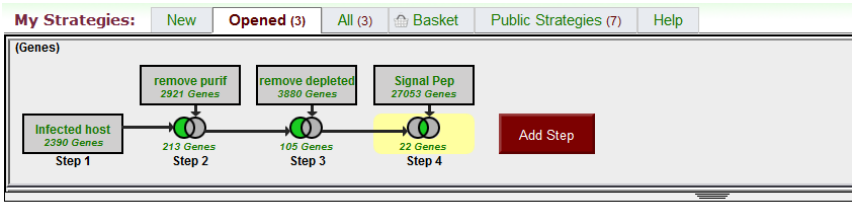


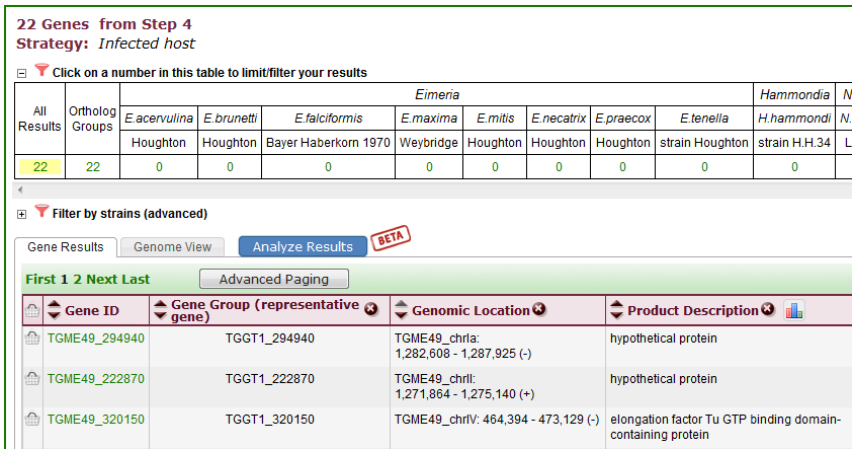
7c. Remove all genes that are also present in the phosphopeptide-depleted fractions (select both intracellular and extracellular).



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.*

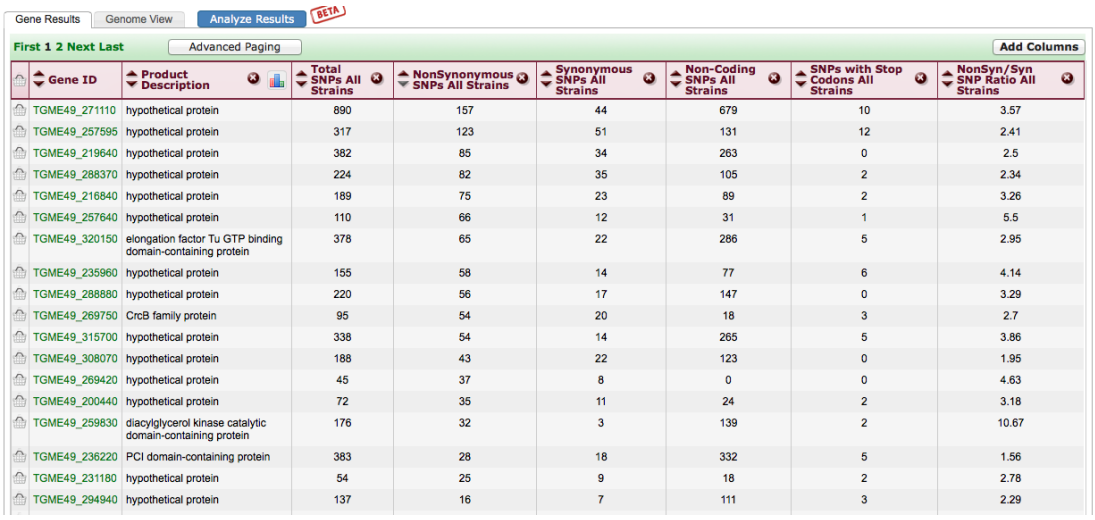
7e. Are any of these genes likely to be secreted? Hint: add a step searching for genes with secretory signal peptides.





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