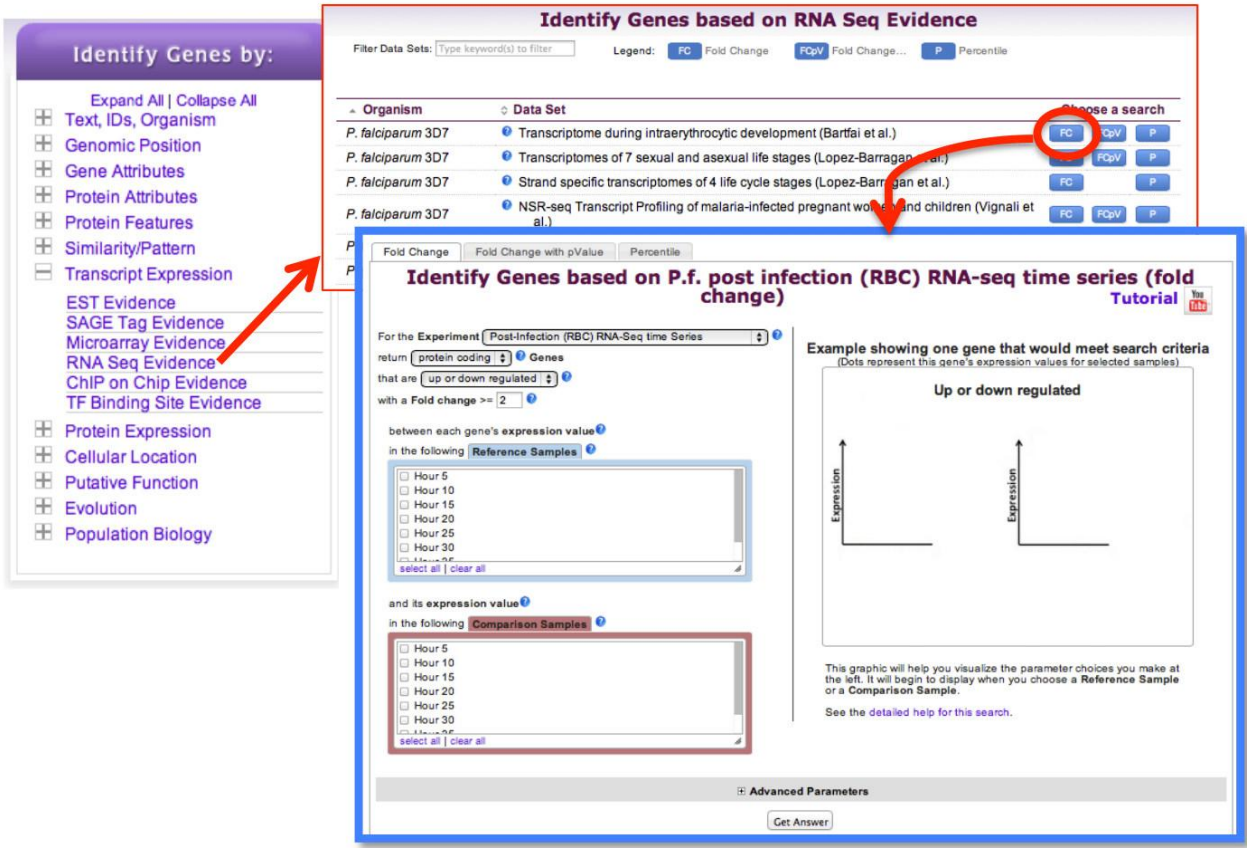
Exploring Transcriptomic data

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.



-

- 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**.

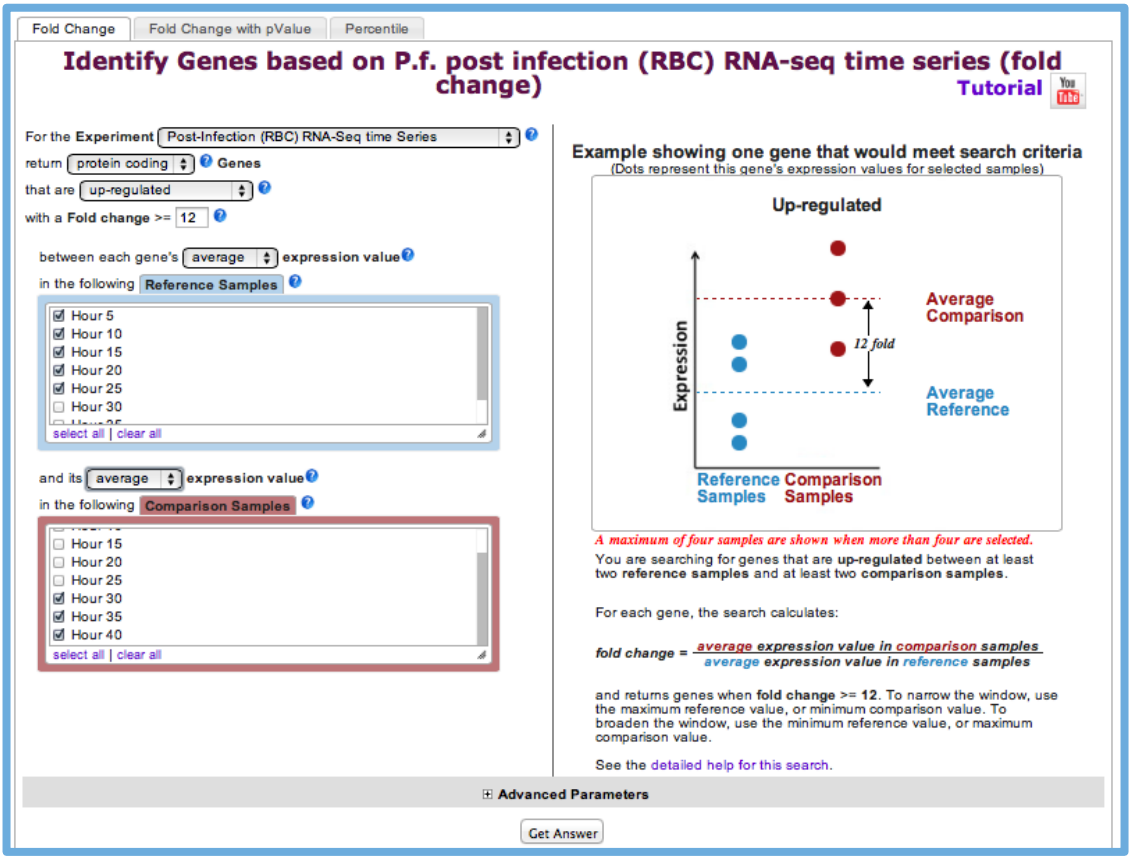
- **Fold Change>=:** the intensity of difference in expression needed before a gene is returned by the search. **Choose 12** but feel free to modify this.

- **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**

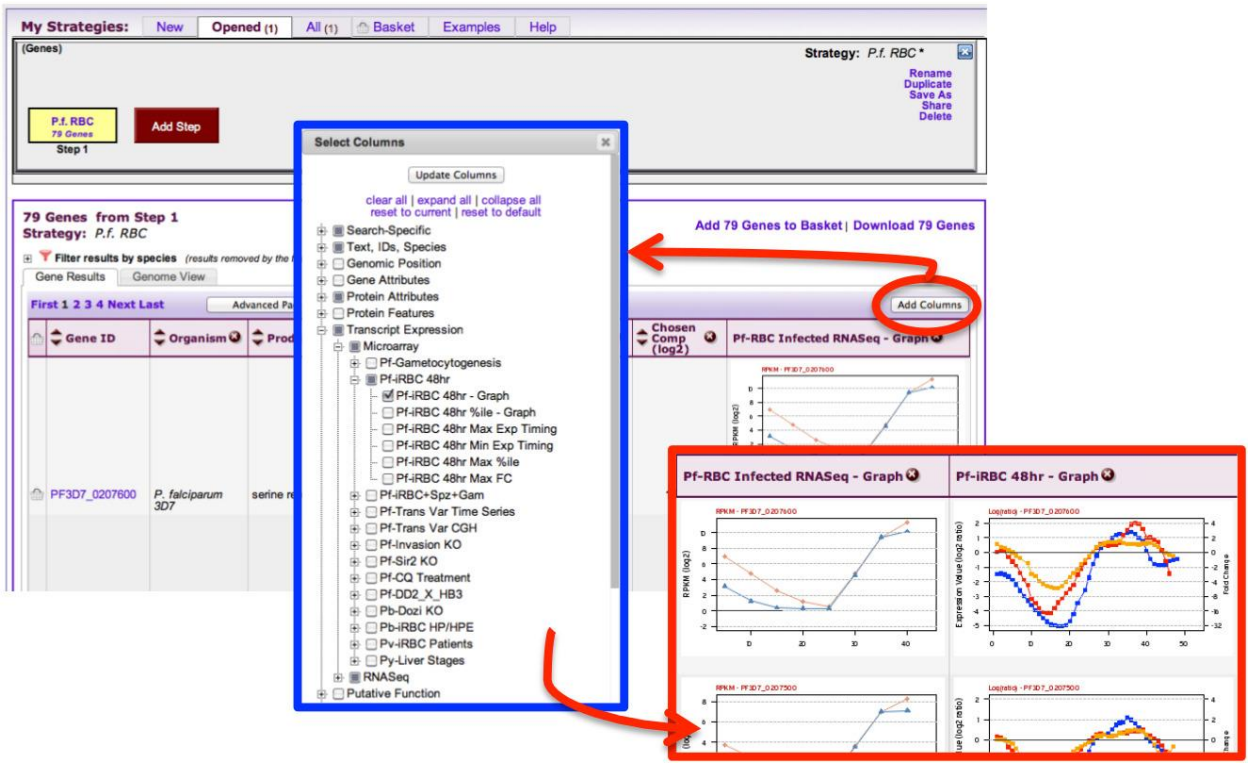
- **And it's AVERAGE expression value:** This is the operation applied to comparison samples. see explanation above. **Choose average**

- **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**

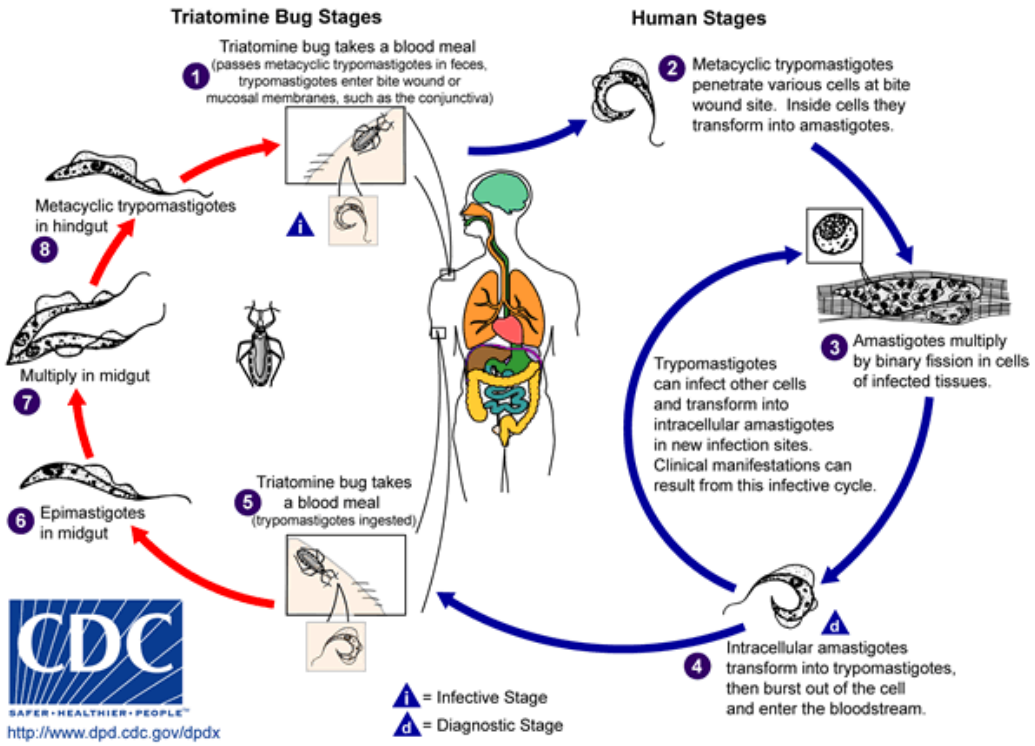


b. For the genes returned by the search, how does the RNA-sequence data compare to microarray data?

- 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".

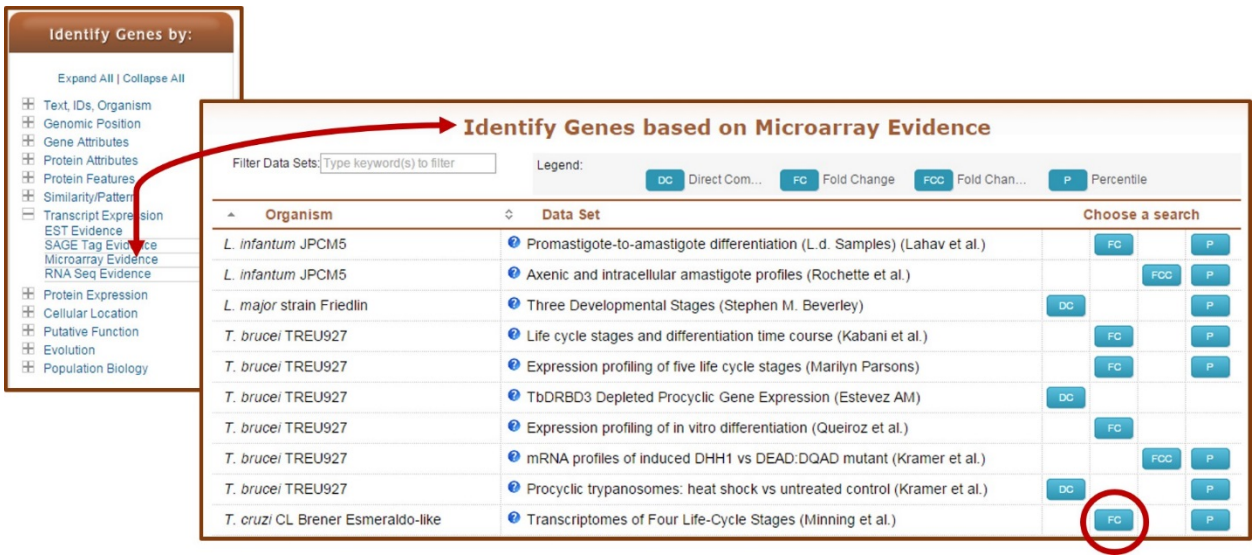


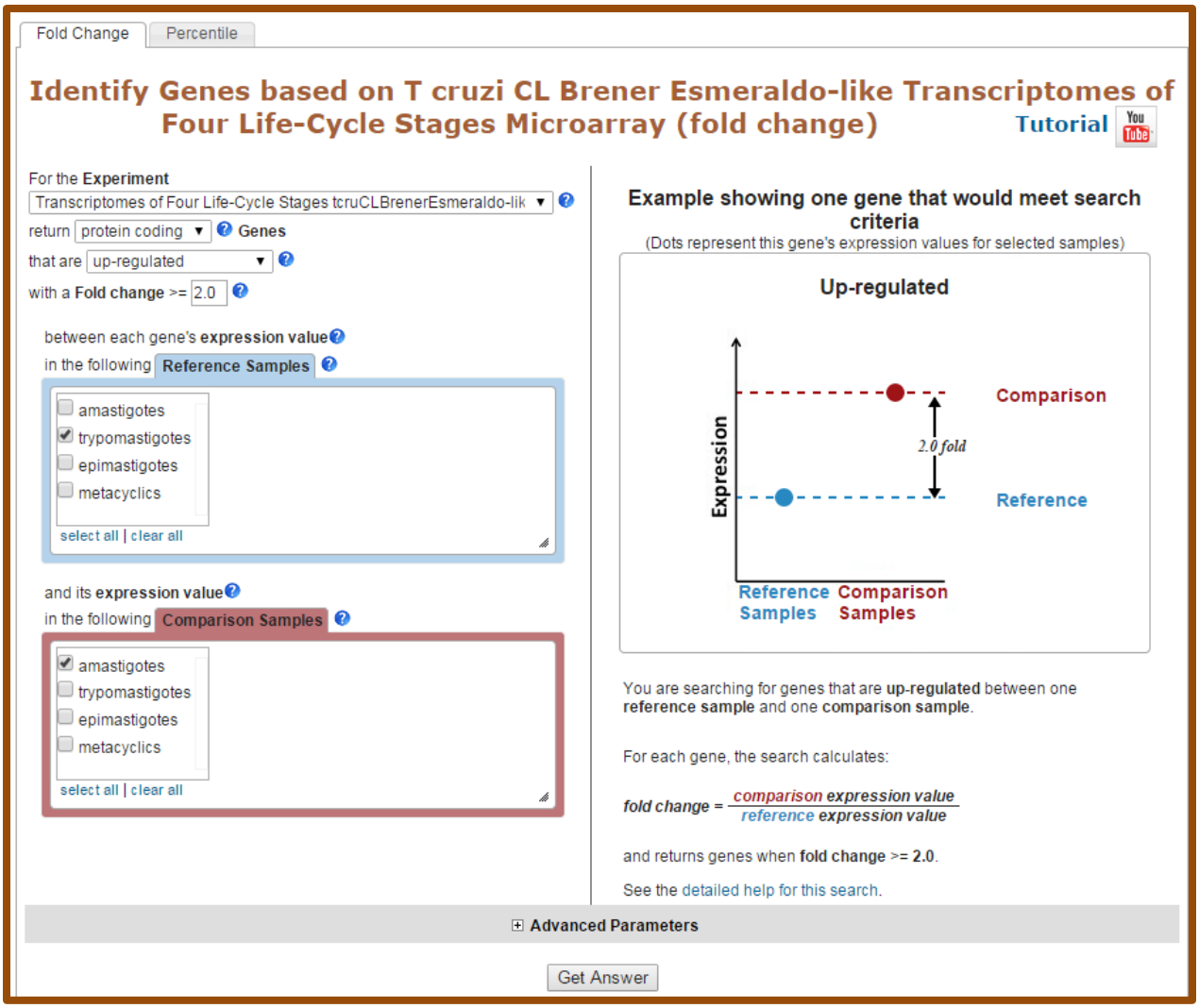
2. Exploring microarray data in TriTrypDB.



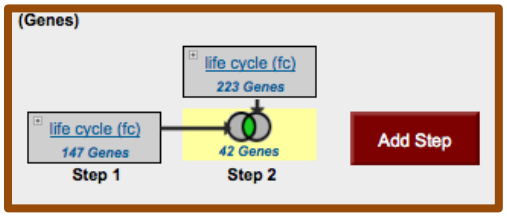
Note: For this exercise use http://www.tritrypdb.org

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.)



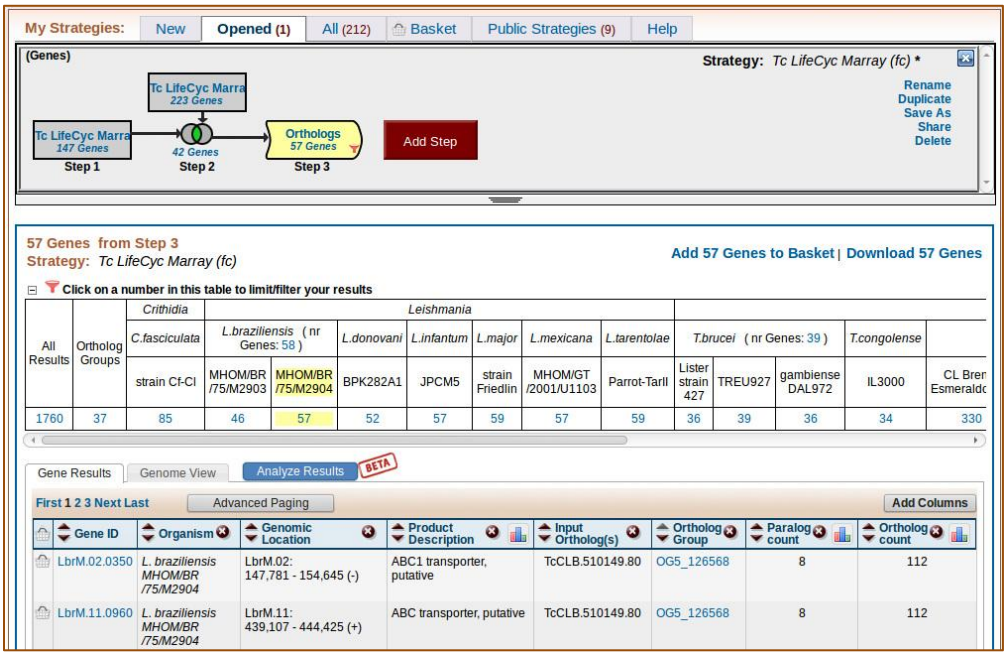


- Select the direction of regulation, your reference sample and your comparison sample. For the fold change keep the default value 2.
- How many genes did you find? Do the results seem plausible?
- Are any of these genes also up-regulated in the replicative insect stage (epimastigotes)? How can you find 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).

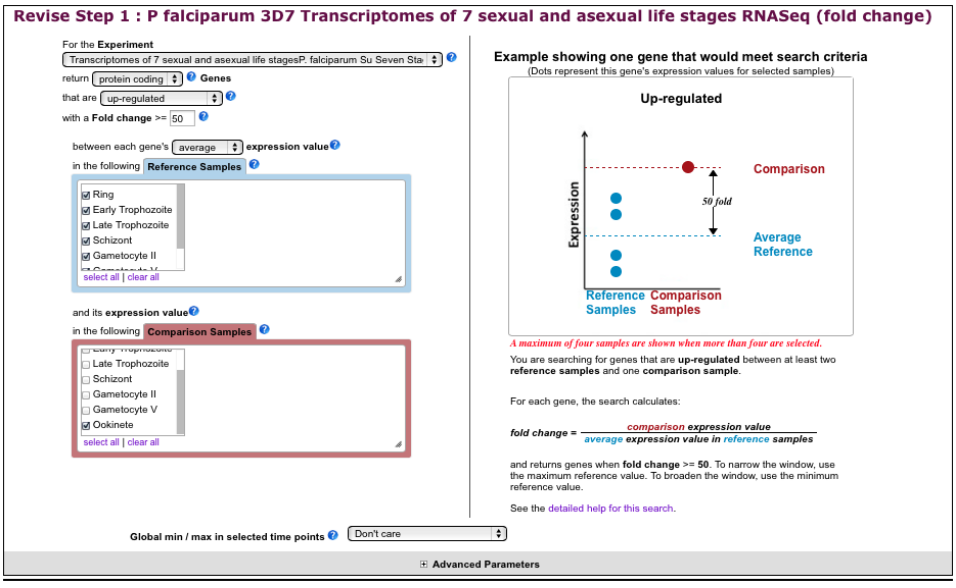
- How many orthologs exist in *L. braziliensis*? (*Hint*: look at the filter table between the strategy panel and your result list. Click on the number in of gene to view results from a specific species).



- Explore your results. Did you find anything interesting?

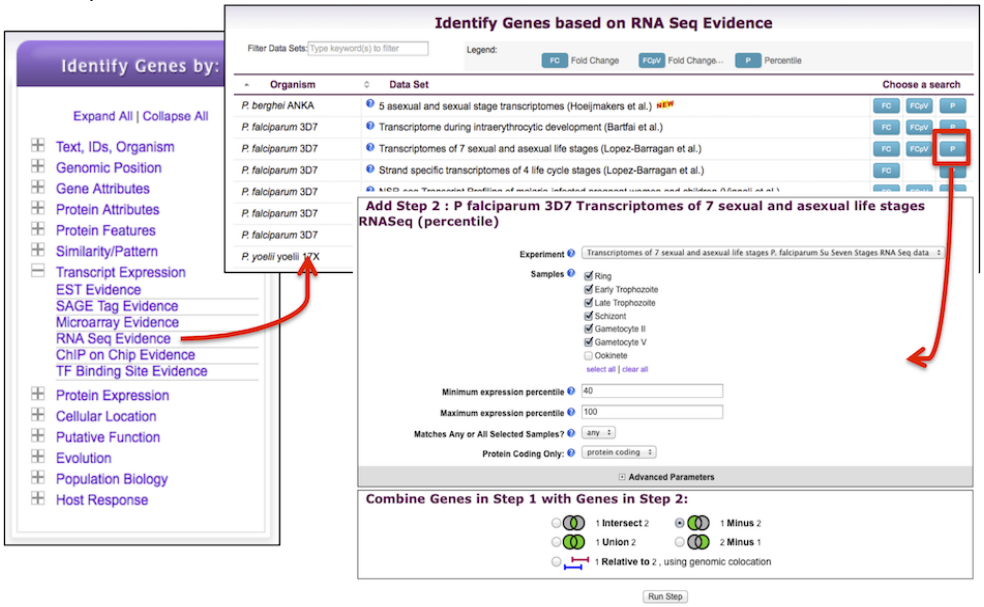
3. Finding genes based on RNAseq evidence and inferring function of hypothetical genes. Note: Use http://plasmodb.org for this exercise.

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.

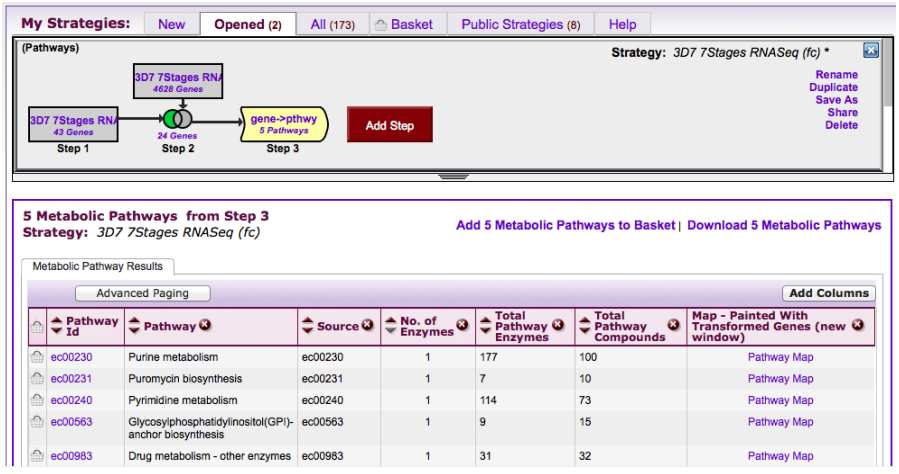


b. The above search will give you all genes that are up-regulated by 50 fold in ookinetes compared to the other stages. However, this does not mean that these genes are not expressed well in the other stages. How can you remove genes from the list that are likely not expressed in the other stages?

- 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).

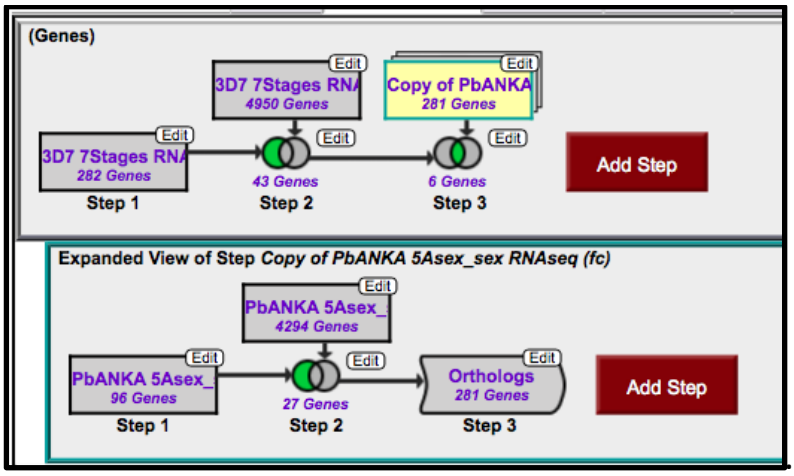


c. Which metabolic pathways are represented in this gene list? (*Hint:* add a step and transform results to metabolic pathways).



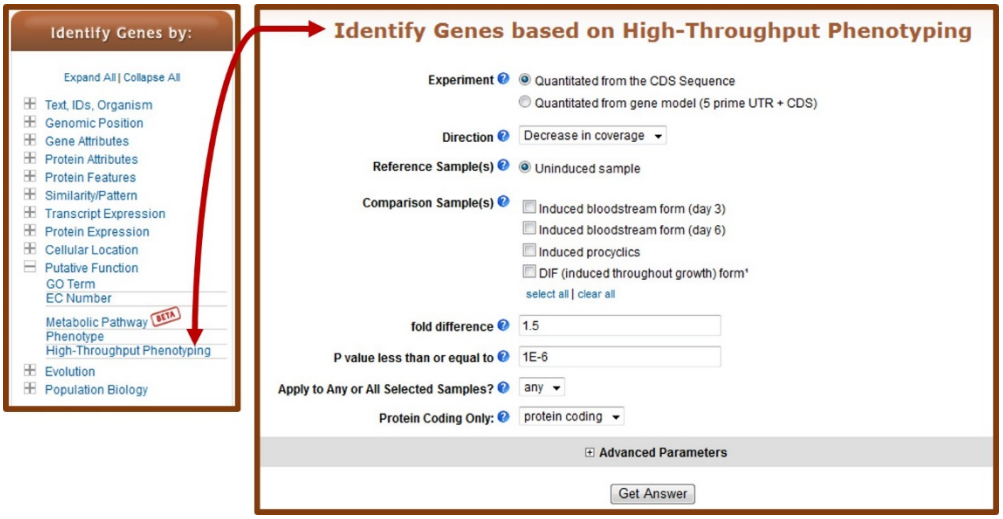
d. What happens is you revise the first step and modify the fold difference to a lower value - 10 for example?

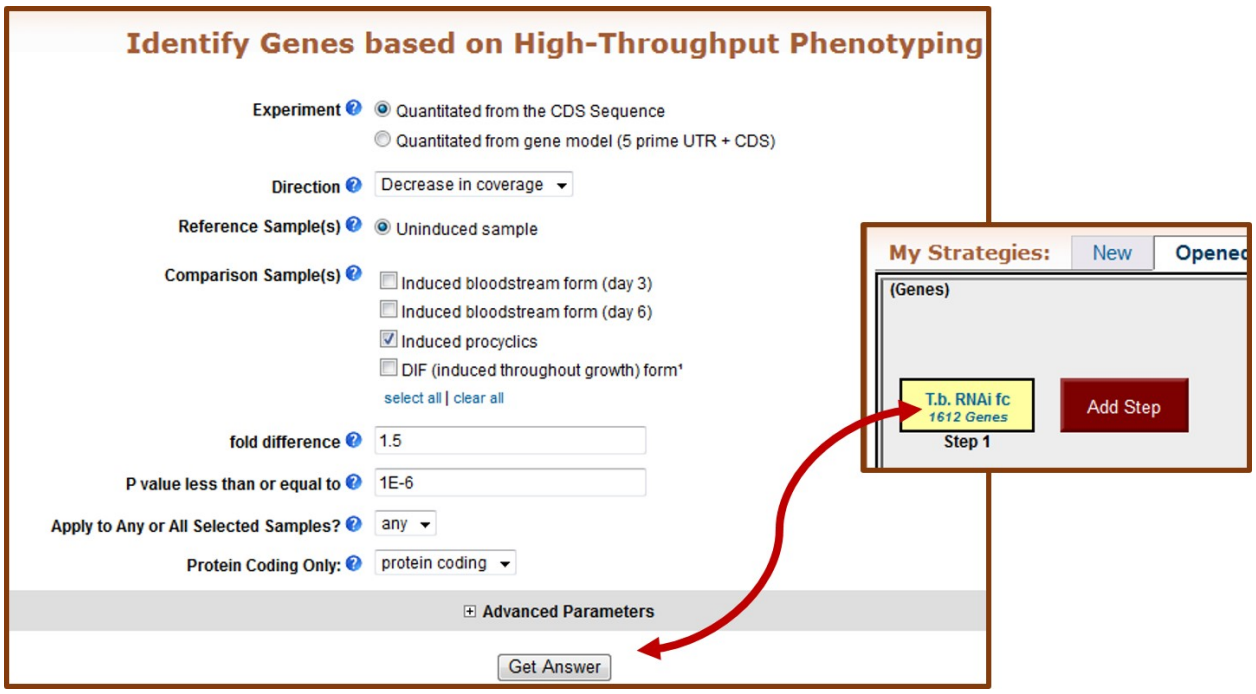
e. PlasmoDB also has an experiment examining gene expression during sexual development in *Plasmodium berghei* (rodent malaria). Can you determine if there are genes that are up-regulated in both human and rodent ookinetes (compared to all other stages)? *Hint:* start by deleting the last step you added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

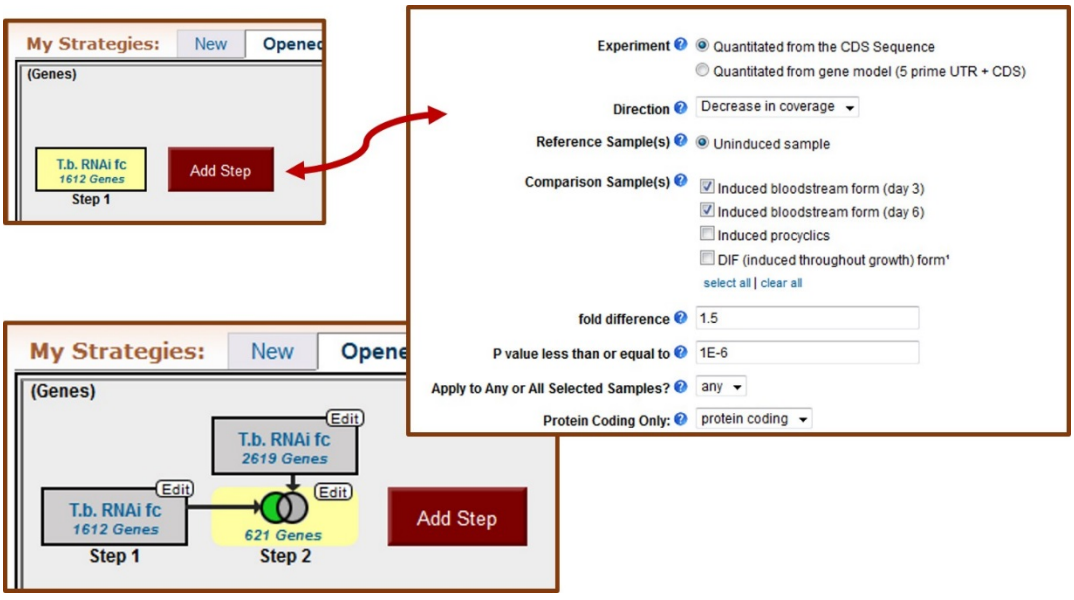
- Find the query for High Throughput Phenotyping. Think about how to set up this query (*Hint*: you will have to set up a two-step strategy). Remember you can play around with the parameters but there is no one correct way of setting them up – try the default parameters first and select the "induced procyclics" as the comparison sample.





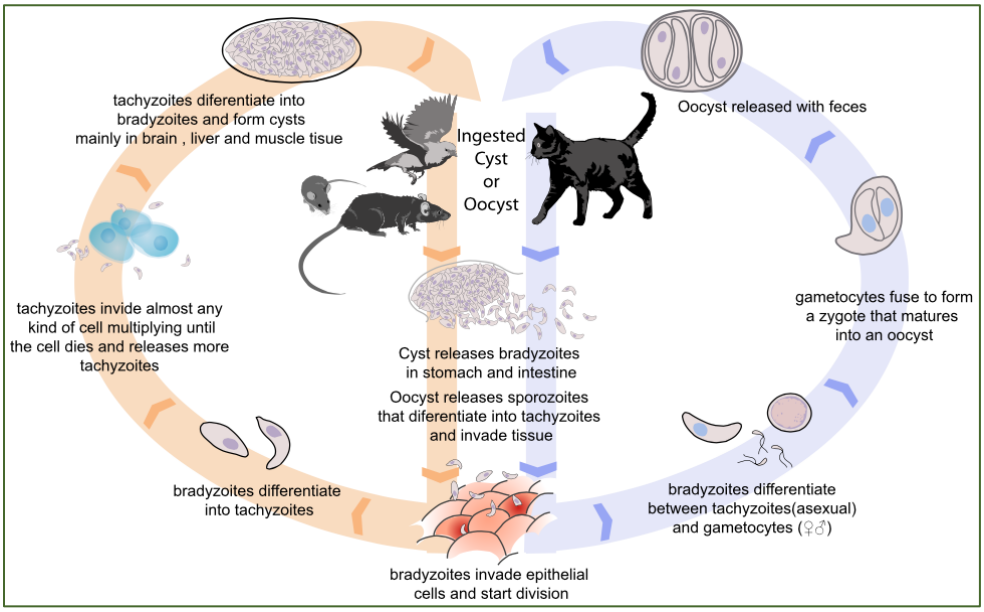
- Next add a step and run the same search except this time select the "induced bloodstream form" samples.

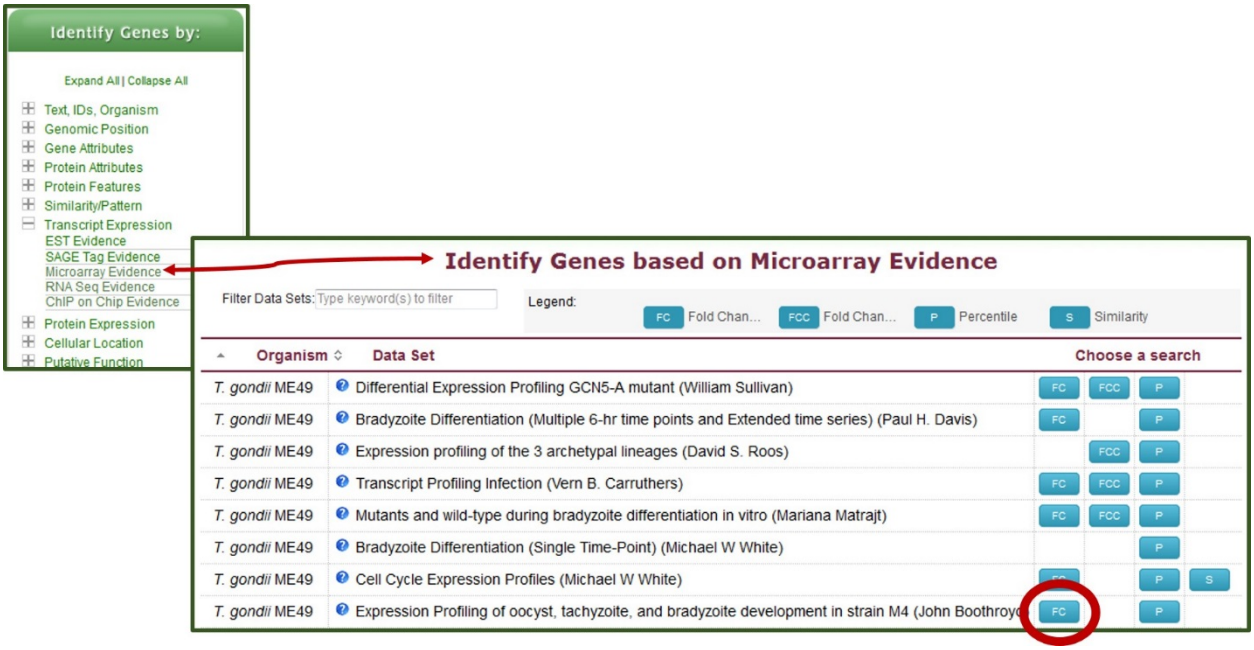
- How did you combine the results? Remember you want to find genes that are essential in procyclics and not in blood form.



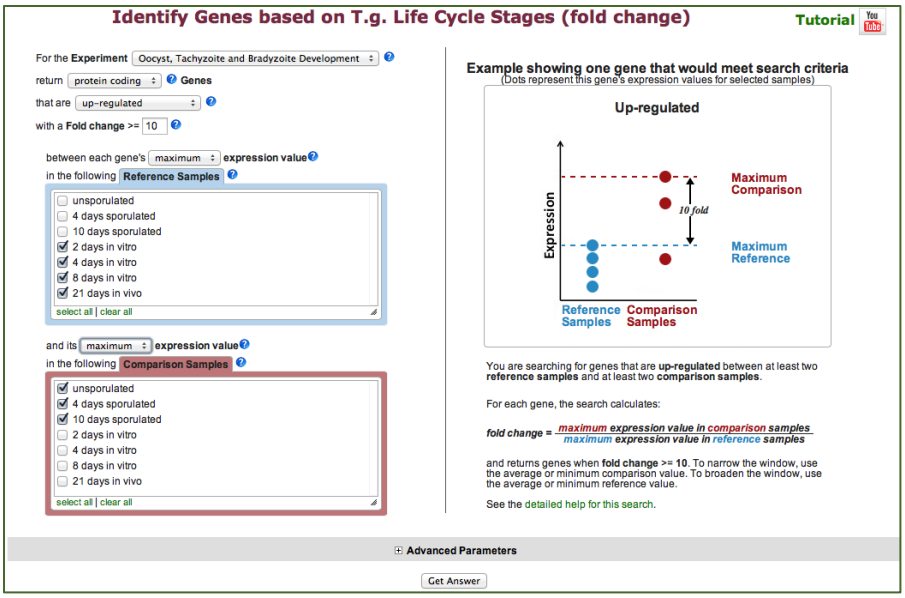
5. Finding oocyst expressed genes in *T. gondii* based on microarray evidence.
Note: For this exercise use http://toxodb.org

a. Find genes that are expressed at 10 fold higher levels in one of the oocyst stages than in any other stage in the "Expression Profiling of oocyst, tachyzoite, and bradyzoite development in strain M4 (John Boothroyd)" microarray experiment. In this experiment,



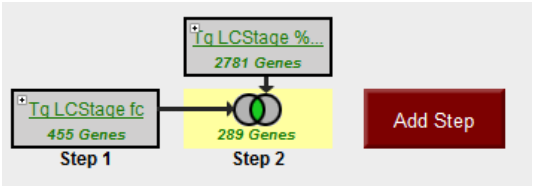


In this example the maximum expression value between genes in the reference and comparison groups was used to determine the fold difference.



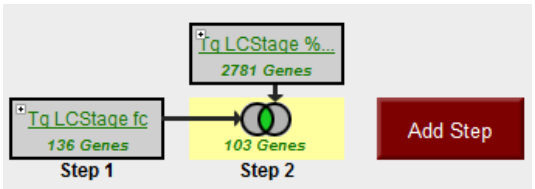
b. Add a step to limit this set of genes to only those for which all the non-oocyst stages are expressed below 50th percentile … ie likely not expressed at those stages. *(Hint*: after you click on add step find the same experiment under microarray expression and chose the percentile search).

- 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***".
- 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.

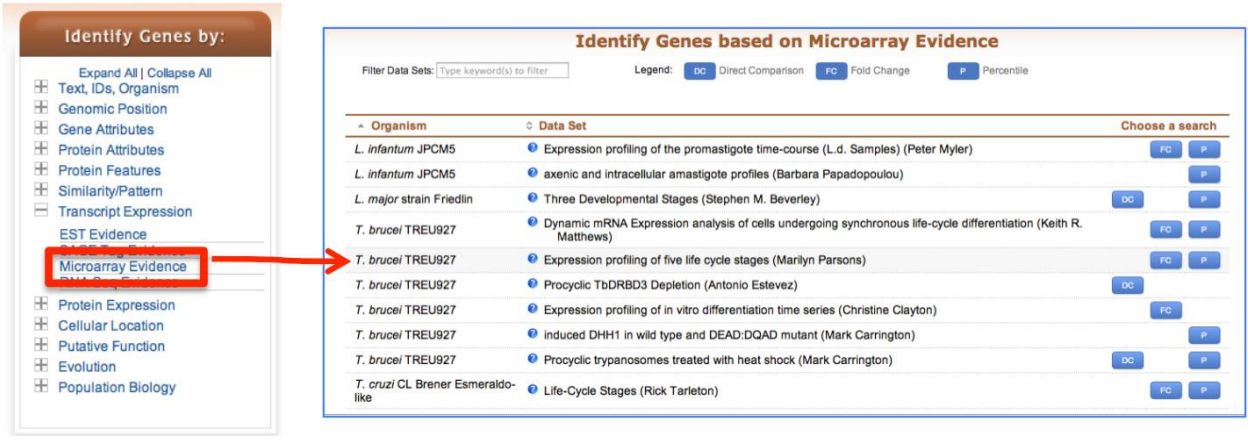
d. Revise the first step of this strategy to find genes that are 3 fold higher in day 4 oocysts than any other life cycle stage in this experiment.

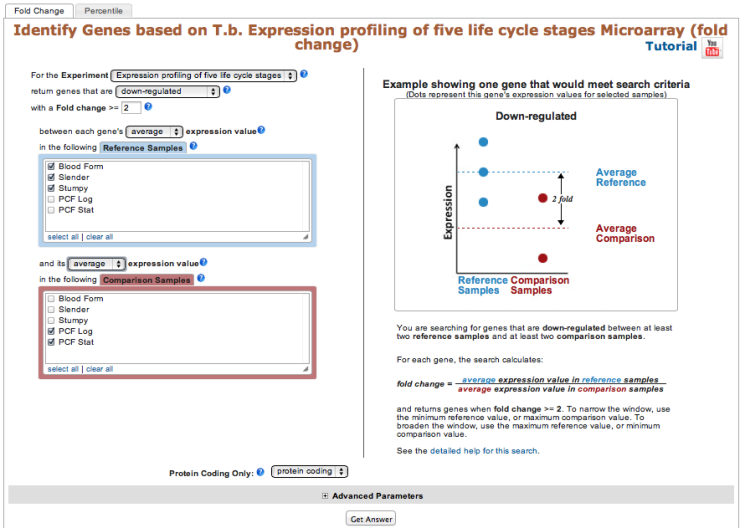
- Do all these genes have day 4 oocysts as the global maximum time point?
- Note that we still have the step to limit the percentile of non-oocyst samples to <= 50th percentile. What happens if you revise this step to also include the unsporulated and day 10 oocyst samples in this percentile range? Do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

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

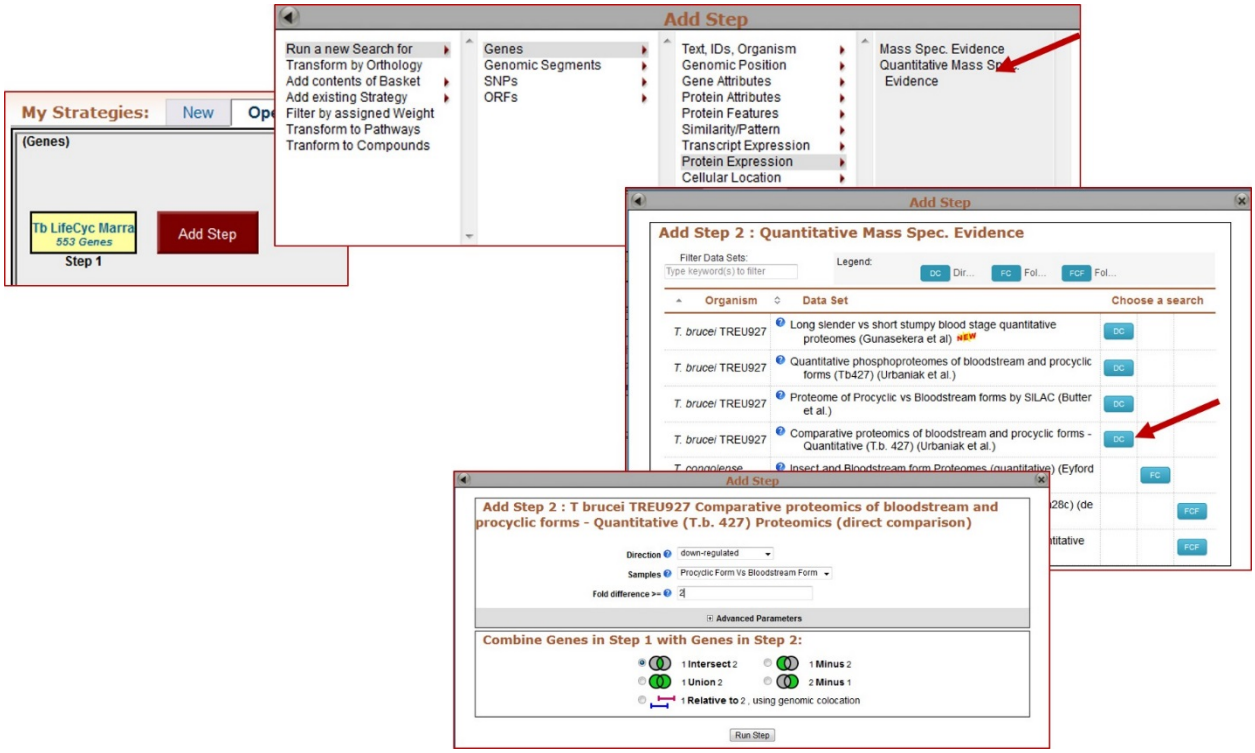
In this exercise we will compare the list of genes that show differential RNA abundance levels between procyclic and blood form stages in *T. brucei* with the list of genes that show differential protein abundance in these same stages.

a. Find genes that are down-regulated 2-fold in procyclyc 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.





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).