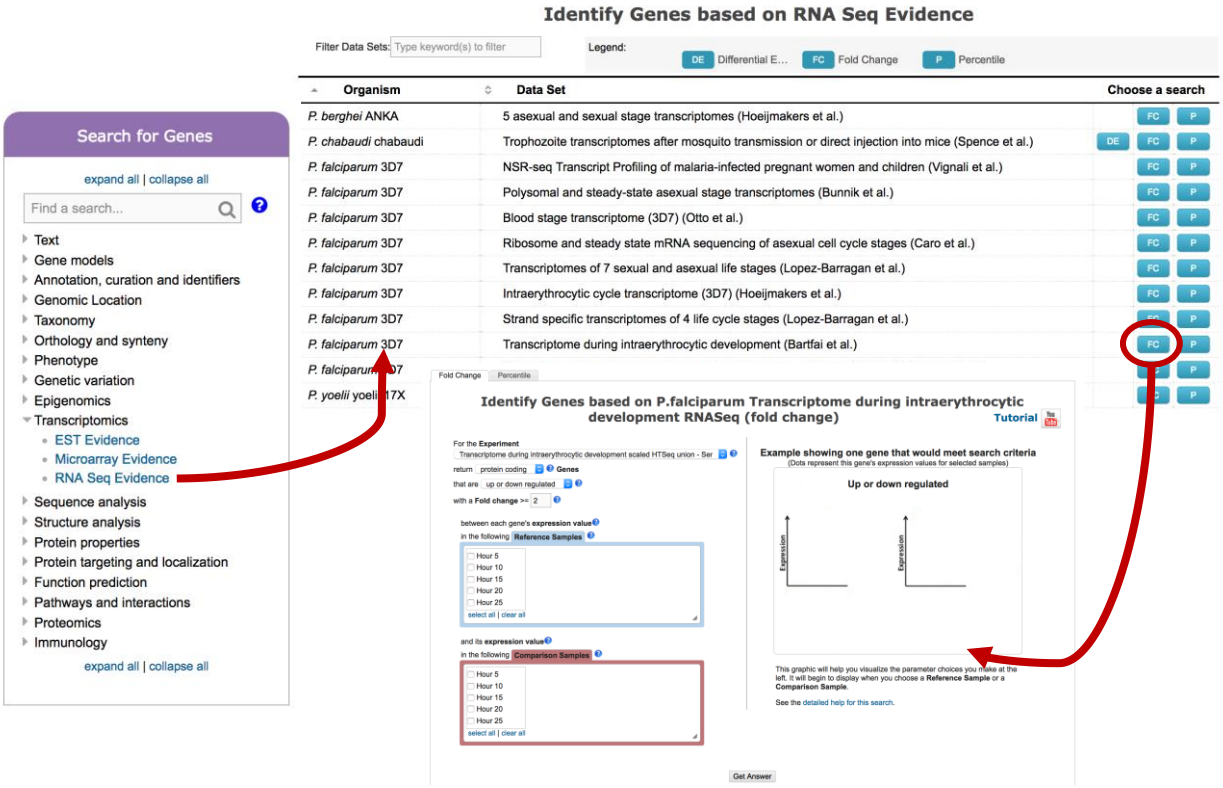
Exploring Transcriptomic data

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.

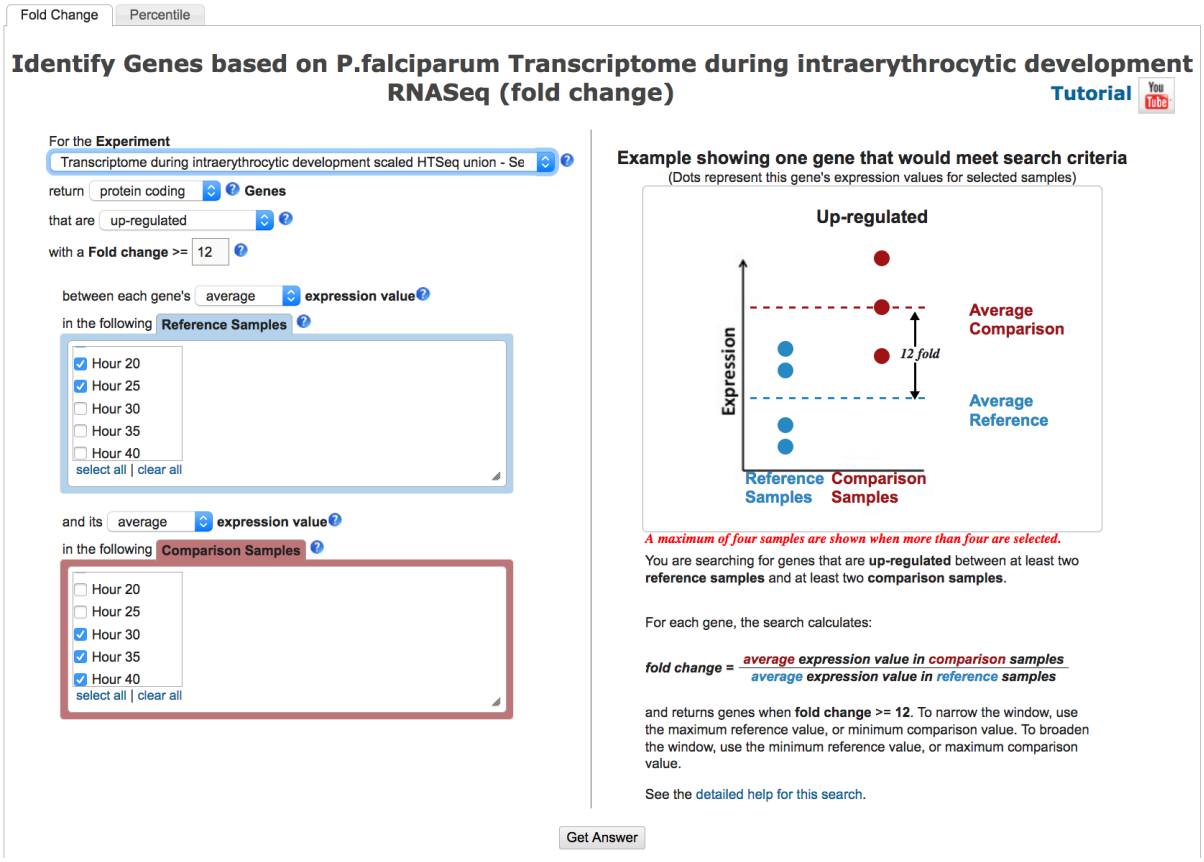
- **Direction**: the direction of change in expression. Choose up-regulated.

- **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 (upregulated ratio = expression in comparison)/(expression in reference). When you choose multiple samples to serve as reference, we generate one number for the fold change calculation by using the minimum, maximum, or average. Choose average

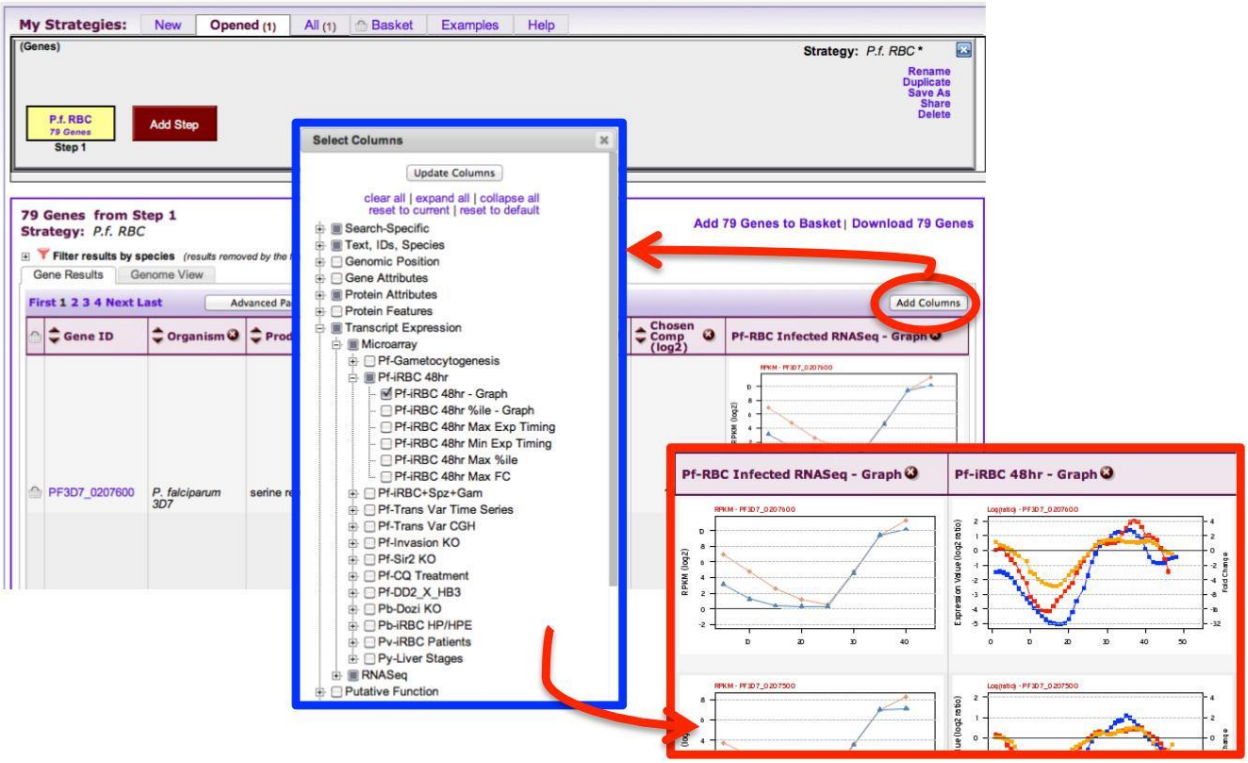
- **Reference Sample**: the samples that will serve as the reference when comparing expression between samples. choose 5, 10, 15, 20, 25
- **And its AVERAGE expression value:** This is the operation applied to comparison samples. see explanation above. Choose average
- **Comparison Sample**: the sample that you are comparing to the reference. In this case you are interested in genes that are up-regulated in later time points choose 30, 35, 40



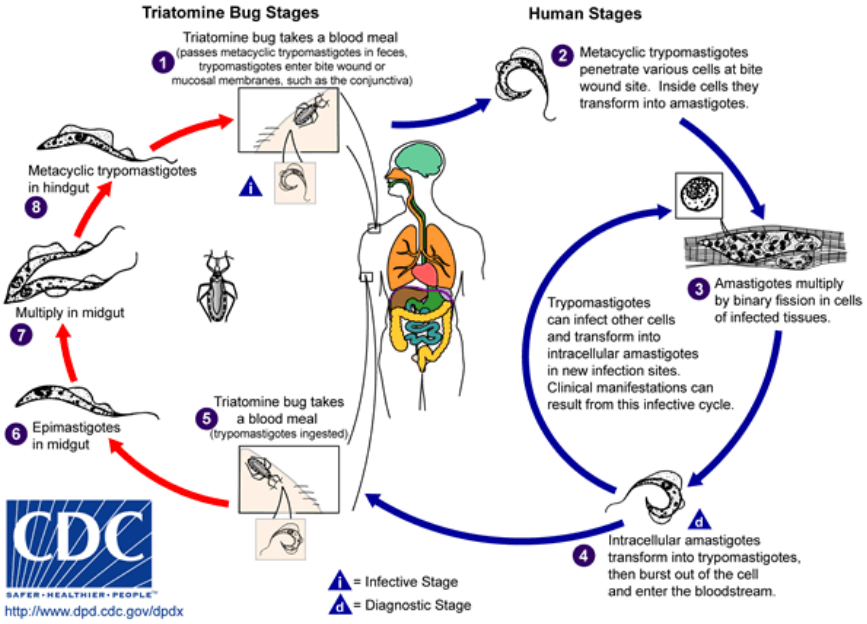
b. For the genes returned by the search, how does the RNA-sequence data compare to microarray data?

- 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".

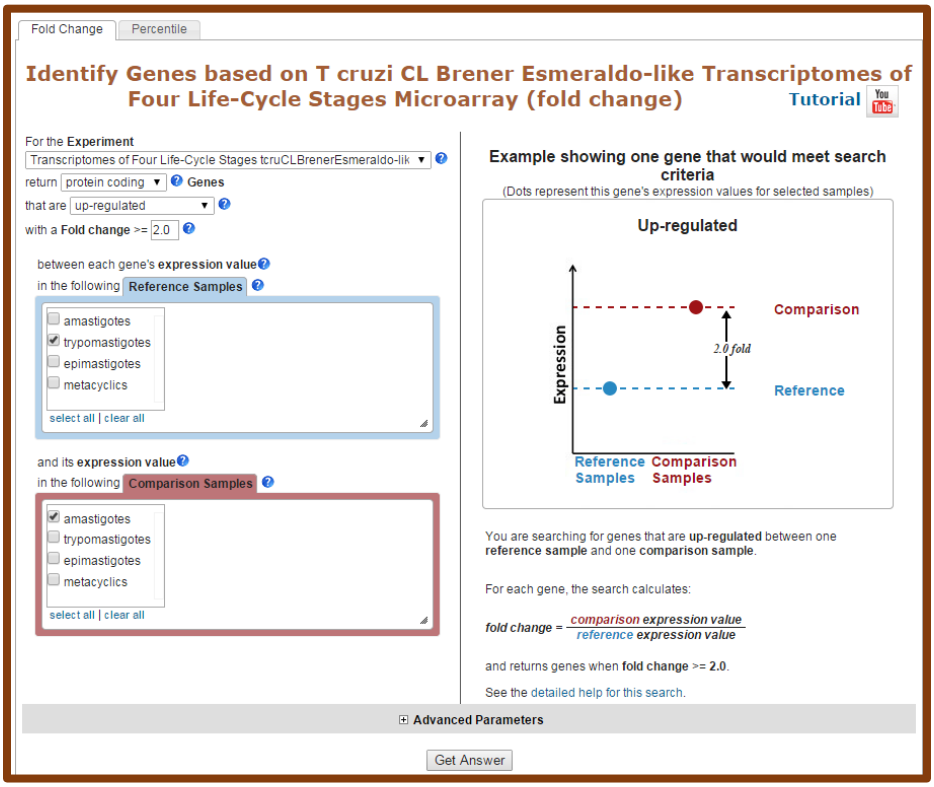
OPTIONAL: You can also run a fold change search using this experiment to compare results on a genome scale. Add a step to your strategy and intersect the results of a fold change search using the "Erythrocytic expression time series (3D7, Dd2, HB3) (Bozdech et al. and Linas et al.)" experiment (under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

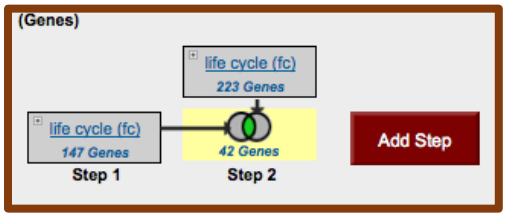


a. Find *T. cruzi* protein coding genes that are upregulated in amastigotes compared to trypomastigotes. Go to the transcript expression section then select microarray. Choose

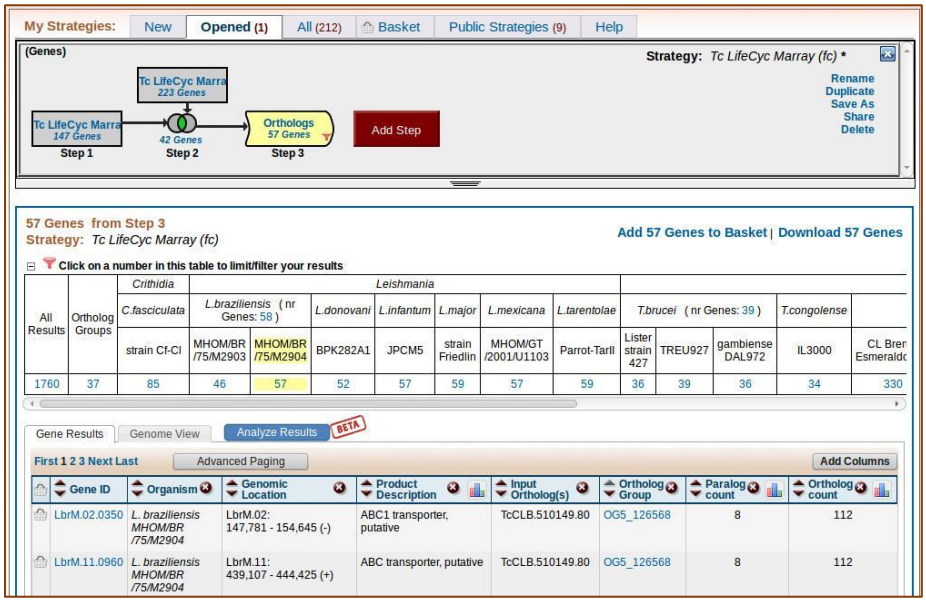


the fold change (FC) search for the data set called: **Transcriptomes of Four Life-Cycle Stages (Minning et al.)**.

- 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).

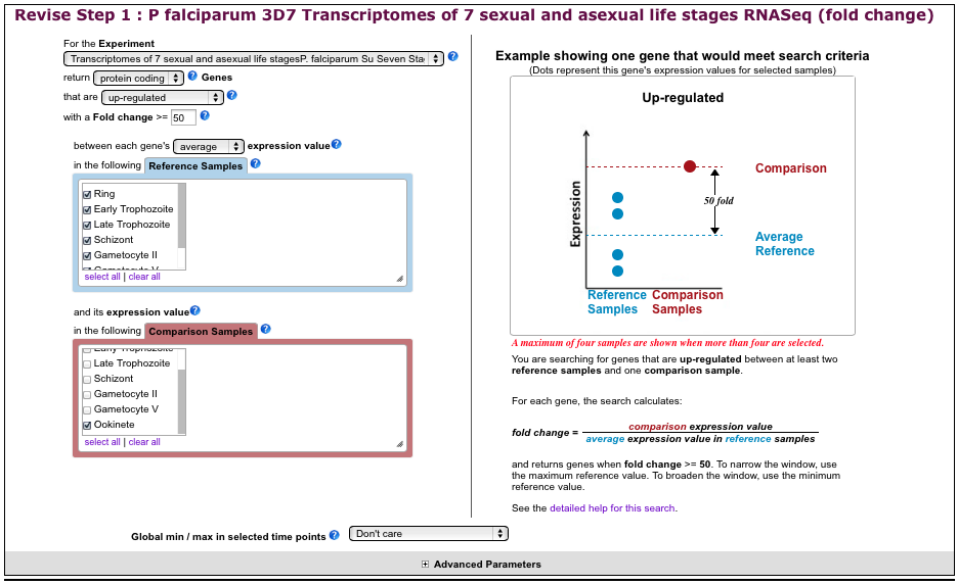


- Do these genes have orthologs in other kinetoplastids? (*Hint*: add a step and run an ortholog transform on your results).
- How many orthologs exist in *L. braziliensis*? (*Hint*: look at the filter table between the strategy panel and your result list. Click on the number in the table under a species to view results from a specific species). Explore your results. Scan the product descriptions for this list of genes. Did you find anything interesting? Perhaps a GO enrichment analysis would support your ideas.



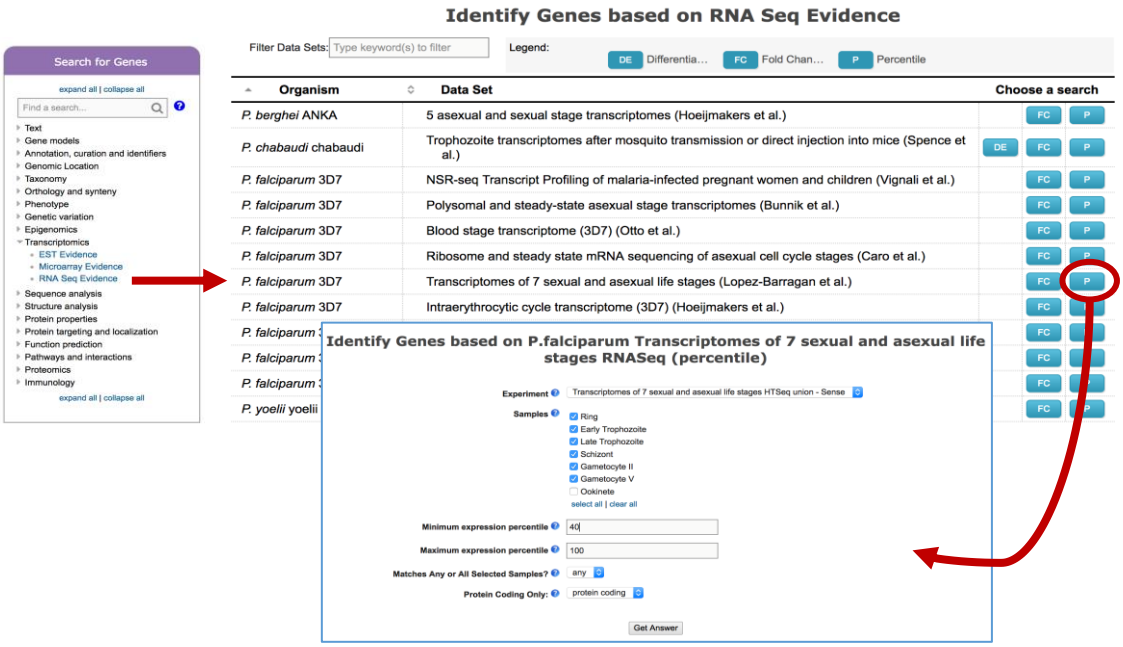
3. Finding genes based on RNAseq evidence and inferring function of hypothetical genes.
Note: Use http://plasmodb.org for this exercise.

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.

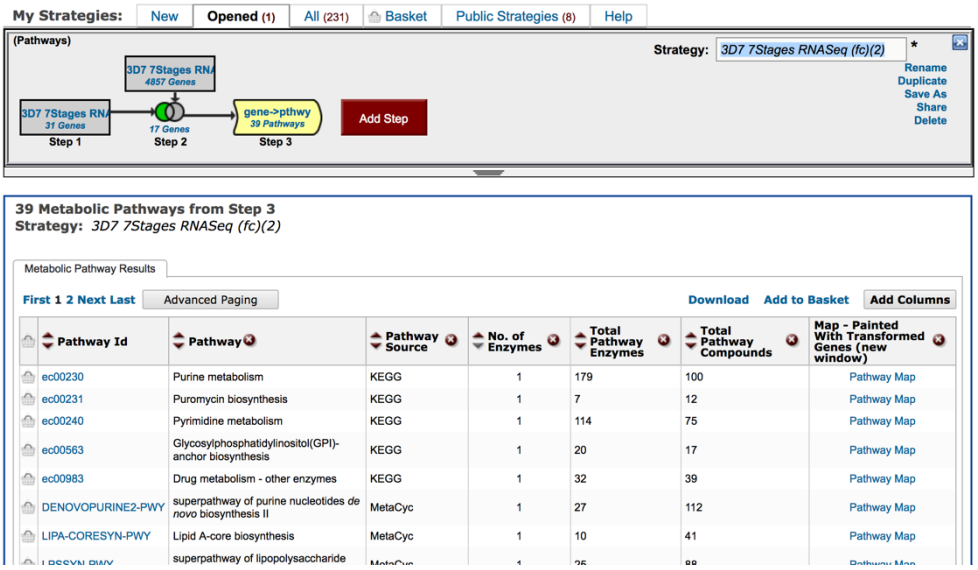


b. The above search will give you all genes that are up-regulated by 50 fold in ookinetes compared to the average expression level of other stages. Despite the high fold change, some genes in the list may be highly expressed in the other stages. How can you remove genes from the list that are highly 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%? How does setting the any / all samples impact the result …. Which would be better in this case?*
- *Hint II: Try changing the operator from average to maximum for the set of non-ookinete stages in your initial fold change search. What does this do? How do the resulting genes compare with the two step strategy you generated in the first hint? Which hint do you think works better?*



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? Compare results when you also modify the "between each genes" parameter. What happens if you set this to maximum? Which value do you think is most stringent for ensuring at 10 fold up regulation compared to the other samples?

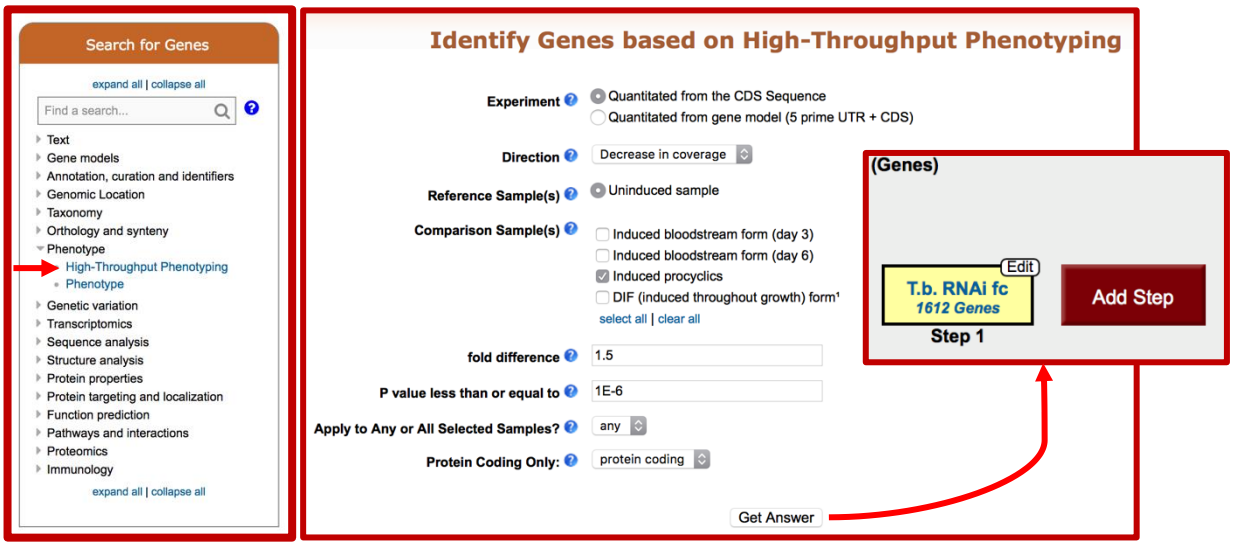
e. PlasmoDB also has an experiment examining gene expression during sexual development in *Plasmodium berghei* (rodent malaria). Can you determine if there are genes that are up-

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.

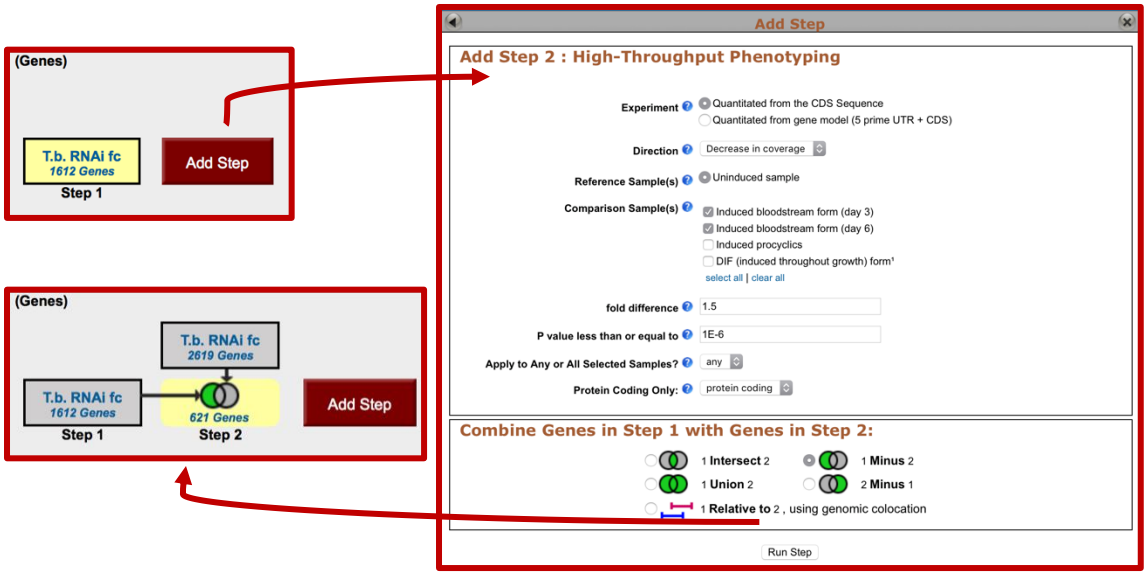
4. Find genes that are essential in procyclics but not in blood form *T. brucei*.
Note: for this exercise use http://TriTrypDB.org.

- 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 –

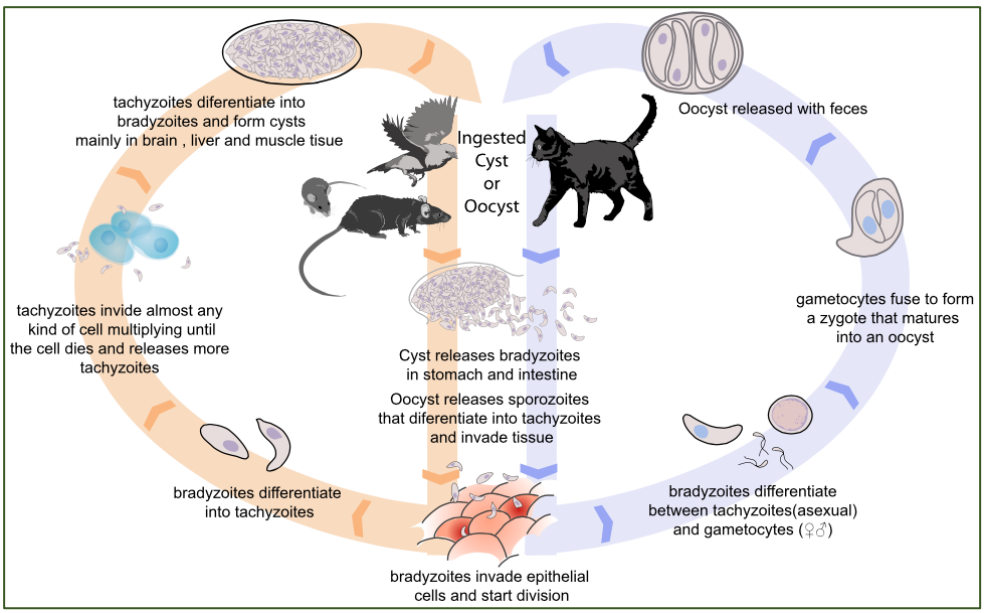


- Next add a step and run the same search except this time select the "induced bloodstream form" samples.

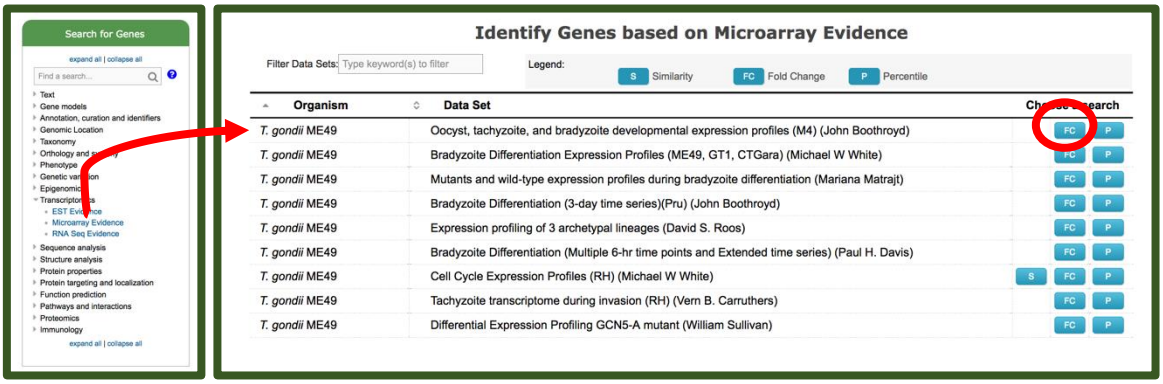
- How did you combine the results? Remember you want to find genes that are essential in procyclics and not in blood form.

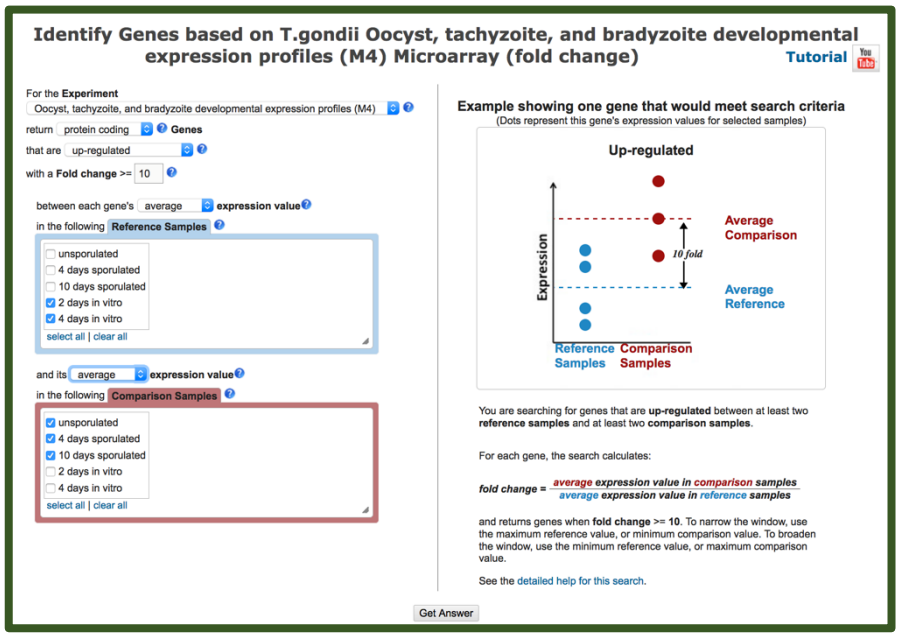


5. Finding oocyst expressed genes in *T. gondii* based on microarray evidence.
Note: For this exercise use http://toxodb.org



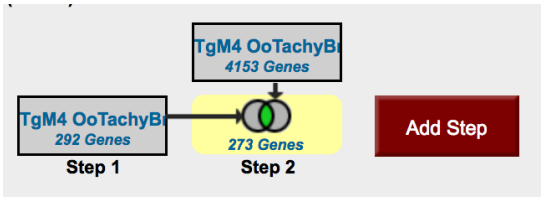
a. Find genes that are expressed at 10 fold higher levels in one of the oocyst stages than in any other stage in the "**Oocyst, tachyzoite, and bradyzoite developmental expression profiles (M4) (John Boothroyd)**" microarray experiment. In this example, the maximum expression value between genes in the 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).

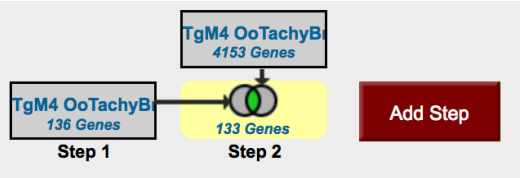
- Select the 4 non-oocyst samples.
- We want all to have less than 50th percentile so set ***minimum percentile*** to 0 and ***maximum percentile*** to 50.
- Since we want all of them to be in this range, choose ALL in the "***Matches Any or All Selected Samples***".



- 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.

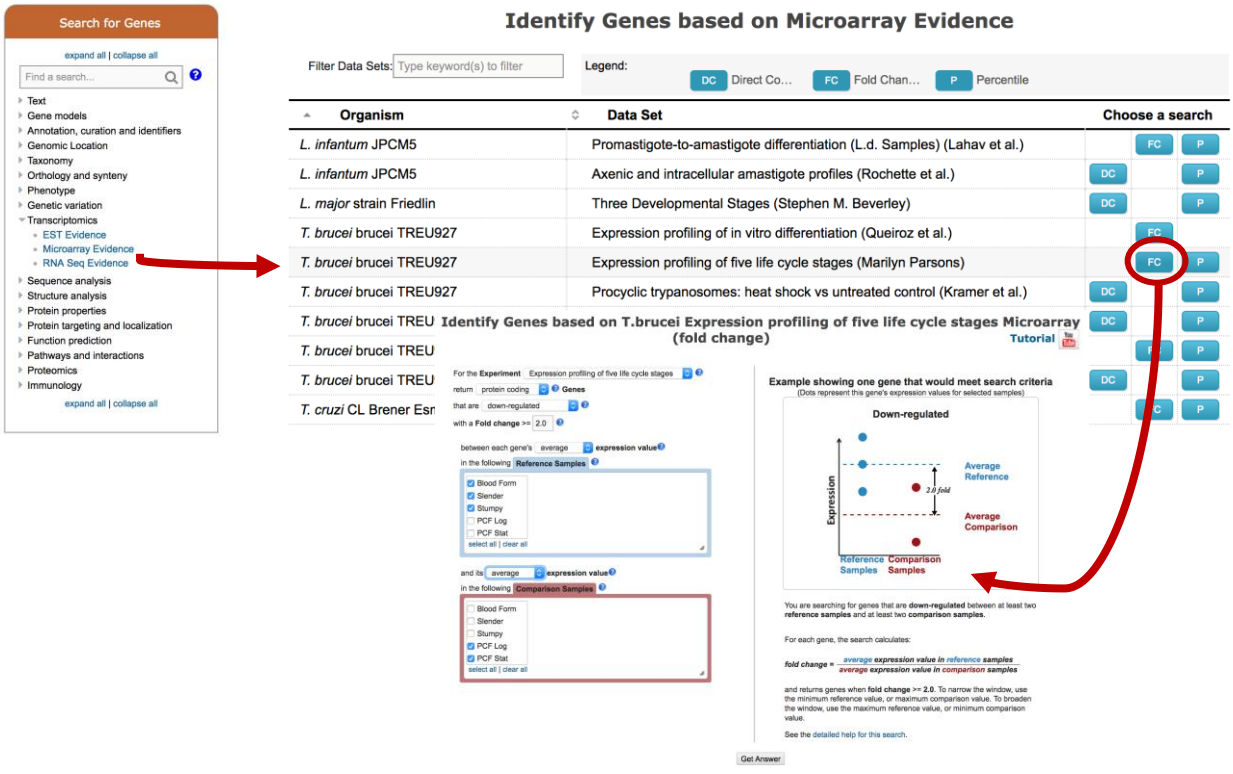
- Does this result look cleaner/more convincing? Why?
- Would you consider these genes to be oocyst specific?



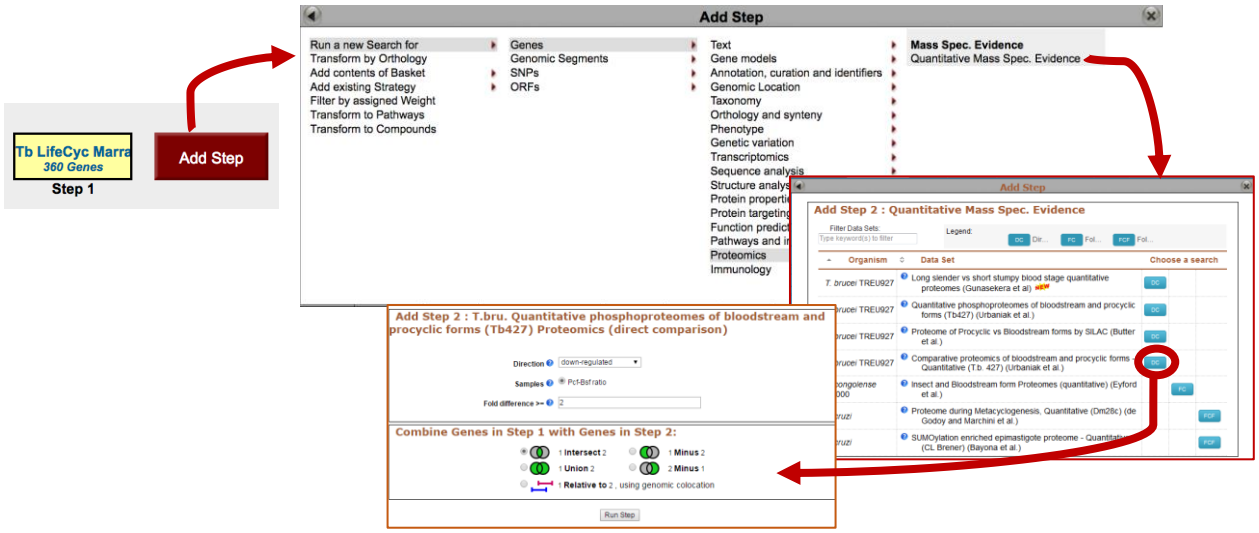
6. Comparing RNA abundance and Protein abundance data.
Note: for this exercise use http://TriTrypDB.org.

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?

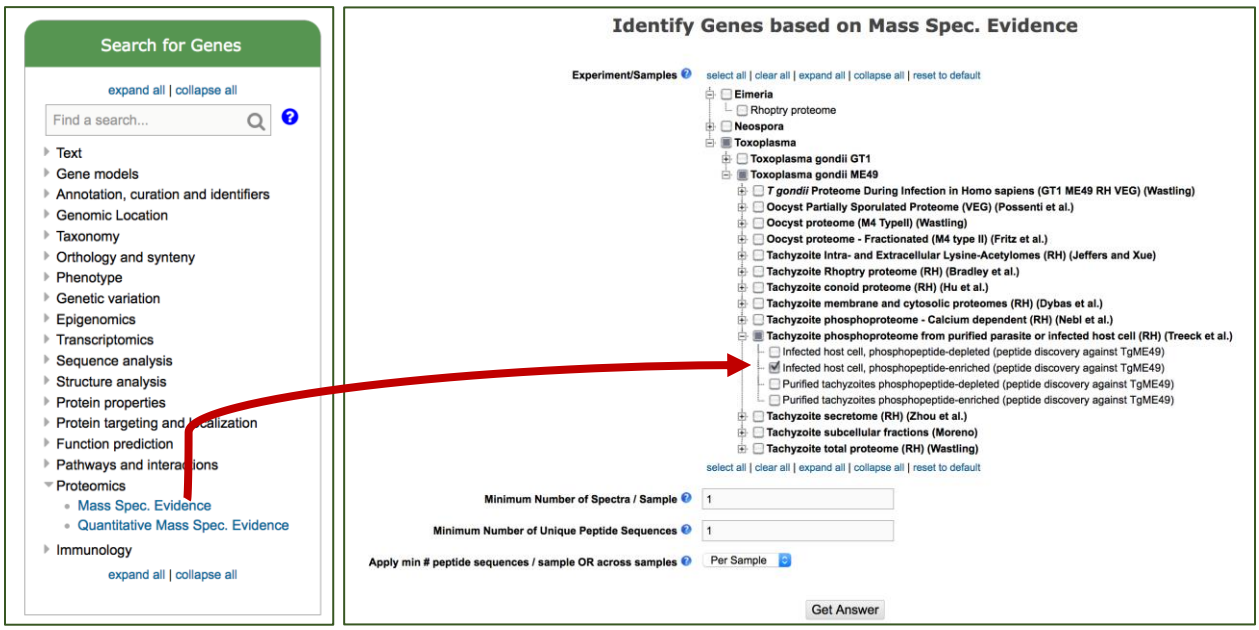
e. Can you think of ways to provide more confidence (or cast a broader net) in the microarray step? (*Hint:* you could insert steps to restrict based on percentile or add a RNA Sequencing step that has the same samples).

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

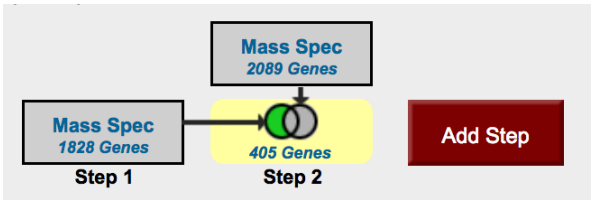
For this exercise use http://www.toxodb.org

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

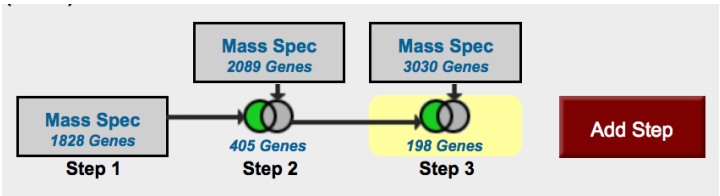
7a. Find all genes with evidence of phosphorylation in intracellular tachyzoites. Select the "**Infected host cell, phosphopeptide-enriched** (peptide discovery against **TgME49**)" sample under the experiment called "**Tachyzoite phosphoproteome from purified parasite or infected host cell (RH) (Treeck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

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

33 Genes from Step 4
Strategy: *Mass Spec*

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

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

Filter by strains (advanced)

Gene Results | Genome View | Analyze Results

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

Gene ID	Transcript ID	Gene Group (representative gene)	Genomic Location (Gene)	Product Description	# Transcripts
TGME49_208830	TGME49_208830-t26_1	TGME49_208830-t26_1	TGME49_chrIb:888,008..891,283(-)	hypothetical protein	1
TGME49_321640	TGME49_321640-t26_1	TGME49_321640-t26_1	TGME49_chrIb:1,665,247..1,675,489(-)	cell division protein CDC48AP	1
TGME49_223140	TGME49_223140-t26_1	TGME49_223140-t26_1	TGME49_chrI:1,469,476..1,475,491(+)	tRNA binding domain-containing protein	1
TGME49_252360	TGME49_252360-t26_1	TGME49_252360-t26_1	TGME49_chrIII:512,377..515,416(+)	rhoptry kinase family protein ROP24 (incomplete catalytic triad)	1
TGME49_288370	TGME49_288370-t26_1	TGME49_288370-t26_1	TGME49_chrIX:2,478,472..2,482,708(-)	hypothetical protein	1

7f. Pick one or two of the hypothetical genes in your results and visit their gene pages. Can you infer anything about their function? Hint: explore the protein and expression sections.

7g. What about polymorphism data? Go back to your strategy and add columns for SNP data found under the population biology section. Explore the gene page for the gene that has the most

Gene Results | Genome View | Analyze Results

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

Gene ID	Transcript ID	Product Description	Total SNPs All Strains	Non-Coding SNPs All Strains	NonSyn/Syn SNP Ratio All Strains	NonSynonymous SNPs All Strains	SNPs with Stop Codons All Strains	Synonymous SNPs All Strains
TGME49_224280	TGME49_224280-t26_1	CPSF A subunit region protein	922	635	0.98	142	0	145
TGME49_202490	TGME49_202490-t26_1	AP2 domain transcription factor AP2VIIa-7	593	328	0.95	129	0	136
TGME49_311080	TGME49_311080-t26_1	transporter, cation channel family protein	551	360	1.51	115	0	76
TGME49_321640	TGME49_321640-t26_1	cell division protein CDC48AP	548	446	1.04	52	0	50
TGME49_205120	TGME49_205120-t26_1	hypothetical protein	447	185	2.2	180	0	82
TGME49_313280	TGME49_313280-t26_1	WD domain, G-beta repeat-containing protein	443	367	1.81	49	0	27
TGME49_286120	TGME49_286120-t26_1	prolyl endopeptidase	427	366	0.79	27	0	34
TGME49_219640	TGME49_219640-t26_1	hypothetical protein	382	263	2.5	85	0	34
TGME49_315700	TGME49_315700-t26_1	hypothetical protein	339	266	1.15	39	0	34
TGME49_239440	TGME49_239440-t26_1	protein kinase (incomplete catalytic triad)	333	215	1.13	62	1	55
TGME49_220350	TGME49_220350-t26_1	tRNA ligases class II (D, K and N) domain-containing protein	317	200	1.79	75	0	42
TGME49_257595	TGME49_257595-t26_1	hypothetical protein	317	131	2.32	130	0	56
TGME49_205625	TGME49_205625-t26_1	hypothetical protein	294	206	1.59	54	0	34
TGME49_288880	TGME49_288880-t26_1	hypothetical protein	231	158	3.29	56	0	17
TGME49_223140	TGME49_223140-t26_1	tRNA binding domain-containing protein	197	172	1.08	13	0	12
TGME49_216840	TGME49_216840-t26_1	hypothetical protein	189	89	1.17	54	0	46
TGME49_288370	TGME49_288370-t26_1	hypothetical protein	189	83	2.42	75	0	31
TGME49_308070	TGME49_308070-t26_1	hypothetical protein	188	123	1.95	43	0	22
TGME49_214080	TGME49_214080-t26_1	toxofilin	177	63	3.67	88	2	24
TGME49_314280	TGME49_314280-t26_1	AAR2 protein	163	113	1.78	32	0	18

number of non-synonymous SNPs. Hint: you can sort the columns by clicking on the up/down arrows next to the column names.