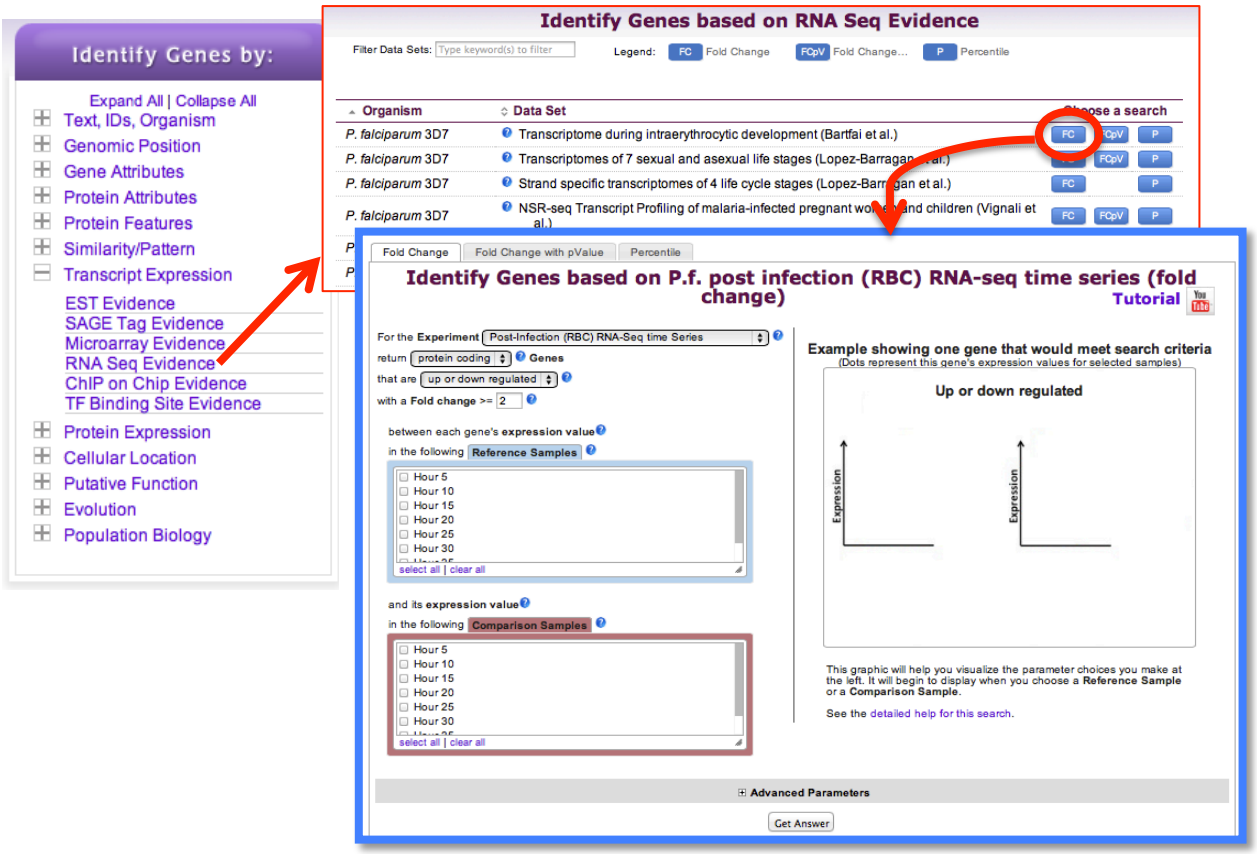
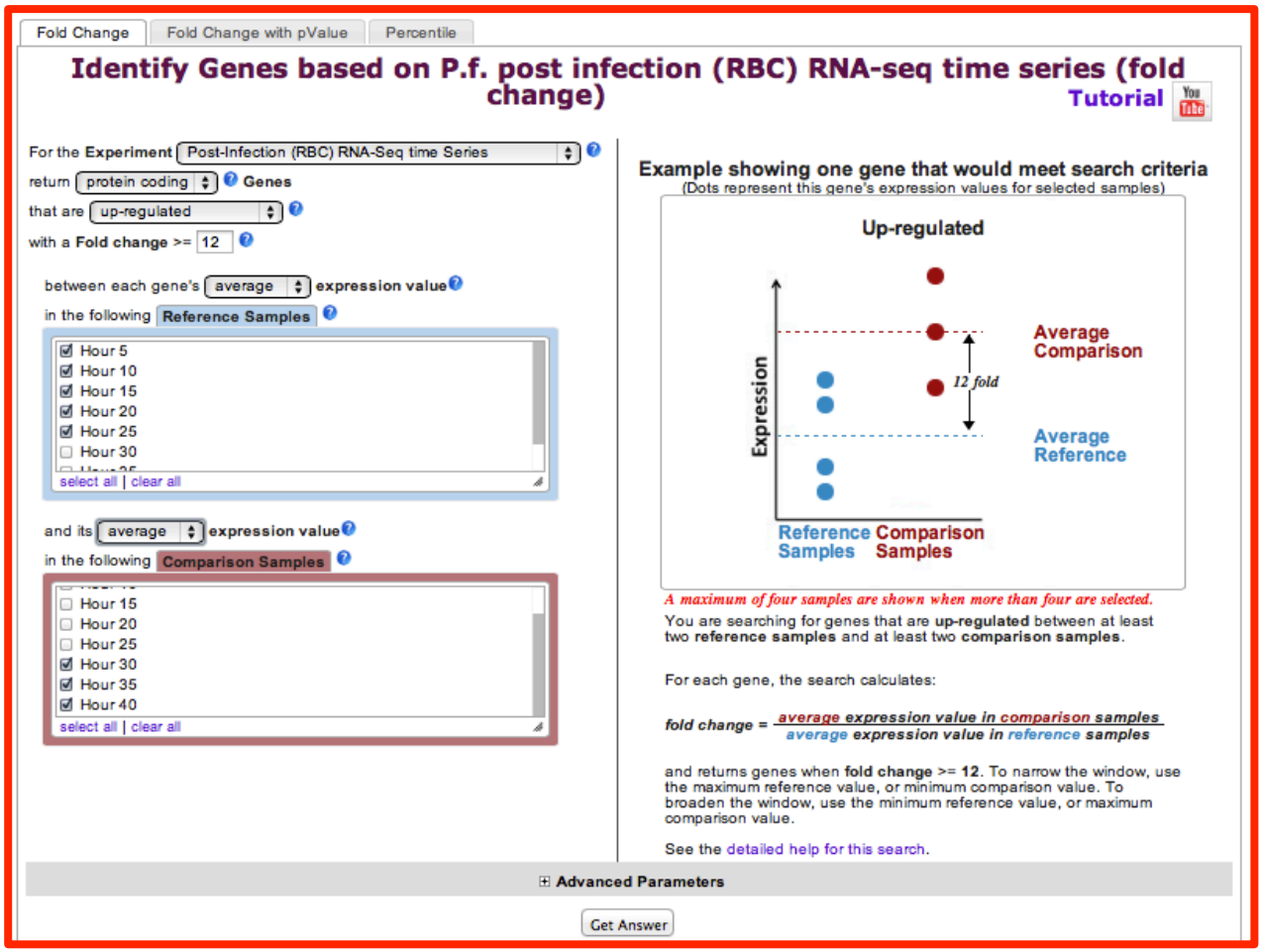
Functional Genomics in EuPathDB
Transcriptomics and Proteomics
Exercise 13

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 upregulated based on RNA-seq data at late time points (30, 35 and 40-hours) compared to early time points in this experiment (1, 10, 15, 20, 25 hrs). *Hint*: for this exercise use a fold change search based on the "Transcriptome during intraerythrocytic development (Bartfai *et al.*)" experiment.

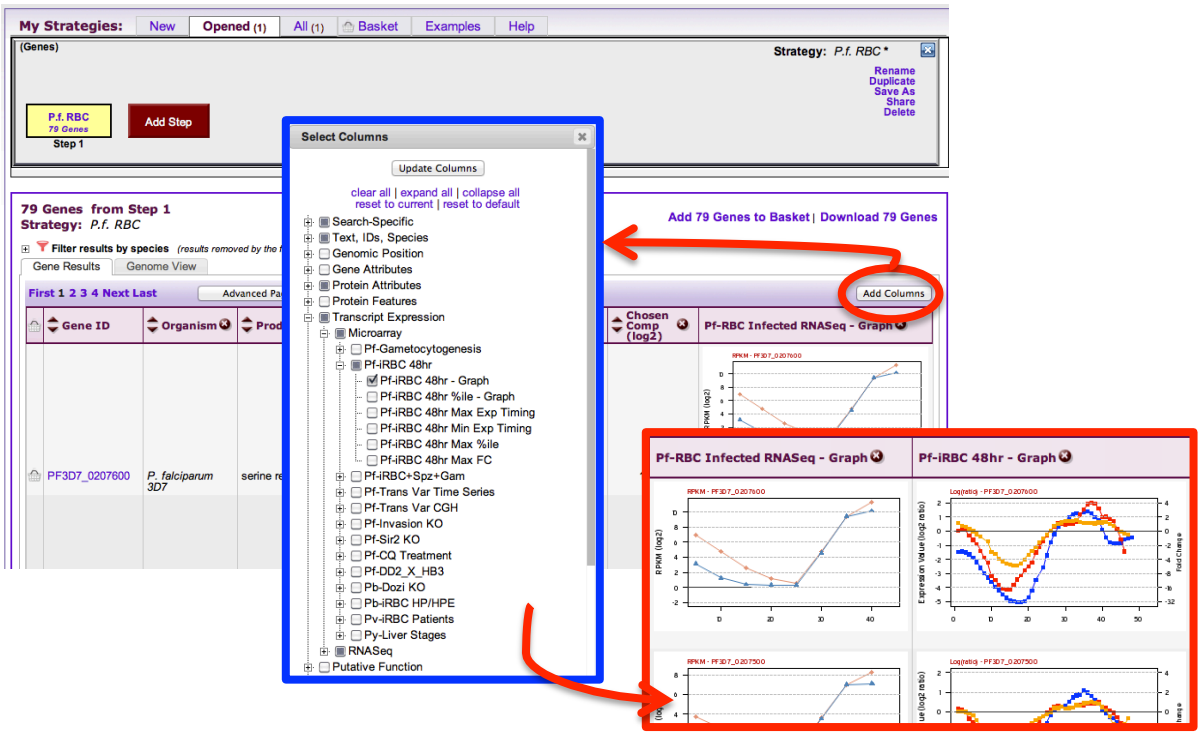


Hint: there are a number of parameters to manipulate in this search. As you modify parameters on the left side note the dynamic help on the right side:

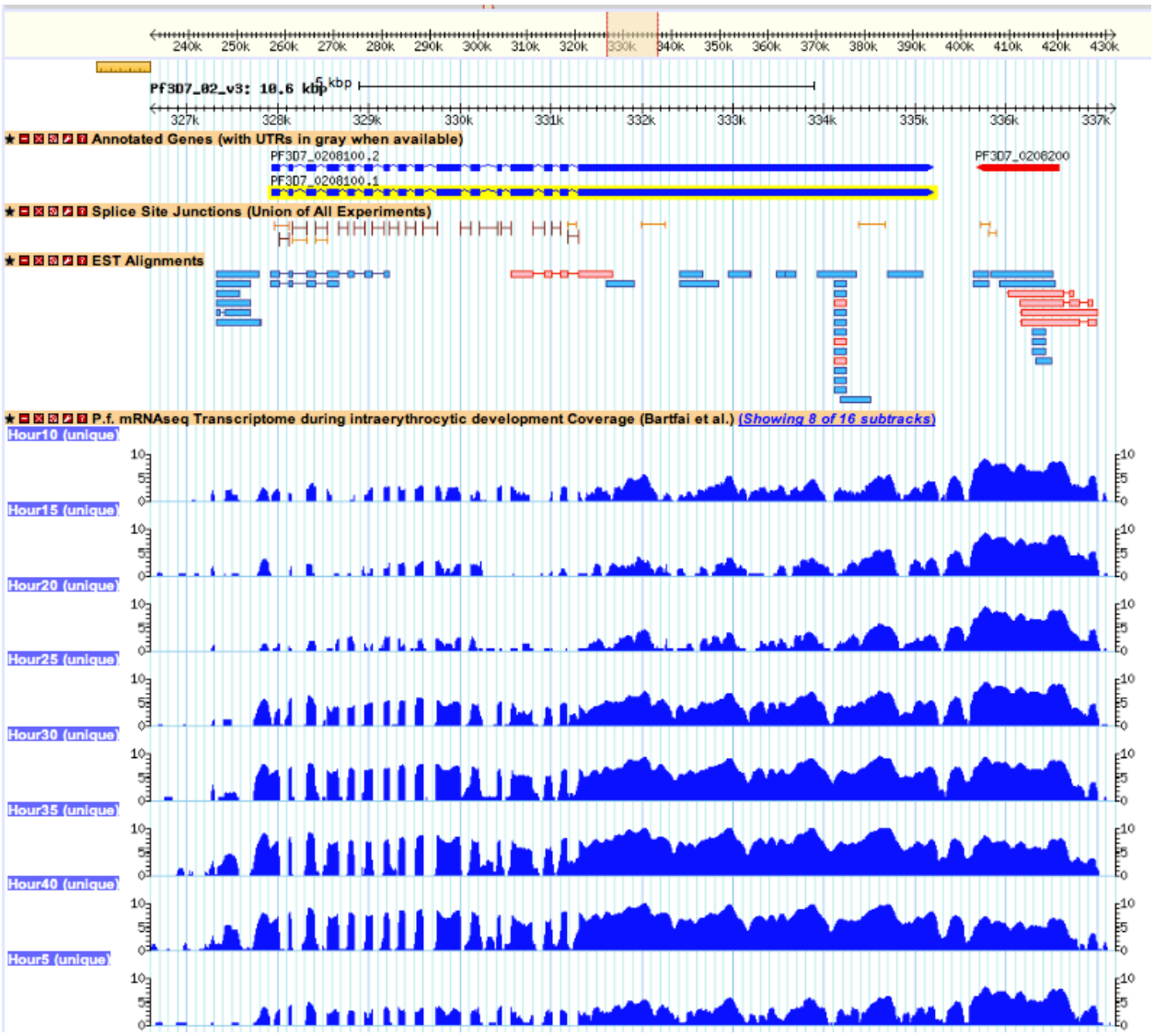


Direction: the direction of change in expression. **Choose up-regulated**.
Reference Sample: the samples that will serve as the reference when comparing expression between samples. **choose 5, 10, 15, 20, 25**
Operation Applied to Reference Samples: fold change is calculated as the ratio of two values (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**
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**
Operation Applied to Comparison Samples: see explanation above. **Choose average**
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.

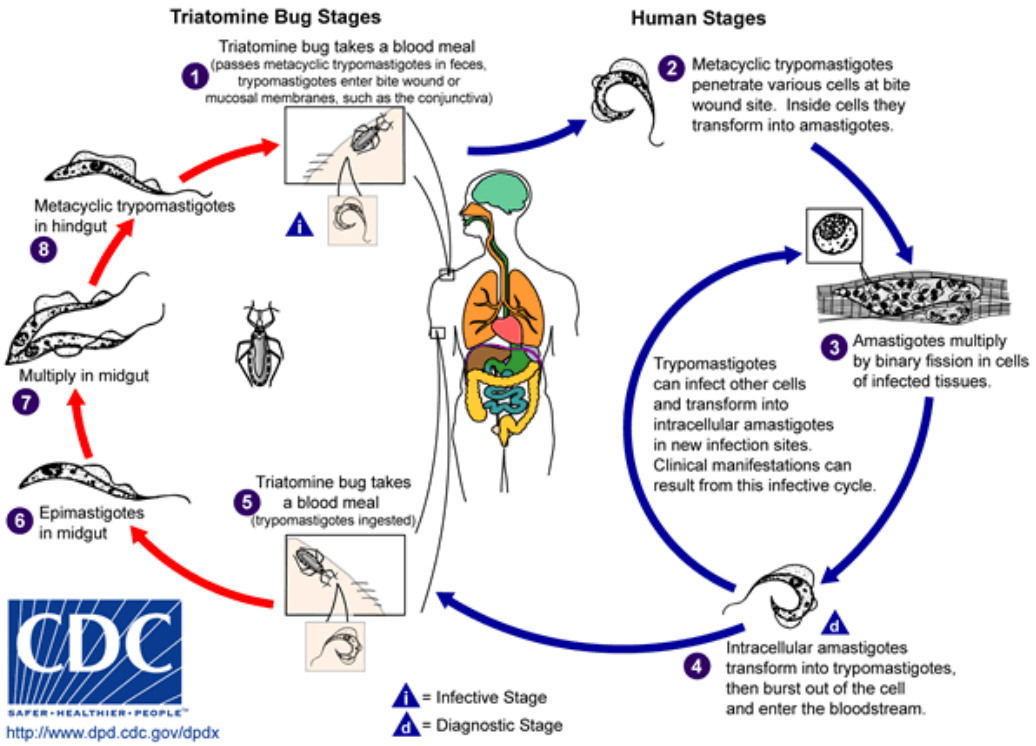
b. For the genes returned by the search, how does the RNA-sequence data compare to microarray data? (*Hint*: add the column called "Pf-iRBC 48hr - Graph" and compare the RNA-seq to the microarray graphs).



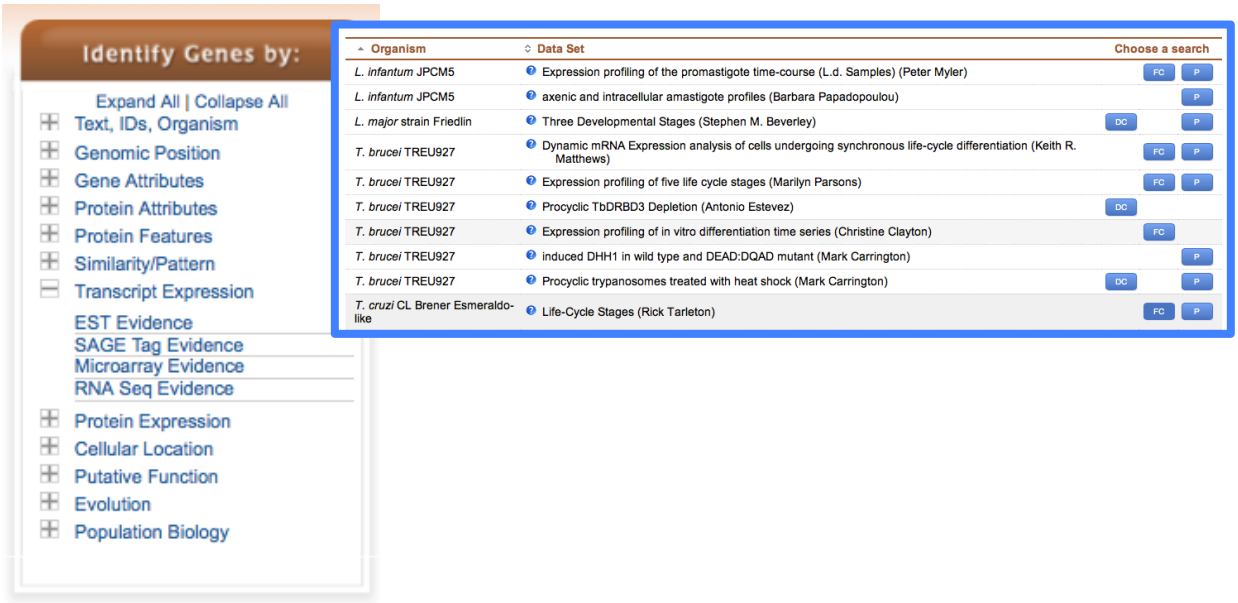
c. Which gene has 16 exons? (*Hint*: add a column for number of exons)
d. Is this gene alternatively spliced? Look at the gene page. Take note of the Gene ID.
e. View this gene in the genome browser and load the RNA-seq tracks for this experiment "P.f. mRNAseq Transcriptome during intraerythrocytic development Coverage (Bartfai *et al.*)". Do these tracks match the results you got above? (ie. is this gene differentially regulated between the early time points and the late ones?)
f. 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)").
g. 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).


240k 250k 260k 270k 280k 290k 300k 310k 320k 330k 340k 350k 360k 370k 380k 390k 400k 410k 420k 430k
PF3D7_02_v3: 10.6 kbp
5 kbp
327k 328k 329k 330k 331k 332k 333k 334k 335k 336k 337k
Annotated Genes (with UTRs in gray when available)
PF3D7_0206100.2
PF3D7_0206100.1
PF3D7_0206200
Splice Site Junctions (Union of All Experiments)
EST Alignments
P.f. mRNAseq Transcriptome during intraerythrocytic development Coverage (Bartfai et al.) (Showing 8 of 16 subtracks)
Hour10 (unique)
Hour15 (unique)
Hour20 (unique)
Hour25 (unique)
Hour30 (unique)
Hour35 (unique)
Hour40 (unique)
Hour5 (unique)

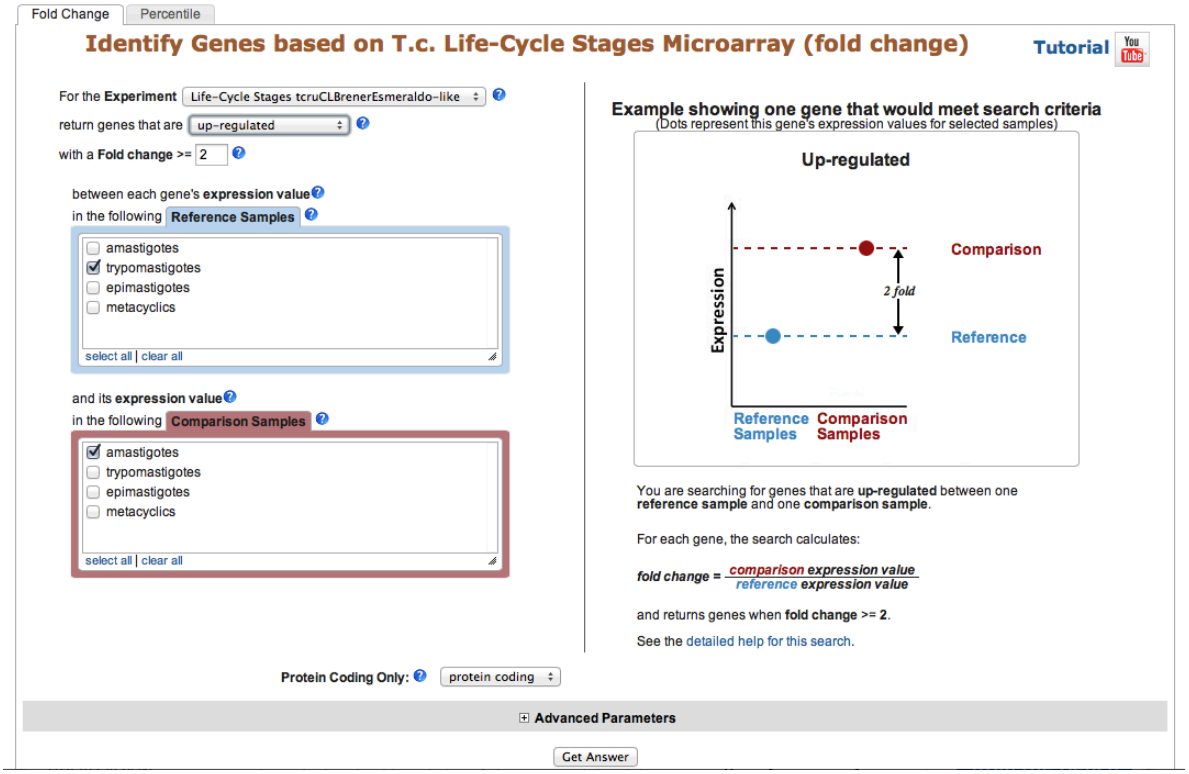
3. **Exploring microarray data in TriTrypDB.**
   Note: For this exercise use http://www.tritrypdb.org



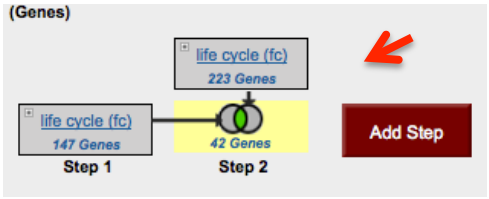
Find T. cruzi genes that are upregulated in amastigotes compared to trypomastigotes. Go to the transcript expression section then select microarray.

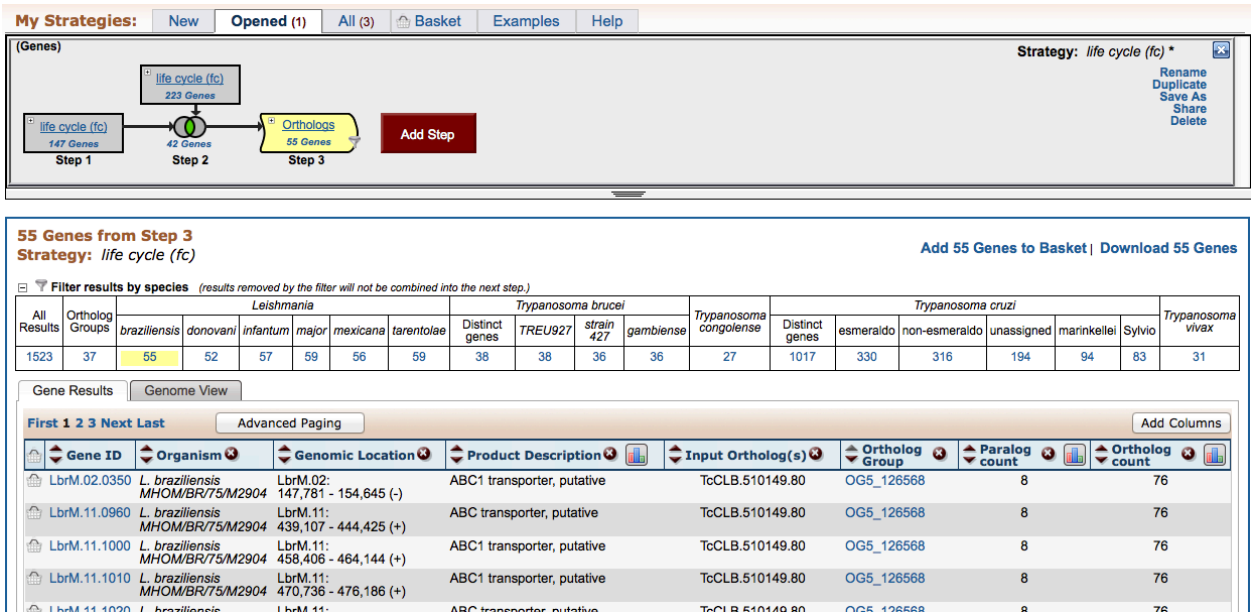


- Select the direction of regulation, your reference sample and your comparison sample. For the fold change keep the default value 2.



- How many genes did you find? Do the results seem plausible?
- Are any of these genes also upreglated 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 and ortholog transform on your results).
- How many ortholgs exist in *L. braziliensis*? (*Hint*: look at the filter table right above your results. Click on the number in of gene to view results from a specific species).

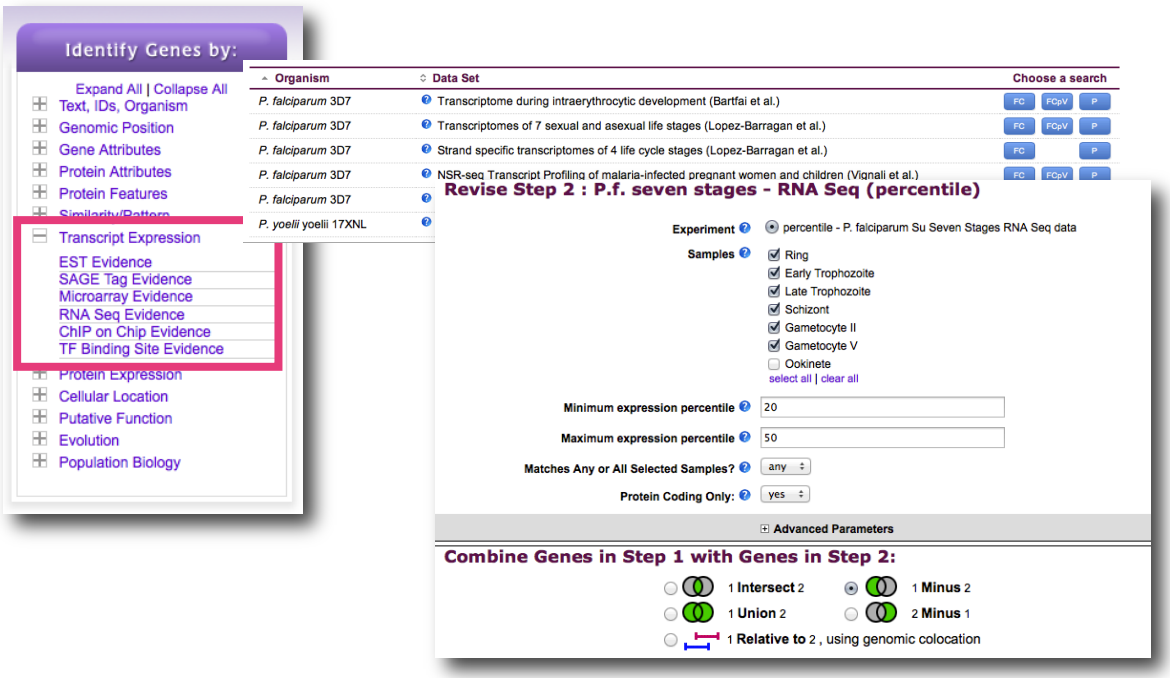




- Explore your results. Did you find anything interesting?

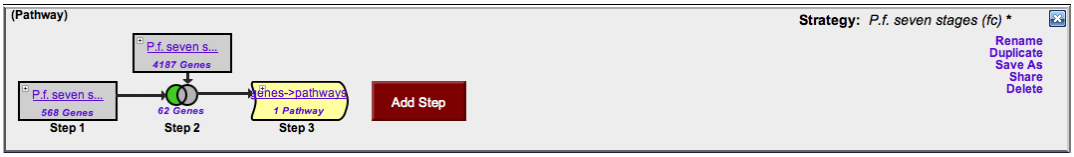
4. **Finding genes based on RNAseq evidence and inferring function of hypothetical genes.**
   Note: Use http://plasmodb.org for this exercise.

   a. Find all genes in *P. falciparum* that are upregulated at least 50-fold in ookinetes compared to other stages: "Transcriptomes of 7 sexual and asexual life stages (Lopez-Barragan et al.)"



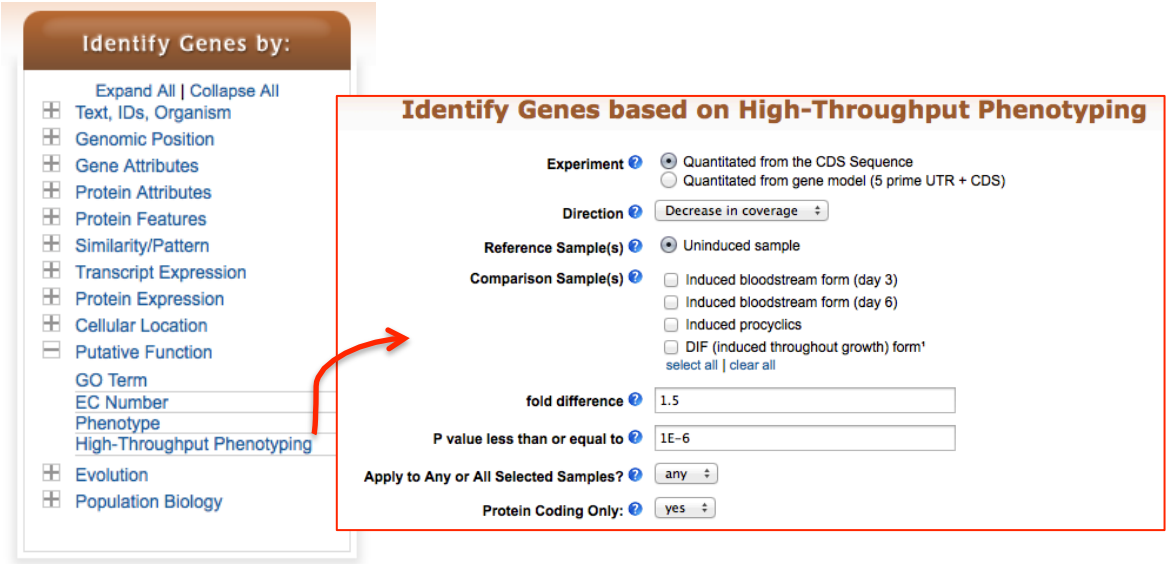
b. The above search will give you all genes that are upregulated by 50 fold in ookinetes compared to the other stages. However, this does not mean that these genes are not expressed in the other stages. How can you remove genes form the list that are likely not expressed in the other stages? (*hint: 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)

c. Which metabolic pathways are represented in this gene list? (*hint:* transform results to metabolic pathways.

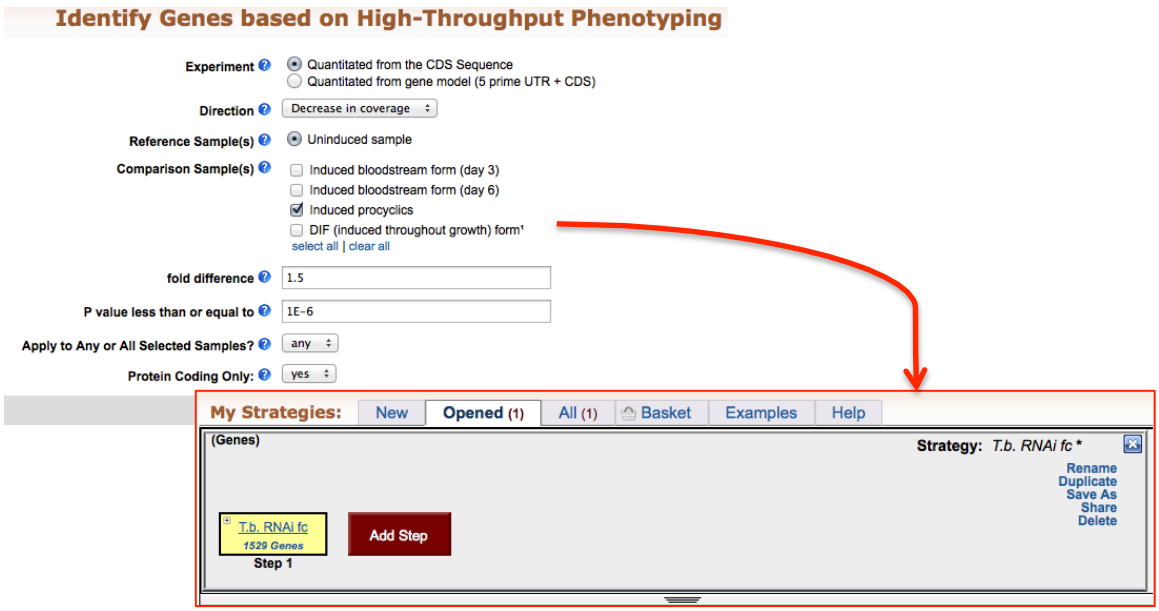


5. Finding genes that are essential in procyclics but not in blood form *T. brucei*. Note: for this exercise use http://TriTrypDB.org.

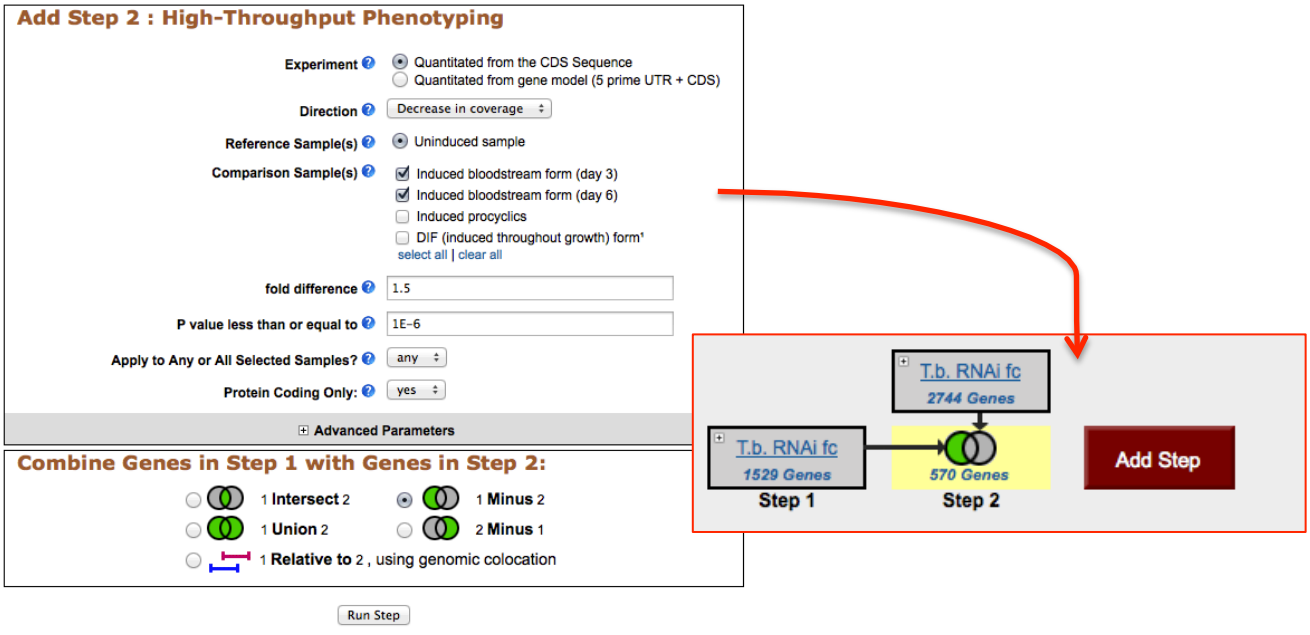
- Find the query for high throughput Phenotyping.



- Think about how to set up this query. (*hint*: you will have to setup 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.



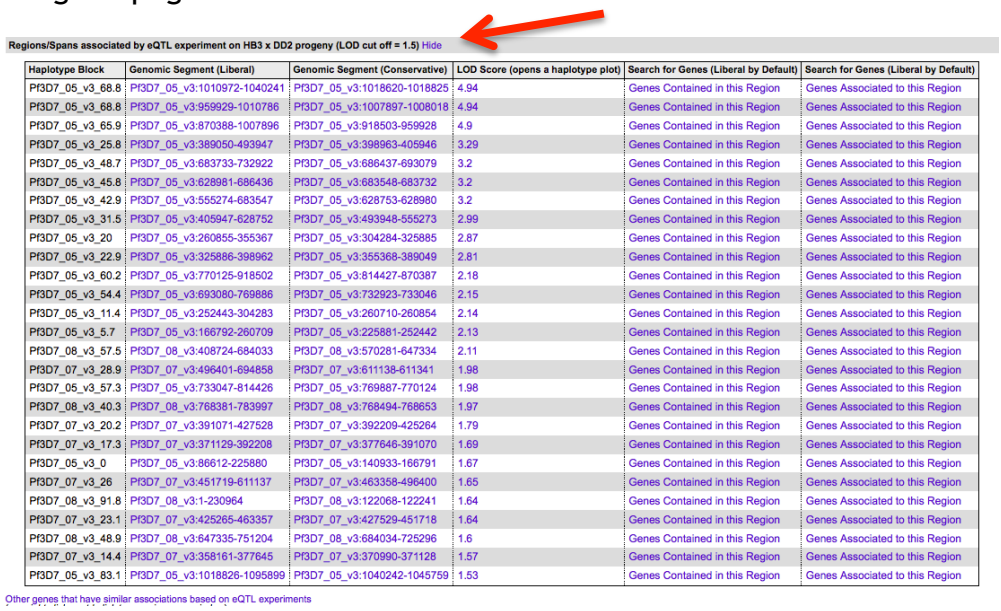
- Next add a step and run the same search except this time select the “induced bloodstream form” samples.
- How did you combine the results? Remember you want to find genes that are essential in procyclics and not in blood form.



6. 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.

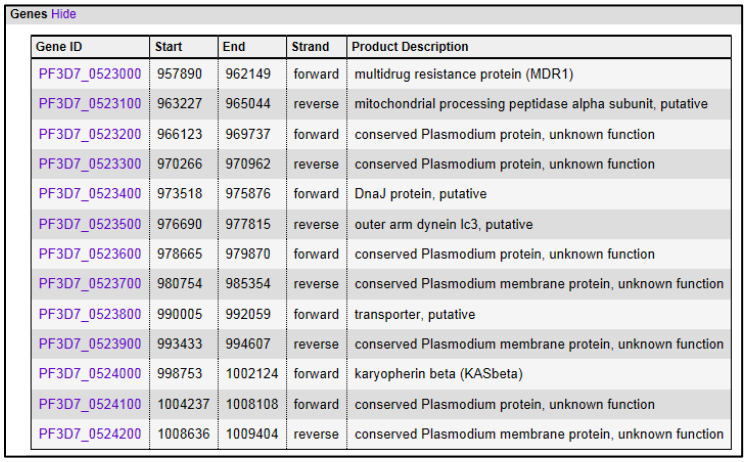


Regions/Spans associated by eQTL experiment on HB3 x DD2 progeny (LOD cut off = 1.5) Hide

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	Pf3D7_05_v3:1010972-1040241	Pf3D7_05_v3:1018620-1018825	4.94	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_05_v3_68.8	Pf3D7_05_v3:959929-1010786	Pf3D7_05_v3:1007897-1008018	4.94	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_05_v3_65.9	Pf3D7_05_v3:870388-1007896	Pf3D7_05_v3:918503-959928	4.9	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_05_v3_25.8	Pf3D7_05_v3:389050-493947	Pf3D7_05_v3:398963-405946	3.29	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_05_v3_48.7	Pf3D7_05_v3:683733-732922	Pf3D7_05_v3:686437-693079	3.2	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_05_v3_45.8	Pf3D7_05_v3:628981-686436	Pf3D7_05_v3:683548-683732	3.2	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_05_v3_42.9	Pf3D7_05_v3:555274-683547	Pf3D7_05_v3:628753-628980	3.2	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_05_v3_31.5	Pf3D7_05_v3:405947-628752	Pf3D7_05_v3:493948-555273	2.99	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_05_v3_20	Pf3D7_05_v3:260855-355367	Pf3D7_05_v3:304284-325885	2.87	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_05_v3_22.9	Pf3D7_05_v3:325886-398962	Pf3D7_05_v3:355368-389049	2.81	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_05_v3_60.2	Pf3D7_05_v3:770125-918502	Pf3D7_05_v3:814427-870387	2.18	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_05_v3_54.4	Pf3D7_05_v3:693080-769886	Pf3D7_05_v3:732923-733046	2.15	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_05_v3_11.4	Pf3D7_05_v3:252443-304283	Pf3D7_05_v3:260710-260854	2.14	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_05_v3_5.7	Pf3D7_05_v3:166792-260709	Pf3D7_05_v3:225881-252442	2.13	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_08_v3_57.5	Pf3D7_08_v3:408724-684033	Pf3D7_08_v3:570281-647334	2.11	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_07_v3_28.9	Pf3D7_07_v3:496401-694858	Pf3D7_07_v3:611138-611341	1.98	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_05_v3_57.3	Pf3D7_05_v3:733047-814426	Pf3D7_05_v3:769887-770124	1.98	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_08_v3_40.3	Pf3D7_08_v3:768381-783997	Pf3D7_08_v3:768494-768653	1.97	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_07_v3_20.2	Pf3D7_07_v3:391071-427528	Pf3D7_07_v3:392209-425264	1.79	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_07_v3_17.3	Pf3D7_07_v3:371129-392208	Pf3D7_07_v3:377646-391070	1.69	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_05_v3_0	Pf3D7_05_v3:86612-225880	Pf3D7_05_v3:140933-166791	1.67	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_07_v3_26	Pf3D7_07_v3:451719-611137	Pf3D7_07_v3:463358-496400	1.65	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_08_v3_91.8	Pf3D7_08_v3:1-230964	Pf3D7_08_v3:122068-122241	1.64	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_07_v3_23.1	Pf3D7_07_v3:425265-463357	Pf3D7_07_v3:427529-451718	1.64	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_08_v3_48.9	Pf3D7_08_v3:647335-751204	Pf3D7_08_v3:684034-725296	1.6	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_07_v3_14.4	Pf3D7_07_v3:358161-377645	Pf3D7_07_v3:370990-371128	1.57	Genes Contained in this Region	Genes Associated to this Region
Pf3D7_05_v3_83.1	Pf3D7_05_v3:1018826-1095899	Pf3D7_05_v3:1040242-1045759	1.53	Genes Contained in this Region	Genes Associated to this Region

Other genes that have similar associations based on eQTL experiments

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 page.

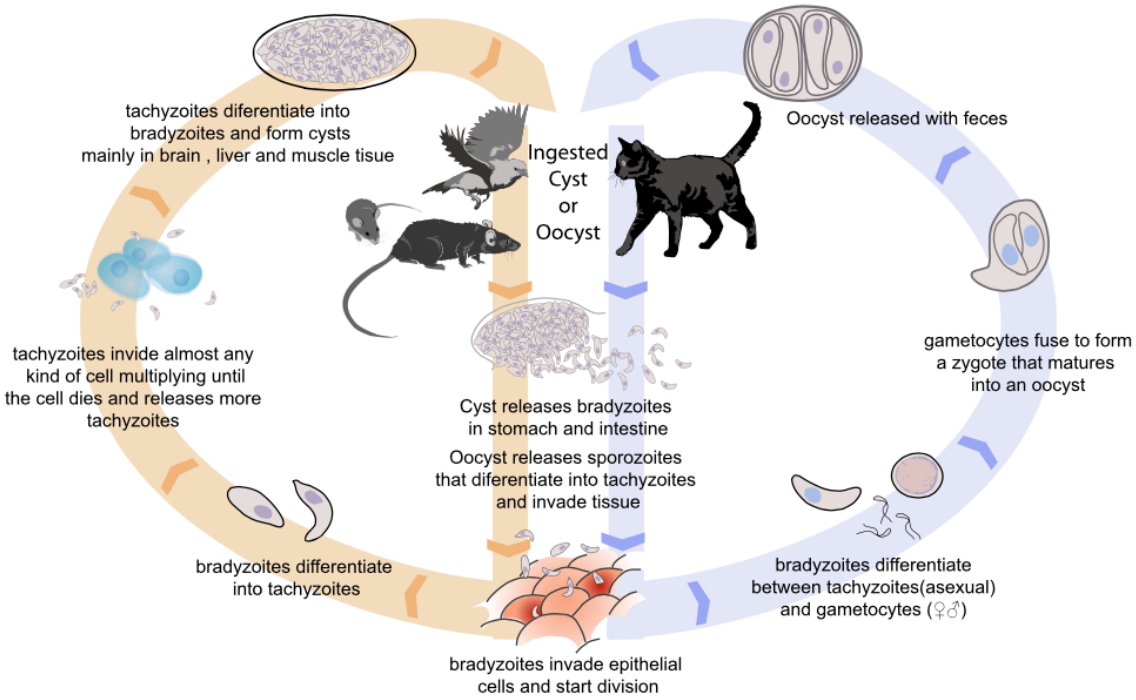


Genes Hide

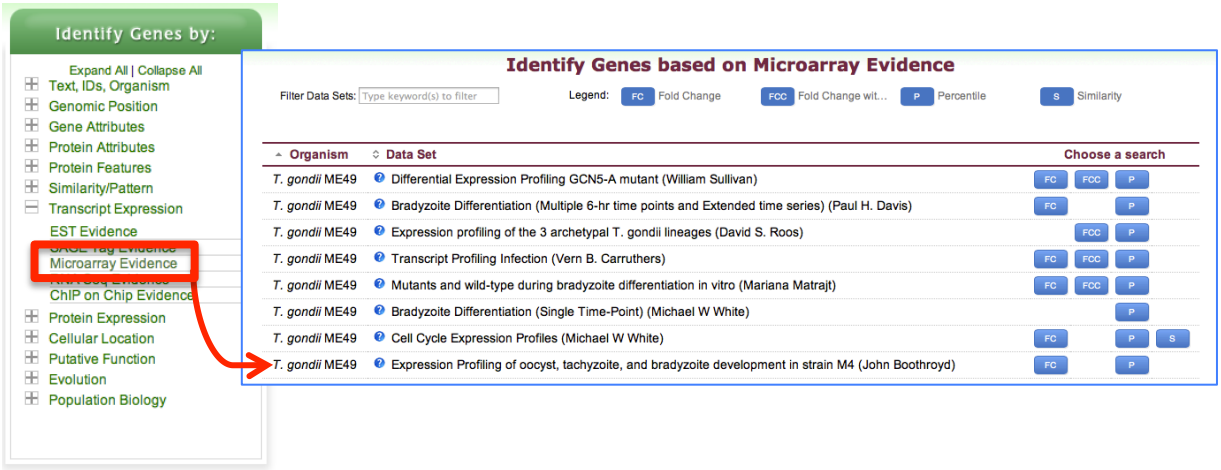
Gene ID	Start	End	Strand	Product Description
PF3D7_0523000	957890	962149	forward	multidrug resistance protein (MDR1)
PF3D7_0523100	963227	965044	reverse	mitochondrial processing peptidase alpha subunit, putative
PF3D7_0523200	966123	969737	forward	conserved Plasmodium protein, unknown function
PF3D7_0523300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

c. What other genes are associated with this block?
(Hint: go back to the gene page eQTL table, 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.)

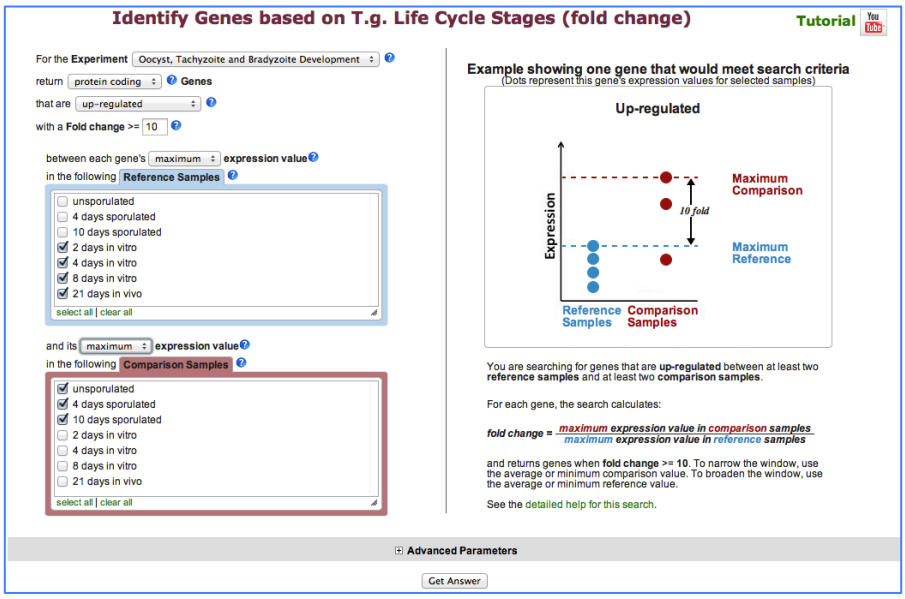
7. Finding oocyst expressed genes in *T. gondii* based on microarray evidence.
Note: For this exercise use http://toxodb.org



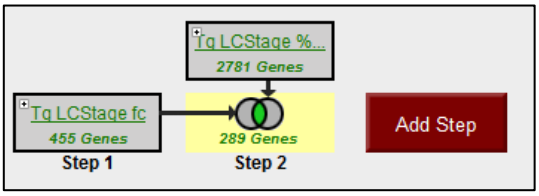
a. Find genes that are expressed at 10 fold higher levels in one of the oocyst stages than in any other stage in the "Expression Profiling of *T. gondii* Oocyst, Tachyzoite, Bradyzoite stages (John Boothroyd)" microarray experiment.



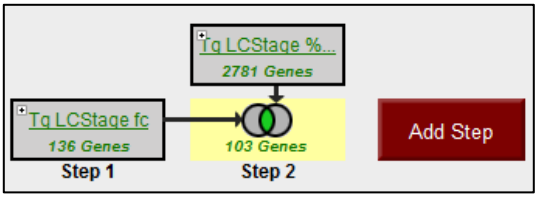
In this example the maximum expression value between genes in the reference and comparison groups was used to determine the fold difference.



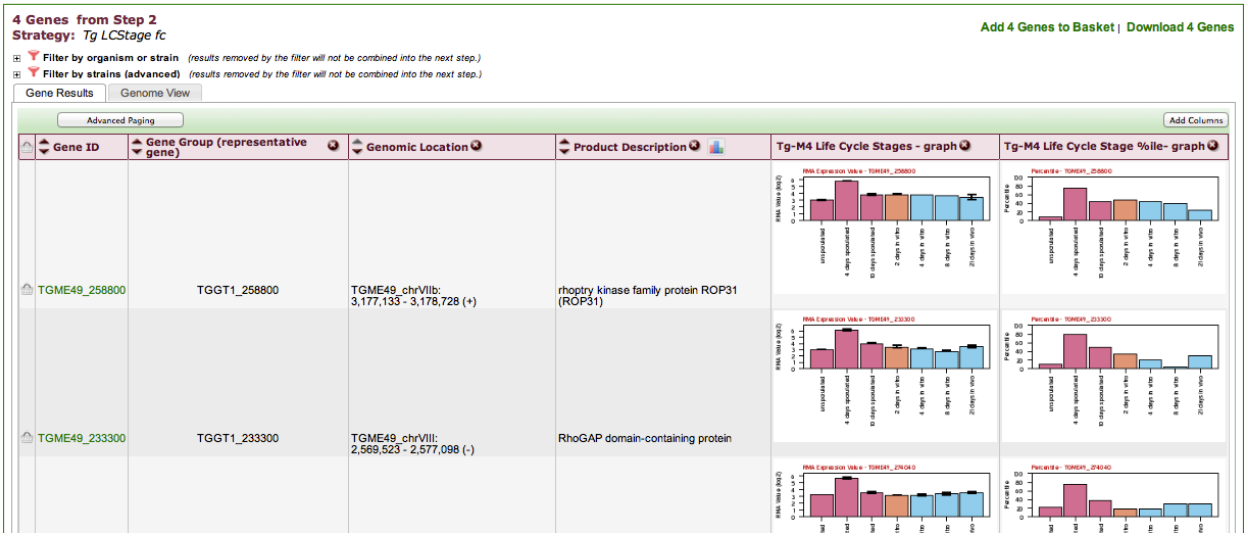
b. Add a step to limit this set of genes to only those for which all the non-oocyst stages are expressed below 50th percentile … ie likely not expressed at those stages.
   - *Hint*: after you click on add step find the same experiment under microarray expression and chose the percentile search.
   - 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***".
   - Note: you can turn on the columns called "Tg-M4 Life Cycle Stages – graph" and "Tg-M4 Life Cycle Stage %ile- graph" 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 as we'll use this strategy for an exercise we are doing later during the course**.



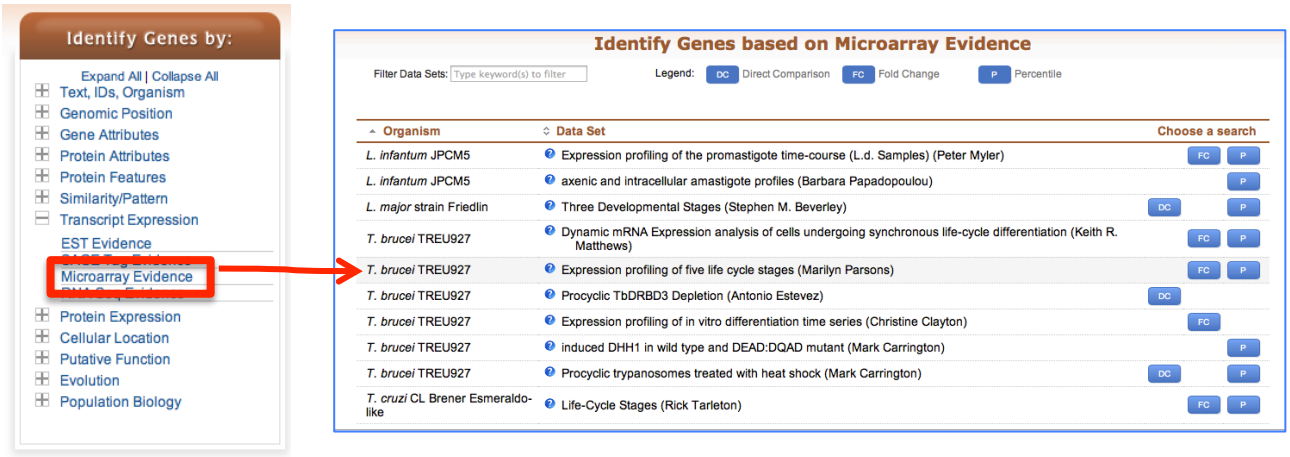
d. Revise the first step of this strategy to find genes that are 3 fold higher in day 4 oocysts than any other life cycle stage in this experiment.
   - Do all these genes have day 4 oocysts as the global maximum time point?
   - Note that we still have the step to limit the percentile of non-oocyst 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?



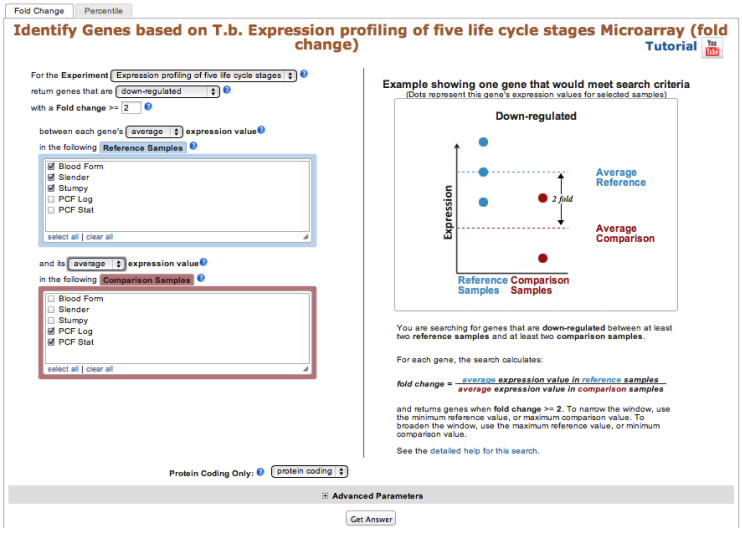
8. **Comparing RNA abundance and Protein abundance data.**
   Note: for this exercise use http://TriTrypDB.org.

   In this exercise we want to 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. Go to the genes by microarray expression and select the fold change search for the "Expression profiling of five life cycle stages (Marilyn Parsons)" experiment.

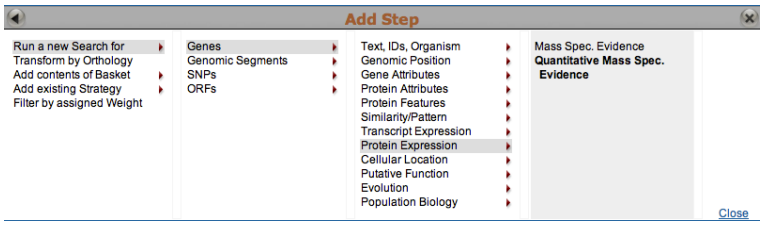


Configure the search to return protein-coding genes that are down-regulated 2 fold in procyclic form (PCF) (I chose both log and Stat and averaged them) relative to the Blood Form reference sample.



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 downregulated 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.



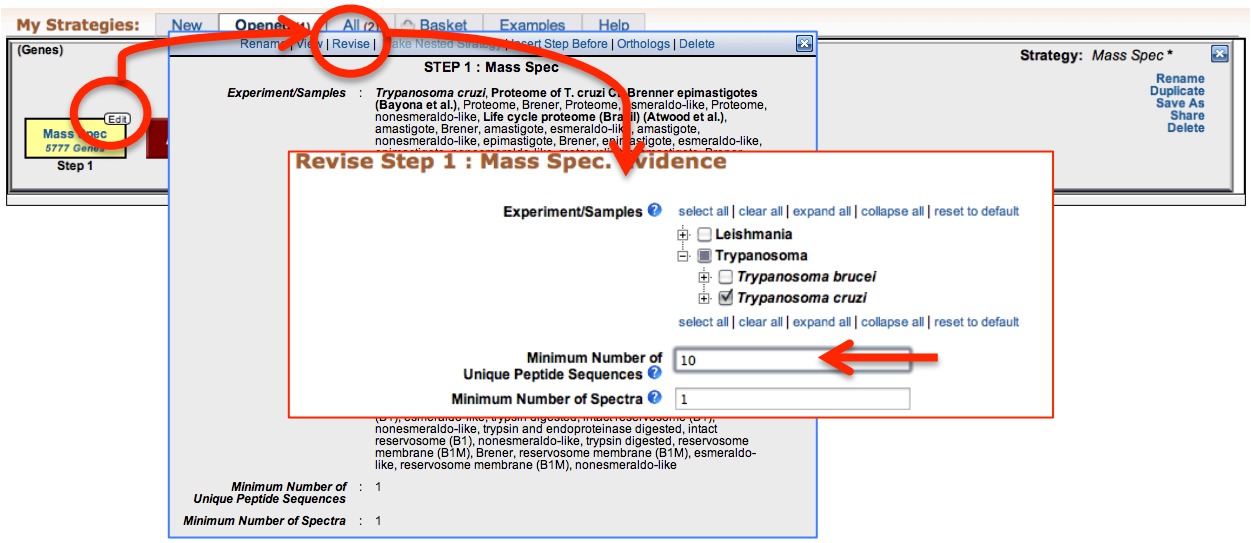
d. Try changing directions and compare up-regulated genes/proteins. (*hint, revise the existing strategy … you might want to duplicate it so you can keep both*). When you change one of the steps but not the other do you have any genes in the intersection? Why might this be??

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 RNASequencing step that has the same samples).*

9. Finding all genes with mass spec evidence in *T. cruzi*.
Note: for this exercise use http://TriTrypDB.org.

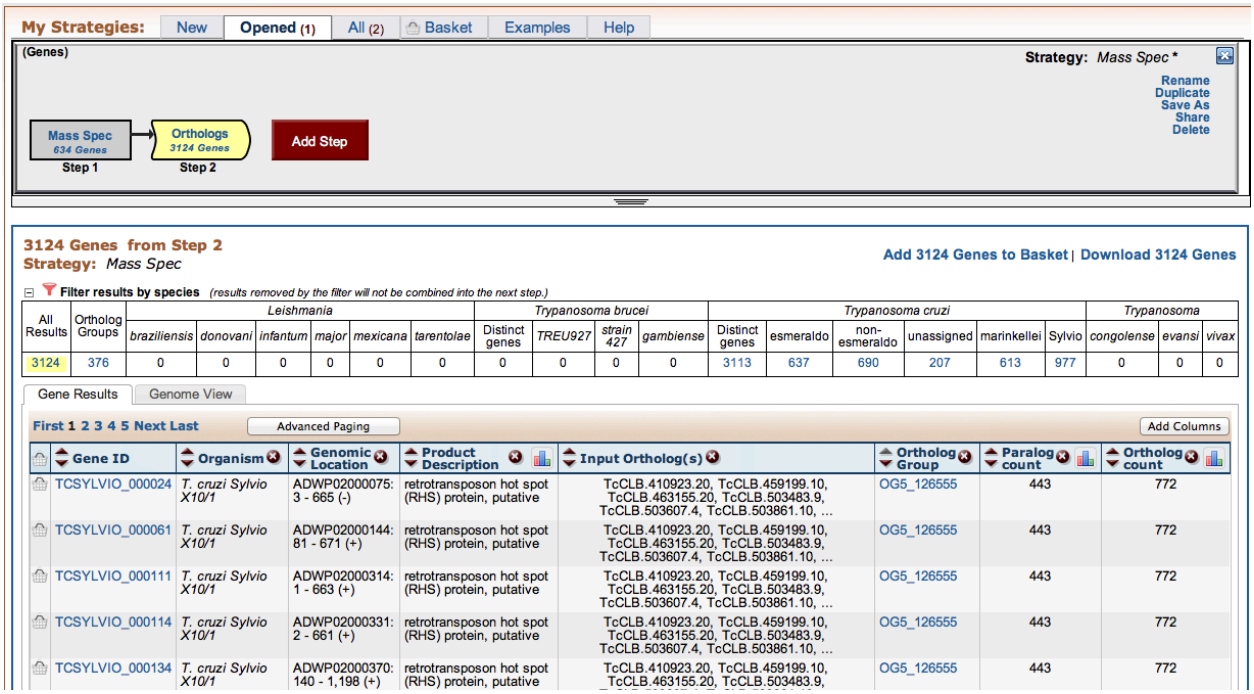
a. How many genes in *T. cruzi* have mass spec evidence?

b. How many genes from the results in a. have at least 10 unique peptide hits? (*hint*: try revising the step in 'a' and change the "minimum number of unique peptide sequences" option to 10.



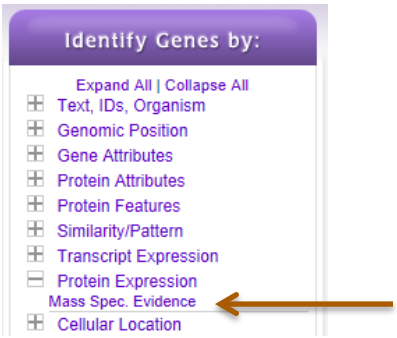
c. Can you expand the list of results in 'b' to include possible orthologs/paralogs in *T. cruzi*?

Hint: you will have to use the ortholog transform option when adding a step and select only *T. cruzi*. Explore the columns in your result set.



10. Finding genes with mass spec evidence in *P. berghei* gametocytes.
Note: For this exercise use http://www.plasmodb.org

a. **Find all *P. berghei* genes that have mass spec evidence in either or both male and female gametocytes.**
(*hint*: mass spec searches are in the "protein expression" expression section. Either or both is the Union of both results, not the intersection).



- How many genes did you get? How did you get to this number?
- Try running this search in two different ways:

i. Select both male and female gametocyte options and run the search.
ii. Select one of them first, run the search then add the other one using the add step button. How did you combine the two steps? Do you get the same results as in (i)?

b. **Find all genes that have mass spec evidence in both male and female gametocytes.**
(*hint*: use the strategy you developed in (ii) to get this answer, but change the union into an intersection).

c. **Find genes that have mass spec evidence only in male gametocytes and not in female ones**.
(*hint*: modify the set operation in b).

d. Find genes that have mass spec evidence only in female gametocytes and not in male ones.
(*hint*: modify the set operation in b).

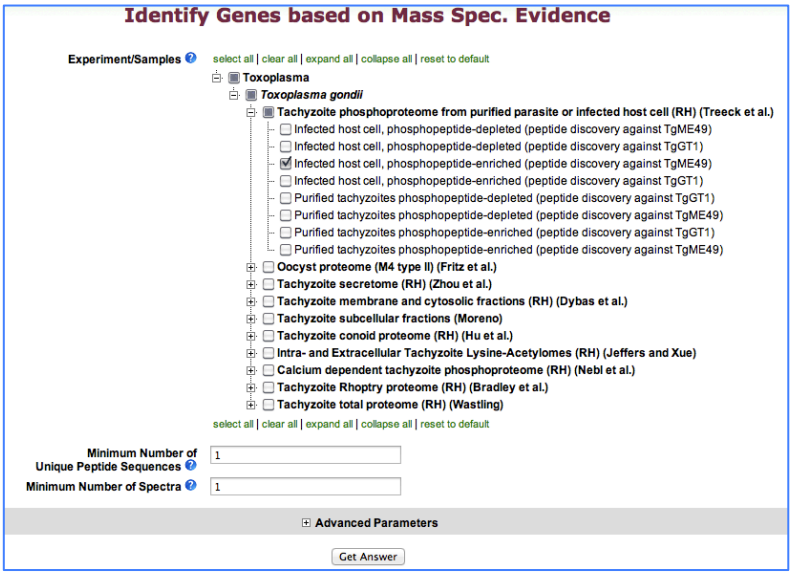
e. **Which female gametocyte gene has the highest number of peptide sequences?**
(*hint*: look at the "number of peptide sequences" column in the list of results).

f. **What does the distribution of peptides in the gene from 'e' look like?**
(*hint*: go to the gene page and look at the "Protein features" section, or go to the genome browser from the gene page and turn on the right tracks).

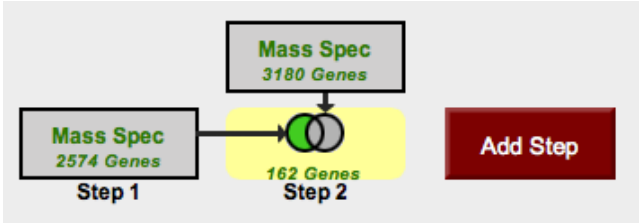
11. Finding genes with evidence of phosphorylation in intracellular *Toxoplasma* tachyzoites.
Note: For this exercise use http://www.toxodb.org

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

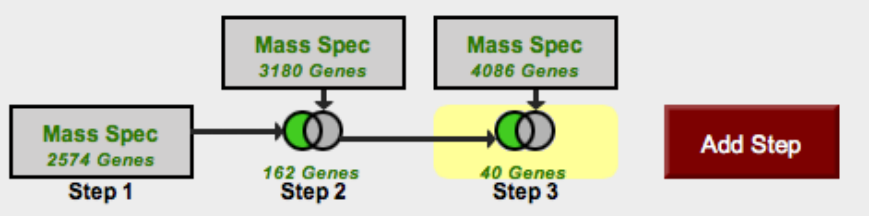
a. Find all genes with evidence of phosphorylation in intracellular tachyzoites. Select the "Infected host cell, phosphopeptide-enriched (peptide discovery against TgME49)" sample under the experiment called "Tachyzoite phosphoproteome from purified parasite or infected host cell (RH) (Treeck et al.)"



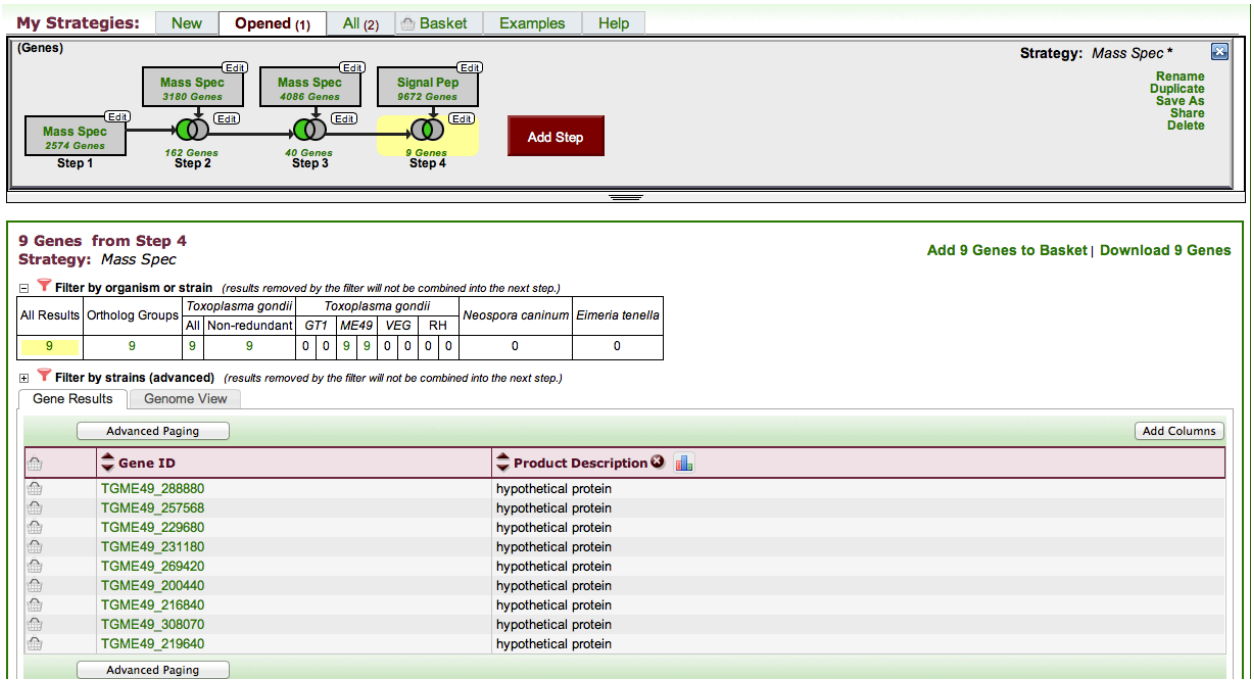
b. Remove all genes with phosphorylation evidence from purified extracellular tachyzoites.



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



d. Explore your results. What kids of genes did you find? Are any of these results to be secreted? (*Hint*: add a step searching for genes with secretory signal peptides).



e. 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).

f. What about polymorphism data? Go back to your strategy and add columns for SNP data found under the population section. Explore the gene page for the gene that has the most number of nonsynonymous SNPs.