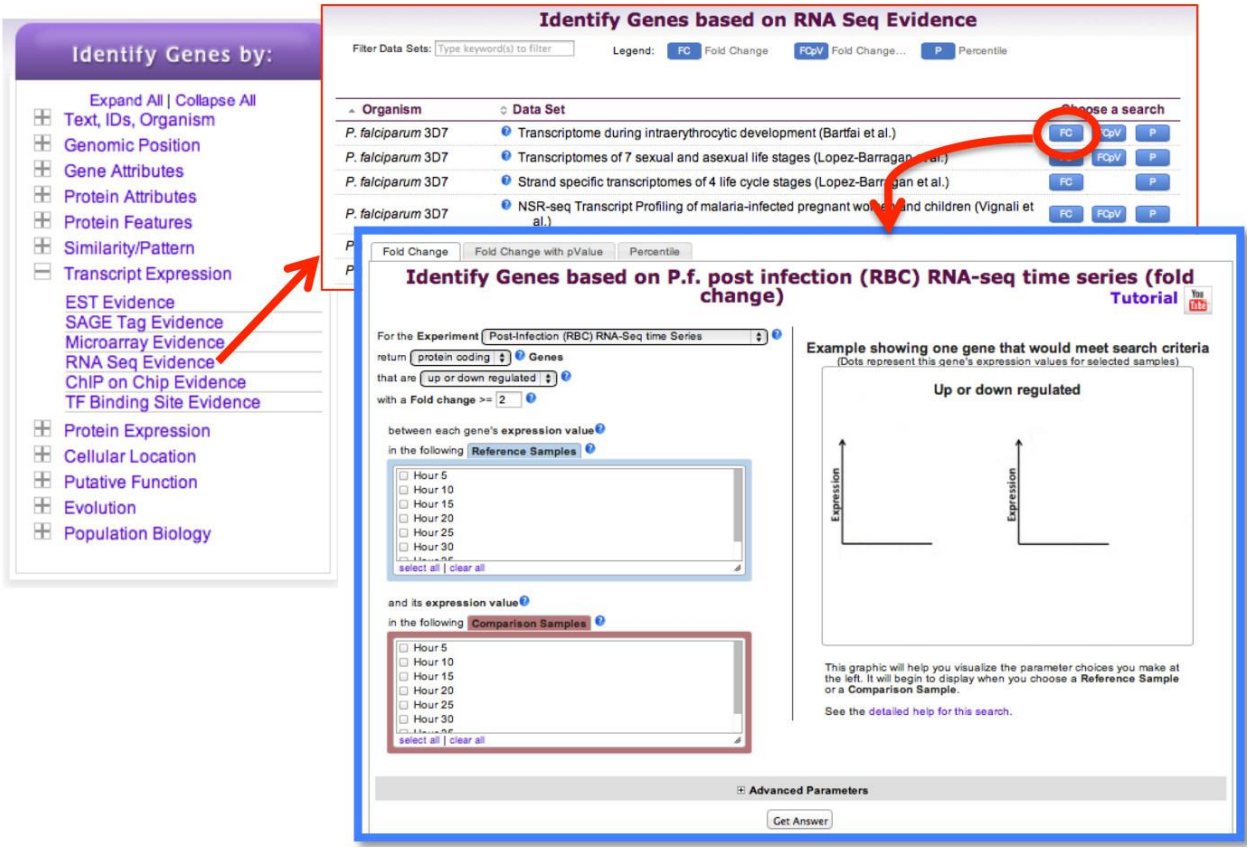
Exploring Transcriptomic and Proteomic data

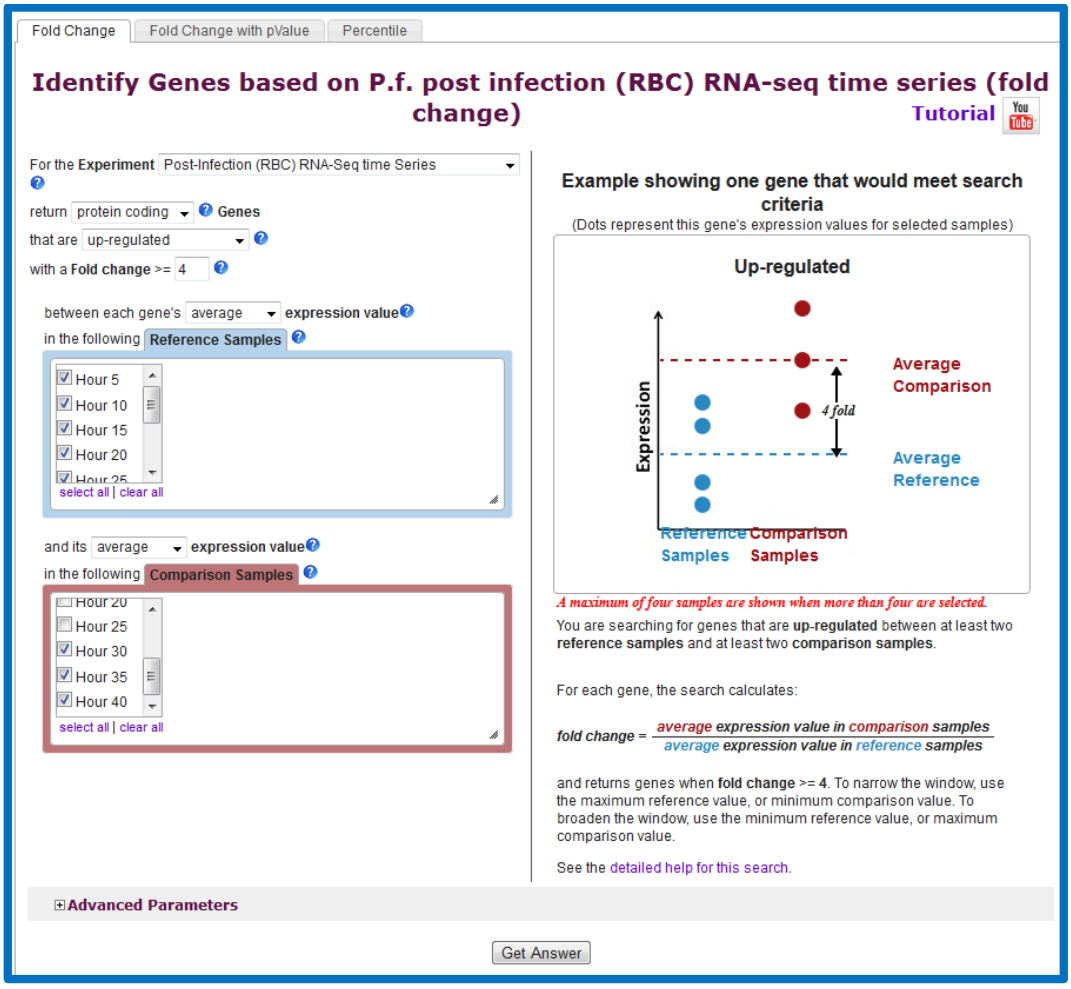
1. Find all *P. falciparum* genes that are up-regulated during the later stages of the intra-erythrocytic cycle.
   Note: For this exercise use http://www.plasmodb.org

a. Use the fold change search for the data set "**Transcriptome during intra-erythrocytic 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.



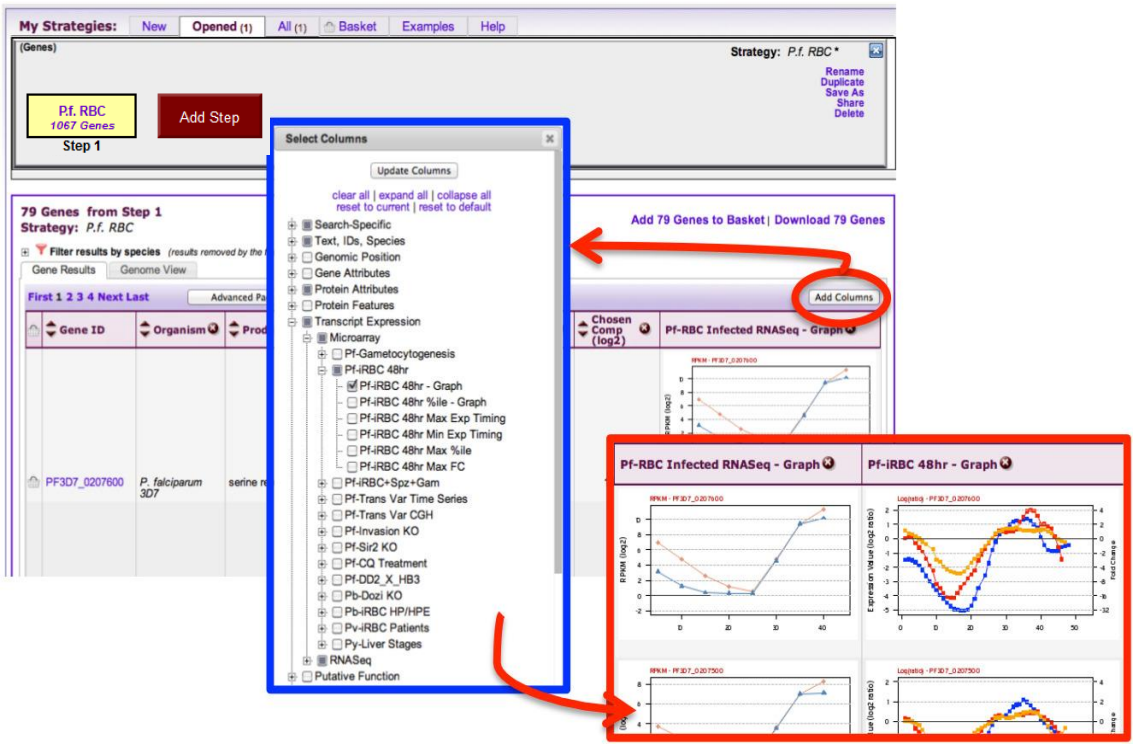
There are a number of parameters to manipulate in this search. As you modify parameters on the left side note the dynamic help on the right side. See screenshots.

- **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 4 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 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. Compare the RNA Sequencing data to similar microarray data on a gene-by-gene basis.

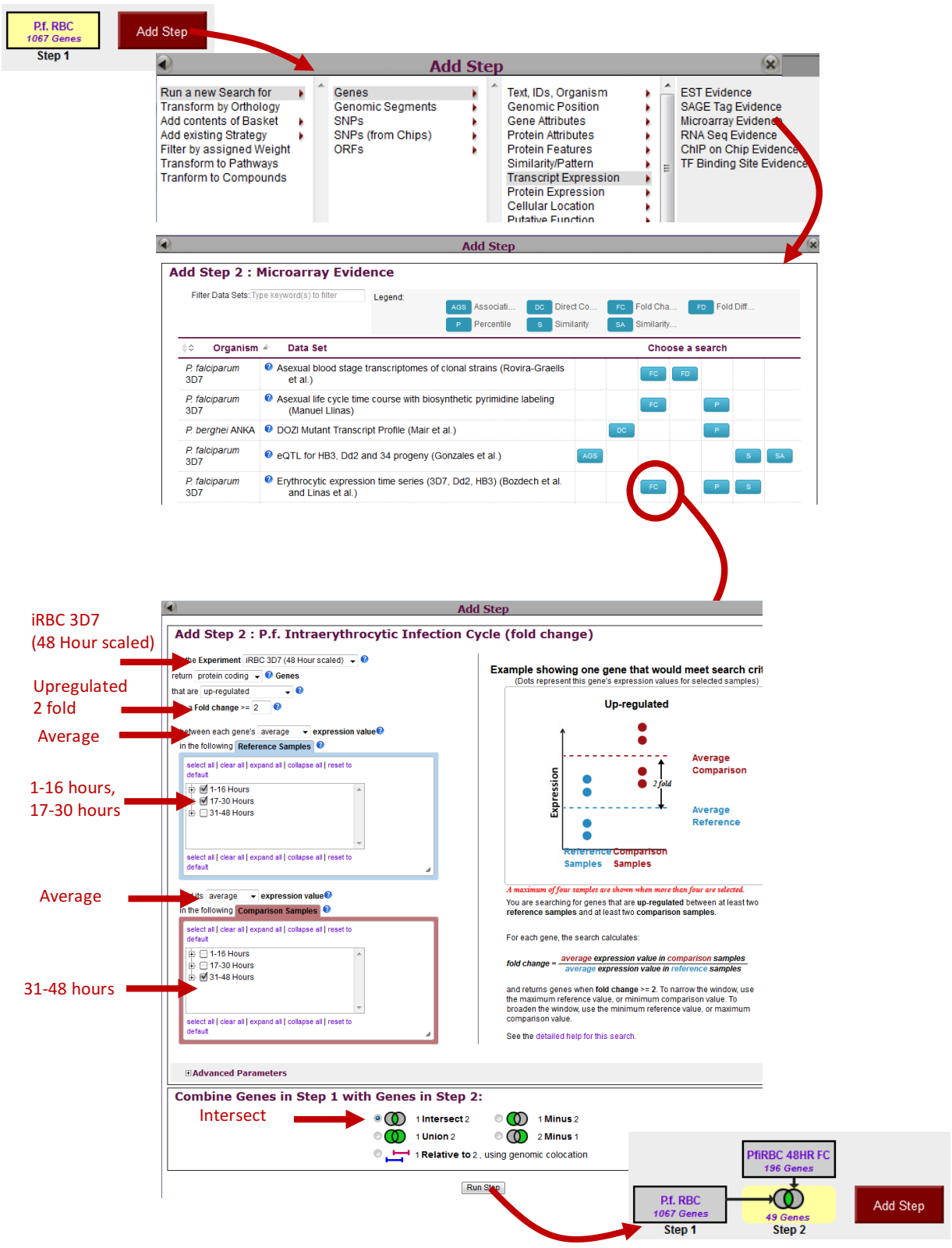
- PlasmoDB contains data from a similar experiment that was analyzed by microarray instead of RNA sequencing. This experiment is called: **Erythrocytic expression time series (3D7, DD2, HB3) (Bozdech et al. and Linas et al.) or Pf-iRBC 48hr for shorter column headings**. To directly compare the data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



c. Compare the RNA-sequencing data to microarray data on a genome scale.

You can also run a fold-change search on the microarray data 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 keeping the fold change >=2 due to the decreased dynamic range of microarrays compared to RNA-seq.

- How many genes are upregulated in the later stages of the erythrocytic cycle based on microarray and RNA sequencing evidence?


P.f. RBC
1067 Genes
Step 1
Add Step
Add Step
Run a new Search for
Transform by Orthology
Add contents of Basket
Add existing Strategy
Filter by assigned Weight
Transform to Pathways
Tranform to Compounds
Genes
Genomic Segments
SNPs
SNPs (from Chips)
ORFs
Text, IDs, Organism
Genomic Position
Gene Attributes
Protein Attributes
Protein Features
Similarity/Pattern
Transcript Expression
Protein Expression
Cellular Location
EST Evidence
SAGE Tag Evidence
Microarray Evidence
RNA Seq Evidence
ChIP on Chip Evidence
TF Binding Site Evidence
Add Step
Add Step 2 : Microarray Evidence
Filter Data Sets: Type keyword(s) to filter
Legend:
Organism
Data Set
Choose a search
P. falciparum 3D7
Asexual blood stage transcriptomes of clonal strains (Rovira-Graells et al.)
Asexual life cycle time course with biosynthetic pyrimidine labeling (Manuel Llinas)
P. berghei ANKA
DOZI Mutant Transcript Profile (Mair et al.)
eQTL for HB3, Dd2 and 34 progeny (Gonzales et al.)
Erythrocytic expression time series (3D7, Dd2, HB3) (Bozdech et al. and Linas et al.)
iRBC 3D7
(48 Hour scaled)
Upregulated
2 fold
Average
1-16 hours,
17-30 hours
Average
31-48 hours
Add Step
Add Step 2 : P.f. Intraerythrocytic Infection Cycle (fold change)
Experiment iRBC 3D7 (48 Hour scaled)
return protein coding Genes
that are up-regulated
Fold change >= 2
between each gene's average expression value
in the following Reference Samples
1-16 Hours
17-30 Hours
31-48 Hours
its average expression value
in the following Comparison Samples
1-16 Hours
17-30 Hours
31-48 Hours
Example showing one gene that would meet search crit
(Dots represent this gene's expression values for selected samples)
Up-regulated
Expression
2 fold
Average Comparison
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 at least two comparison samples.
For each gene, the search calculates:
fold change = average expression value in comparison samples / average expression value in reference samples
and returns genes when fold change >= 2. 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
Combine Genes in Step 1 with Genes in Step 2:
Intersect
1 Intersect 2
1 Minus 2
1 Union 2
2 Minus 1
1 Relative to 2, using genomic colocation
Run Step
PfiRBC 48HR FC
196 Genes
P.f. RBC
1067 Genes
49 Genes
Step 1
Step 2
Add Step

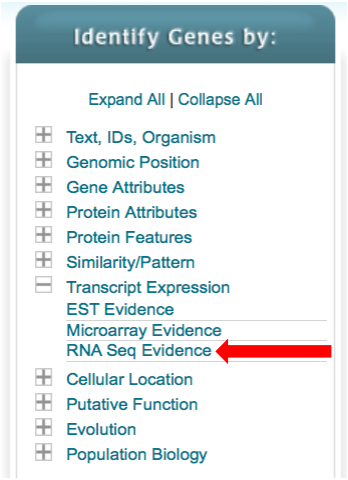
2. Exploring RNA-seq data in fungi

For this exercise use fungidb.org

C. albicans is a commensal fungal organism that inhabits skin, mucosal surfaces, and gut of humans. It can cause superficial and subcutaneous infections that may become life threatening in immunocompromised individuals. *C. albicans* has a yeast like non-invasive form, and a hyphal form associated with increased virulence and systemic infections. Let's look at the genes that are up-regulated in response to serum and high oxidative stress, conditions similar to those that are encountered by *C. albicans* during infection.

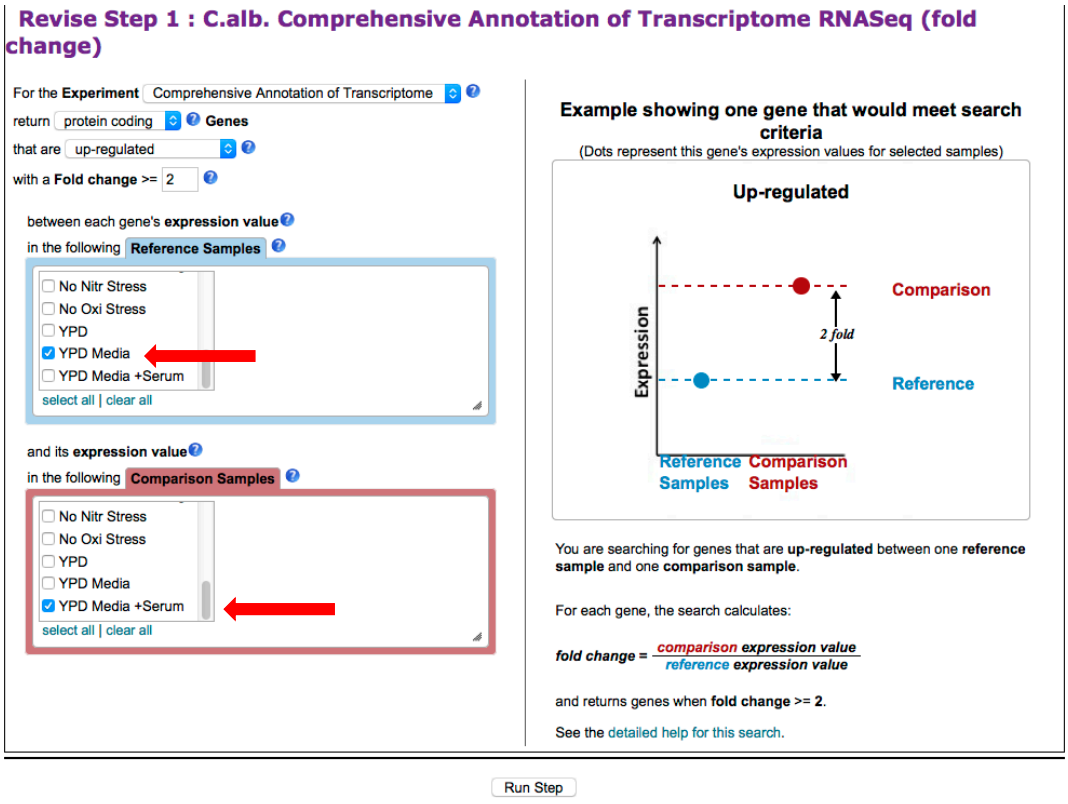
a. Identify genes that are up-regulated in YPD media supplied with serum.

- Navigate to the transcript expression data using the "Transcript Expression" menu in the "Identify Genes by" panel.



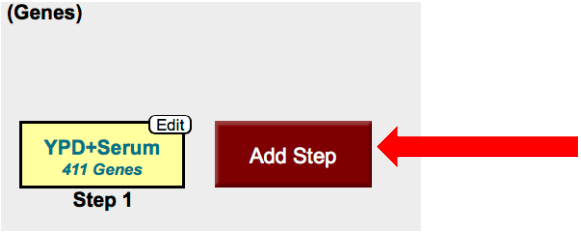
- Select the "Comprehensive Annotation of Transcriptome" by Michael Snyder.

- Find protein coding genes that are up-regulated by 2 fold in response to serum. Compare "YPD media + Serum" (comparison sample) to YPD Media (Reference control sample).



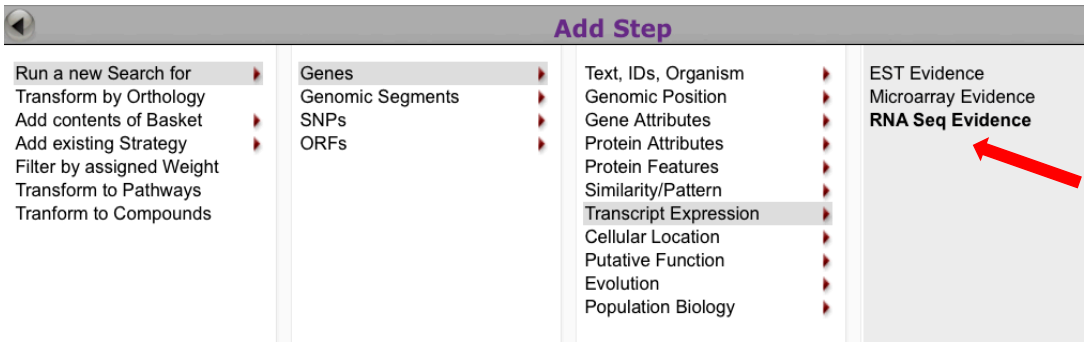
b. Identify those genes that are also up-regulated in response to high oxidative stress conditions.

- Click “Add step” and navigate to the same RNA-seq data set.

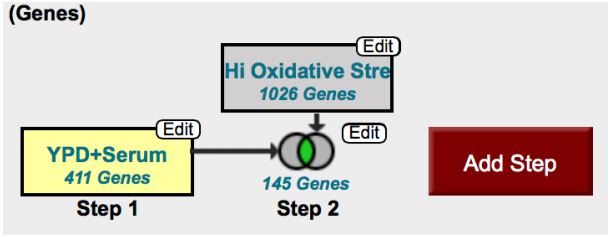


Hint: To change the name of the first step to “YPD+Serum” access the “Edit” option in the search box and select “Rename”.

- Set up a search using “Hi Oxi stress” data set this time using the no oxidative stress control (“No Oxi”) as a reference sample.



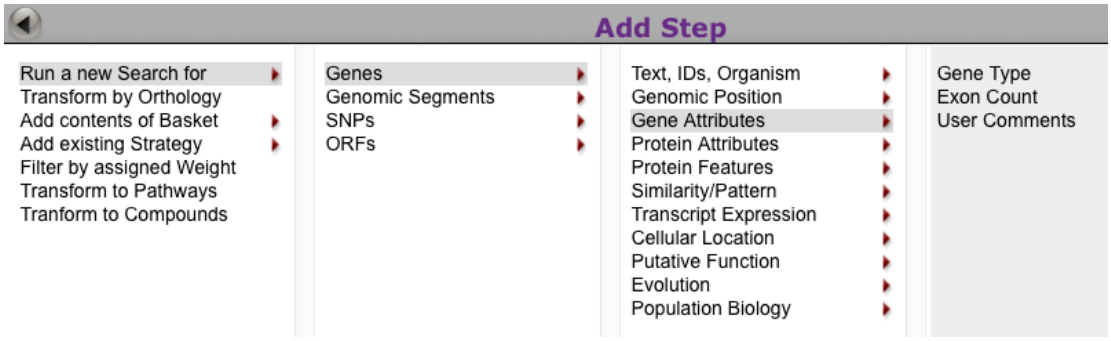
- Your search should return 145 genes:



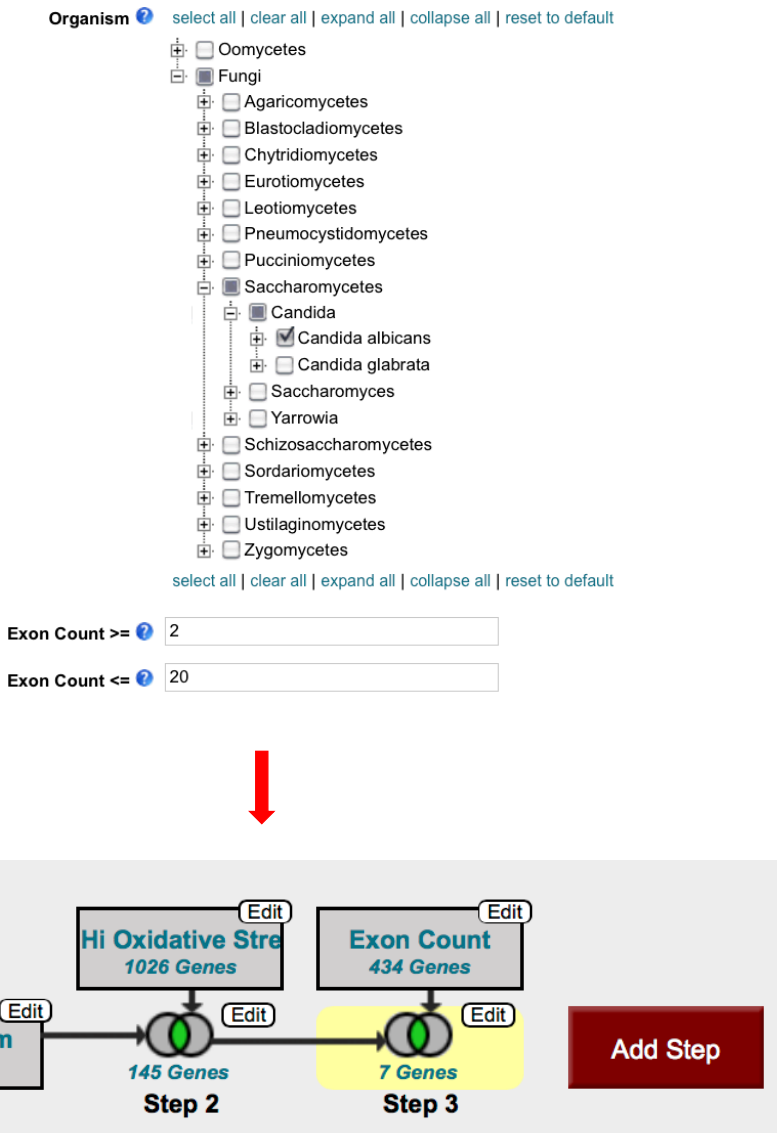
c. Identify genes with more than one exon.

The Gene Attributes function offers an option to further customize your search to identify genes with more than one exon.

- Click “Add Step” and navigate to “Exon count” option:



- For this exercise leave the parameters of the search in *C. albicans* at default (between 2 and 20 exons):



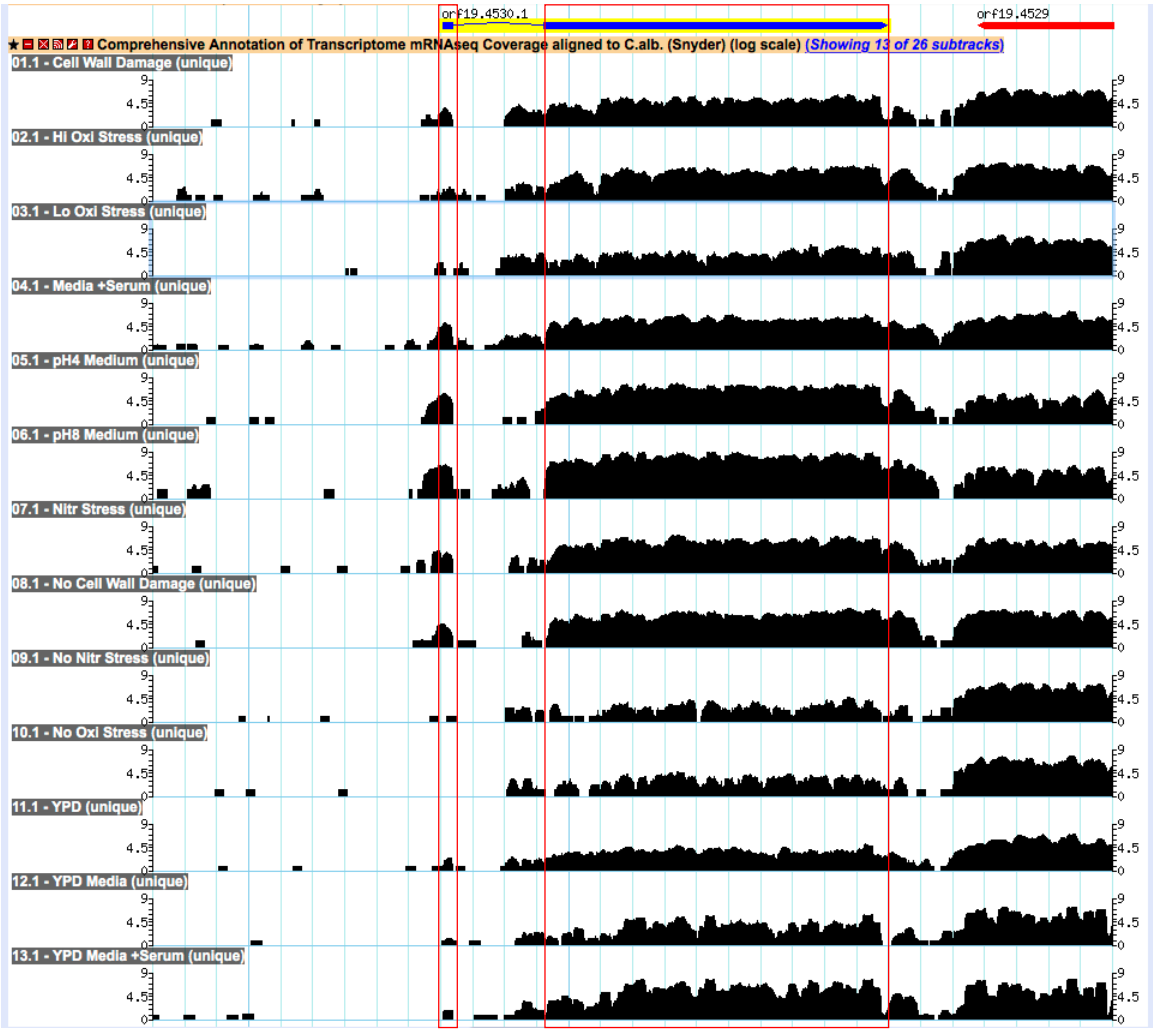
- Explore the first gene in the list - orf19.4530.1.

d. Explore RNA-seq tracks and genome annotations using Genome Browser.

- From the gene page navigate to the Genome Browser window.

- Turn on the track called “Comprehensive Annotation of Transcriptome mRNAseq Coverage aligned to C.alb. (Snyder) (log scale)” by navigating to the Select tracks tab in the GBrowse window.

- Examine how well RNA-seq tracks correspond to current gene model.

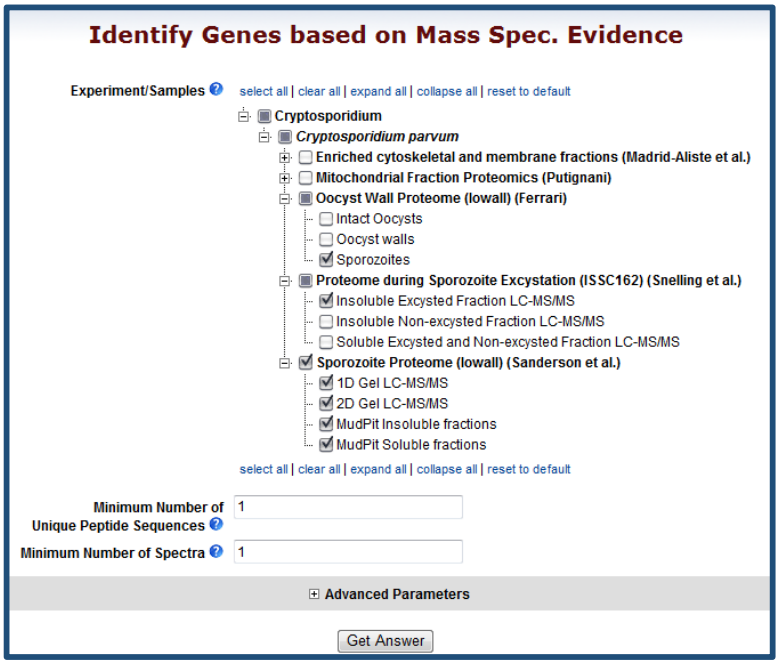


- Do these results make sense?
- Do you think there is evidence for alternative splicing or intron read through due to various growth conditions?
- Can you explain transcript mapping outside the current annotation coordinates?

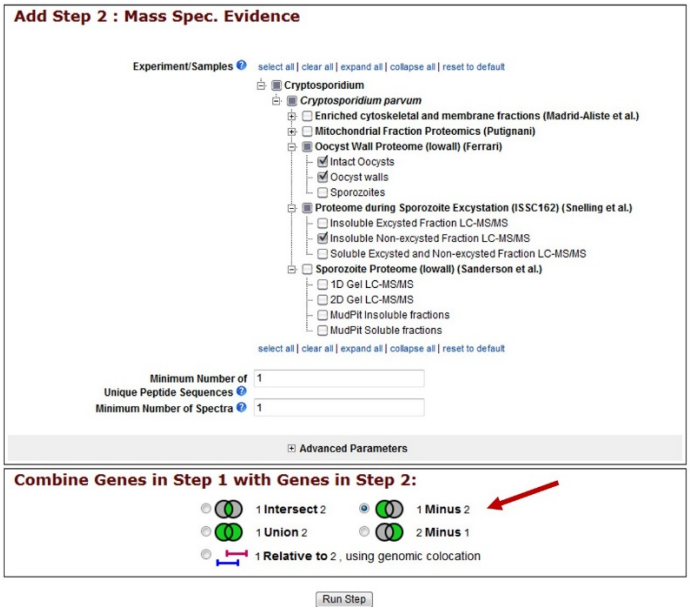
    Hint: You can choose which tracks to view (turn off / turn on tracks) by clicking on the blue link at the top “Showing 13 of 26 sub-tracks” and further adjusting the view options.

3. Find sporozoite-specific *Cryptosporidium* genes that are expressed at the protein level and are likely secreted.
For this exercise use http://cryptodb.org

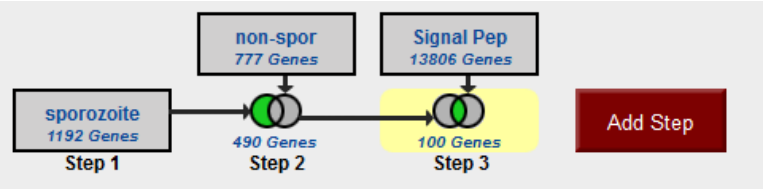
3 a. Find *Cryptosporidium* genes that are expressed at the protein level with evidence from any of the sporozoite proteomics experiments available in CryptoDB. Explore the available proteomics data and select samples that make sense. You may need to click on the '+' sign to expand experiments to see the underlying samples.



3 b. Remove any gene with peptide evidence from non-sporozoite samples
Hint: add a step for mass spec data and think about how you will combine your results.

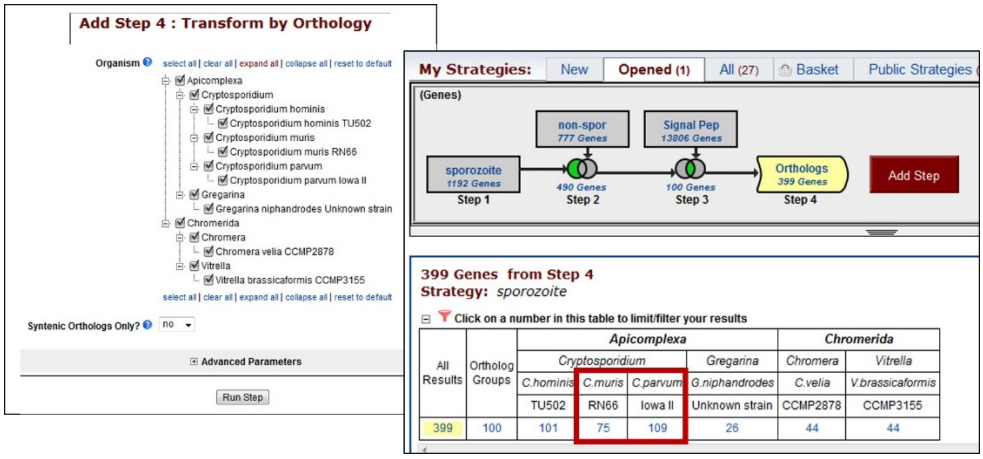


3 c. How many of these genes are also predicted to be secreted?



3 d. So far you have been searching for *C. parvum* genes because we only have proteomics data from this species. However, what if you are studying *C. muris*? How can you garner information about the protein expression of *C. muris* genes from your *C. parvum* results? (Hint: add a step then select the "Transform by Orthology" option).

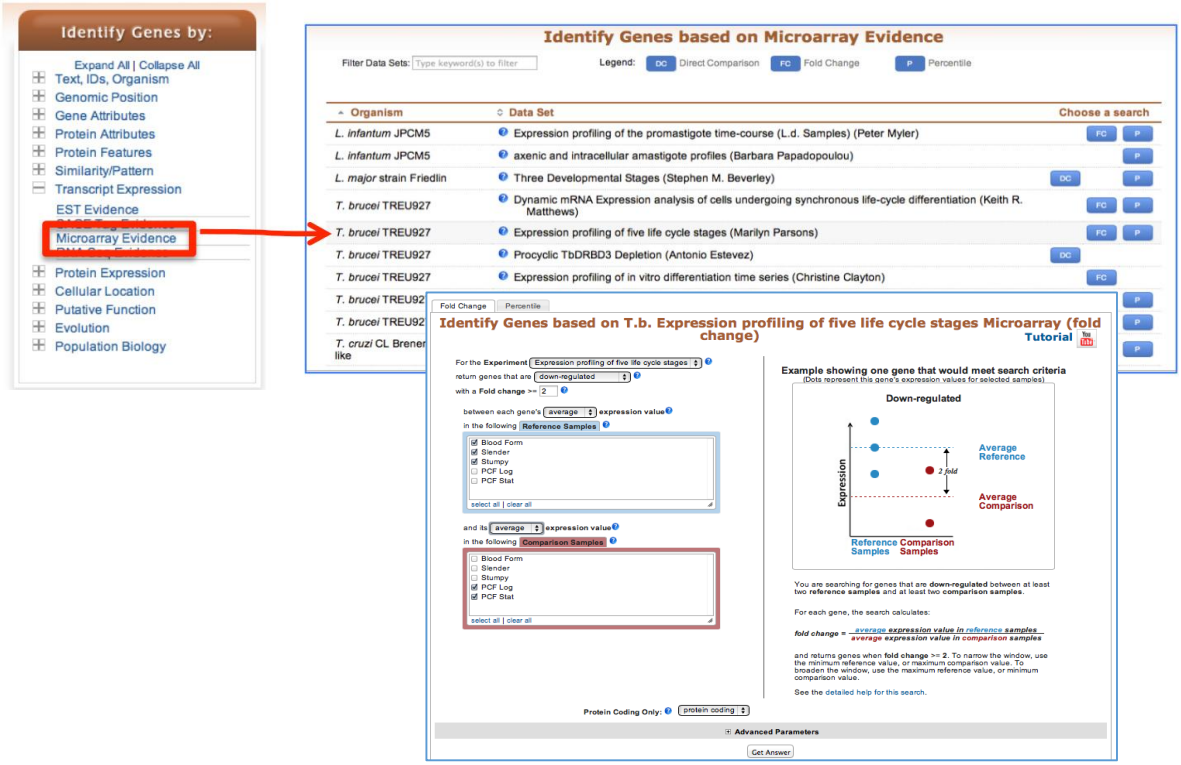
- Did the number of *C. parvum* genes increase or decrease? Why?



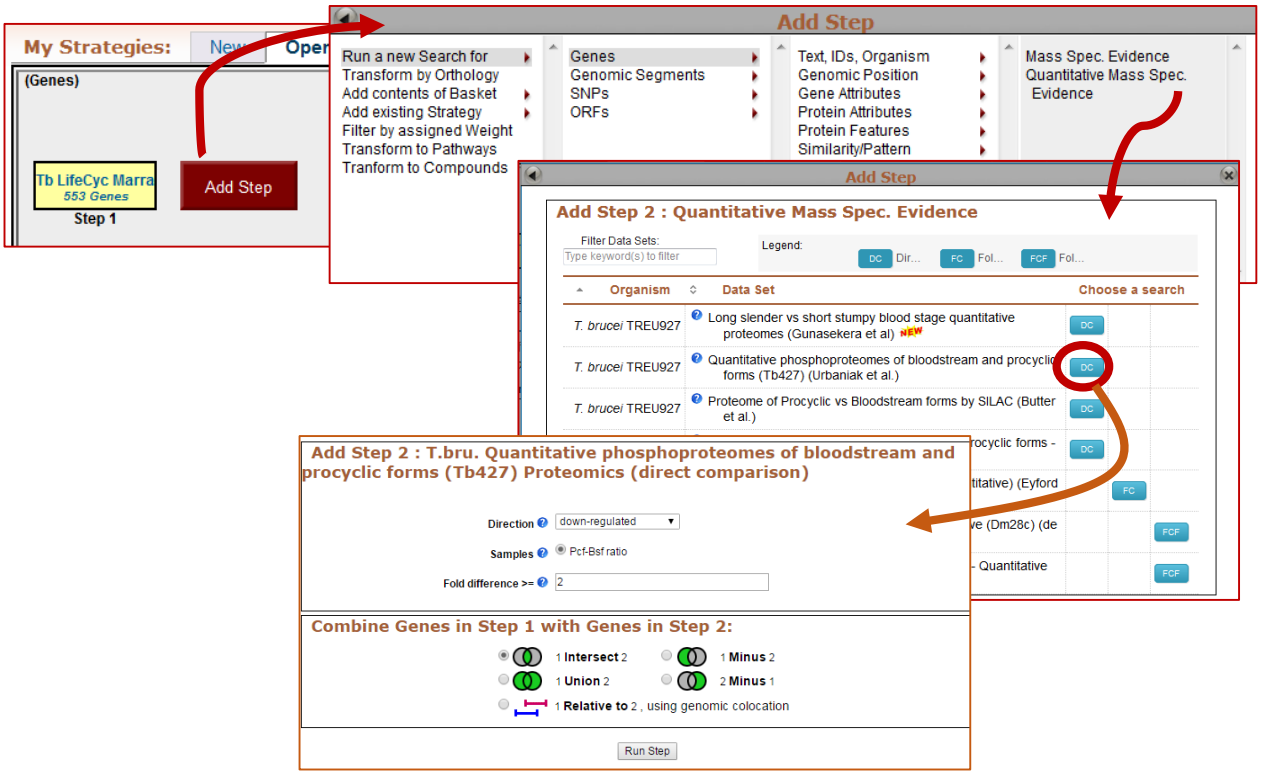
4. Comparing RNA abundance and Protein abundance data.
Note: for this exercise use http://TriTrypDB.org.

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?

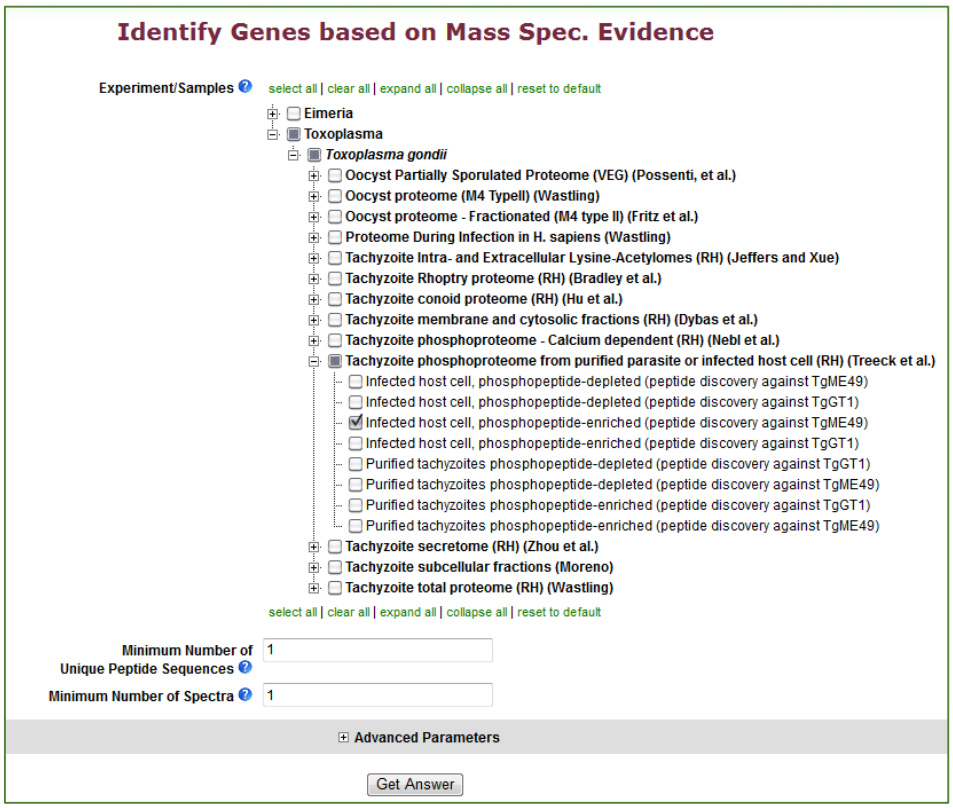
e. Can you think of ways to provide more confidence (or cast a broader net) in the microarray step? (*Hint:* you could insert steps to restrict based on percentile or add a RNA Sequencing step that has the same samples).

5. OPTIONAL: Find genes with evidence of phosphorylation in intracellular *Toxoplasma* tachyzoites.

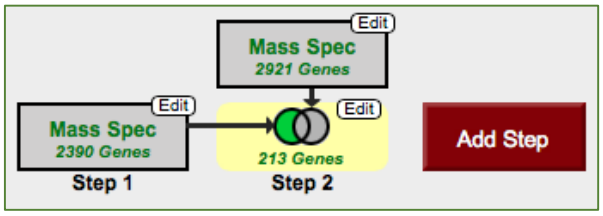
For this exercise use http://www.toxodb.org

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

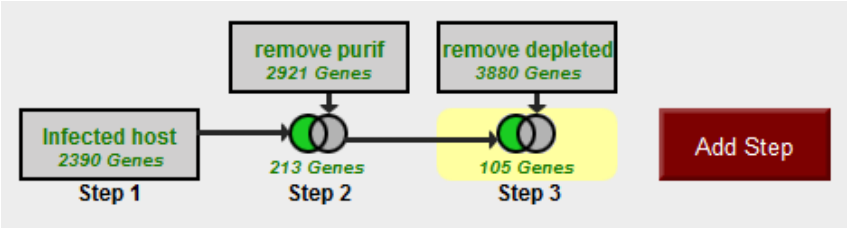
5a. Find all genes with evidence of phosphorylation in intracellular tachyzoites. Select the "**Infected host cell, phosphopeptide-enriched** (peptide discovery against **TgME49**)" sample under the experiment called "**Tachyzoite phosphoproteome from purified parasite or infected host cell (RH) (Treeck et al.)**"



5b. Remove all genes with phosphorylation evidence from purified tachyzoites.

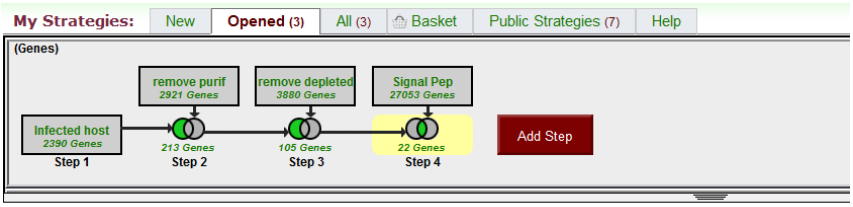


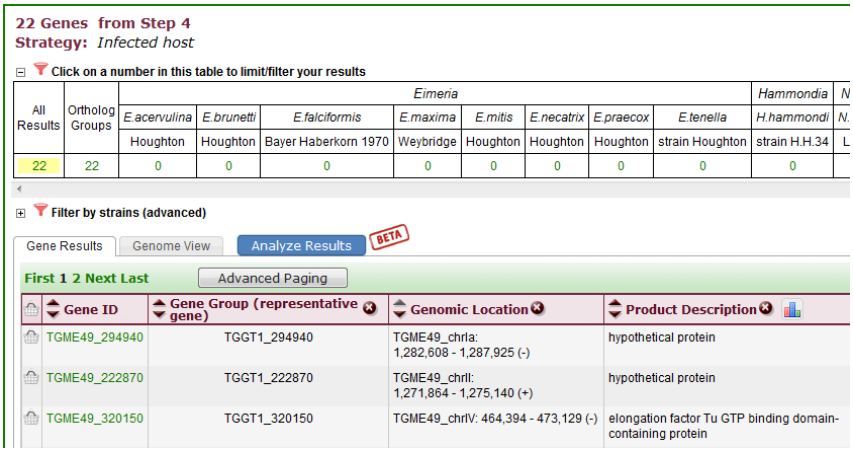
5c. Remove all genes that are also present in the phosphopeptide-depleted fractions (select both intracellular and extracellular).



5d. 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.*

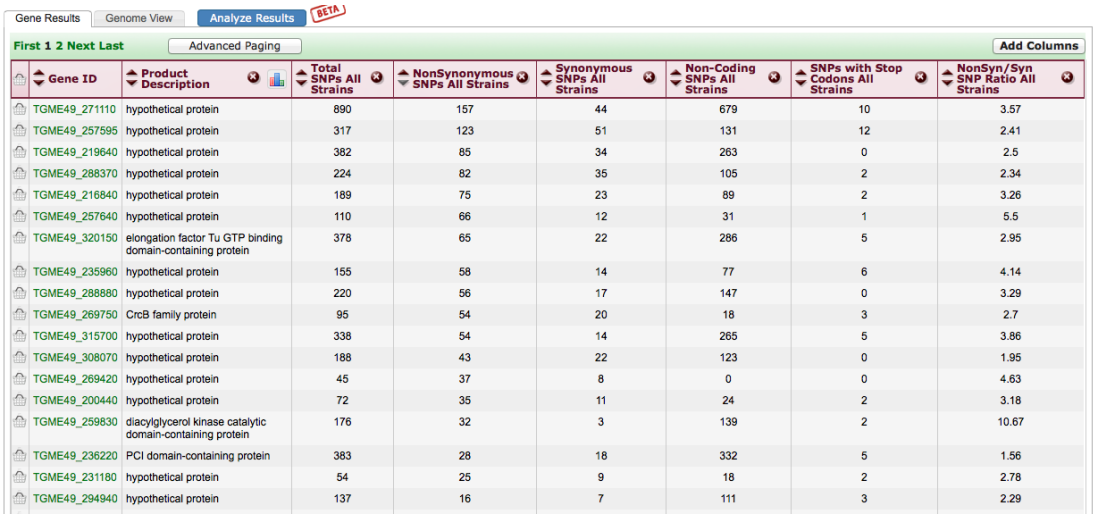
5e. Are any of these genes likely to be secreted? Hint: add a step searching for genes with secretory signal peptides.





5f. 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.

5g. 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