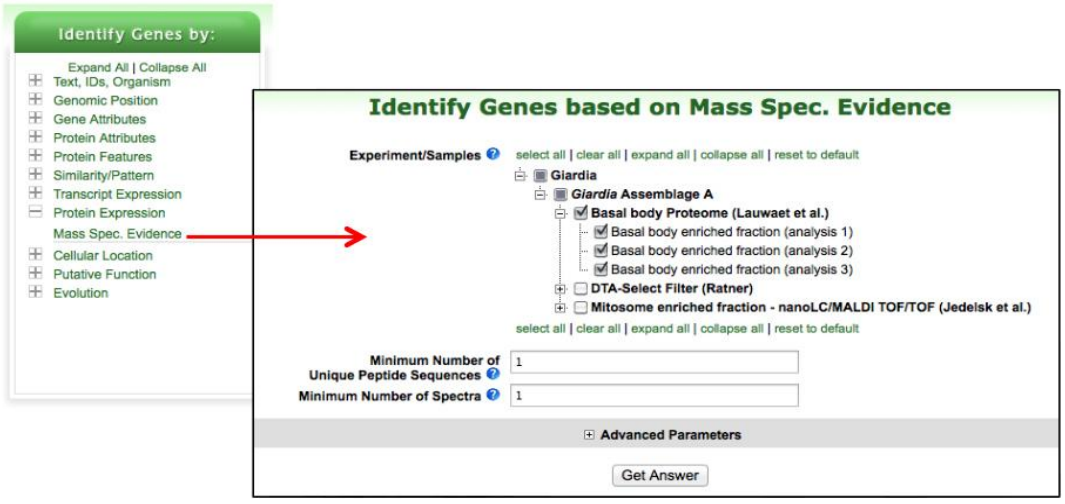
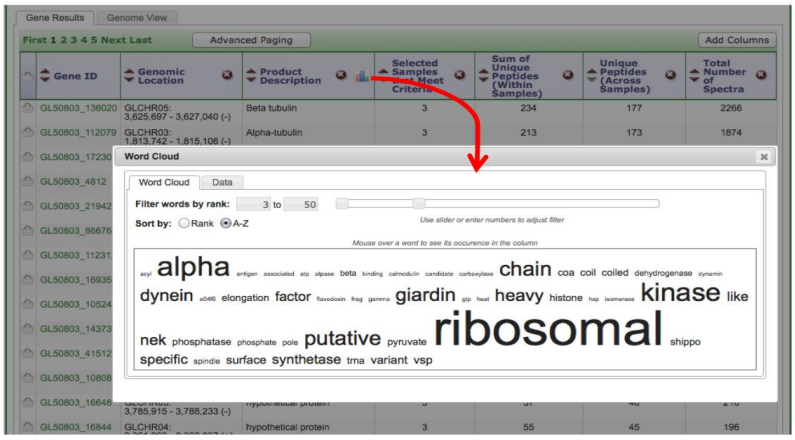
Proteomics

1. Find all Giardia genes with evidence of basal body expression based on mass spec/proteomics data. For this exercise use http://giardiadb.org

   The proteomics search is available under the heading "Protein Expression" in the "Identify gene by" section.

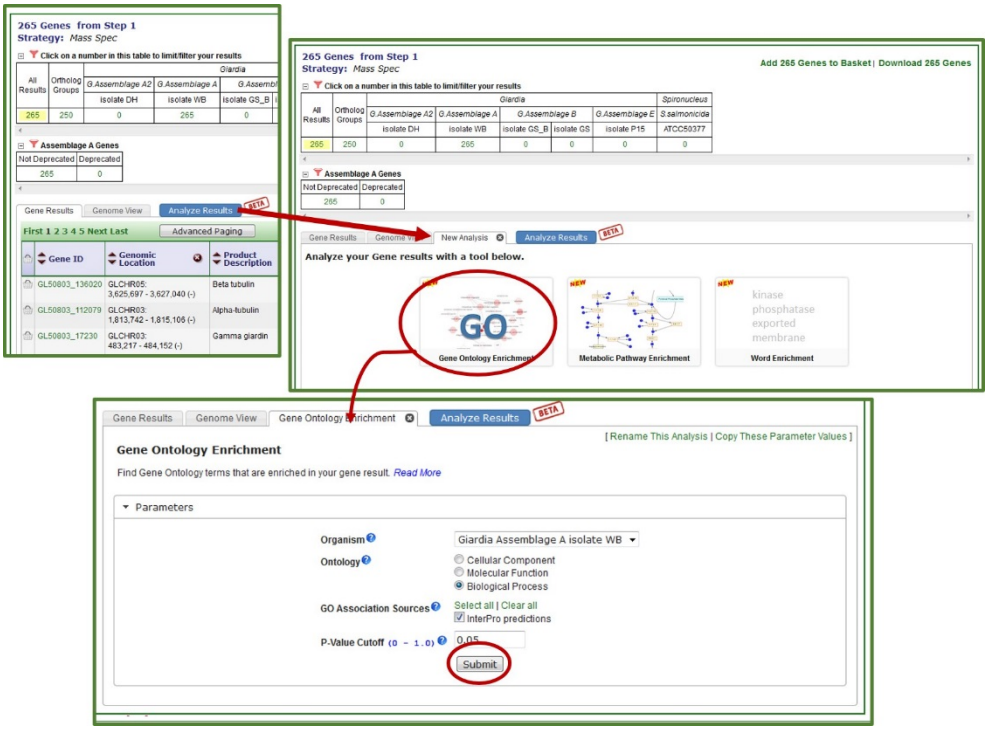


   a. How many genes did you identify?
   b. What kinds of genes are in your result list? Are there genes with similar functions?
      Hint: analyze the Product Description column.
      The **word cloud tool** counts the number of times a word appears in the column and then draws a word cloud in which the size of the word reflects how many times the word appears in the product description column. Click on the little graphic icon next to the column called "Product description".

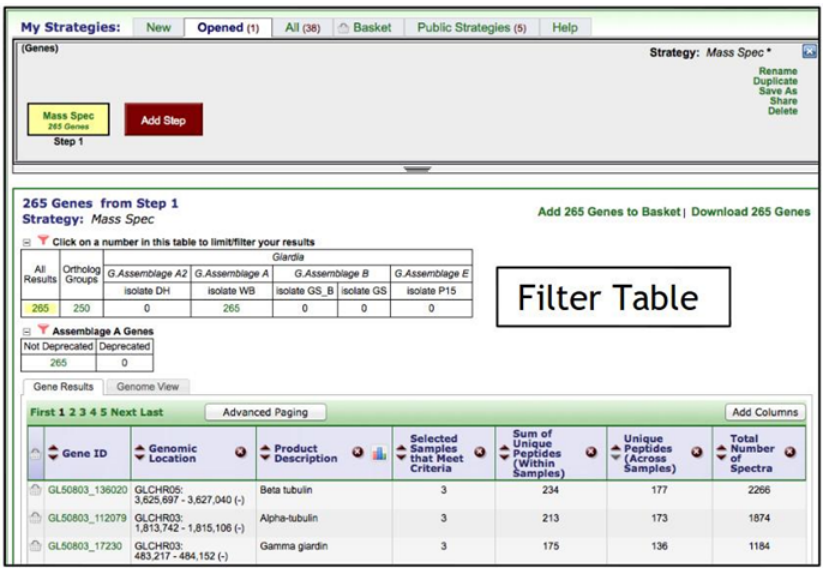


   Hint: analyze the Gene Ontology terms assigned to the genes in your result list.

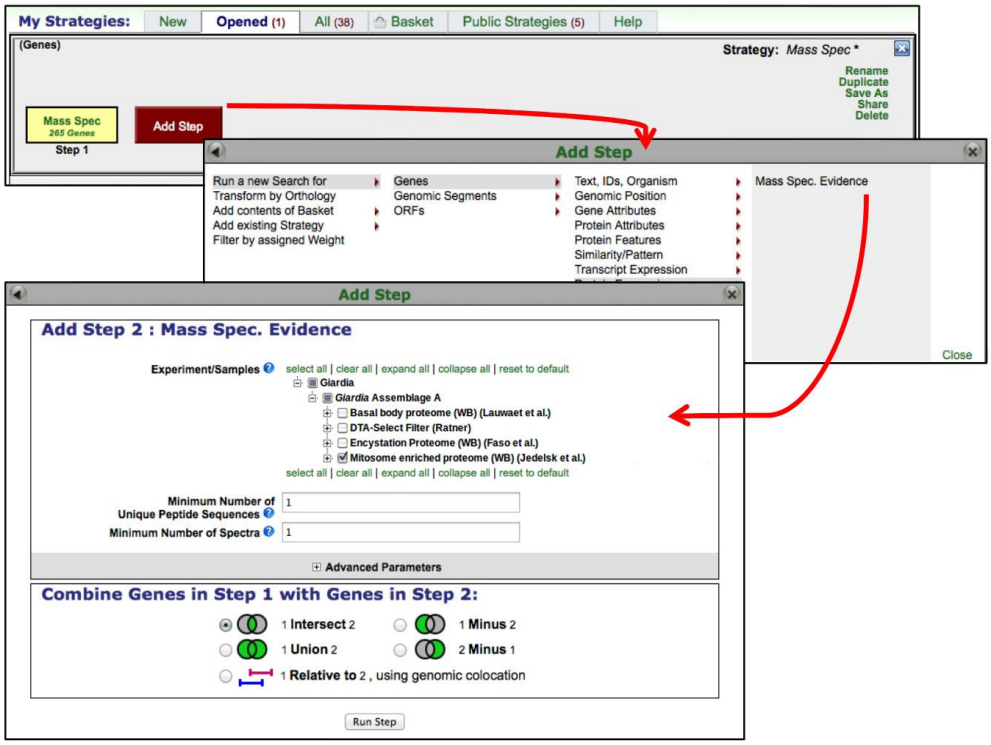
The Result Analysis/Enrichment tool applies the Fischer's Exact test to compare your gene result to the entire genome. Use the Gene Ontology Enrichment to find Biological Process ontology terms that are enriched in your gene result.



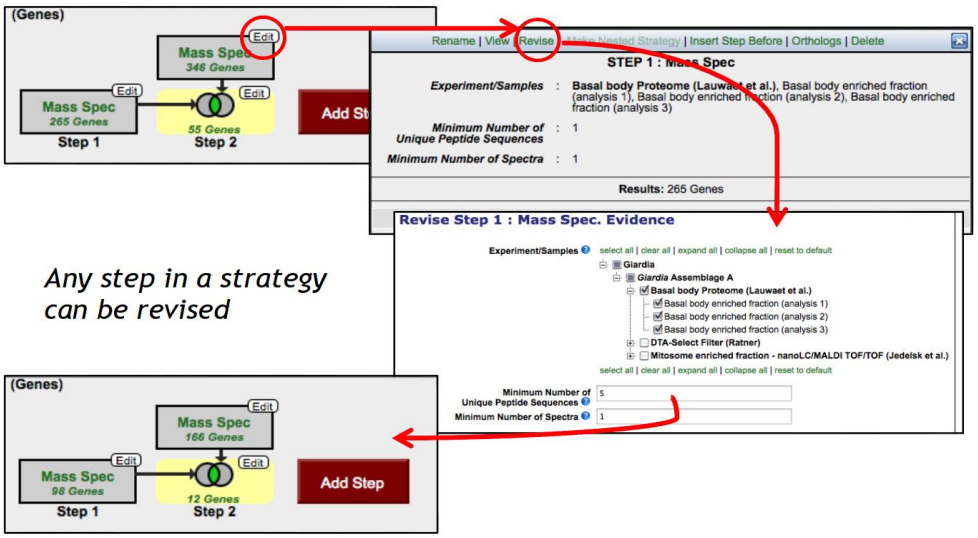
c. Examine the filter table below the strategy section. Why is the number in the ortholog groups column less than the number in the Assemblage A column? Why do the other cells have a zero?



d. How many genes with mass spec data from the basal proteome experiment also have mass spec evidence from the "Mitosome enriched proteome (WB) (Jedelsk et al.)"?

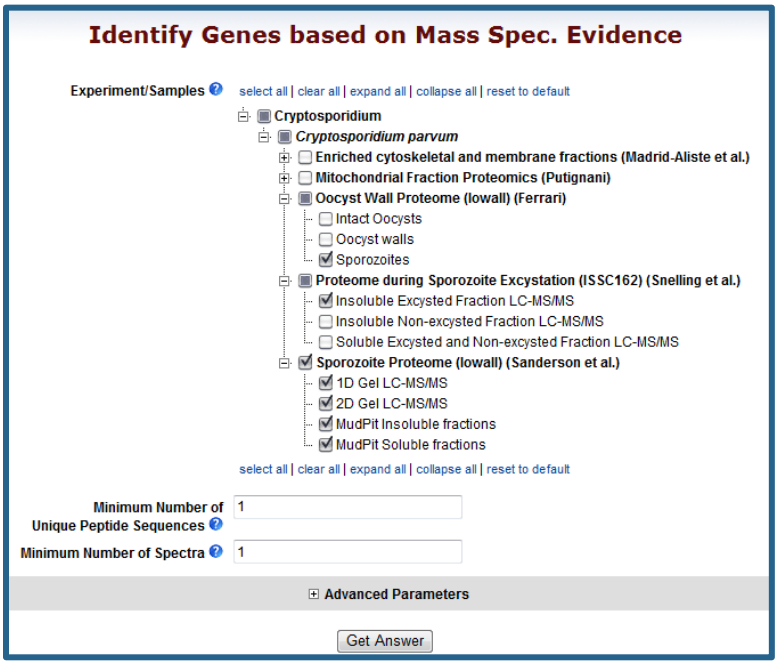


e. The default parameters of the mass spec search are set to identify any gene with at least 1 mapped peptide identified. How will your results change if you **revise both steps** in your search strategy to only **return genes with at least 5 peptides identified**?

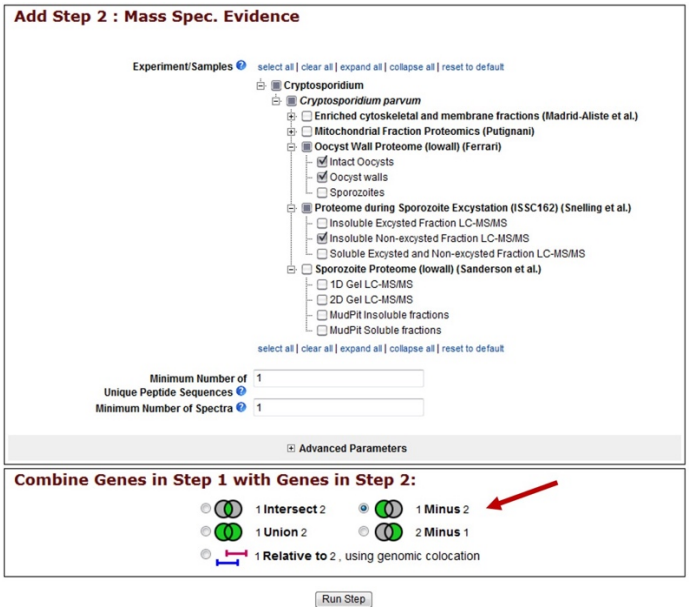


2. Find genes in Cryptosporidium that have mass spec evidence from any of the sporozoite proteomics experiments available in CryptoDB.
For this exercise use http://cryptodb.org

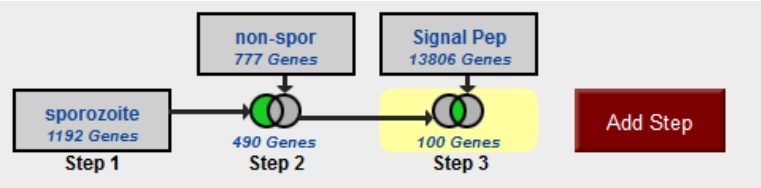
a. Explore the available proteomics data and select samples that make sense. You may need to click on the '+' sign to expand experiments to see the underlying samples.



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.

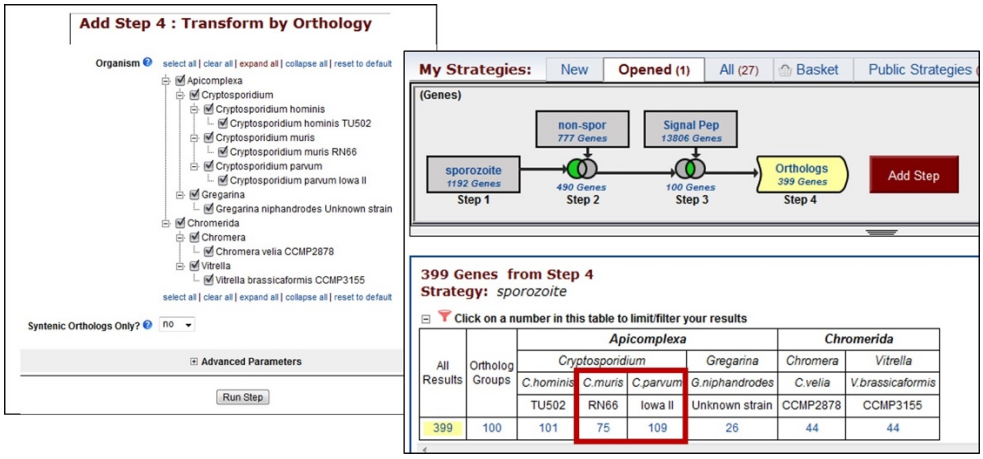


c. How many of these genes are also predicted to be secreted?



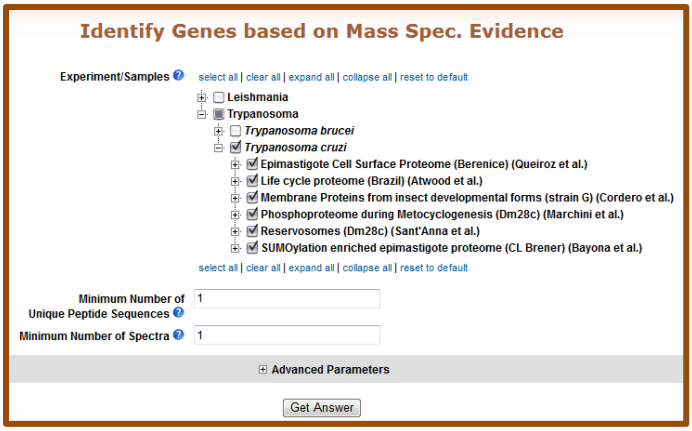
d. So far you have been searching for *C. parvum* genes because we only have proteomics data from this species. However, what if you are studying *C. muris*? How can you garner information about the protein expression of *C. muris* genes from your *C. parvum* results? (Hint: add a step then select the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

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

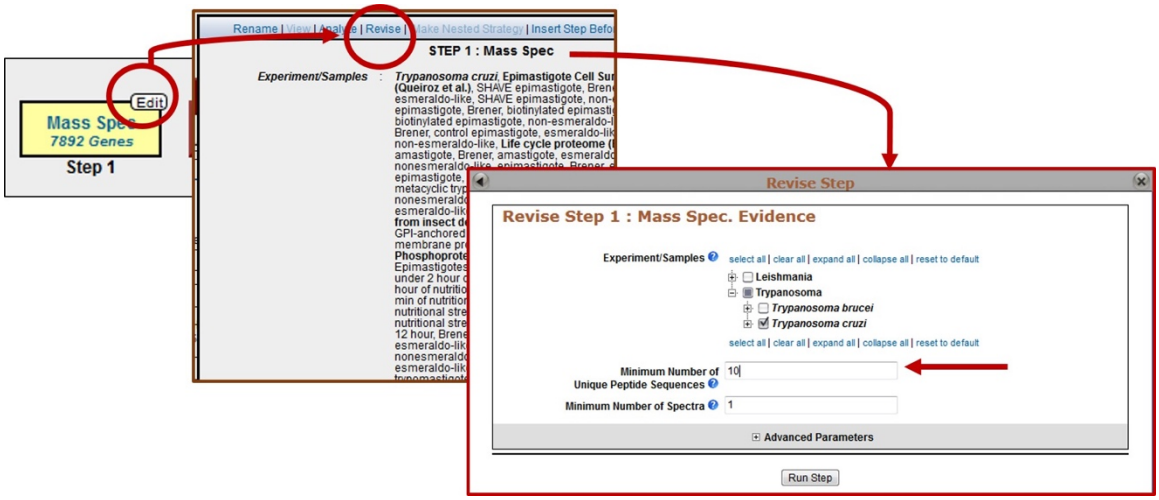


3. Finding all genes with mass spec evidence in *T. cruzi*.
For this exercise use http://TriTrypDB.org

a. How many genes in *T. cruzi* have expression evidence based on mass spec data?

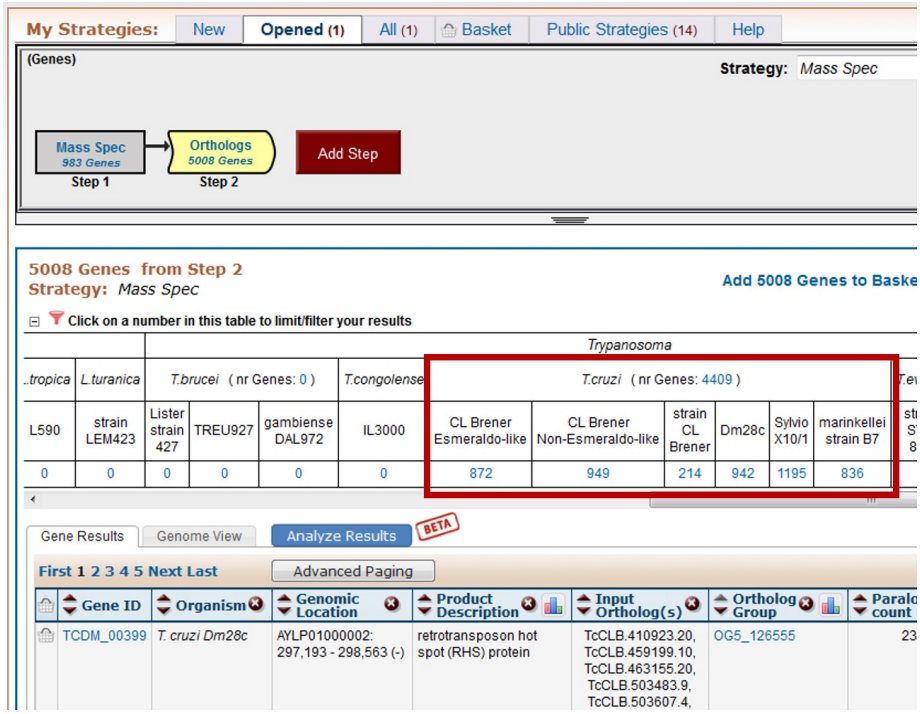


b. How many genes from the results in **'a'** have at least **10 uniquely mapped peptides**? Hint: try revising the step in 'a' and change the "minimum number of unique peptide sequences" option to 10.



c. Expand the list of results in **'b'** to include possible orthologs/paralogs in *T. cruzi*.

Hint: use the ortholog transform option when adding a step and select only *T. cruzi*. Explore the columns in your result set. Pay close attention to the organism filter table.



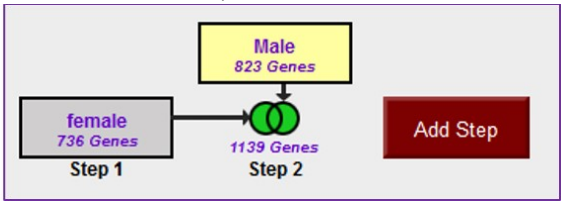
4. Finding genes with evidence for protein level expression in *P. berghei* gametocytes.

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

a. Find all *P. berghei* genes that have mass spec evidence in either or both male and female gametocytes.

- What proteomics experiment and samples did you search? 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)?

Hint: When using a two-step search to find genes that have either or both characteristics, take the union of the two searches, not the intersection.

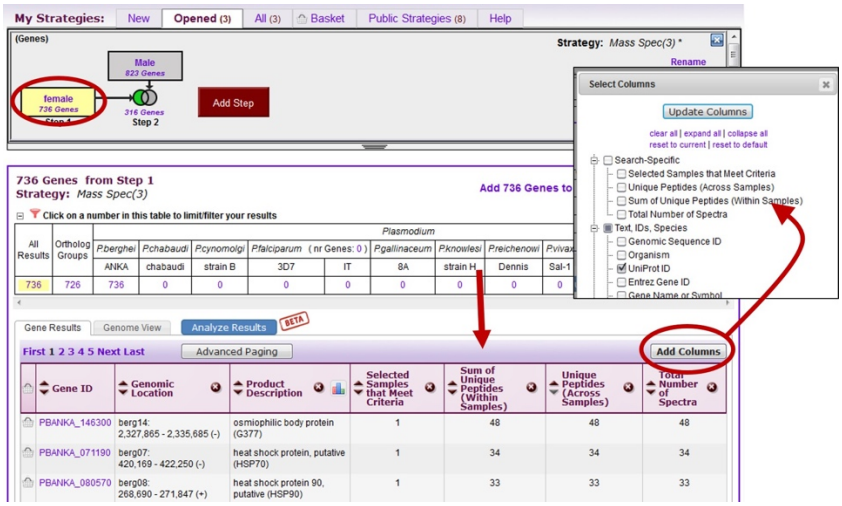


b. Find all genes that have mass spec evidence in both male and female gametocytes. Hint: use the strategy you developed in (ii) to get this answer, but change the union into an intersection.

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? Focus on the female gametocyte search step. Hint: look at the "Sum of Unique Peptides (Within Samples)" column in the list of results.



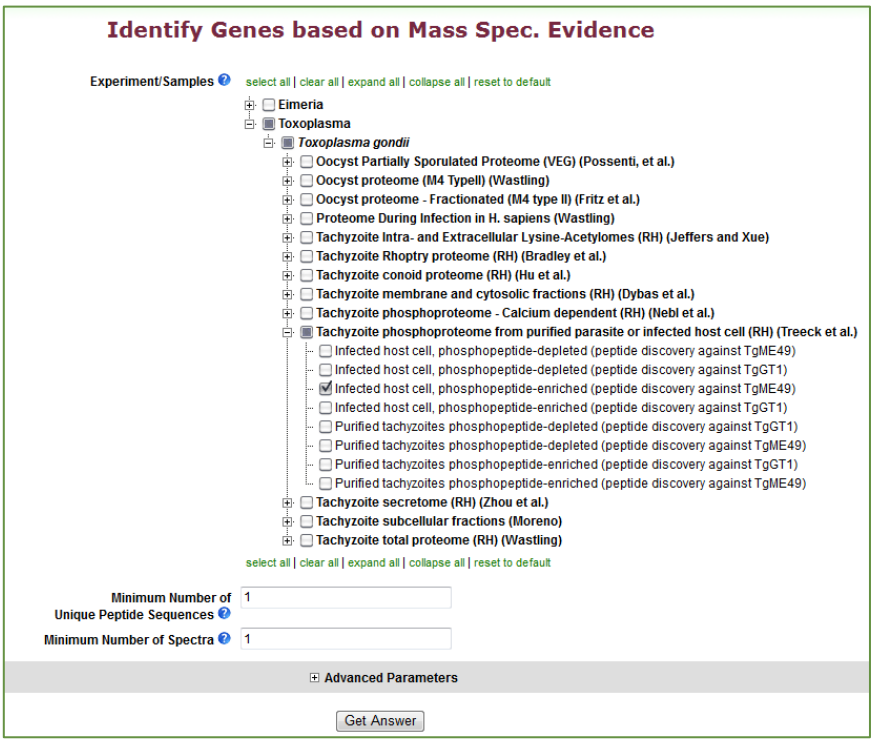
f. What does the distribution of peptides in the gene from 'e' look like? Hint: go to the gene page and look at the "Protein features" section, or go to the genome browser from the gene page and turn on the right tracks.

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

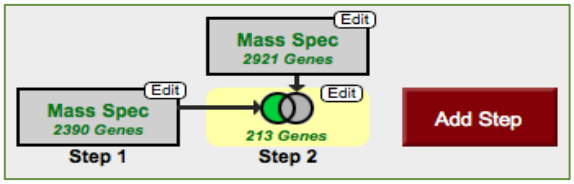
For this exercise use http://www.toxodb.org

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

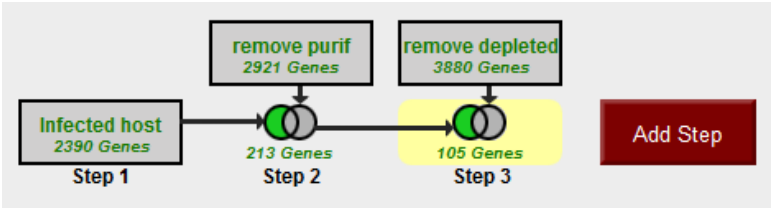
a. Find all genes with evidence of phosphorylation in intracellular tachyzoites. Select the "**Infected host cell, phosphopeptide-enriched** (peptide discovery against **TgME49**)" sample under the experiment called "**Tachyzoite phosphoproteome from purified parasite or infected host cell (RH) (Treeck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



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

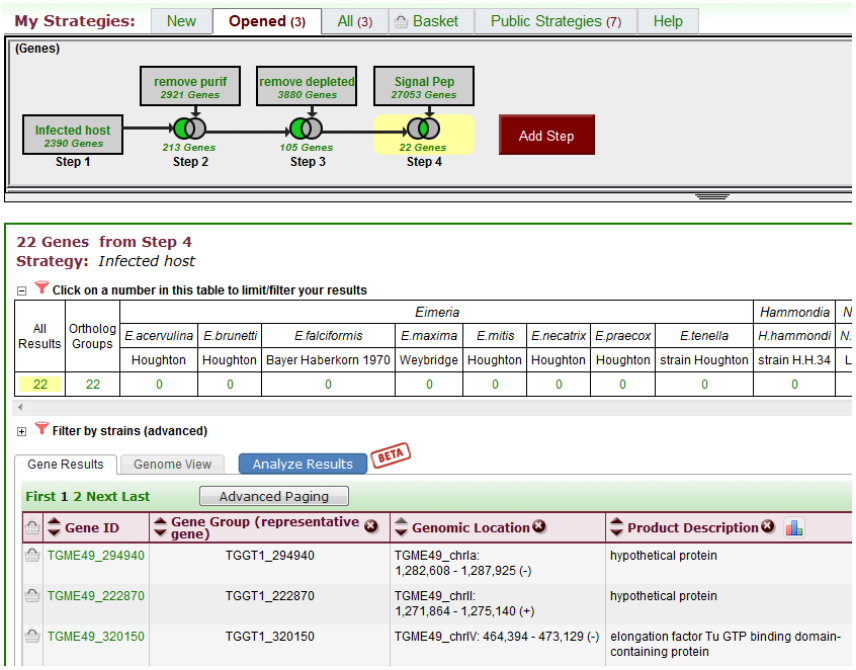


c. Remove all genes that are also present in the phosphopeptide-depleted fractions (select both intracellular and extracellular).



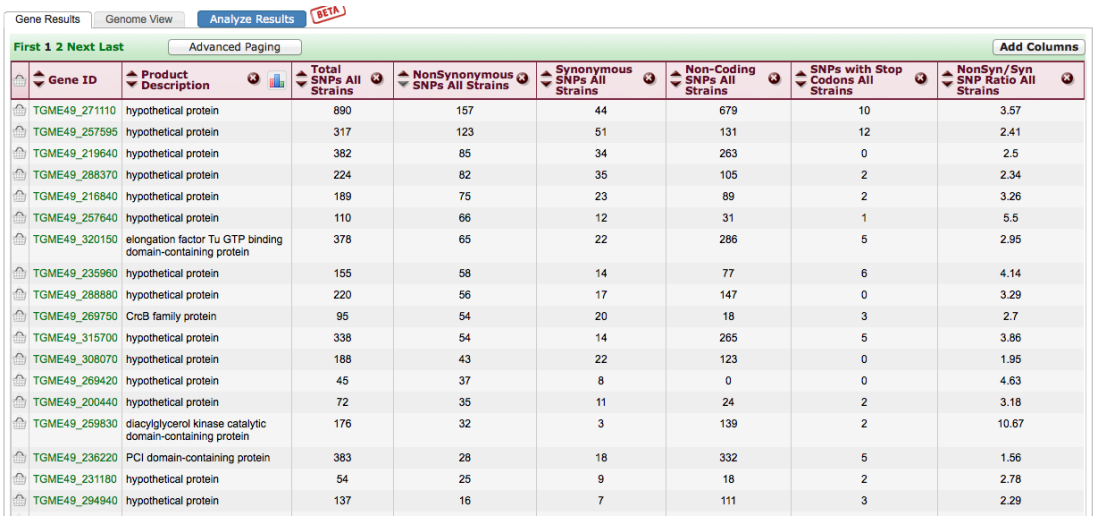
d. Explore your results. What kinds of genes did you find? Hint: use the Product description word column or perform a GO enrichment analysis of your results.

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



f. 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.

g. 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 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