Expression Searches: Transcriptomics and Proteomics

Learning Objectives

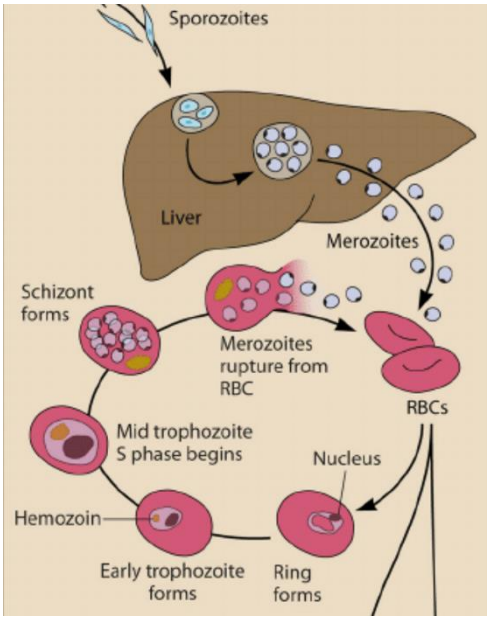
- Review the types of expression searches in VEuPathDB.
- Use the RNA-Seq differential expression, fold change and percentile searches to explore gene expression in liver stage *Plasmodium berghei* infections.
- Use the Proteomics peptides and quantitative searches to explore expression at the protein level.

Transcript expression or the abundance of an mRNA, can be determined in the laboratory with several different techniques including RNA-sequence, microarray, and RT-PCR. VEuPathDB supports these data types with several searches (see table below). For RNA sequence data, expression values are graphed on gene pages and mapped reads can be visualized in the genome browser. Proteomics experiments can determine the peptide sequence by mass spectrometry or the abundance value via e.g. isobaric tagging for a peptide in a sample. Each data type is available on gene record pages but only the mass spec peptides have data that can be viewed in JBrowse. Using the search strategy system, it's easy to delve deep into a specific data set and to take advantage of several types of data when combining search results in the strategy system.

Search	Description	RNA-seq	Micro-array	Proteomics
Differential Expression	Statistical analysis of studies whose experimental design includes biological replicates. A differential expression search finds genes based on fold change difference between two samples with a user defined p-value cutoff. Only pairwise comparisons can be made with this search.	✔		
Fold Change	Expression differences between samples are calculated but statistical analyses are not performed. A fold change search finds genes whose expression value differs between samples without considering statistical parameters. This search offers a form of differential expression analysis when the experimental design did not include replicates and allows for comparing groups of samples, e.g. find genes whose expression is up-regulated in the liver time course (2, 24, 36, and 54 hours) vs the control (0 hours).	✔	✔	
Percentile	For each sample in an experiment, each genes' expression value is sorted from lowest to highest and a percentile rank is determined. For example, a percentile search can find genes whose expression is in the highest 10% of expression values within a sample.	✔	✔	
Sense/Antisense	For strand-specific RNA sequence, expression values are determined in the sense and antisense direction. This search finds genes that exhibit simultaneous changes in sense and antisense transcripts. For example you can look for genes with increasing antisense transcripts and decreasing sense transcripts, as might occur when antisense transcription suppresses sense transcription.	✔		
Splice-site Location	This trypanosome-specific search takes advantage of the 'splice-leader' RNA-seq data which determines transcript abundance within the polycistronic mRNA using splice-leader specific primers. This search identified genes whose 5' splice site location varies between samples.	✔		

Metacycle	The MetaCycle package detects rhythmic signals from large scale time-series data, such as circadian rhythms within expression time courses, using either ARSER or JTK-Cycle. This search returns genes whose rhythmic signals match the conditions (period and amplitude range) you specify. The search will return the corresponding period, amplitude and p-value of genes that meet your search criteria.	✔	✔	
Similarity	The similarity search returns genes whose expression profile within the experiment follow a similar pattern as the gene you specify.	✔	✔	
Direct Comparison	Microarray data for two samples is often collected on the same glass slide. For these experiments, the direct comparison search returns genes whose expression varies between samples in pairwise comparisons.		✔	
Co-expression	Meta-analysis across multiple microarray experiments defined a co-expression network. This search returns genes within the co-expression network of your gene(s) of interest.		✔	
Mass spec. evidence	Peptides from proteomics experiments are mapped to a reference genome enabling searches for genes based on the mapping			✔
Quantitative mass spec. evidence	Quantitative proteomic experiments produce abundance values for proteins identified in the sample analyzed. (excel spreadsheets of gene/protein IDs and expression intensity values)			✔
Post-translational modification (PTM)	PTM data from proteomics experiments, excel spreadsheets of gene/protein IDs, location and type of modification, are associated with genes in VEuPathDB, enabling searches for genes based on the type and number of the PTM.			✔

1. **Find genes that are up-regulated in the later liver stages of *Plasmodium* infection.** PlasmoDB.org



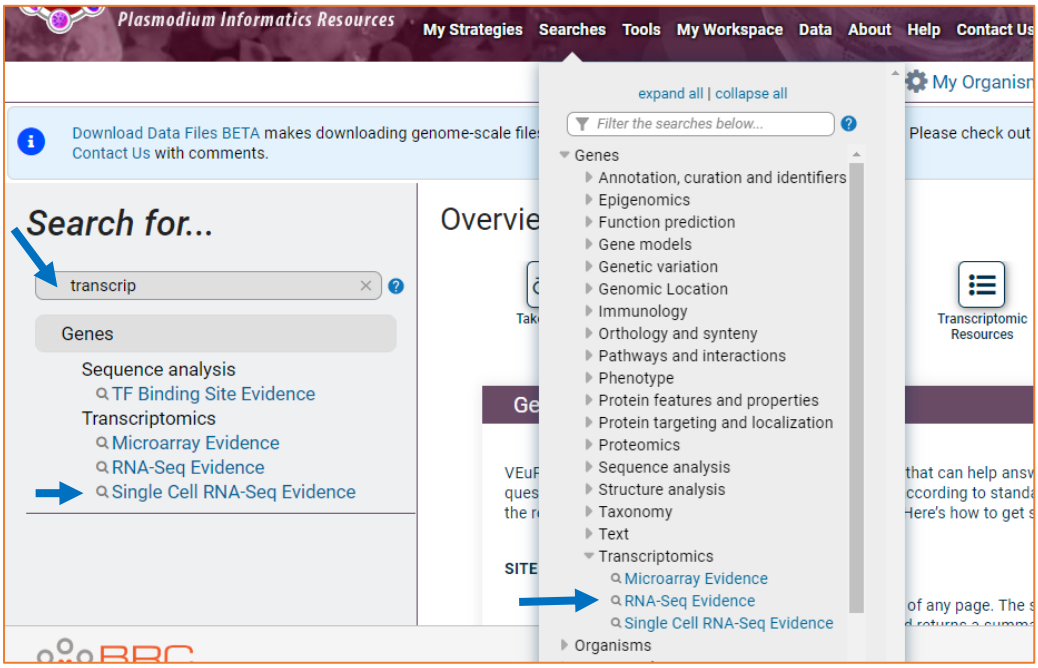
The life cycle of *Plasmodium* is split between the sexual mosquito stage and the asexual host phase. The host phase includes a 6-7-day asymptomatic liver stage which ends with the release of merozoites into the bloodstream where they infect erythrocytes. The erythrocytic stages are well studied compared to the liver stages.

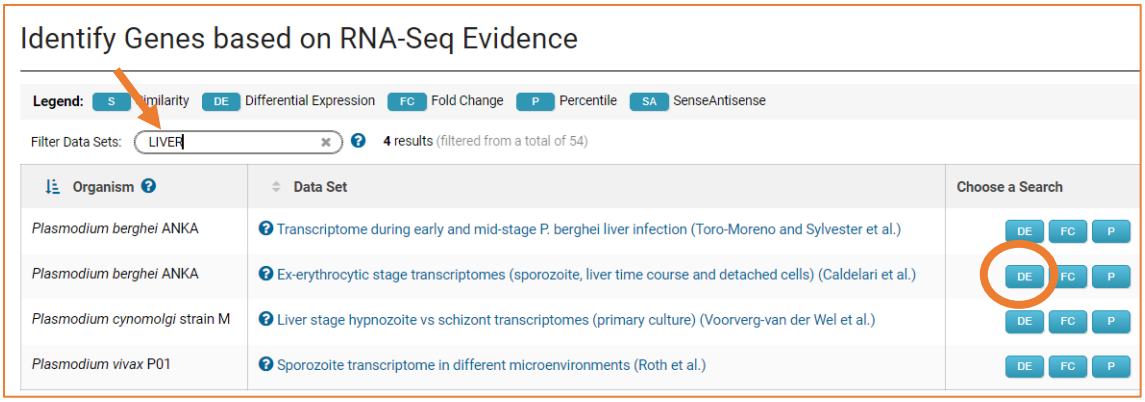
PlasmoDB contains RNA seq data from a study in the rodent model *Plasmodium berghei*, that includes a time course of liver infection as well as sporozoite and merozoite samples for comparison. (Caledlari et al. 2019) Seven samples were assayed in triplicate for RNA sequence:

1. Sporozoites
2. 6 hr liver infection
3. 24 hr liver infection
4. 48 hr liver infection
5. 54 hr liver infection
6. 60 hr liver infection
7. Merozoites (detached cells).

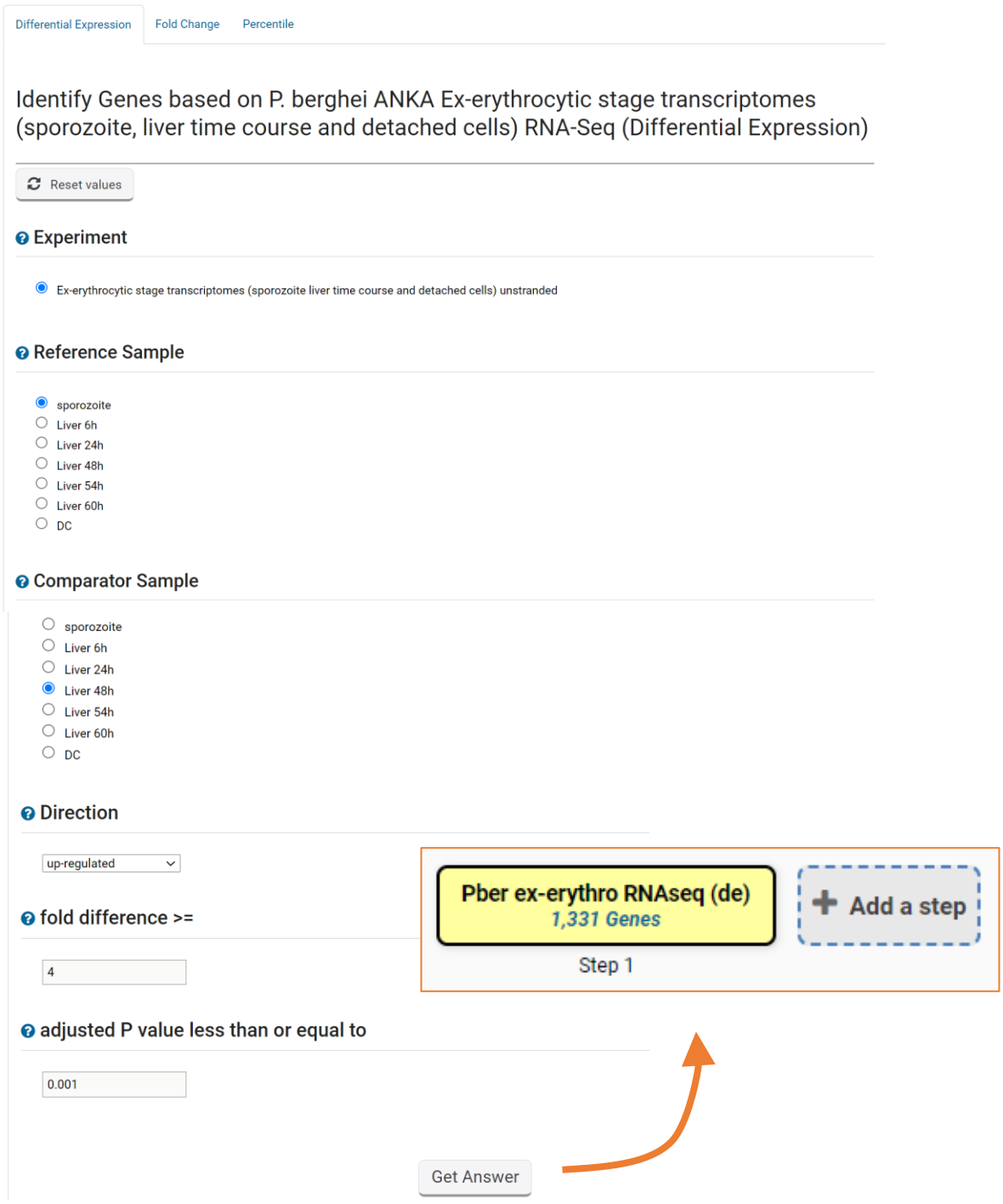
Use this data set to determine what genes are upregulated at least 4 fold (p-value <= 0.001) at 48 hr post infection vs the sporozoite stage.

a. Navigate to the RNA seq search page and find the data set called **Ex-erythrocytic stage transcriptomes (sporozoite, liver time course and detached cells) (Caldelari et al.**). Searches are available from the Search For… menu on the left side of the home page, as well as the Searches drop down menu in the header.



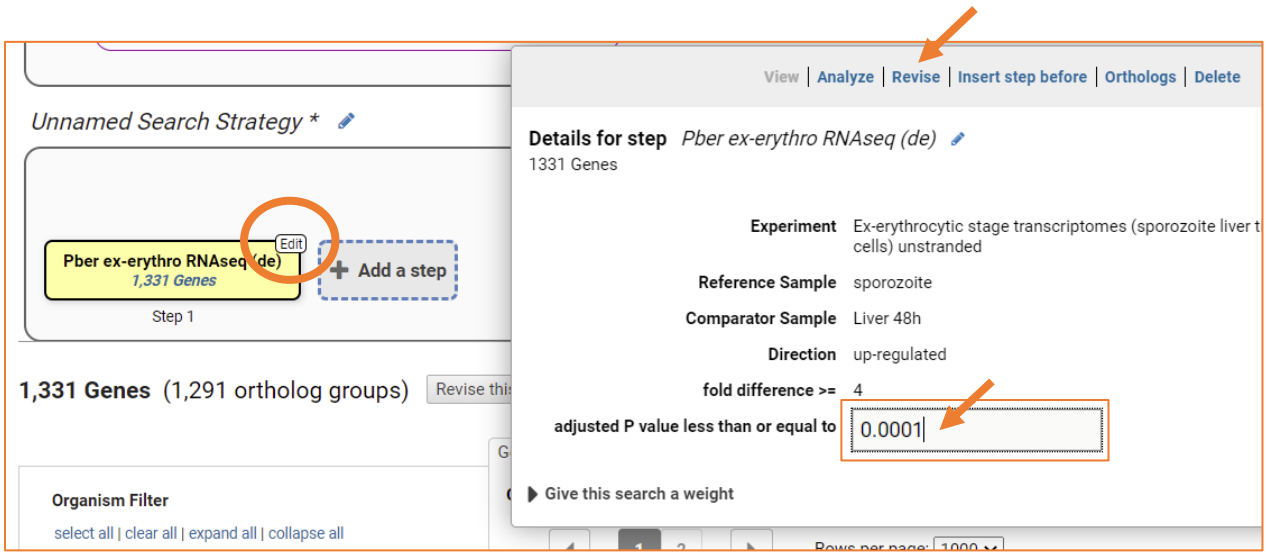


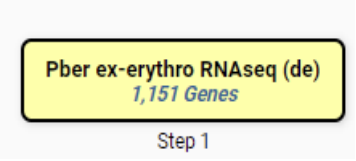
b. Arrange the differential expression search to return genes that are at least 4 fold up-regulated in the 48-hour liver infection compared to sporozoites with a p-value of p<0.001.



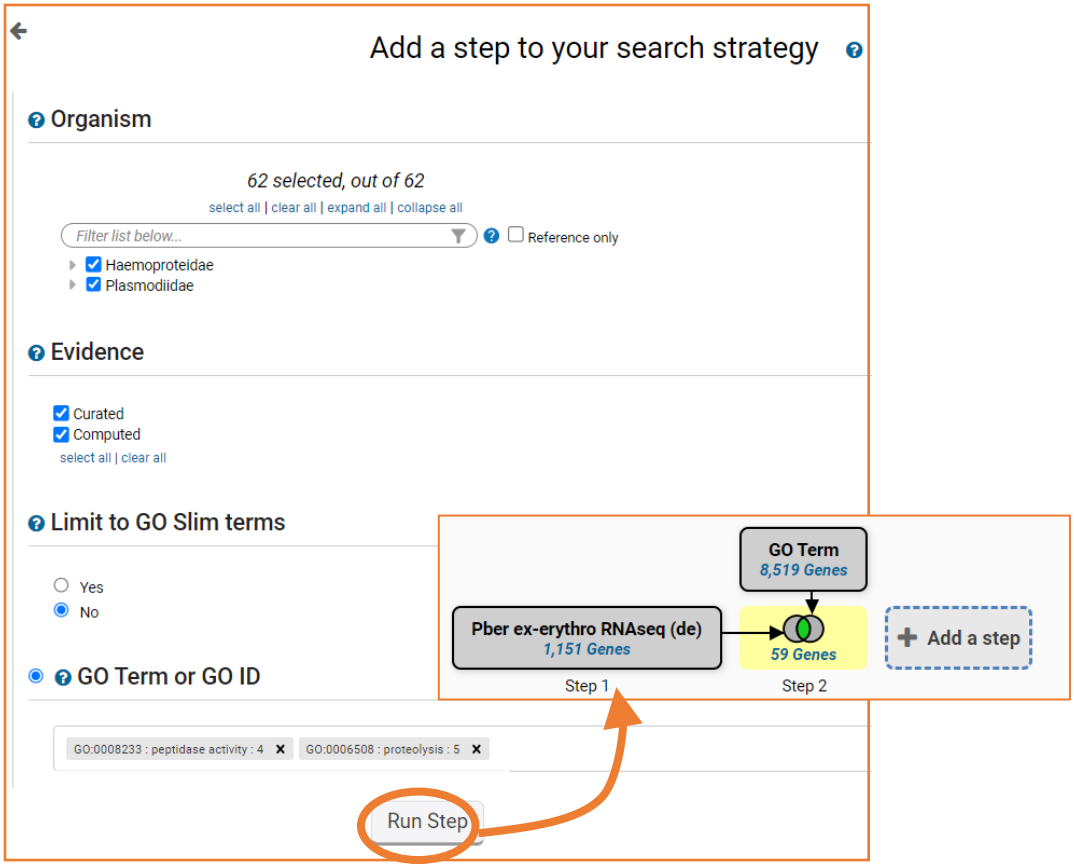
c. How many genes were returned by the search? Do you believe these results? To convince yourself, you could browse the product description column. Are there clues that these genes are liver-specific?

d. Increase the statistical stringency of the search from p <= 0.001 to p<0.0001. How many genes are returned by the search now? Hint: revise the search and change the p-value. Hover over the yellow search box until the Edit icon appears. Click the Edit icon and choose revise from the options panel.





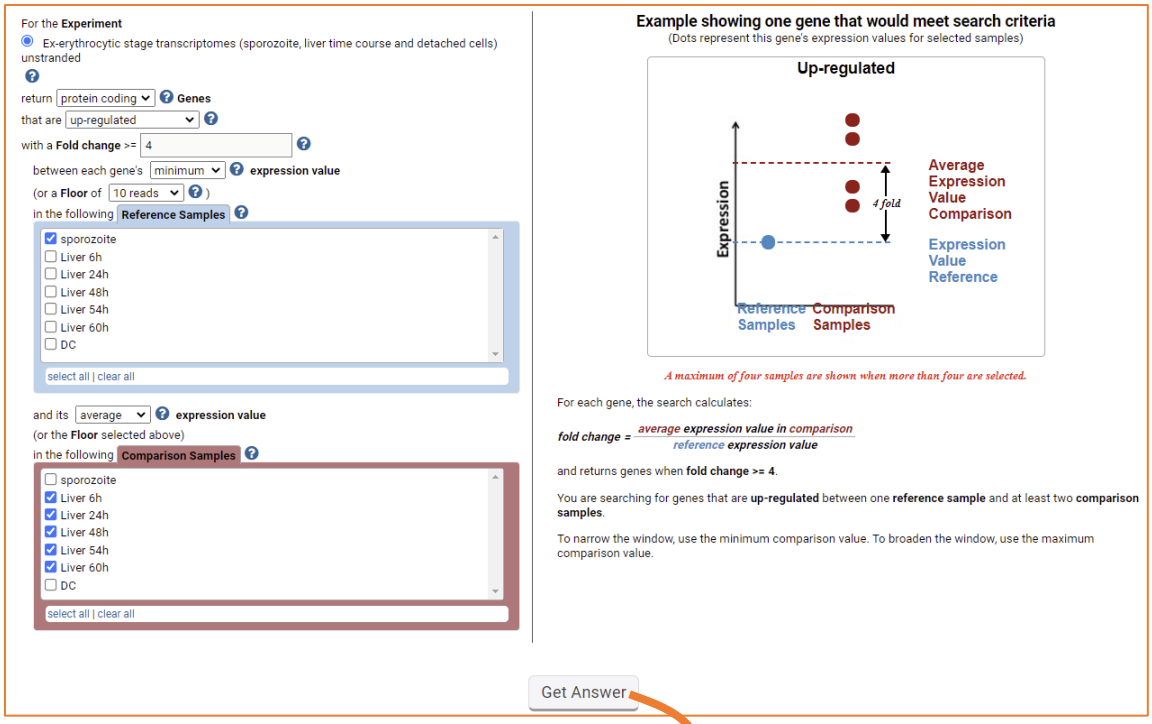
e. What other properties would you expect of a late liver stage gene/protein? Since the next step is to emerge from the hepatocyte, these genes may have proteolytic activity. Intersect your RNA seq search with a GO term search to see if any of your genes are annotated with proteolytic or peptidase activity. (GO:0008233 peptidase activity GO:0006508 proteolysis) How many genes have these activities?



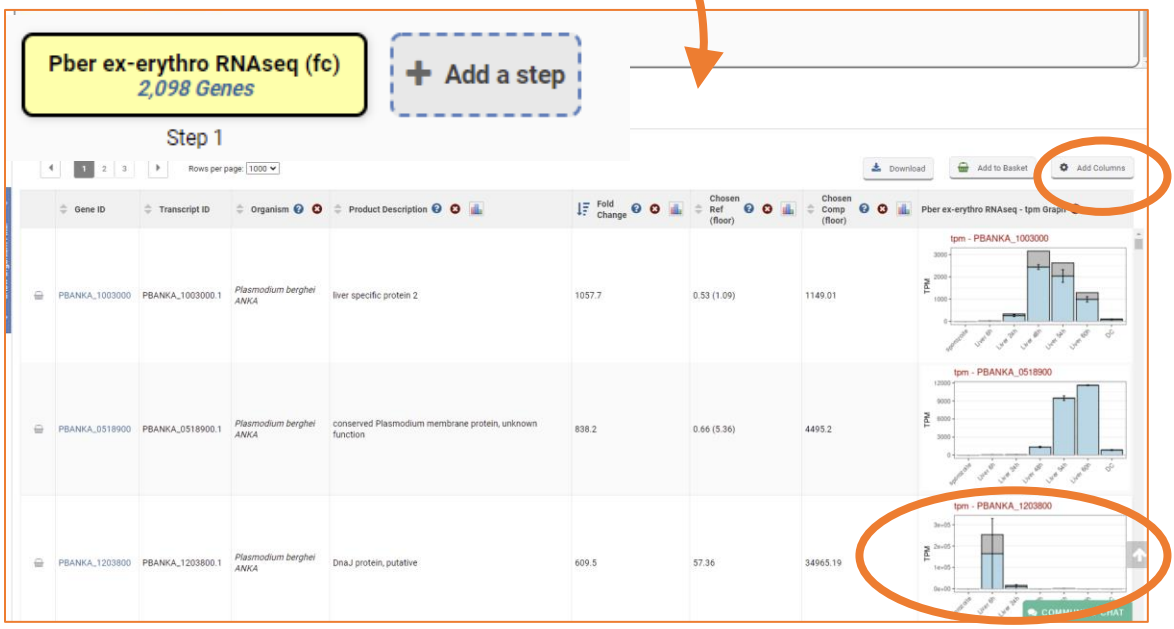
2. **Using the same data set, find genes that are upregulated 4 fold in any liver stage compared to sporozoites**. Hint: use the Fold change search to compare the 6, 24, 48, 54 and 60-hour time points to sporozoites.

a. Navigate to the RNA Seq search page and choose the Fold Change search for the **Ex-erythrocytic (Caldelari et al 2019)** data set as in 1a above.

b. Arrange the fold change search to return genes that are up-regulated in the average expression across the liver stages compared to the sporozoites. (do not use the DC sample)



Get Answer

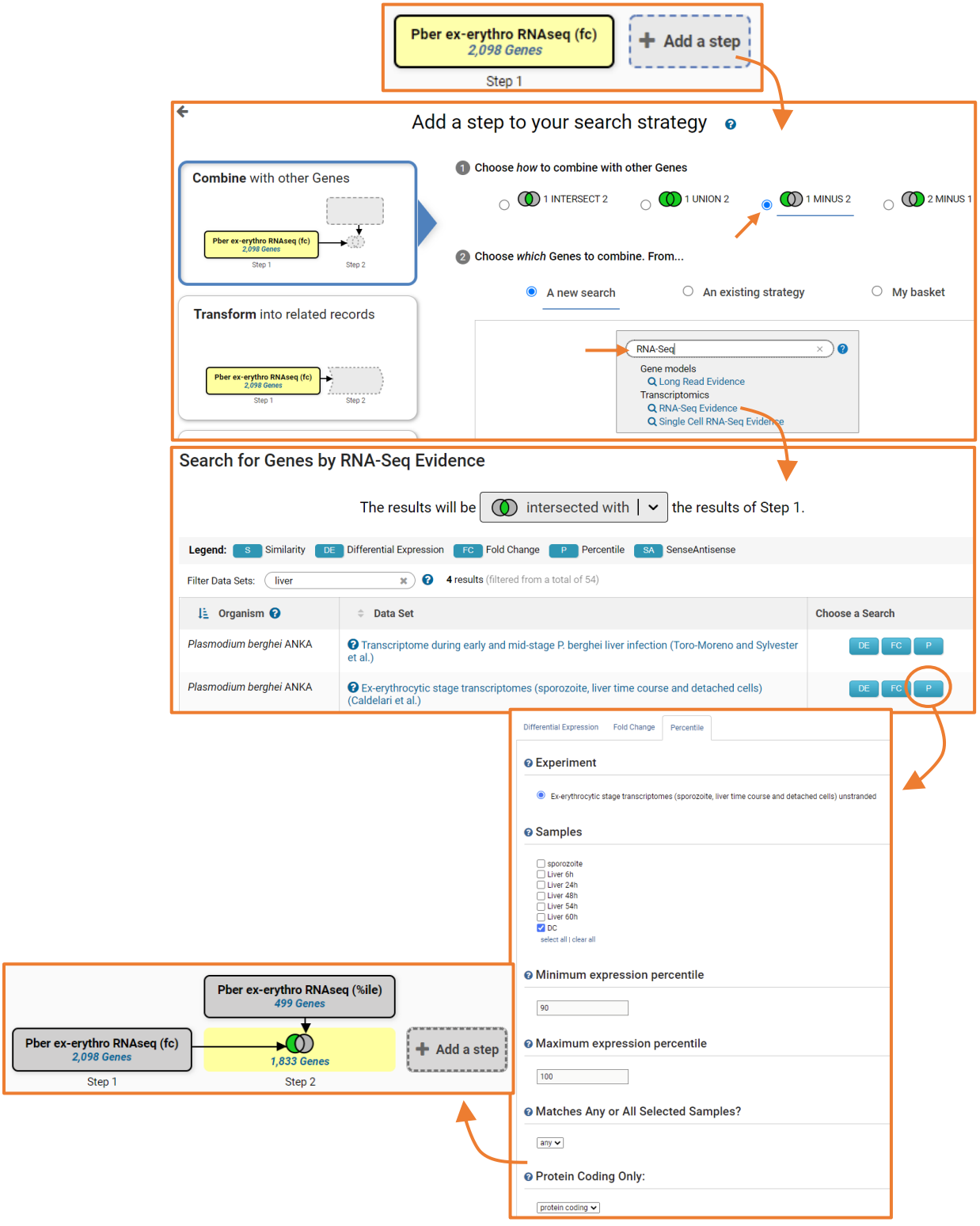


c. Explore your results. Did the search return more genes or fewer genes than the differential expression search that employed a statistical analysis along with fold change values?

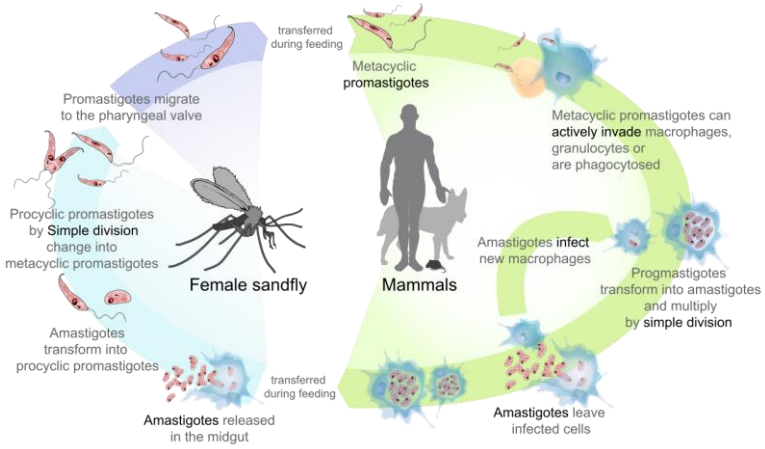
d. Use the Add Columns to turn on the TPM graph for the 'Ex-erythrocytic stages' data set. Notice the error bars for the DNAJ protein PBANKA_1203800. Would this gene be returned by the Differential Expression search that applies statistics before returning genes?

e. Use the **Percentile search** to determine which genes in this result are also expressed in the top 10% of genes in the merozoite (detached cells) sample AND remove them from the result? Hint: Add a

step to the strategy that intersects your current result with search that returns the 90-100th percentile genes of the merozoite sample and choose 1 minus 2 for the operator.



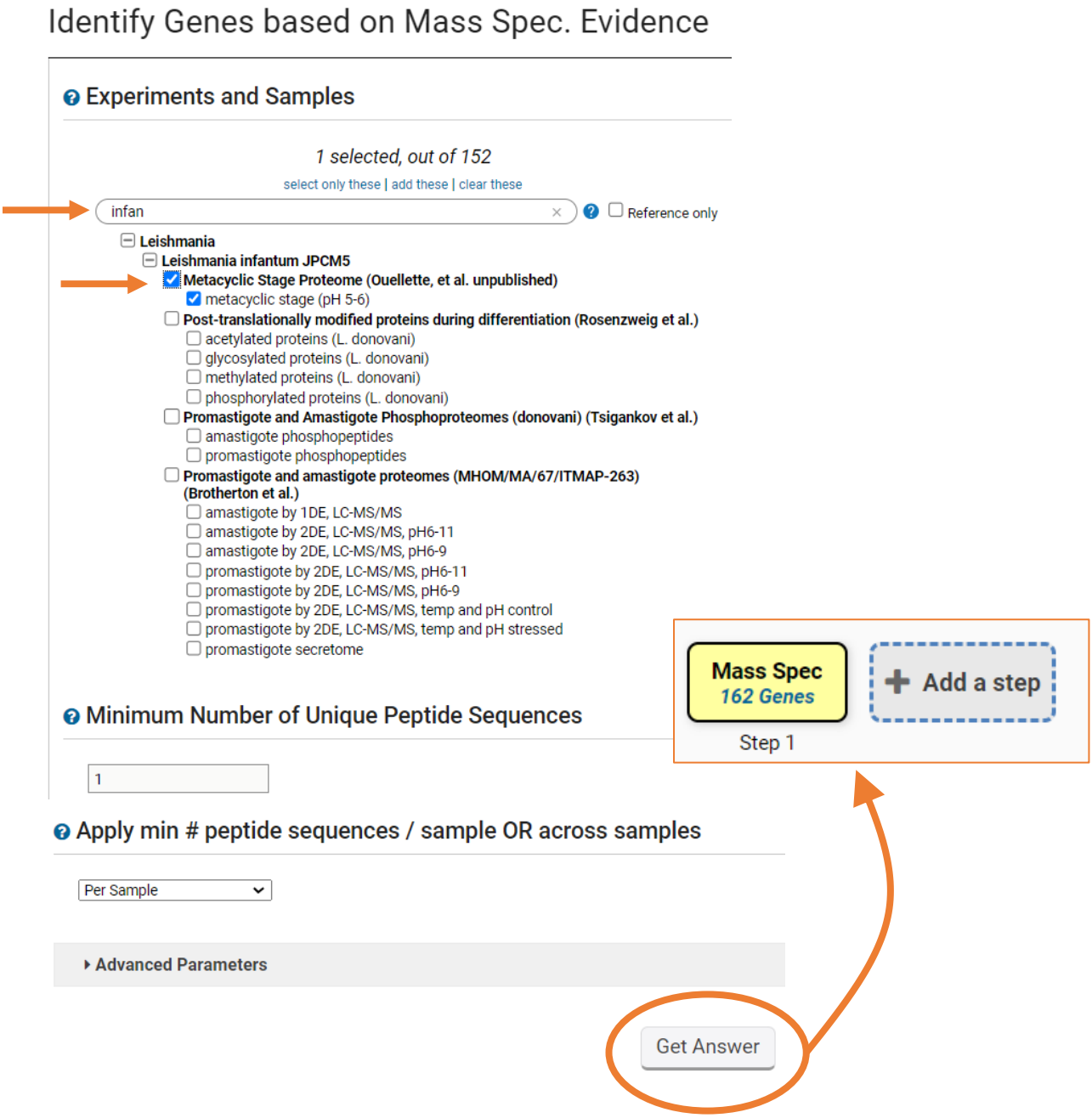
3. **Find genes that have peptide evidence from metacyclic stages but not amastigote or promastigote stages of *Leishmania infantum*. Note: for this exercise use http://tritrypdb.org**



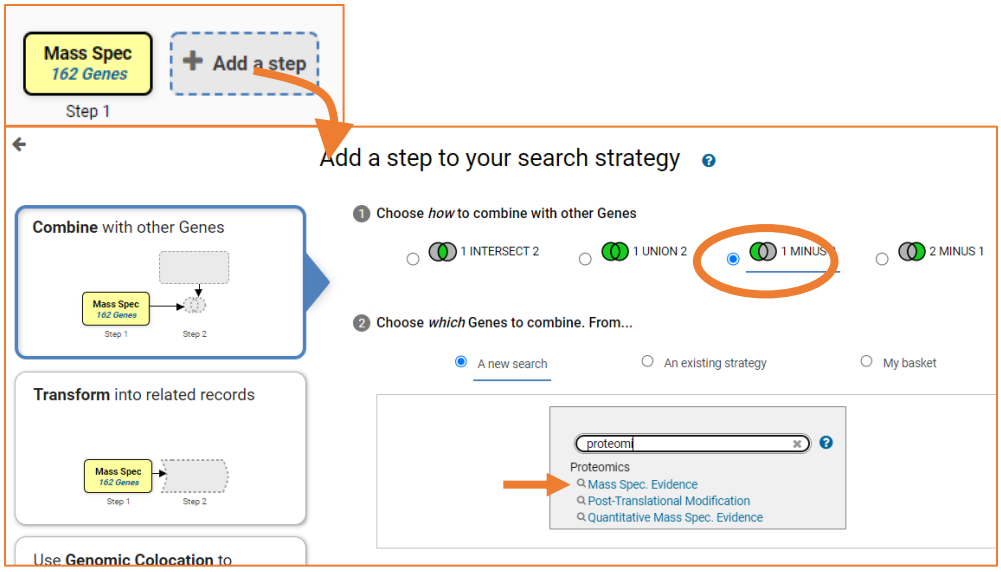
Life cycle of Leishmania. https://commons.wikimedia.org/wiki/File:Leishmaniasis_life_cycle_diagram_en.svg

a. Navigate to the mass spec. evidence search. This search returns genes whose protein products mapped to peptides found in proteomics experiments.

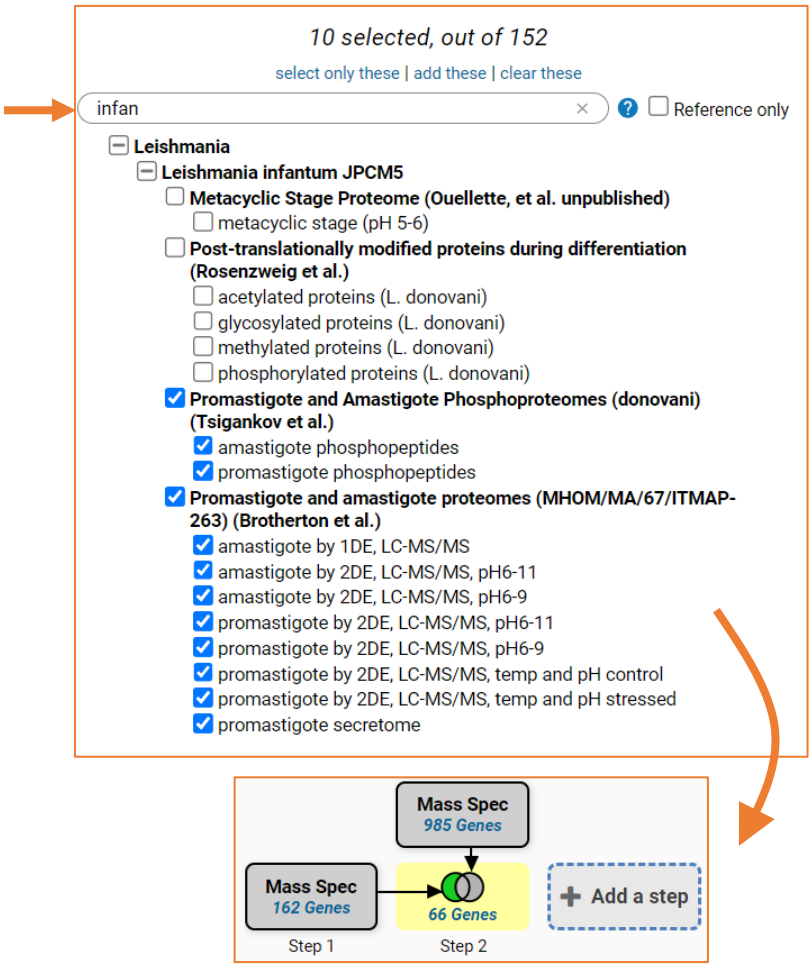
b. Filter the experiment and sample tree by typing a word in the filter box. **Select all *L. infantum* samples that come from the metacyclic stages.** Keep the default search parameters and click on the Get Answer button.



c. How many genes did you get?
d. Now subtract the genes that have protein expression in the amastigote and promastigote stages. Add a step to your strategy that returns amastigote and promastigote genes and choose the 1 minus 2 operator to combine the searches.



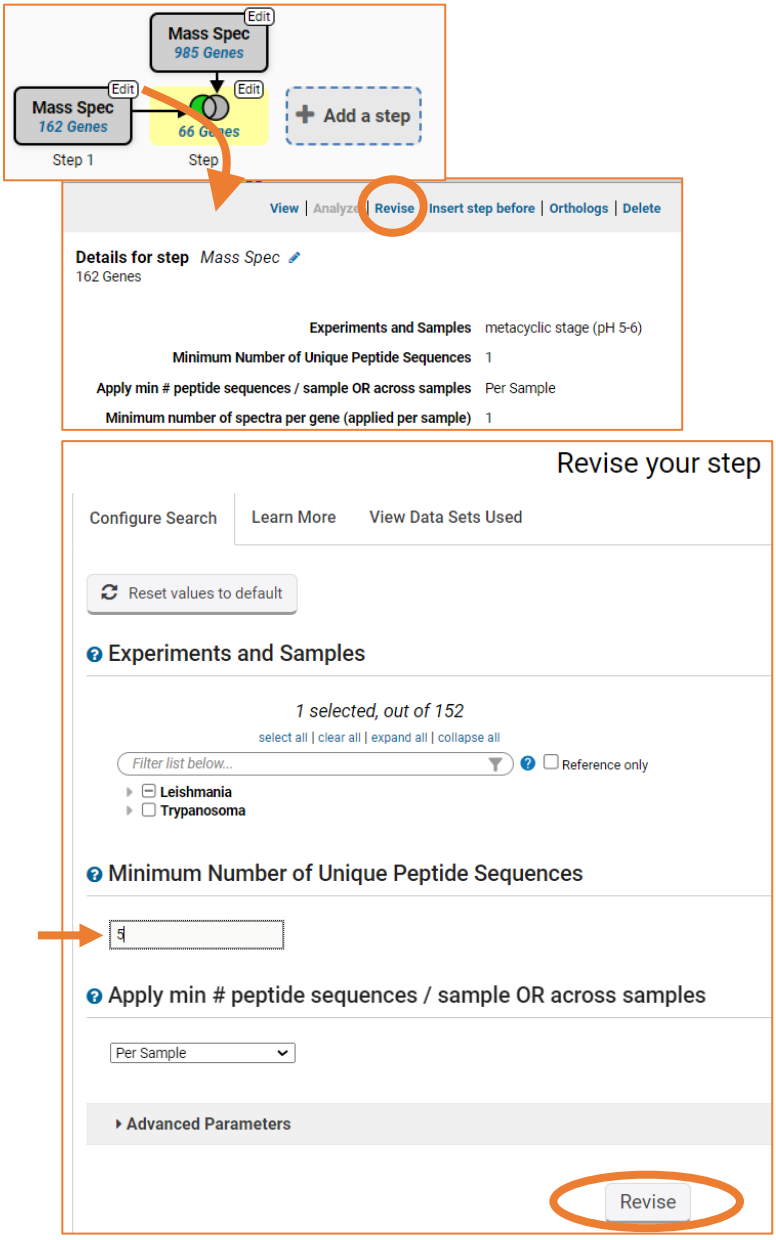
e. Choose all the *L. infantum* samples labeled amastigote and promastigote and run the search



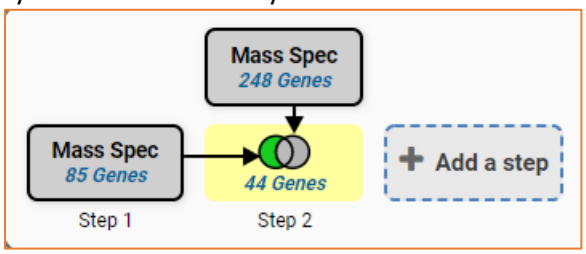
f. Visit the gene pages of some of your results. There you can view mapped peptides and data from other experiments. (https://tritrypdb.org/tritrypdb/app/record/gene/LINF_340044200#ProteinExpressionPBrowse)

g. How can you increase the stringency of your results? One way is to increase the minimum number of unique peptides that are required to map to a gene before it is returned by the search. The default settings that we used above return any gene with a minimum of one peptide.

-Click on the edit button
-Click on the revise option in the popup
-Change the value from 1 to 5 and click on the Revise button.
-Remember you need to do this for each step.



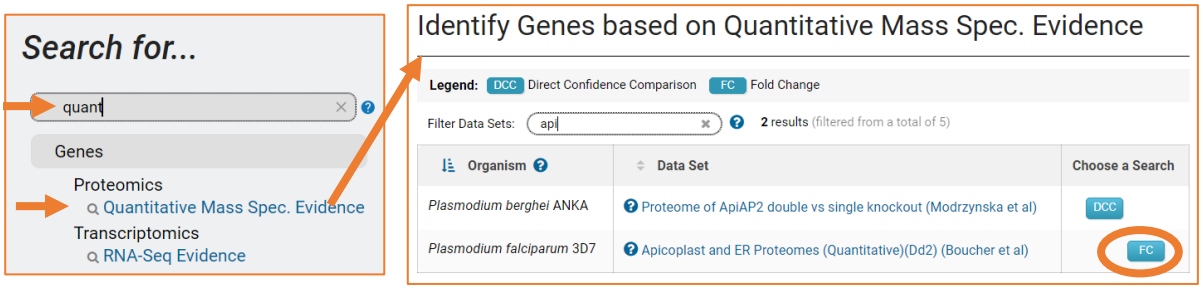
h. How did this change your results? Would you consider these results more stringent?



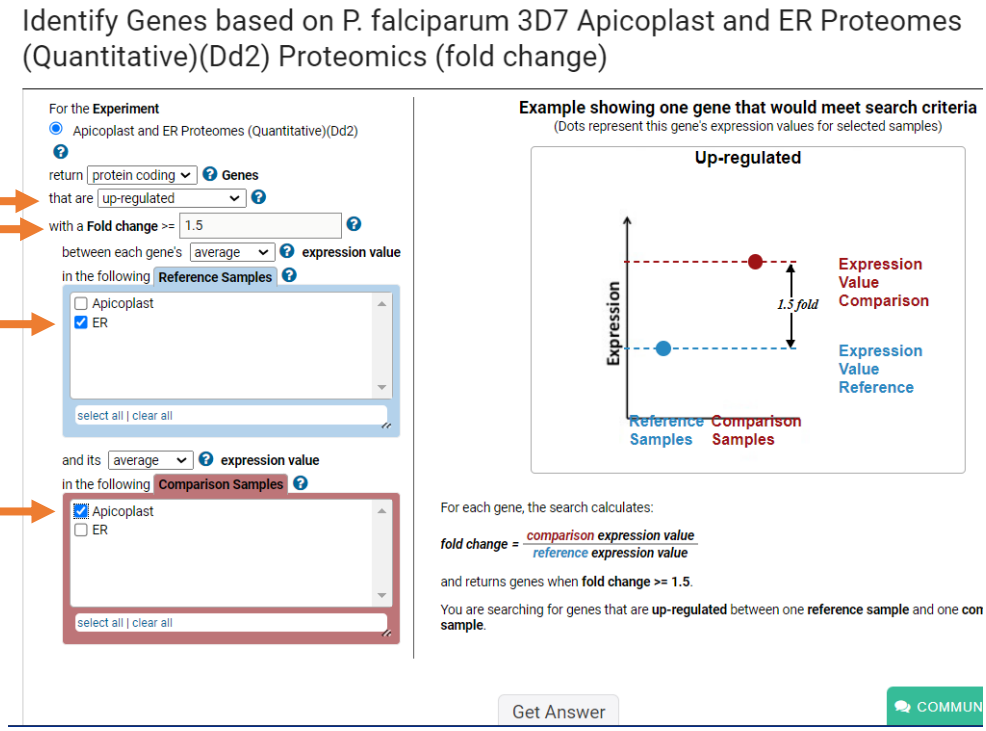
4. Find genes in *Plasmodium falciparum* that are present at a higher concentration in the apicoplast compared to the endoplasmic reticulum (ER). Note for this exercise use https://plasmodb.org

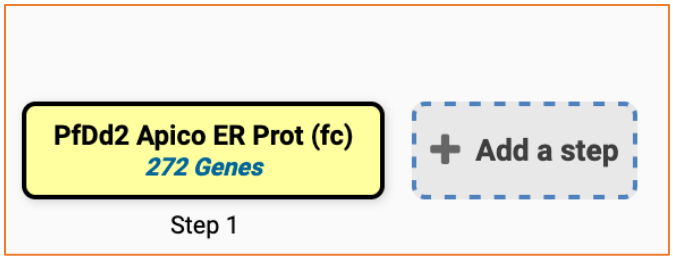
a. Go to the quantitative mass spec evidence searches

b. Select the experiment called Apicoplast and ER Proteomes (Quantitative)(Dd2) (Boucher et al)

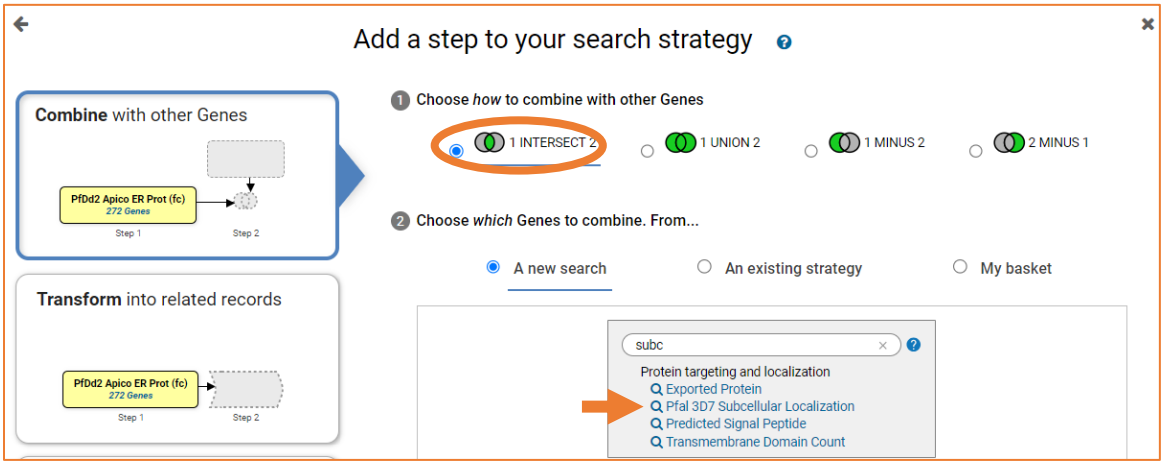


c. Configure this search to return all genes that are upregulated by 1.5 fold in the apicoplast sample compared to the ER sample.

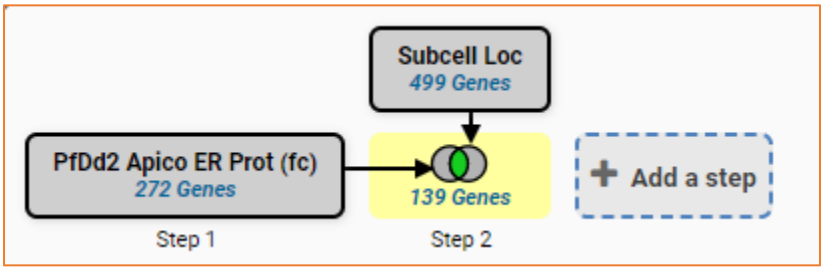




d. Can you leveraging other data about apicoplast biology to validate your results? For example, it is well known that proteins with transit peptides are targeted to the apicoplast. PlasmoDB has a search that returns genes with the transit peptides. Add a step to your strategy that increases the stringency of evidence for these genes being apicoplast genes. The search is called 'Pfal 3D7 Subcellular Localization'



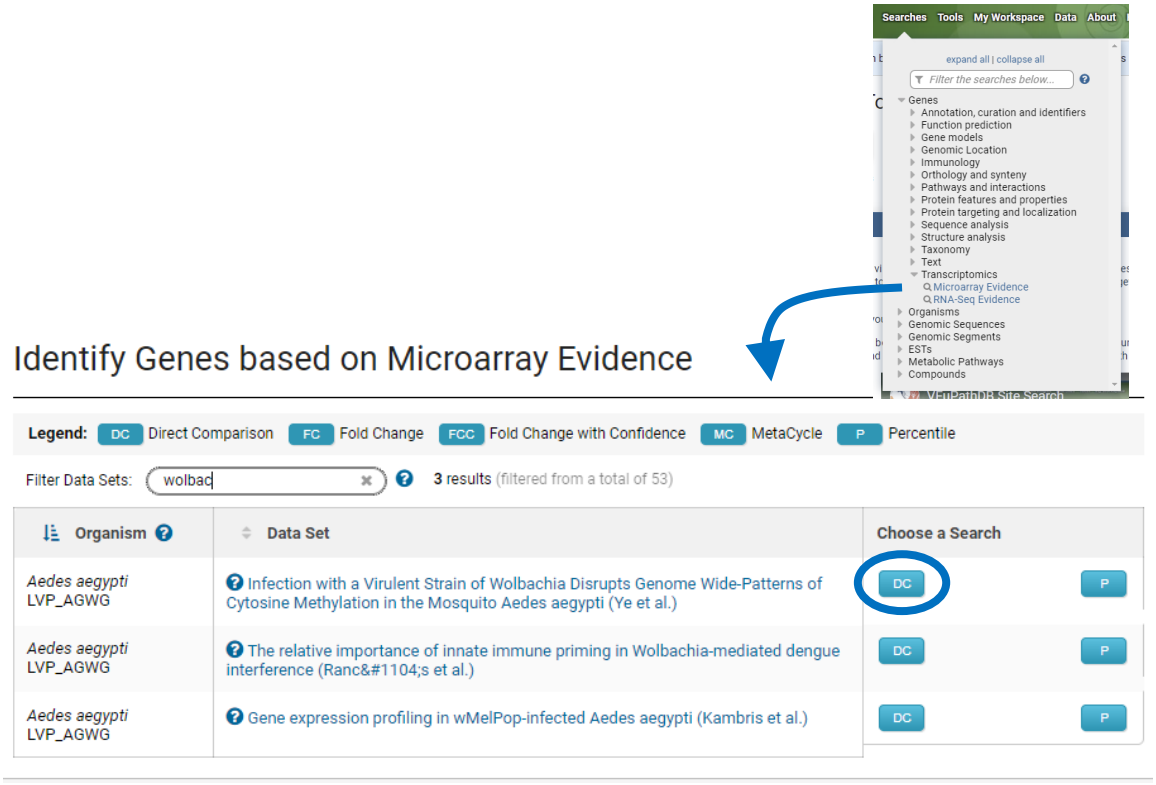
e. Make sure Apicoplast localization is selected and click on the Run Step button. How many genes did you identify? Are you more confident that these genes are apicoplast genes? How would you use the PlasmoDB tools to boost your confidence in these so called apicoplast genes?



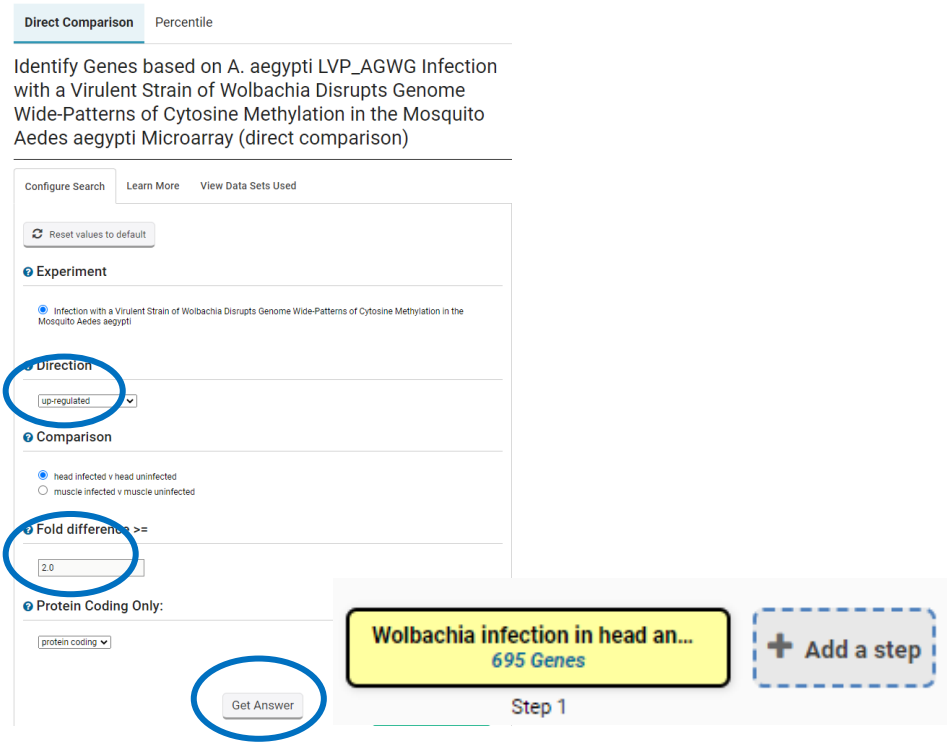
OPTIONAL exercises below concerning transcriptomics and proteomics. Result numbers in images may not match the current database

5. **Find Aedes aegypti genes that are upregulated in both head and muscle during infection with *Wolbachia***. The *Wolbachia* strain wMelPop, which reduces longevity in *Drosophila melanogaster*, has been introduced into the Dengue virus mosquito vector, *Aedes aegypti* as a strategy to reduce disease transmission. VectorBase has a microarray data set that compared *Wolbachia* infected and uninfected mosquito head and muscle. This exercise uses VectorBase.org.

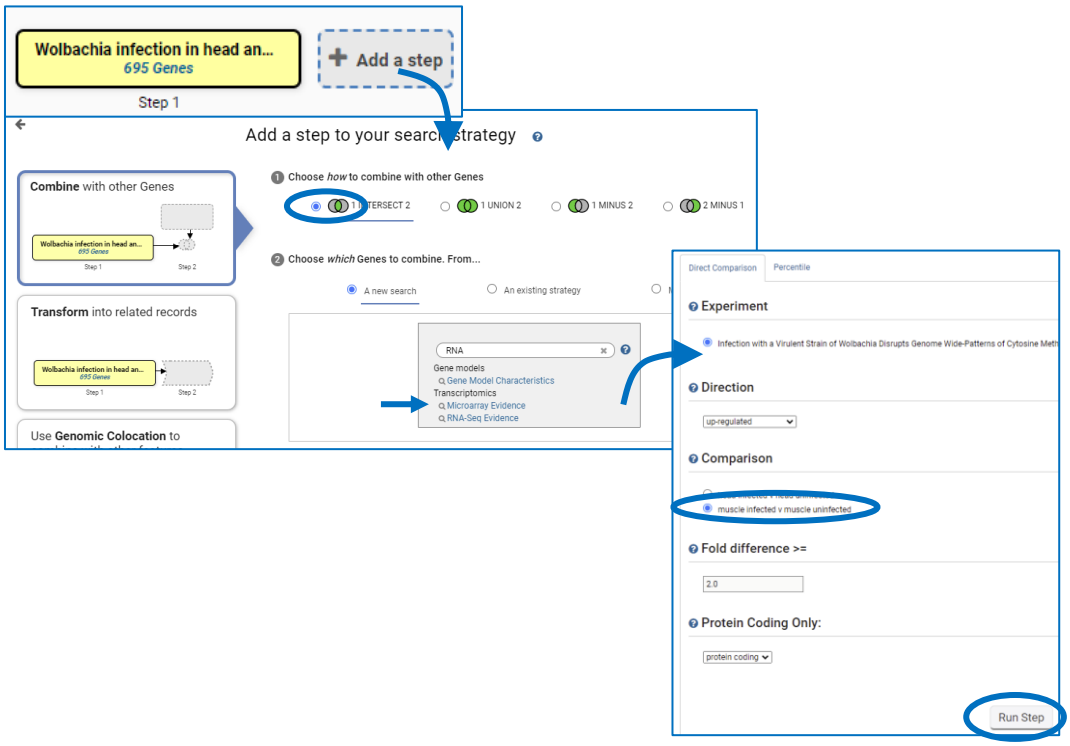
a. Navigate to the microarray search and choose the Direct Comparison search for the dataset titled **'Infection with a Virulent Strain of Wolbachia Disrupts Genome Wide-Patterns of Cytosine Methylation in the Mosquito Aedes aegypti (Ye et al.)'**

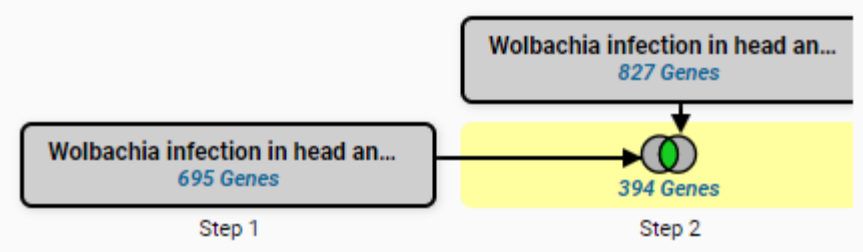


b. Initiate a search that returns genes that **are upregulated 2 fold in infected head vs uninfected**.

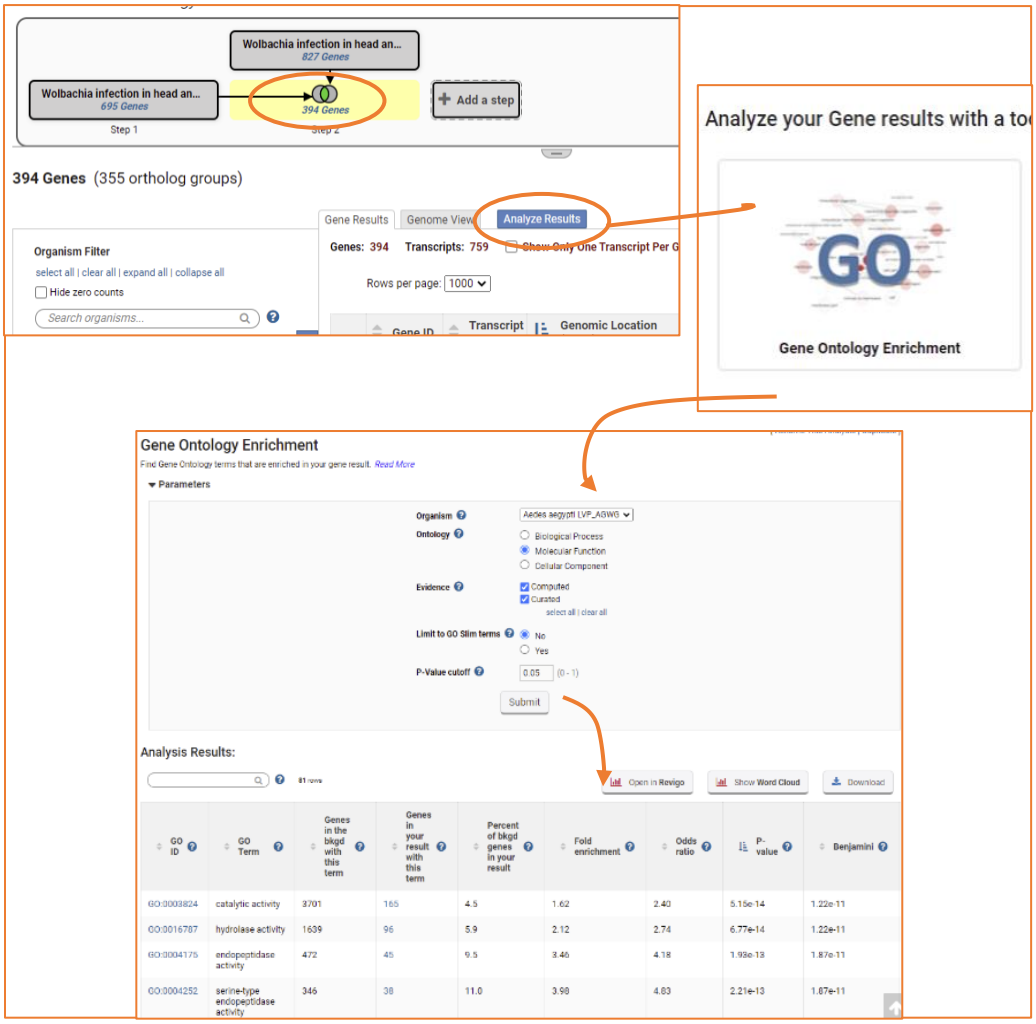


c. Intersect your search result with another search that returns genes upregulated 2 fold in muscle vs uninfected. Your combined result will be genes that are upregulated in head and muscle in response to *Wolbachia* infection.





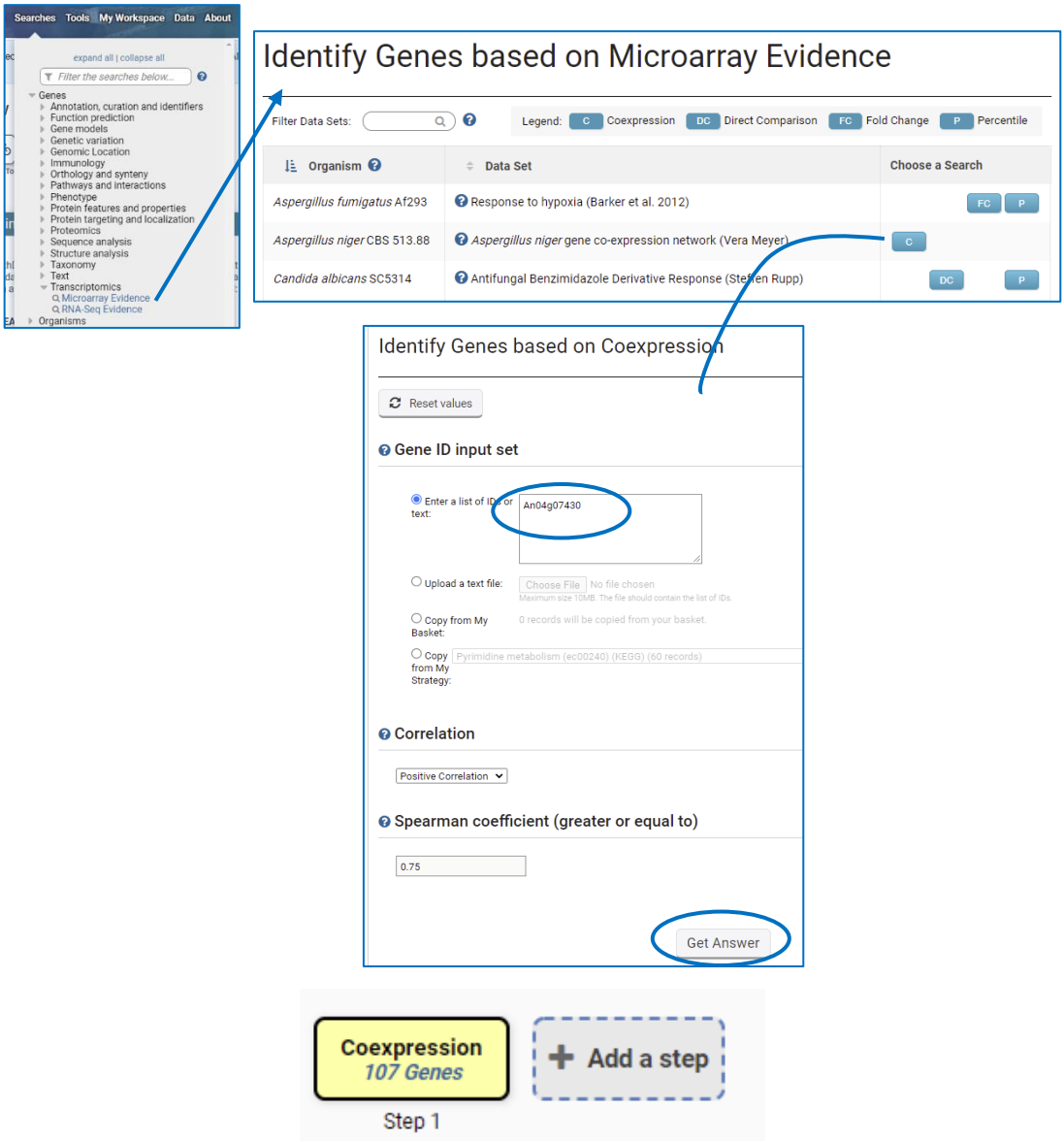
d. Determine enriched Molecular Function GO terms for the upregulated genes. Make sure you are viewing the combined result (the Step 2 result will be highlighted yellow) and click Analyze Result to open the Enrichment Tool. What gene functions are shared by the combined result? What biological role can you envision for these mosquito genes during the *wolbachia* infection?



3. **Find genes that are likely co-expressed with An04g07430, an *Aspergillus niger* protein coding gene with little functional annotation.** By finding genes that are expressed at the same time as An04g07430, we may find clues about its function and the biological processes that it participates in. This exercise uses FungiDB.

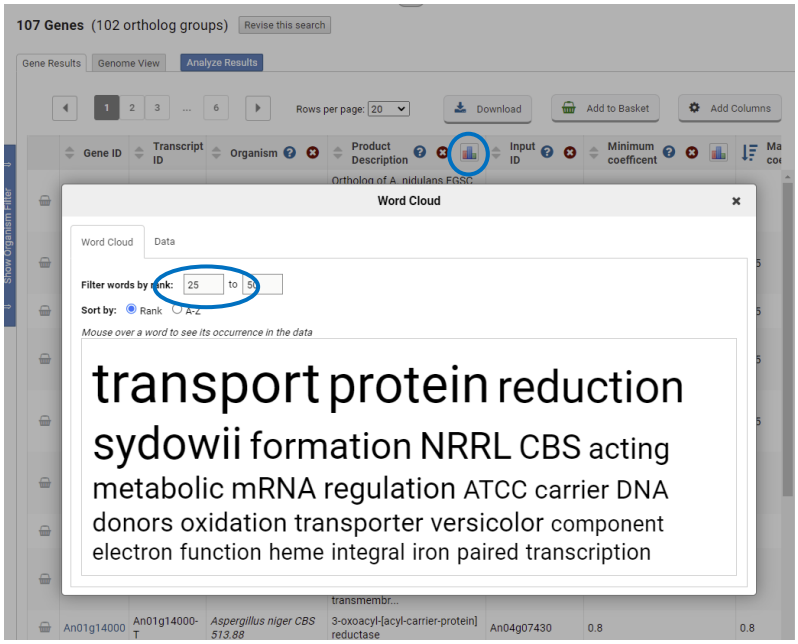
b. Navigate to the microarray searches in FungiDB and choose the Coexpression search for the data set titled ***Aspergillus niger* gene co-expression network (Vera Meyer).** Schape et al Nucleic Acids Research 2019. This data are the results of a meta-analysis of 155 publicly available transcriptomics analyses for *A. niger*, which were used to generate a genome-level co-expression network and sub-networks for >9,500 genes.

c. Run the search to find the co-expression network for An04g07430.



d. What genes share the co-expression profile of An04g07430? Several genes have a correlation coefficient of 0.85. What are these genes? Visit their gene pages to learn more.

e. Scan the product description column for genes with known functions. Use the Column Histogram tool to view a word cloud of the product descriptions in the column. Set the rank range to 25-50. What words occur most often in the product descriptions of An04g07430 co-expressed genes?



f. Run the enrichment analyses for Molecular Function, Cellular Component and Biological Processes. Do these provide information about what this group of co-expressed genes might be doing?