Advanced Search Strategies

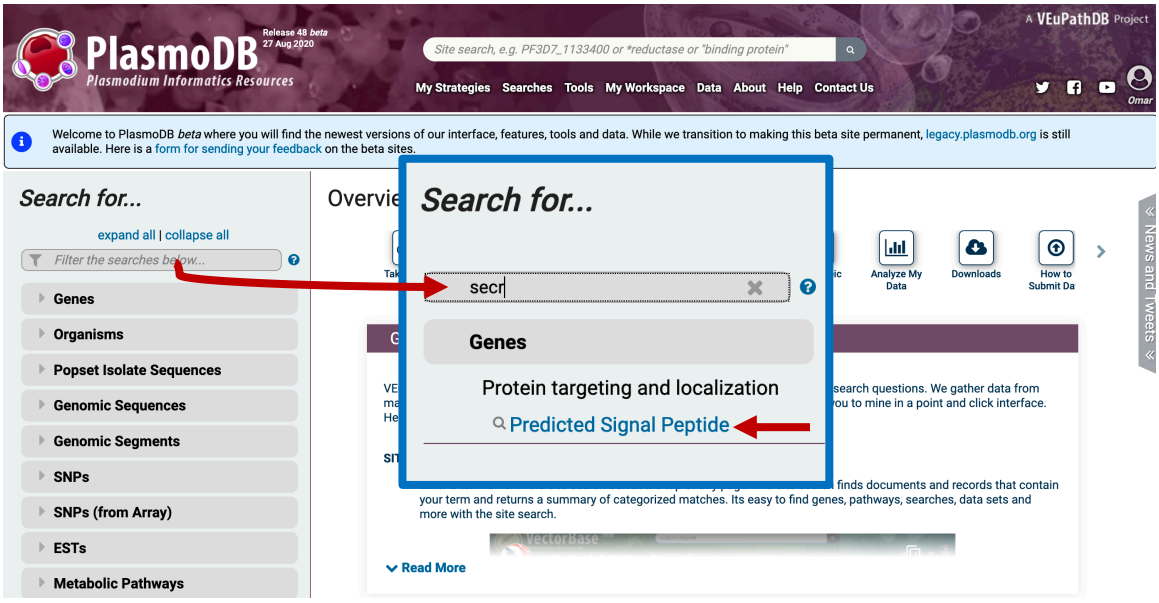
Note: *this exercise uses PlasmoDB.org as an example database, but the same functionality is available on all VEuPathDB resources.*

Learning objectives:

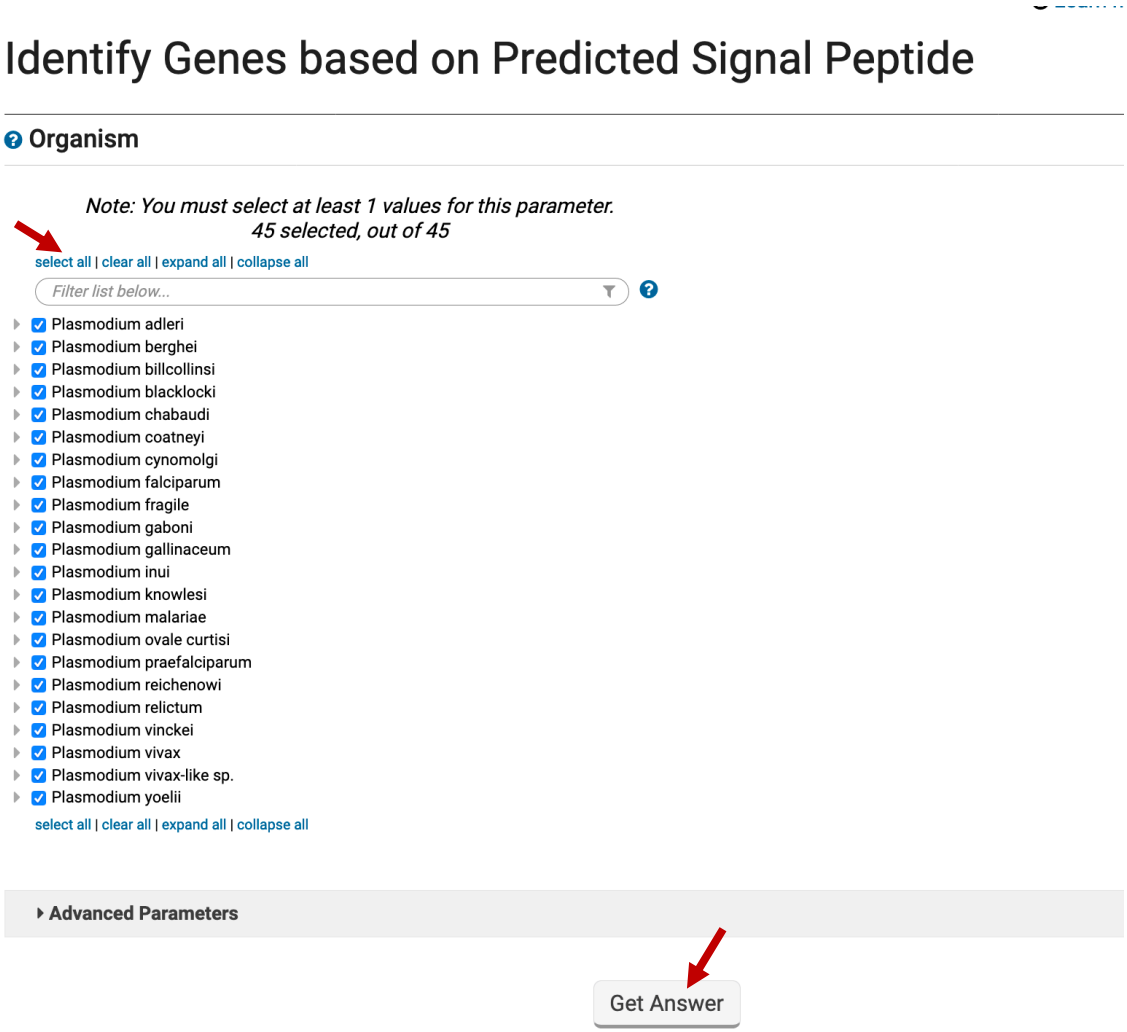
- Integrate diverse datatypes in a search strategy
- Leverage orthology and phylogenetic profile searches

This exercise walks you through the process of building a multi-step strategy, integrating different datatypes. The final search strategy identifies plasmodium genes that are likely secreted, or membrane bound, highly polymorphic, “essential” for parasite survival, not conserved in mammals and expressed in liver stages of the Plasmodium life cycle. There are many ways to build these strategies and order the steps to reach a similar answer.

1. Identify all genes in PlasmoDB that are predicted to have a secretory signal peptide as defined by SignalP. An easy way to identify a search type is to filter the searches on the left of the home page. Start typing a word to identify the search type. For example, start typing the word "secreted”, you should see the searches being filtered even before you finish typing the complete word.

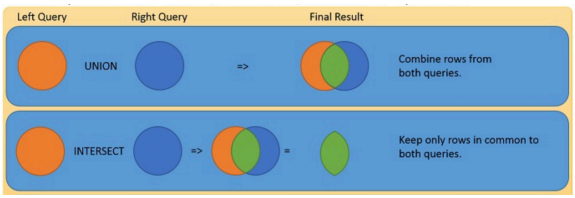


2. Click on the search for genes by predicted signal peptide. On the next page select all organisms and click on the get answer button at the bottom of the page.

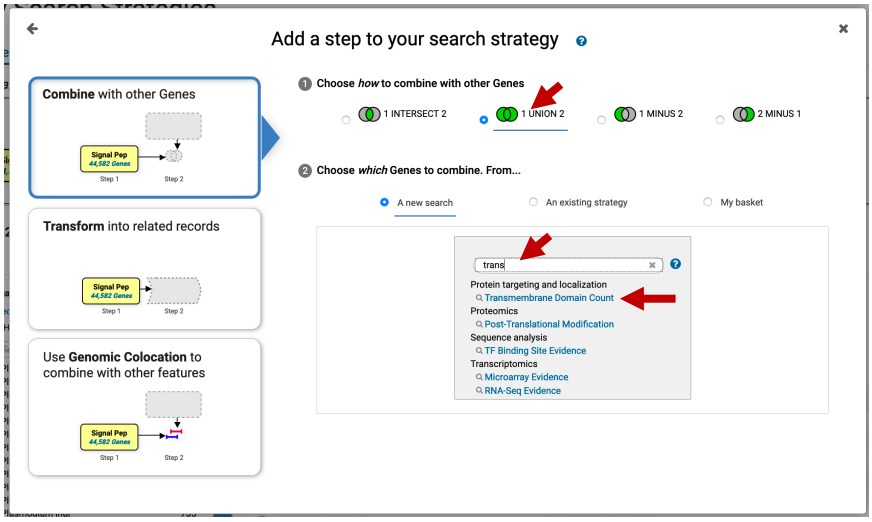


3. The next step is to combine the signal peptide results with results of genes that are predicted to have at least one transmembrane domain (TM). Click on the add step button in the search strategy panel.

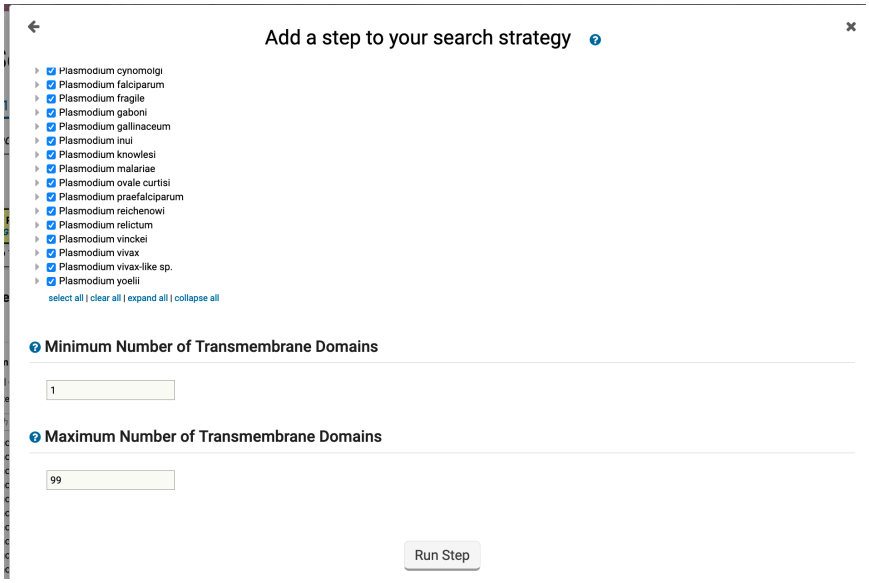
The popup window offers you option to add additional steps and ways to combine the searches (intersect, union, minus). For this exercise we are interested in finding genes that a signal peptide or a TM domain or both. What operation will you use to combine the searches – Union or Intersect?



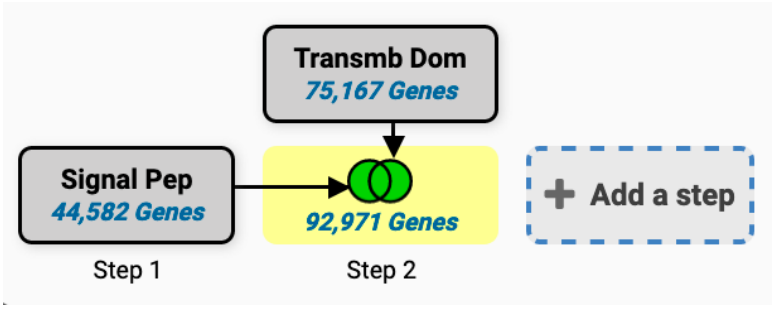
Once you select the option for combining the searches, find the search for transmembrane domain count. Notice that you can use the same query filtering mechanism as before. Start typing transmembrane to find this search. Once you find it click on to open the search parameters.



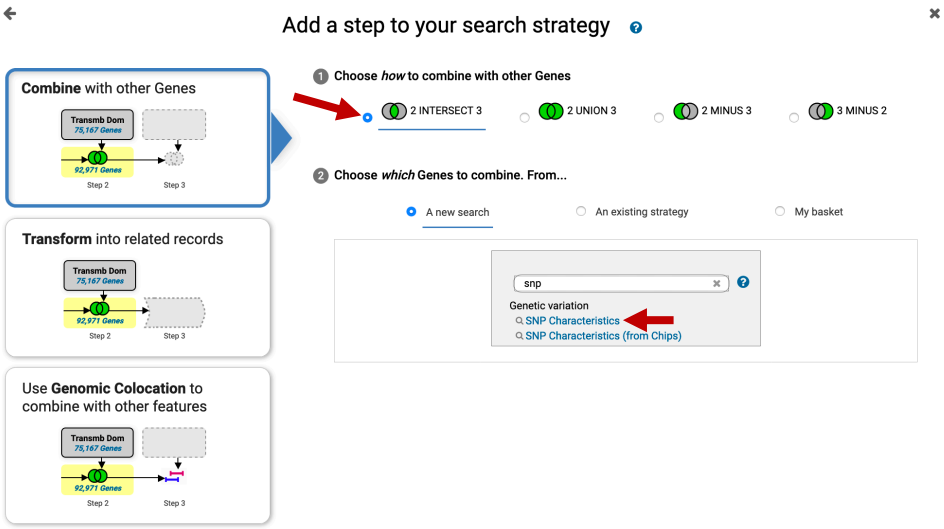
4. For the TM search, again select all organisms, use the default parameters and click on the get answer button.



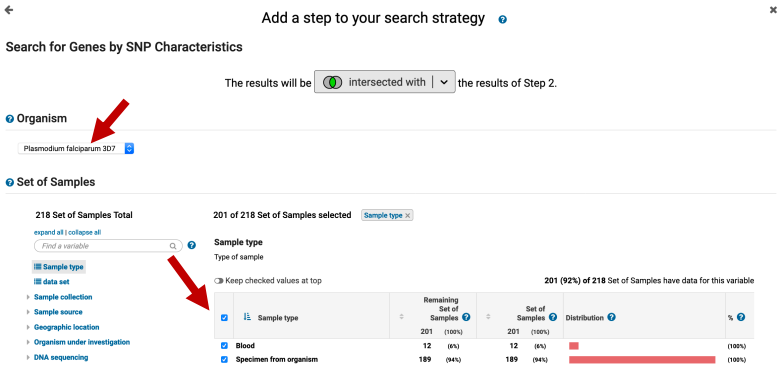
5. How many genes did you get? Since you used a union the number of results should be more that each of the individual steps that were combined.



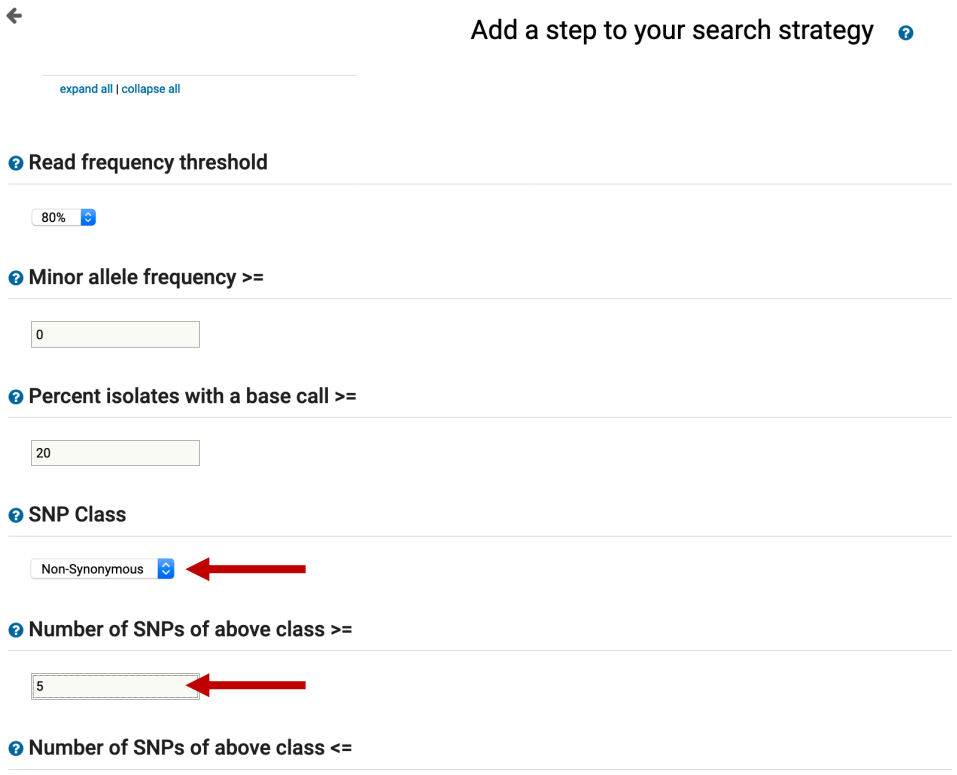
6. Next, identify genes from step 2 that contain at least 5 non-synonymous SNPs (non-synonymous SNPs are single nucleotide polymorphisms that result in an amino acid change). Were you able to find the SNP search by clicking on add step and filtering the searches with a keyword? Which operation will you select to combine the searches?



7. On the Genes by SNP characteristics search popup, select Plasmodium falciparum from the drop down and select all available isolates by selecting the checkbox at the top of the filter panel (See image below).

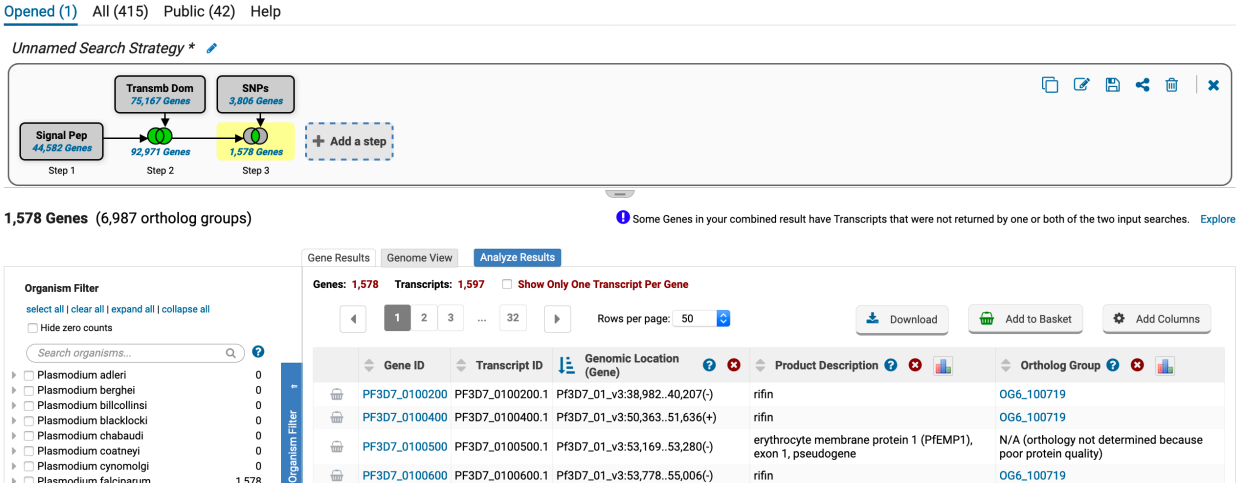


8. Next scroll down and select the following parameters. SNP class = Non-synonymous. Number of SNPs of above class >/= 5. After you seelct these parameters, scroll down to the bottom and click on Run Step.

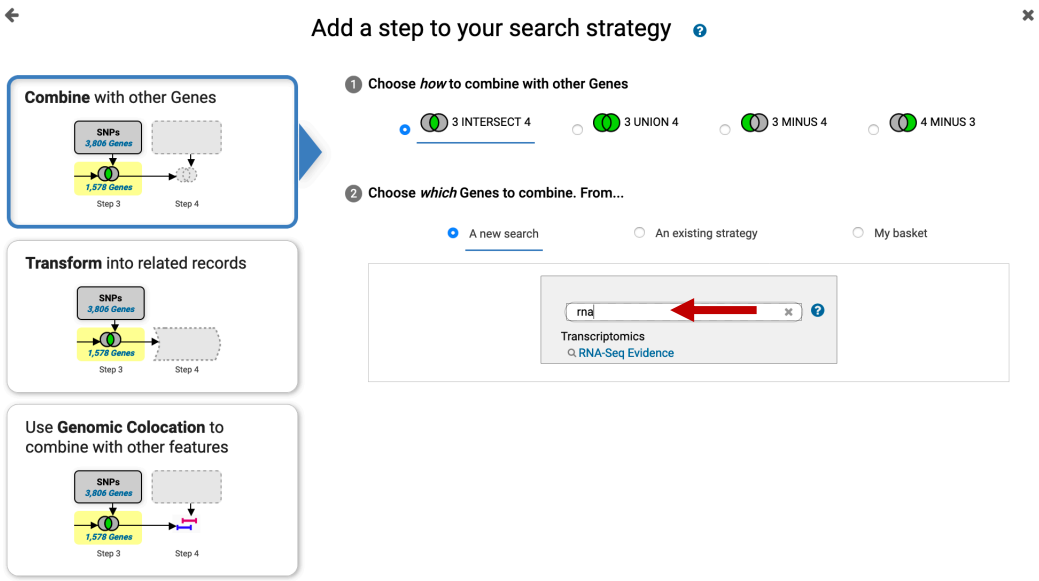


What do the results look like? What species are represented in the results? Is this surprising? Remember that your last search only queried *P. falciparum* data.

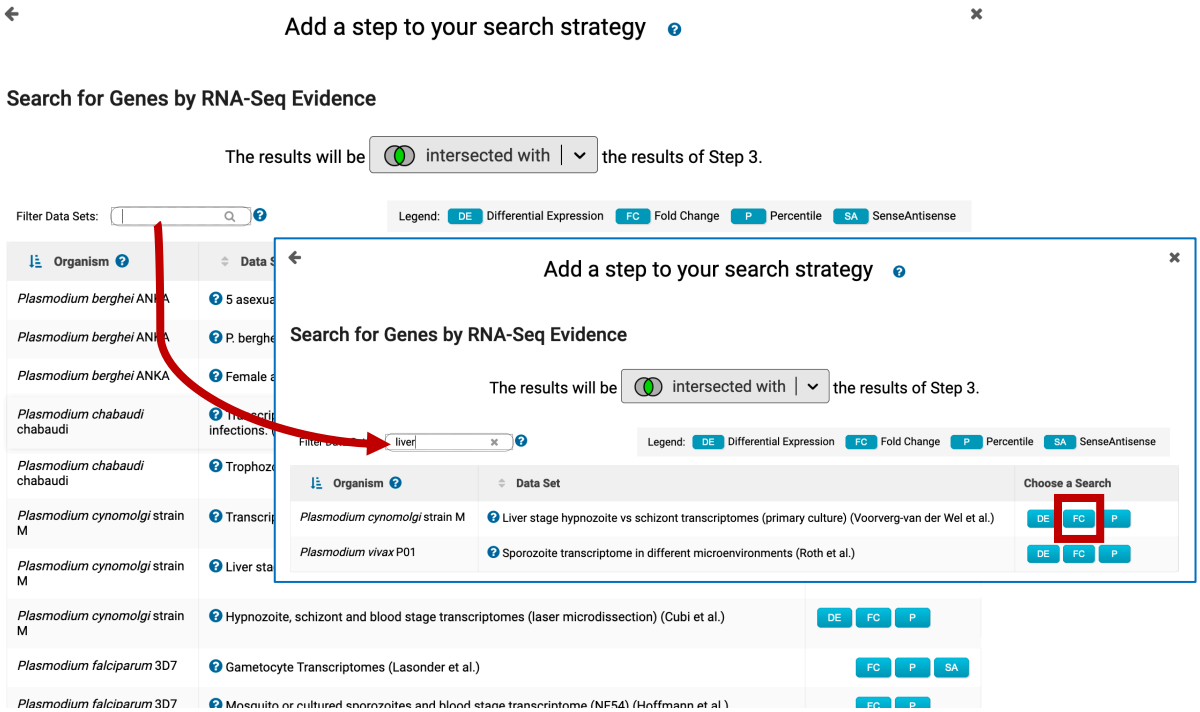
My Search Strategies



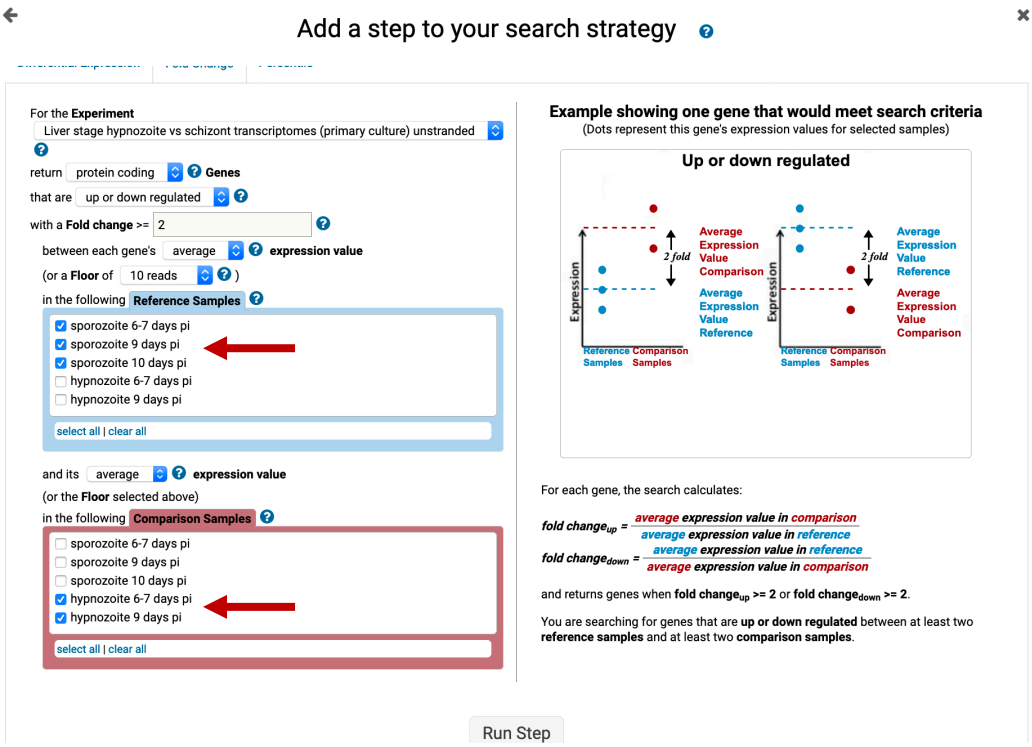
9. Determine how many of these genes are also differentially expressed in liver stages. Click on add step then search for the RNA-seq search. Type RNA in the search filter in the popup.



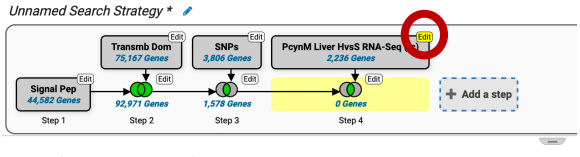
10. On the next page find data that queries liver stages. You can filter the data by typing the word liver in the filter box at the top of the page. This should yield two datasets from *P. cynomolgi* and *P. viviax*. For this exercise, select the fold change query for the *P. cynomolgi* dataset: Liver stage hypnozoite vs schizont transcriptomes (primary culture) (Voorverg-van der Wel et al.).



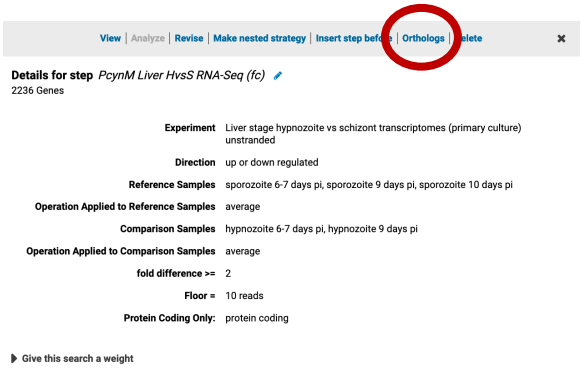
11. Configure the RNA-Seq search to identify genes that are differentially regulated by at least 2-fold between all the hypozoite stages and the sporozoite statges. For example, select the hypnozoite stages in the reference selection box and the sporozoite samples in the comparator selection box, then click on run step.



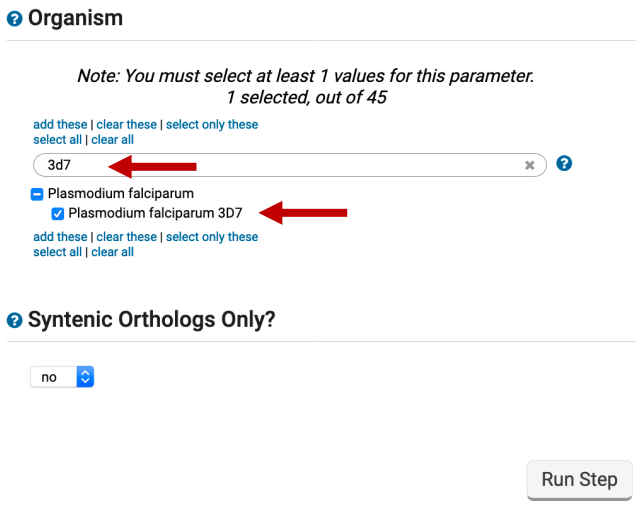
12. How many results did you get? Why did you get 0 results? How can you change this? *Remember that the previous search was a list of P. falciparum genes and this RNA-Seq was from P. cynomolgy. What you would like to do is convert the P. cynomolgy genes into P. falciparum genes.* To do this folloow these steps:
    a. hover your mouse of the RNA-seq step then click on the edit option on that step.



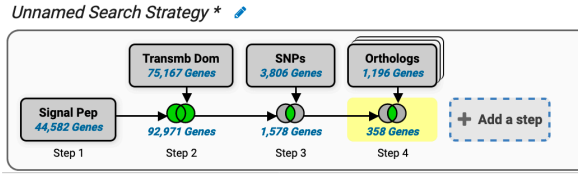
    b. In the popup window, click on the **orthologs** link.



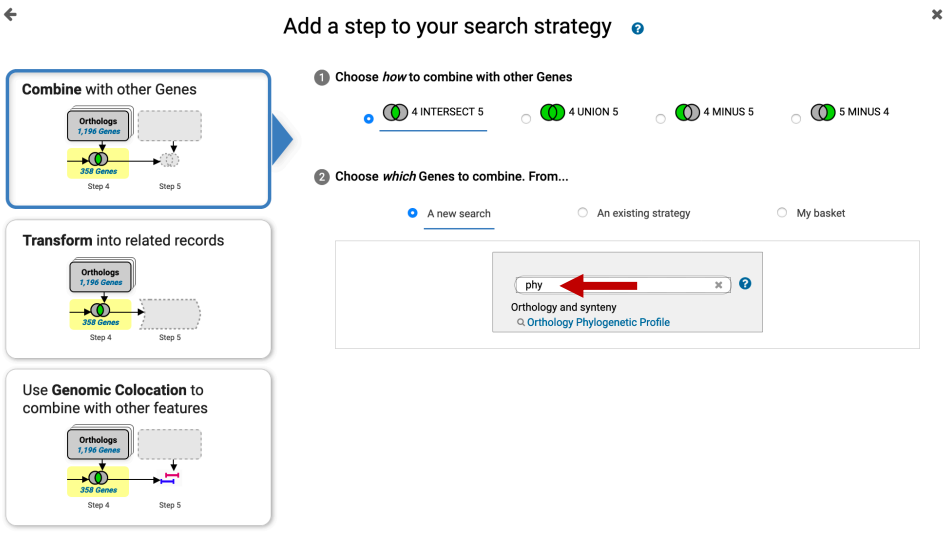
c. In the next window select which organism(s) you would like to transform to. For this exercise select *P. falciparum* 3D7 and click on run step.



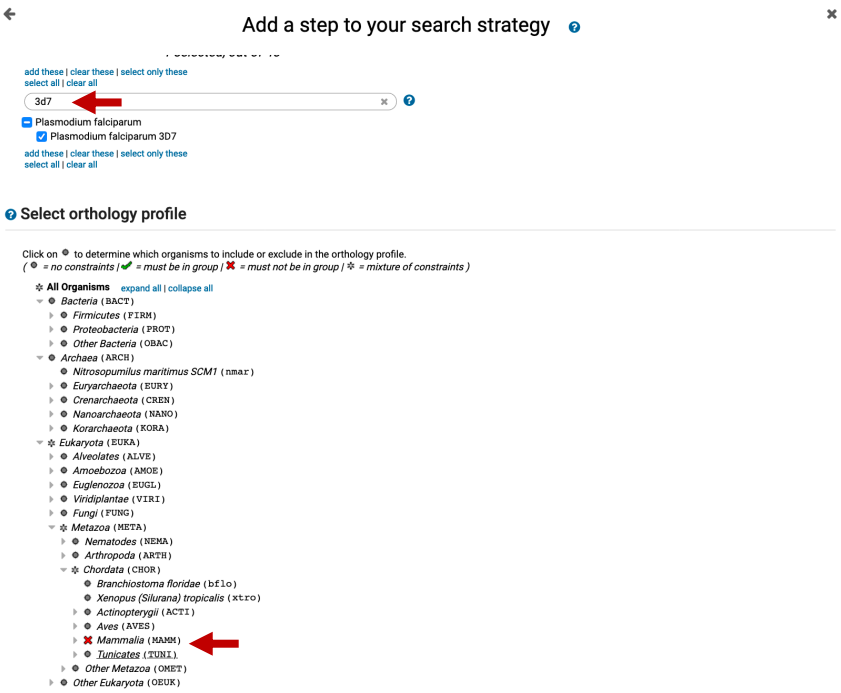
d. Did you get results now?



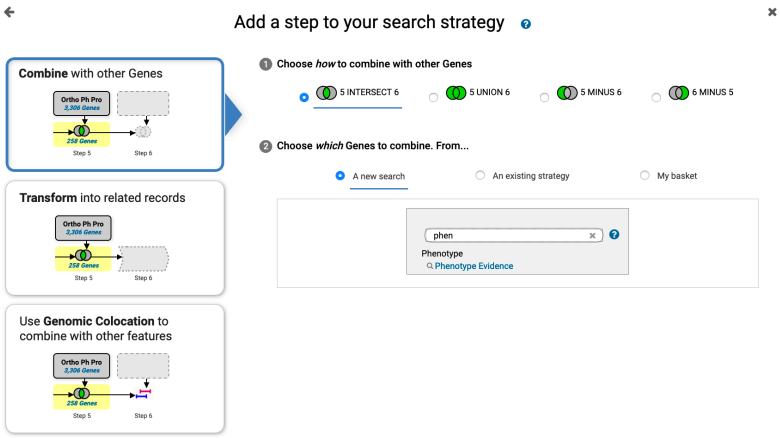
13. Next identify how many of these genes do not have orthologs in mammals. To do this add a step for genes based on orthology phylogenetic profile. Again you can filter the searches by typing the word “phylogenetic”.



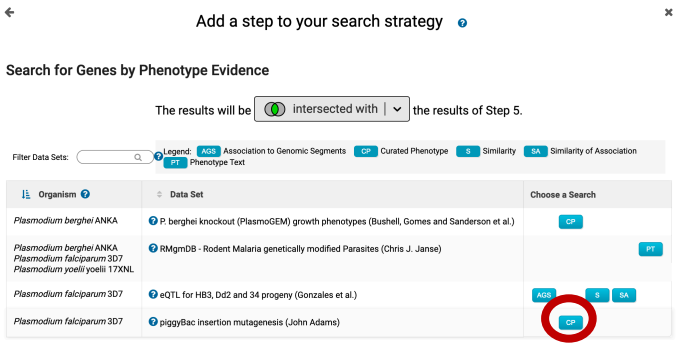
On the next page select *P. falciparum* 3D7 the configure the phylogenetic profile by finding Mammalia under Chordata which are under Metazoa. Click twice on the circle next to Mammalia – it should become a red x (See image below).



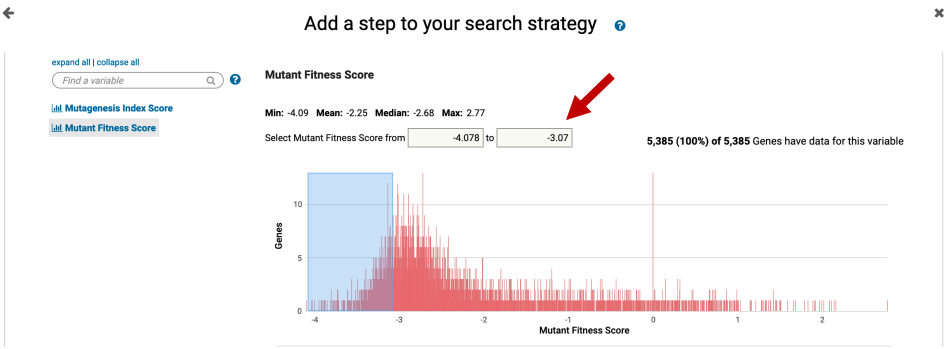
14. Determine if a mutation in any of these genes affects fitness. Click on add step and find the search for phenotye evidence.



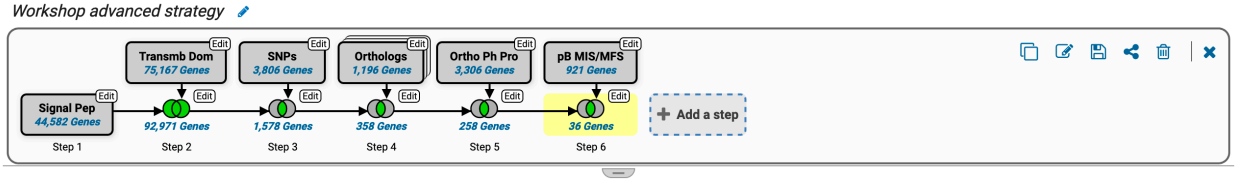
15. Select the P. falciparum piggyBac insertion mutagenesis (John Adams) experiment.



16. On the next page select the Mutan Fitness Score (MFS) option and choose any score range – generally the more negative the bigged the effect is on fitness. For this example a score range of -4.078 to -3.07 was chosen.



Explore your final results. Do they make sense/plausible? Note that you can revise any of the steps in the strategy to explore the data further. You can also save your strategy and share it with others or make it public. Here is a link to this search stragey:



https://plasmodb.org/plasmo/app/workspace/strategies/import/fd387e8d3acda856