

Advanced Search Strategies

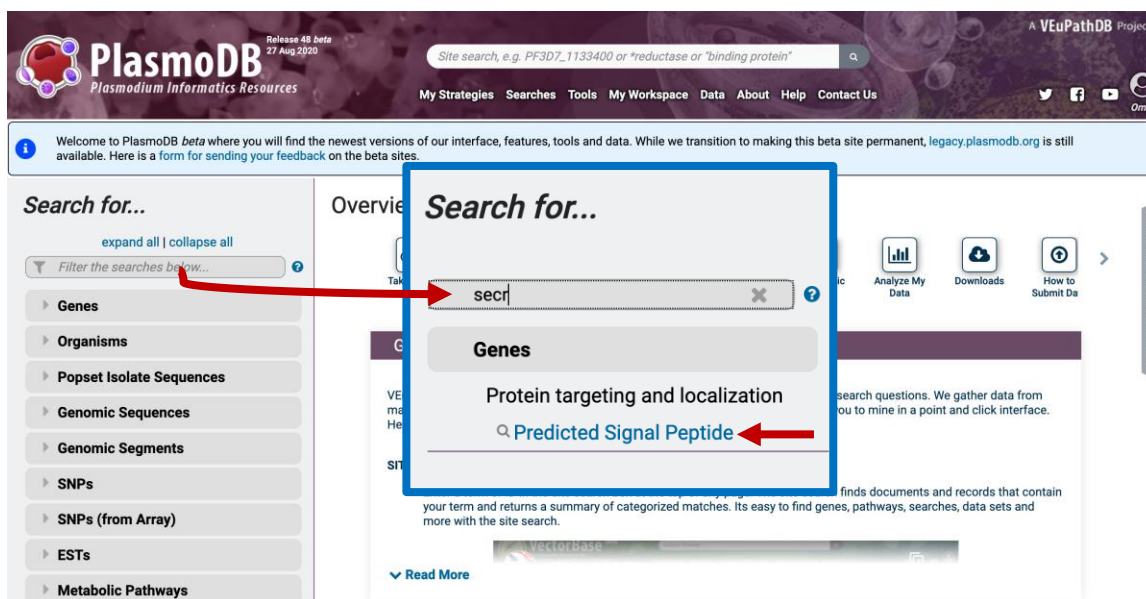
Note: this exercise uses PlasmoDB.org as an example database, but the same functionality is available on all VEuPathDB resources.

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.

Identify Genes based on Predicted Signal Peptide

Reset values

Organism

0 selected, out of 57

select all | clear all | expand all | collapse all

Filter list below... ?

- ▶ ☐ Haemoproteidae
- ▶ ☐ Plasmodiidae

select all | clear all | expand all | collapse all

Get Answer

Build a Web Services URL from this Search >>

? Give this search a name (optional)
? Give this search a weight (optional)

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.

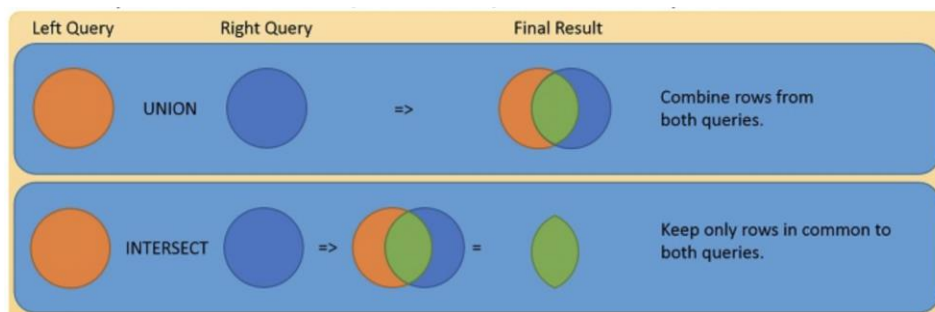
Unnamed Search Strategy

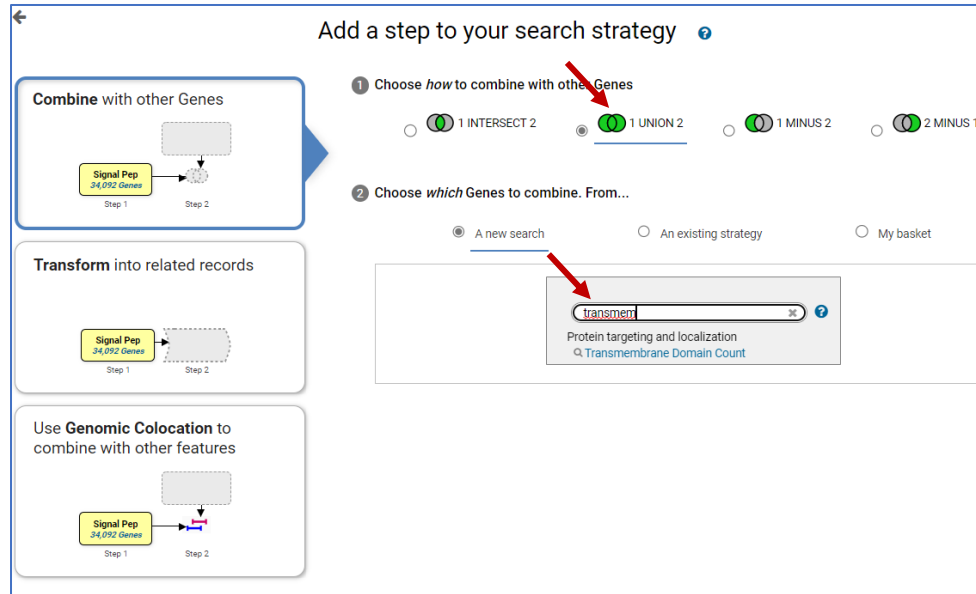
📄 ✎ 💾 🔗 🗑 ✕

Signal Pep
 34,092 Genes
 Step 1

+ Add a step

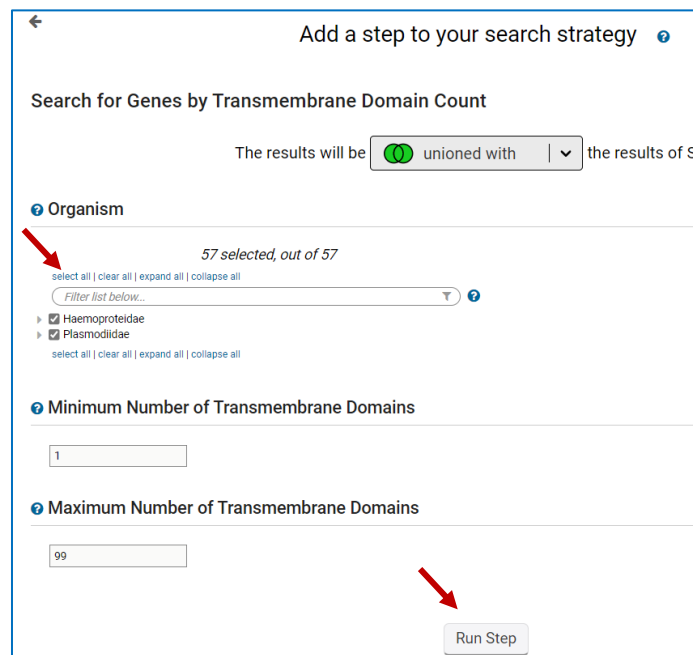
The popup window offers you option to add additional steps and ways to combine the searches (intersect, union, minus). For this exercise we are interested in finding genes that a signal peptide or a TM domain or both. What operation will you use to combine the searches – Union or Intersect?





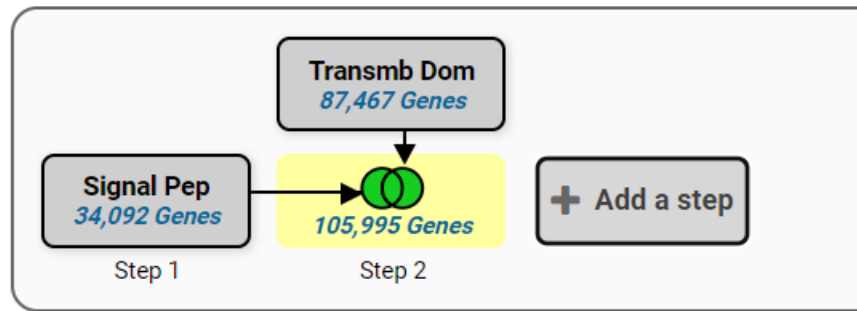
Once you select the option for combining the searches, find the search for transmembrane domain count. Notice that you can use the same query filtering mechanism as before. Start typing transmembrane to find this search. Once you find it click on to open the search parameters.

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 than each of the individual steps that were combined.

Unnamed Search Strategy *



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?

← Add a step to your search strategy ? ×

1 Choose *how* to combine with other Genes

☒ 2 INTERSECT 3 ☐ 2 UNION 3 ☐ 2 MINUS 3 ☐ 3 MINUS 2

2 Choose *which* Genes to combine. From...

☒ A new search ☐ An existing strategy ☐ My basket

Combine with other Genes

Transmb Dom
75,167 Genes

82,971 Genes

Step 2

Step 3

Transform into related records

Transmb Dom
75,167 Genes

82,971 Genes

Step 2

Step 3

Use Genomic Colocation to combine with other features

Transmb Dom
75,167 Genes

82,971 Genes

Step 2

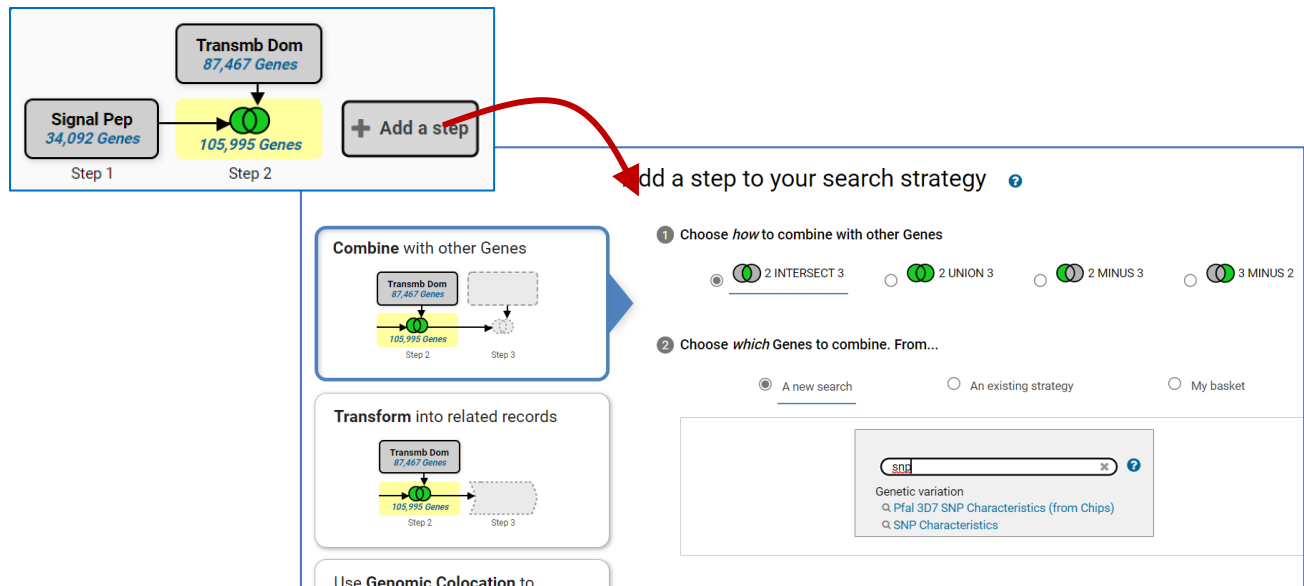
Step 3

snp

Genetic variation

SNP Characteristics

SNP Characteristics (from Chips)



- 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).

Add a step to your search strategy

Search for Genes by SNP Characteristics

The results will be intersected with the results of Step 2.

Organism

Plasmodium falciparum 3D7

Set of Samples

219 Set of Samples Total

Find a variable

Proportion mapped reads

Average mapping coverage

Data Set

Parasite organism

No filters applied

Check items below to apply this filter

	Remaining Set of Samples	Set of Samples	Distribution
<input type="checkbox"/> Data Set	219 (100%)	219 (100%)	
<input type="checkbox"/> Aligned genomic sequence reads - 8 strains	8 (4%)	8 (4%)	
<input type="checkbox"/> Aligned genomic sequence reads - 65 Gambian Isolates	65 (30%)	65 (30%)	
<input type="checkbox"/> Aligned genomic sequence reads - 150 lab strains or field isolates	69 (32%)	69 (32%)	
<input type="checkbox"/> Aligned genomic sequence reads - KADdd2 (sensitive) and KAD707A (resistant)	2 (1%)	2 (1%)	

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

?

SNP Class

Non-Synonymous

?

Number of SNPs of above class >=

5

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

Opened (1) All (1) Public (50) Help

Unnamed Search Strategy *

Signal Pep
34,092 Genes
Step 1

Transmb Dom
87,467 Genes
Step 2

SNPs
3,805 Genes
Step 3

+ Add a step

1,419 Genes (1,075 ortholog groups)

Some Genes in your combined result have Transcripts that were not

Gene Results Genome View Analyze Results

Genes: 1,419 Transcripts: 1,436 ☐ Show Only One Transcript Per Gene

1 2 3 ... 72 Rows per page: 20

	Gene ID	Transcript ID	Genomic Location (Gene)	Product Descri
	PF3D7_0100200	PF3D7_0100200.1	Pf3D7_01_v3:38,982..40,207(-)	rifin
	PF3D7_0100400	PF3D7_0100400.1	Pf3D7_01_v3:50,363..51,636(+)	rifin
	PF3D7_0100600	PF3D7_0100600.1	Pf3D7_01_v3:53,778..55,006(-)	rifin
	PF3D7_0100800	PF3D7_0100800.1	Pf3D7_01_v3:59,772..61,003(+)	rifin

Organism Filter

select all | clear all | expand all | collapse all

☐ Hide zero counts

Search organisms...

☐ Haemoproteidae 0

☐ Plasmodiidae 1,419

select all | clear all | expand all | collapse all

☐ Hide zero counts

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 and choose the RNA-Seq evidence search.

10. The next page lists all RNA-Seq data sets (studies) available to search. To find data that queries liver stages use the data set filter. Type the word liver in the filter box at the top of the page to filter the list to data sets that include liver in the title or description. This should yield datasets from *P. berghei*, *P. cynomolgi* and *P. vivax*. 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.).

Organism	Data Set	Choose a Search
<i>Plasmodium berghei</i> ANKA	Transcriptome during early and mid-stage <i>P. berghei</i> liver infection (Toro-Moreno and Sylvester et al.)	DE FC P
<i>Plasmodium berghei</i> ANKA	Ex-erythrocytic stage transcriptomes (sporozoite, liver time course and detached cells) (Caldelari et al.)	DE FC P
<i>Plasmodium cynomolgi</i> strain M	Liver stage hypnozoite vs schizont transcriptomes (primary culture) (Voorverg-van der Wel et al.)	DE FC P
<i>Plasmodium vivax</i> P01	Sporozoite transcriptome in different microenvironments (Roth et al.)	DE FC P

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 stages. For example, select the hypozoite stages in the reference selection box and the sporozoite samples in the comparator selection box, then click on run step.

Add a step to your search strategy ?

For the Experiment
Liver stage hypozoite vs schizont transcriptomes (primary culture) unstranded

return protein coding ? Genes
that are up or down regulated ?
with a Fold change ≥ 2
between each gene's average expression value
(or a Floor of 10 reads ?)
in the following Reference Samples ?

☒ sporozoite 6-7 days pi
☒ sporozoite 9 days pi
☒ sporozoite 10 days pi
☐ hypozoite 6-7 days pi
☐ hypozoite 9 days pi

select all | clear all

and its average expression value
(or the Floor selected above)
in the following Comparison Samples ?

☐ sporozoite 6-7 days pi
☐ sporozoite 9 days pi
☐ sporozoite 10 days pi
☒ hypozoite 6-7 days pi
☒ hypozoite 9 days pi

select all | clear all

Run Step

Example showing one gene that would meet search criteria
(Dots represent this gene's expression values for selected samples)

Up or down regulated

For each gene, the search calculates:

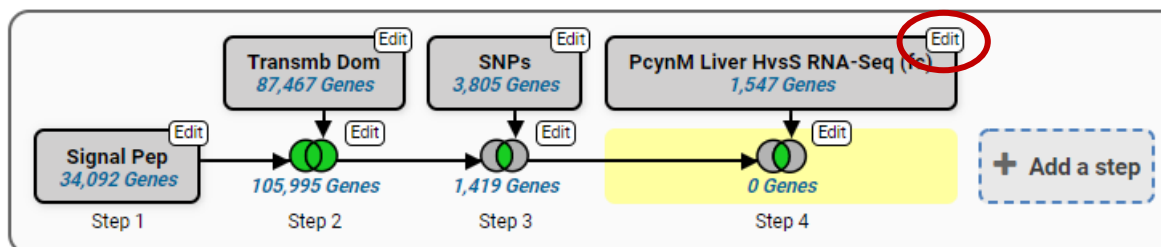
$$\text{fold change}_{\text{up}} = \frac{\text{average expression value in comparison}}{\text{average expression value in reference}}$$

$$\text{fold change}_{\text{down}} = \frac{\text{average expression value in reference}}{\text{average expression value in comparison}}$$

and returns genes when $\text{fold change}_{\text{up}} \geq 2$ or $\text{fold change}_{\text{down}} \geq 2$.

You are searching for genes that are up or down regulated between at least two reference samples and at least two comparison samples.

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 follow these steps:
- 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 to open the Ortholog transform tool.

View | Analyze | Revise | Make nested strategy | Insert step before | **Orthologs** | Delete

Details for step *PcynM Liver HvsS RNA-Seq (fc)* [✎](#)
1547 Genes

Experiment Liver stage hypnozoite vs schizont transcriptomes (primary culture) unstranded

Direction up or down regulated

Reference Samples sporozoite 6-7 days pi, sporozoite 9 days pi, sporozoite 10 days pi

Operation Applied to Reference Samples average

Comparison Samples hypnozoite 6-7 days pi, hypnozoite 9 days pi

Operation Applied to Comparison Samples average

fold difference >= 2

Floor = 10 reads

Protein Coding Only: protein coding

► Give this search a weight

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

← Add a step to your search strategy ?

Your Genes from Step 1 will be converted into Orthologs

? Organism

1 selected, out of 57

select only these | add these | clear these

3d7 ✕ ?

Plasmodiidae

Plasmodium

Plasmodium falciparum

→ ☒ Plasmodium falciparum 3D7 [Reference]

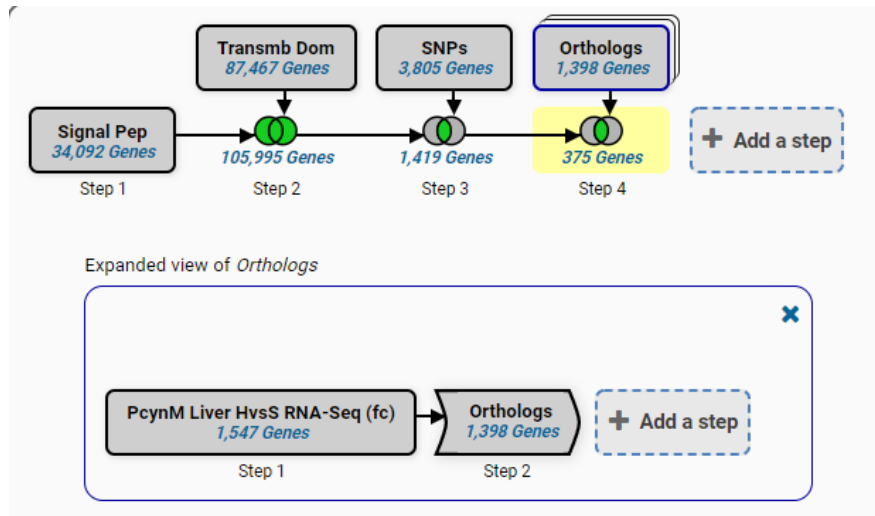
select only these | add these | clear these

? Syntenic Orthologs Only?

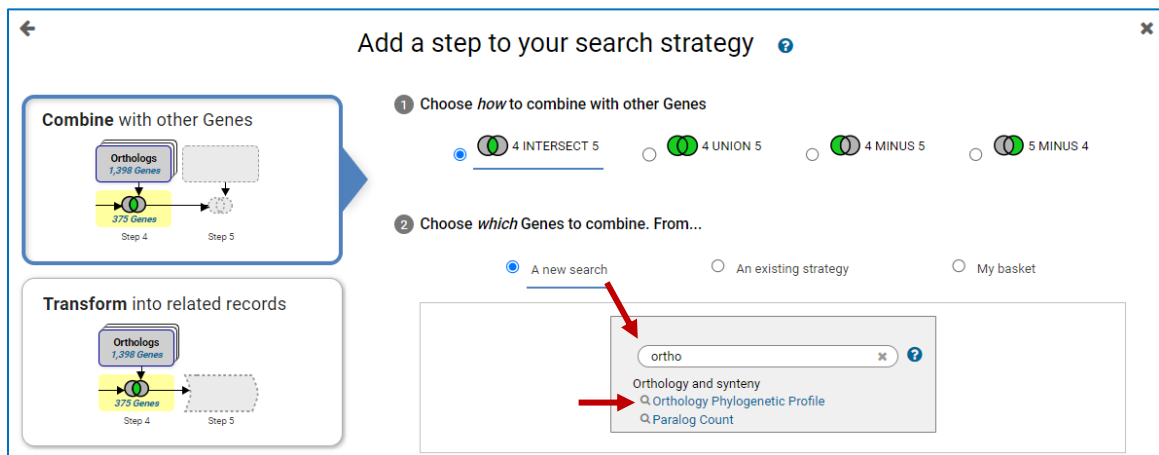
no ▾

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 search that returns genes based on orthology phylogenetic profile and intersect that with your previous results. Again you can filter the searches by typing the word “orthology” or “phylogenetic”.



Configure the search to return *P. falciparum* 3D7 genes and the phylogenetic profile that excludes Mammalia under Chordata which are under Metazoa. Click twice on the circle next to Mammalia – it should become a red x (See image below).

?

[select only these](#) | [add these](#) | [clear these](#)

3d7

Plasmodium

- Plasmodium falciparum

[select only these](#) | [add these](#) | [clear these](#)

?

(● = no constraints | ✓ = must be in group | ✗ = must not be in group | * = mixture of constraints)

* **All Organisms** [expand all](#) | [collapse all](#)

▼ ● *Bacteria* (BACT)

- ▶ ● *Firmicutes* (F)

- ▶ ● *Proteobacteria* (PF)

- ▶ ● *Other Bacteria* (OBAC)

- *Archaea* (ARCH)

● *Nitrosopumilus maritimus* (strain SCM1) (nmar)

- ▶ ● *Euryarchaeota* (EURY)

- ▶ ● *Crenarchaeota* (CREN)

- ▶ ● *Nanoarchaeota* (NANO)

► ● *Korarchaeota* (KORA)

▼ ✱ *Eukaryota* (EUKA)

- ▶ ● *Alveolates* (ALVE)

- ▶ ● *Amoebozoa* (AMOE)

- ▶ ● *Euglenozoa* (EUGL)

- ▶ ● *Viridiplantae* (VIR)

► ● *Fungi* (FUNG)

▼ ✱ Metazoa (META)

- ▶ ● *Nematodes* (NE)

► ● *Arthropoda* (AR)

▼ * Chordata (CHOR)

- *Branchiostoma floridae* (Florida lancelet) (Amphioxus) (bf1o)

- *Xenopus tropicalis* (W)

▶ ● *Actinopterygii*

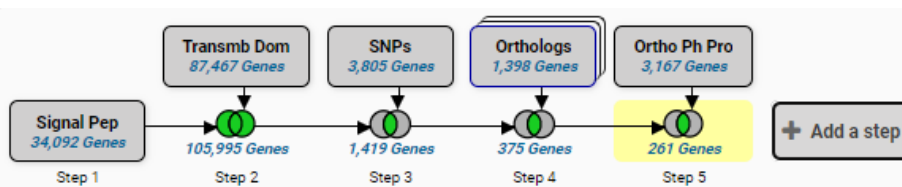
► ● **Aves (AVES)**

➡ ~~Mammalia~~ (Mammals)

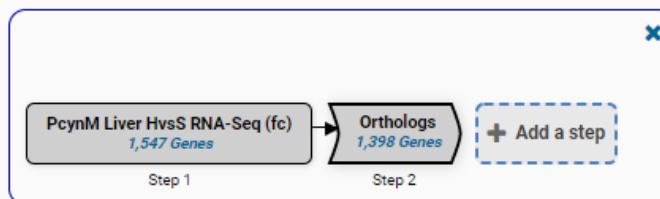
- ▶ ● *Tunicates* (Tunicata)

► ● *Other Metazoa* (

Run Step





Expanded view of *Orthologs*



14. Determine if a mutation in any of these genes affects fitness based on the phenotype screening data in PlasmoDB. Click on add step and find the search for phenotype evidence. Select the *P. falciparum* **piggyBac insertion mutagenesis (John Adams)** experiment.













← Add a step to your search strategy ⓘ

Search for Genes by Phenotype Evidence

The results will be  intersected with  the results of Step 5.

Legend: AGS Association to Genomic Segments CP Curated Phenotype S Similarity SA Similarity of Association PT Phenotype Text


Filter Data Sets: ⓘ 5 rows

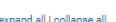
Organism ⓘ	Data Set	Choose a Search
<i>Plasmodium berghei</i> ANKA	 P. berghei knockout (PlasmoGEM) growth phenotypes (Bushell, Gomes and Sanderson et al.)	
<i>Plasmodium berghei</i> ANKA <i>Plasmodium falciparum</i> 3D7 <i>Plasmodium yoelii</i> yoelii 17XNL	 RMgMD - Rodent Malaria genetically modified Parasites (Chris J. Janse)	
<i>Plasmodium falciparum</i> 3D7	 eQTL for HB3, Dd2 and 34 progeny (Gonzales et al.)	  
<i>Plasmodium falciparum</i> 3D7	 piggyBac insertion mutagenesis (John Adams)	
<i>Plasmodium falciparum</i> 3D7	 Heat-shock response phenotypic screen (Zhang et al. 2021)	


15. Configure the piggyBac search to return genes whose Mutant Fitness Score (MFS) is between -4.094 to -3.07, or something similar. Generally the more negative the MFS score, the bigger the effect is on fitness.

Curated Phenotype

Genes

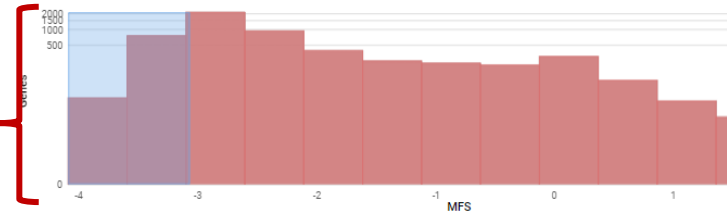
5,385 Genes Total 922 of 5,385 Genes selected 

 Find a variable ⓘ

→ Select MFS from to 5,385 (100%) of 5,385

Type your chosen MFS scores in the box or drag an area on the graph.



Plot Settings

y-axis

Scale counts: ☐ linear ☒ log₁₀

Range: to (1 - 166)



x-axis

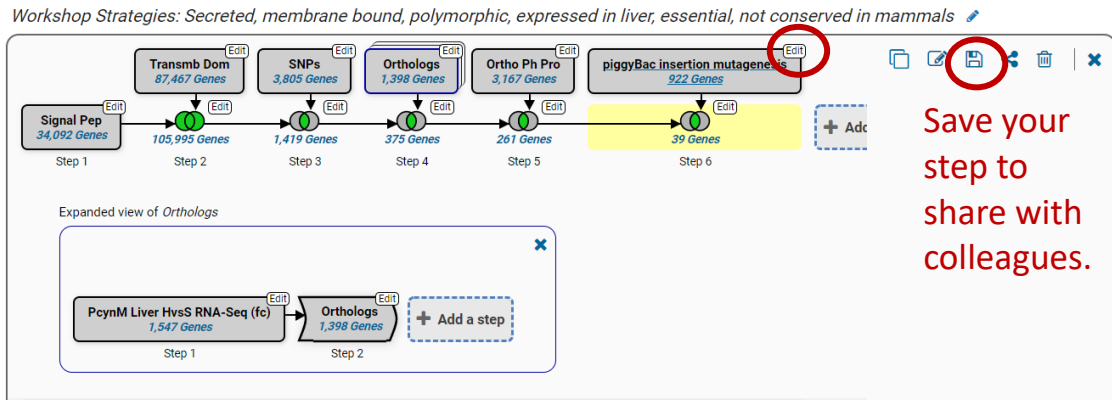
Bin width: When bin size = 0, the c

Range: to (-



Explore your final results. Do they make sense and seem 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.

Revise any step
by clicking edit.



Save your
step to
share with
colleagues.

Here is a link to this search stragey:

<https://plasmodb.org/plasmo/app/workspace/strategies/import/db454a408ec99e98>

