

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.

The screenshot shows the PlasmoDB beta homepage. On the left, there is a sidebar with a 'Search for...' section containing a dropdown menu with options like 'expand all' and 'collapse all'. Below this is a list of categories: Genes, Organisms, Popset Isolate Sequences, Genomic Sequences, Genomic Segments, SNPs, SNPs (from Array), ESTs, and Metabolic Pathways. A red arrow points from the 'expand all' button in the sidebar to the 'Filter the searches below...' dropdown in the main search area. The main search area has a large input field labeled 'Search for...' with the prefix 'secr' typed into it. This triggers a search results panel for 'Genes' which is highlighted with a blue box. Inside this panel, under the heading 'Protein targeting and localization', there is a link labeled 'Predicted Signal Peptide' with a red arrow pointing to it. The background of the page features a dark banner with the text 'A VEuPathDB Project' and social media icons.

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

Organism

Note: You must select at least 1 values for this parameter.
45 selected, out of 45

[select all](#) | [clear all](#) | [expand all](#) | [collapse all](#)

Filter list below... ?

- ▶ Plasmodium adleri
- ▶ Plasmodium berghei
- ▶ Plasmodium bilcollinsi
- ▶ Plasmodium blacklocki
- ▶ Plasmodium chabaudi
- ▶ Plasmodium coatneyi
- ▶ Plasmodium cynomolgi
- ▶ Plasmodium falciparum
- ▶ Plasmodium fragile
- ▶ Plasmodium gaboni
- ▶ Plasmodium gallinaceum
- ▶ Plasmodium inui
- ▶ Plasmodium knowlesi
- ▶ Plasmodium malariae
- ▶ Plasmodium ovale curtisi
- ▶ Plasmodium praefalciparum
- ▶ Plasmodium reichenowi
- ▶ Plasmodium relictum
- ▶ Plasmodium vinckeii
- ▶ Plasmodium vivax
- ▶ Plasmodium vivax-like sp.
- ▶ Plasmodium yoelii

[select all](#) | [clear all](#) | [expand all](#) | [collapse all](#)

▶ Advanced Parameters Get Answer

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.

My Search Strategies

Opened (1) All (415) Public (42) Help

Unnamed Search Strategy * ✎

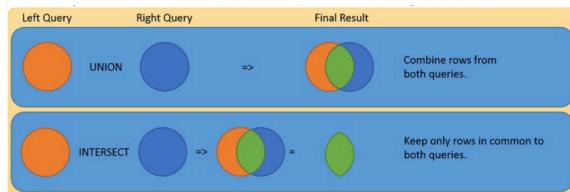
Signal Pep 44,582 Genes

+ Add a step

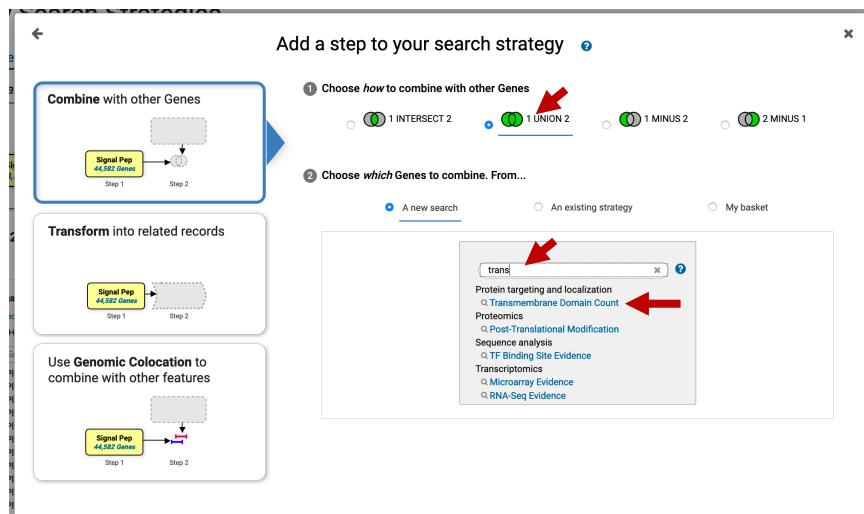
Step 1

✖ ✖ ✖ ✖ ✖ ✖ | ✖

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 have a signal peptide or a TM domain or both. What operation will you use to combine the searches – Union or Intersect?

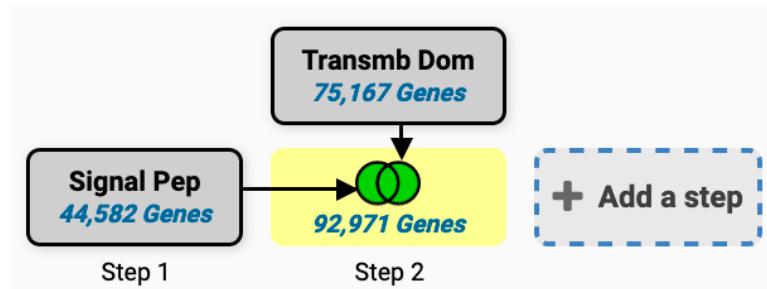


Once you select the option for combining the searches, find the search for transmembrane domain count. Notice that you can use the same query filtering mechanism as before. Start typing transmembrane to find this search. Once you find it click on to open the search parameters.



- 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.



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?

Combine with other Genes

① Choose how to combine with other Genes

② Choose which Genes to combine. From...

2 INTERSECT 3

A new search An existing strategy My basket

Genetic variation
SNP Characteristics (from Chips)

Transform into related records

Use Genomic Colocation to combine with other features

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).

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

218 Set of Samples Total 201 of 218 Set of Samples selected Sample type

Sample type	Type of sample	Remaining Set of Samples	Set of Samples	Distribution	%
<input checked="" type="checkbox"/> Blood	<input checked="" type="checkbox"/> Specimen from organism	201 (100%)	201 (100%)		(100%)
		12 (6%)	12 (6%)		
		189 (94%)	189 (94%)		

expand all/collapse all Find a variable

Sample type
Type of sample
Keep checked values at top

201 (92%) of 218 Set of Samples have data for this variable

Sample collection
Sample source
Geographic location
Organism under investigation
DNA sequencing

8. 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.

← Add a step to your search strategy ?

[expand all](#) | [collapse all](#)

② Read frequency threshold

80% ↕

② Minor allele frequency \geq

0

② Percent isolates with a base call \geq

20

② SNP Class

Non-Synonymous ↕ ←

② Number of SNPs of above class \geq

5 ←

② Number of SNPs of above class \leq

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 (415) Public (42) Help

Unnamed Search Strategy * ↗

1,578 Genes (6,987 ortholog groups)

Some Genes in your combined result have Transcripts that were not returned by one or both of the two input searches. [Explore](#)

Organism Filter		Gene Results		Genome View		Analyze Results	
<input type="checkbox"/> select all <input type="checkbox"/> clear all <input type="checkbox"/> expand all <input type="checkbox"/> collapse all <input type="checkbox"/> Hide zero counts <input type="checkbox"/> Search organisms...		Genes: 1,578 Transcripts: 1,597 <input type="checkbox"/> Show Only One Transcript Per Gene		Rows per page: 50		Download Add to Basket Add Columns	
Organism Filter <input type="checkbox"/> Plasmodium adleri <input type="checkbox"/> Plasmodium berghei <input type="checkbox"/> Plasmodium billrothi <input type="checkbox"/> Plasmodium blacklocki <input type="checkbox"/> Plasmodium chabaudi <input type="checkbox"/> Plasmodium coatneyi <input type="checkbox"/> Plasmodium cynomolgi <input type="checkbox"/> Plasmodium falciparum 1,578		Gene ID Transcript ID Genomic Location (Gene) Product Description Ortholog Group					
		PF3D7_0100200 PF3D7_0100200.1 Pf3D7_01_v3:38,982..40,207(-) PF3D7_0100400 PF3D7_0100400.1 Pf3D7_01_v3:50,363..51,636(+) PF3D7_0100500 PF3D7_0100500.1 Pf3D7_01_v3:53,169..53,280(-) PF3D7_0100600 PF3D7_0100600.1 Pf3D7_01_v3:53,778..55,006(-)					

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. 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
<i>Plasmodium berghei</i> ANKA	5 asexual stages
<i>Plasmodium berghei</i> ANKA	P. berghei
<i>Plasmodium berghei</i> ANKA	Female asexual stages
<i>Plasmodium chabaudi</i> chabaudi	Transcriptomes from infections.
<i>Plasmodium chabaudi</i> chabaudi	Trophozoites
<i>Plasmodium cynomolgi</i> strain M	Transcriptomes
<i>Plasmodium cynomolgi</i> strain M	Liver stages
<i>Plasmodium cynomolgi</i> strain M	Hypnozoite, schizont and blood stage transcriptomes (laser microdissection) (Cubi et al.)
<i>Plasmodium falciparum</i> 3D7	Gametocyte Transcriptomes (Lasonder et al.)
<i>Plasmodium falciparum</i> 3D7	Mosquito or cultured sporozoites and blood stage transcriptome (NF54) (Hoffmann 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 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

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:

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

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

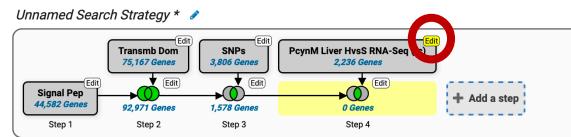
and returns genes when $fold\ change_{up} \geq 2$ or $fold\ change_{down} \geq 2$.

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

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 follow these steps:

- hover your mouse over the RNA-seq step then click on the edit option on that step.



- In the popup window, click on the orthologs link.

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

Details for step PcynM Liver HvsS RNA-Seq (fc)

Experiment: Liver stage hypozoite 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: hypozoite 6-7 days pi, hypozoite 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. 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.

Organism

Note: You must select at least 1 values for this parameter.
1 selected, out of 45

add these | clear these | select only these
select all | clear all

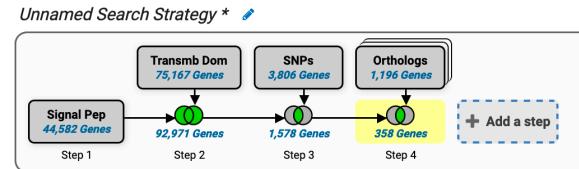
3d7 Plasmodium falciparum
 Plasmodium falciparum 3D7

add these | clear these | select only these
select all | clear all

Syntenic Orthologs Only?

no

- 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”.

Add a step to your search strategy

Combine with other Genes

Choose how to combine with other Genes
4 INTERSECT 5 4 UNION 5 4 MINUS 5 5 MINUS 4

Choose which Genes to combine. From...
A new search An existing strategy My basket
phy

Transform into related records

Use Genomic Colocation to combine with other features

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).

Add a step to your search strategy ×

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

3d7 ←

Plasmodium falciparum
 Plasmodium falciparum 3D7

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

Select orthology profile

Click on + to determine which organisms to include or exclude in the orthology profile.
* = no constraints / * = must be in group / * = must not be in group / * = mixture of constraints

All Organisms expand all | collapse all

- Bacteria (BACT)
- Firmicutes (FIRM)
- Proteobacteria (PROT)
- Other Bacteria (OBAC)
- Archaea (ARCB)
 - Nitrosopumilus maritimus SCMT1 (nmar)
 - Euryarchaeota (EURY)
 - Crenarchaeota (CREN)
 - Nanoarchaeota (NANO)
 - Korarchaeota (KORA)
- Eukaryota (EUKA)
 - Alveolates (ALVE)
 - Amoebozoa (AMOE)
 - Euglenozoia (EUGL)
 - Viridiplantae (VIRI)
 - Fungi (FUNG)
 - Metazoa (META)
 - Nemata (NEMA)
 - Arthropoda (ARTH)
 - Chordata (CHOR)
 - Branchiostoma floridae (bflo)
 - Xenopus (Silurana) tropicalis (xtro)
 - Actinopterygii (ACTI)
 - Aves (AVES)
 - Mammalia (MAMM)
 - Tunicates (TUNI)
 - Other Metazoa (OMET)
 - Other Eukaryota (OEUK)

← **Mammalia (MAMM)**

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

Add a step to your search strategy ×

Combine with other Genes

OrthoPh Pro 3,306 Genes Combine with 250 Genes Step 5 Step 6

Transform into related records

OrthoPh Pro 3,306 Genes Transform into 250 Genes Step 5 Step 6

Use Genomic Colocation to combine with other features

OrthoPh Pro 3,306 Genes Use Genomic Colocation with 250 Genes Step 5 Step 6

Choose how to combine with other Genes

5 INTERSECT 6 5 UNION 6 5 MINUS 6 6 MINUS 5

Choose which Genes to combine. From...

A new search An existing strategy My basket

phen Phenotype Phenotype Evidence

15. 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 (A) intersected with (B) the results of Step 5.

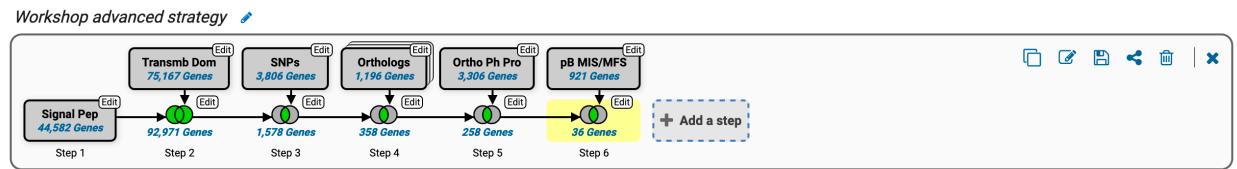
Filter Data Sets: (A) (B) Legend: (A) Association to Genomic Segments (B) Curated Phenotype (C) Similarity (D) Similarity of Association (E) Phenotype Text

Organism	Data Set	Choose a Search
<input type="checkbox"/> Plasmodium berghei ANKA	<input type="checkbox"/> P. berghei knockout (PlasmoGEM) growth phenotypes (Bushell, Gomes and Sanderson et al.)	OP
<input type="checkbox"/> Plasmodium berghei ANKA Plasmodium falciparum 3D7 Plasmodium yoelii yoelii 17XNL	<input type="checkbox"/> RMgM08 - Rodent Malaria genetically modified Parasites (Chris J. Janse)	PT
<input type="checkbox"/> Plasmodium falciparum 3D7	<input type="checkbox"/> eQTL for HB3, Dd2 and 34 progeny (Gonzales et al.)	AS
<input type="checkbox"/> Plasmodium falciparum 3D7	<input type="checkbox"/> piggyBac insertion mutagenesis (John Adams)	CP CP

16. On the next page select the Mutan Fitness Score (MFS) option and choose any score range – generally the more negative the bigger 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>