

Advanced Search Strategies

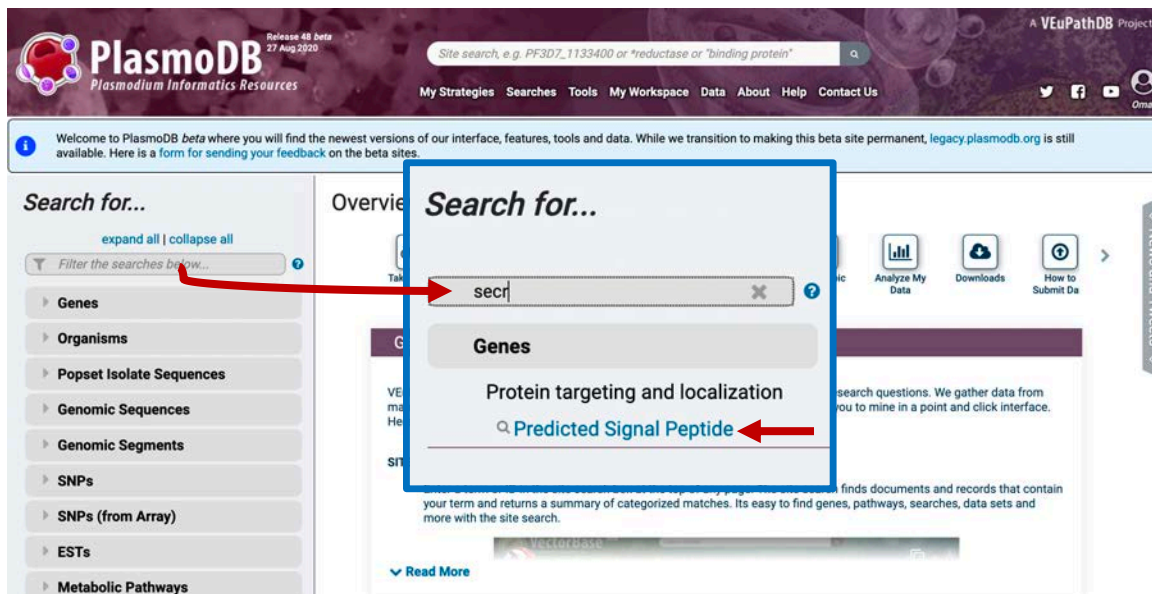
Note: this exercise uses PlasmoDB.org as an example database, but the same functionality is available on all VEuPathDB resources.

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.



- 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 billcollinsi
- ☒ 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 vinckei
- ☒ Plasmodium vivax
- ☒ Plasmodium vivax-like sp.
- ☒ Plasmodium yoelii

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

Advanced Parameters

Get Answer

- 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

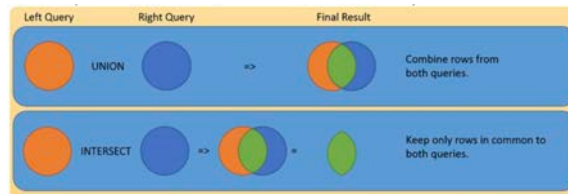
[Opened \(1\)](#) All (415) Public (42) Help

Unnamed Search Strategy *

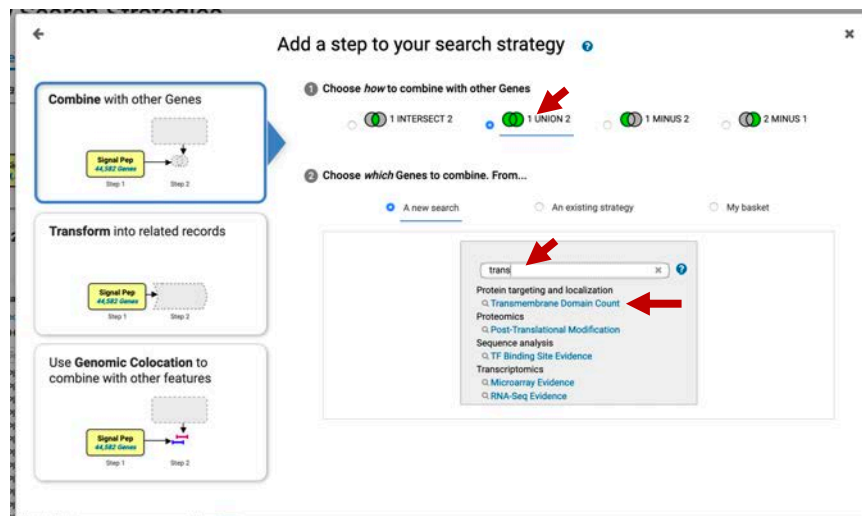
Signal Pep
44,582 Genes

+ Add a step

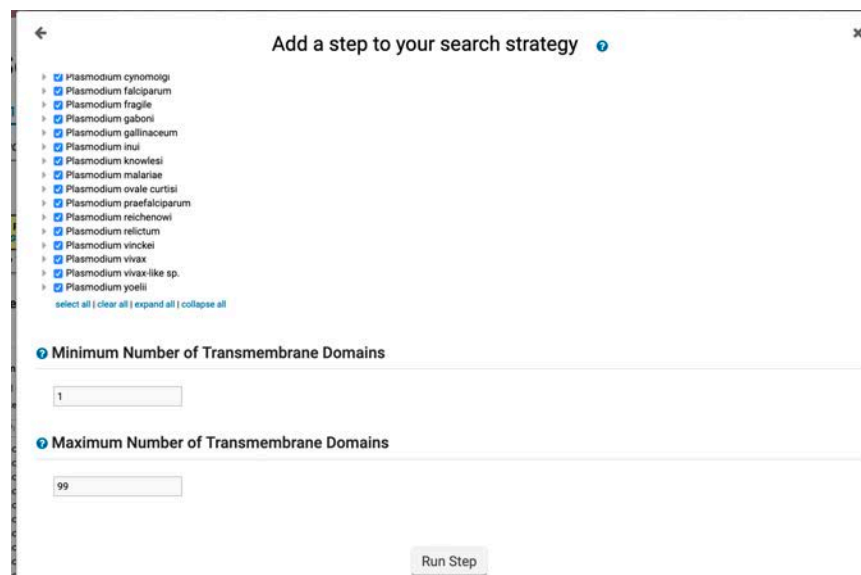
The popup window offers you option to add additional steps and ways to combine the searches (intersect, union, minus). For this exercise we are interested in finding genes that a signal peptide or a TM domain or both. What operation will you use to combine the searches – Union or Intersect?



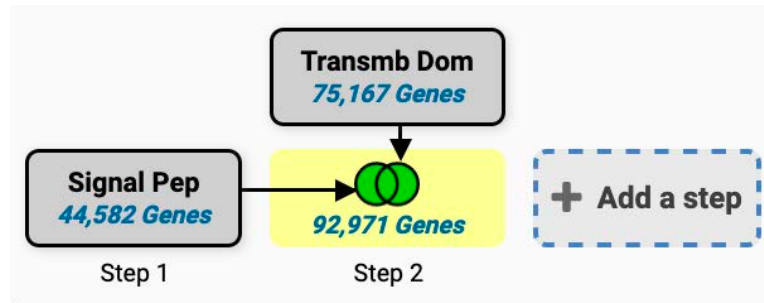
Once you select the option for combining the searches, find the search for transmembrane domain count. Notice that you can use the same query filtering mechanism as before. Start typing transmembrane to find this search. Once you find it click on to open the search parameters.



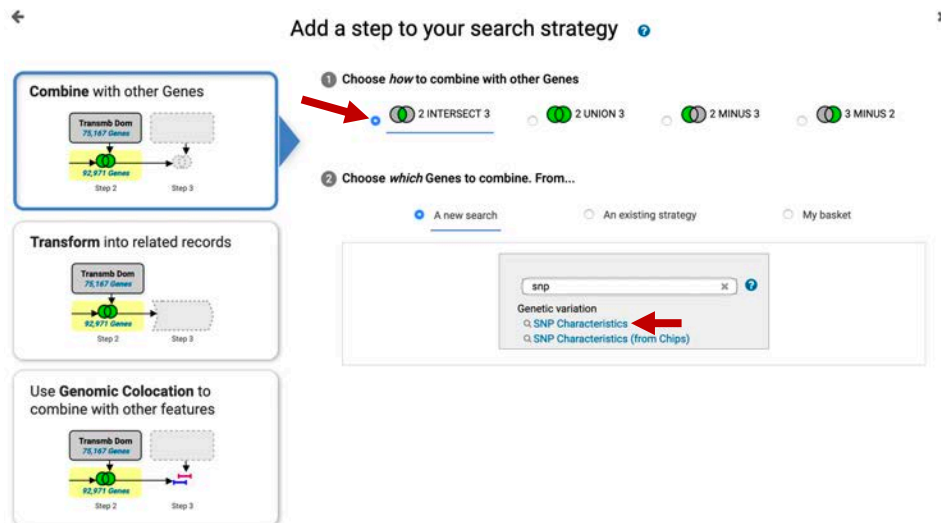
4. For the TM search, again select all organisms, use the default parameters and click on the get answer button.



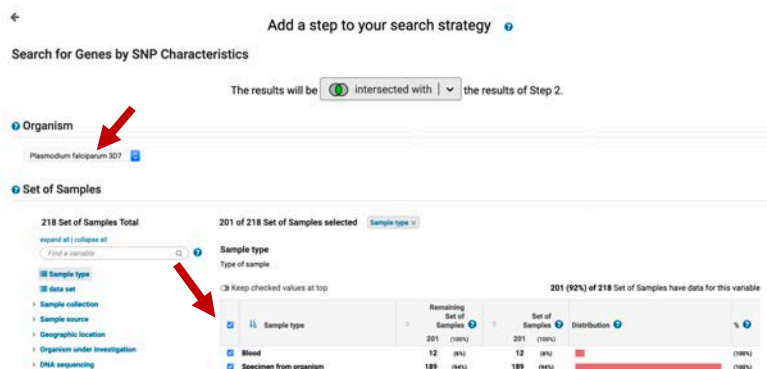
5. How many genes did you get? Since you used a union the number of results should be more than each of the individual steps that were combined.



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).



- 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

Read frequency threshold

80%

Minor allele frequency \geq

0

Percent isolates with a base call \geq

20

SNP Class

Non-Synonymous

Number of SNPs of above class \geq

5

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](#)

Gene Results Genome View Analyze Results

Genes: 1,578 Transcripts: 1,597 Show Only One Transcript Per Gene

select all | clear all | expand all | collapse all

Hide zero counts

Search organisms...

Organism Filter

Plasmodium adleri 0

Plasmodium berghei 0

Plasmodium bilcollinsi 0

Plasmodium blacklocki 0

Plasmodium chabaudi 0

Plasmodium coatneyi 0

Plasmodium cynomolgi 0

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(-)	rifin	OG6_100719
PF3D7_0100400	PF3D7_0100400.1	PF3D7_01_v3:50,363..51,636(+)	rifin	OG6_100719
PF3D7_0100500	PF3D7_0100500.1	PF3D7_01_v3:53,169..53,280(-)	erythrocyte membrane protein 1 (PIEMP1), exon 1, pseudogene	N/A (orthology not determined because poor protein quality)
PF3D7_0100600	PF3D7_0100600.1	PF3D7_01_v3:53,778..55,006(-)	rifin	OG6_100719

Download Add to Basket Add Columns

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.).

The results will be ☒ intersected with ☐ the results of Step 3.

Filter Data Sets:

Legend: **DE** Differential Expression **FC** Fold Change **P** Percentile **SA** SenseAntisense

Organism	Data Set	Choose a Search
<i>Plasmodium berghei</i> ANKA	5 asexual	DE FC P
<i>Plasmodium berghei</i> ANKA	P. bergh	DE FC P
<i>Plasmodium berghei</i> ANKA	Female	DE FC P
<i>Plasmodium chabaudi</i> chabaudi	Transcript	DE FC P
<i>Plasmodium chabaudi</i> chabaudi	Trophozo	DE FC P
<i>Plasmodium cynomolgi</i> strain M	Transcript	DE FC P
<i>Plasmodium cynomolgi</i> strain M	Liver sta	DE FC P
<i>Plasmodium cynomolgi</i> strain M	Hypnozoite, schizont and blood stage transcriptomes (laser microdissection) (Cubi et al.)	DE FC P
<i>Plasmodium falciparum</i> 3D7	Gametocyte Transcriptomes (Lasender et al.)	FC P SA
<i>Plasmodium falciparum</i> 3D7	Merozoite or cultured merozoites and blood stage transcriptome (NFS4) (Hoffmann et al.)	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:

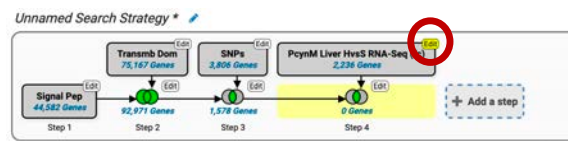
$$\text{fold change}_{\text{up}} = \frac{\text{average expression value in comparison}}{\text{average expression value in reference}}$$

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

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.



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

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

Details for step PcymM Liver HvsS RNA-Seq (fc)
2,236 Genes

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

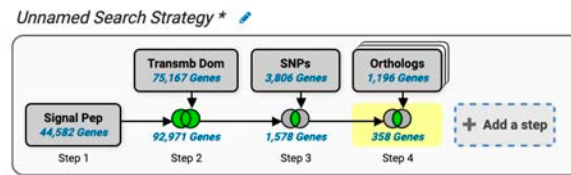
add these | clear these | select only these
select all | clear all

Syntenic Orthologs Only?

no

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”.

Add a step to your search strategy

Combine with other Genes

1 Choose how to combine with other Genes

☒ 4 INTERSECT 5 ☐ 4 UNION 5 ☐ 4 MINUS 5 ☐ 5 MINUS 4

2 Choose which Genes to combine. From...

☒ A new search ☐ An existing strategy ☐ My basket

phy

Orthology and synteny
Orthology Phylogenetic Profile

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 (ARCH)
 - 🔍 Nitrosopumilus maritimus SCM1 (nmar)
 - 🔍 Euryarchaeota (EURY)
 - 🔍 Crenarchaeota (CREN)
 - 🔍 Nanoarchaeota (NANO)
 - 🔍 Korarchaeota (KORA)
 - 🔍 Eukaryota (EUKA)
 - 🔍 Alveolates (ALVE)
 - 🔍 Amoebozoa (AMOE)
 - 🔍 Euglenozoa (EUGL)
 - 🔍 Viridiplantae (VIRI)
 - 🔍 Fungi (FUNG)
 - 🔍 Metazoa (META)
 - 🔍 Nematodes (NEMA)
 - 🔍 Arthropoda (ARTH)
 - 🔍 Chordata (CHOR)
 - 🔍 Branchiostoma floridae (bf1o)
 - 🔍 Xenopus (Silurana) tropicalis (xtro)
 - 🔍 Actinopterygii (ACTI)
 - 🔍 Aves (AVES)
 - 🔍 Mammalia (MAMM)
 - 🔍 Tunicates (TUN1)
 - 🔍 Other Metazoa (OMET)
 - 🔍 Other Eukaryota (OEUK)

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

Ortho Pk Pro 2,304 genes
23d Genes
Step 5

Transform into related records

Ortho Pk Pro 2,304 genes
23d Genes
Step 5

Use Genomic Colocation to combine with other features

Ortho Pk Pro 2,304 genes
23d Genes
Step 5

1 Choose how to combine with other Genes

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

2 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 5 intersected with the results of Step 5.

Filter Data Sets: Legend Add Association to Genomic Segments CP Curated Phenotype S Similarity SA Similarity of Association
Phenotype Text

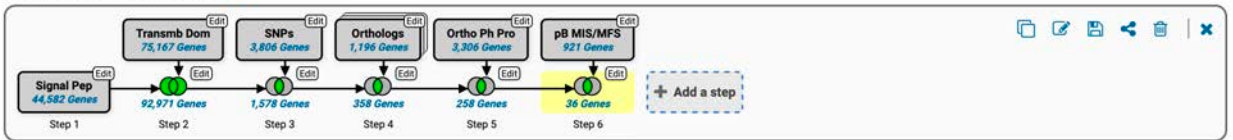
Organism	Data Set	Choose a Search
Plasmodium berghei ANKA	P. berghei knockout (PlasmoGEM) growth phenotypes (Bushell, Gomes and Sanderson et al.)	CP
Plasmodium berghei ANKA Plasmodium falciparum 3D7 Plasmodium yoelii yoelii 17XNL	RMgmDB - Rodent Malaria genetically modified Parasites (Chris J. Janse)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

16. On the next page select the Mutant 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 strategy:

Workshop advanced strategy



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