

Strategies Tutorial

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






Learning objectives:

- Build a multistep strategy
- Use the Text, GO Term, RNA-Seq, and SNP searches
- Combine search results using Boolean operators and the colocation tool
- Transform genes of one organism into their orthologs in another organism
- Infer expression timing from a well-studied organism onto another organism that lacks data.

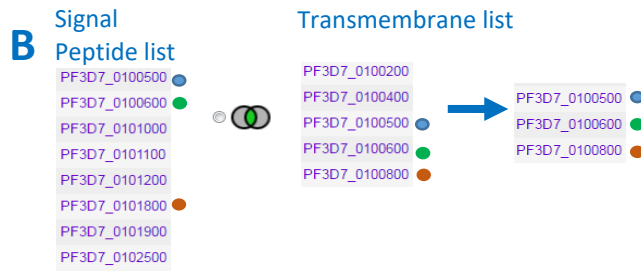
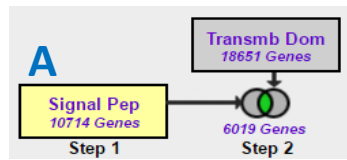
In this tutorial you will find *P. vivax* genes that are likely expressed in gametocytes, act as proteases and have variation in their upstream regions, possibly affecting promoter regions and other gene features. The strategy you build will take advantage of the data rich organism of *P. falciparum* 3D7 to perform three different searches against data from *P. falciparum*. You will take advantage of the orthology profiles to transform the *P. falciparum* genes into their *P. vivax* orthologs and then search for SNPs in the upstream regions of the *P. vivax* genes. The ortholog transform enables you make inferences about genes in *P. vivax*, an organism with limited functional data, based on existing data in the closely related and well-studied *P. falciparum*. The *P. vivax* genes returned by the search are likely to share two biological properties, proteolytic activity and expression in gametocytes, and show variation in their upstream regions, possibly affecting promoter regions or other gene features.

Before we get started... a few words about combining search results:

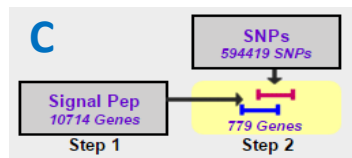
Each search returns a list of IDs. When two searches are combined, the two result sets (list of IDs) are merged. The table shows the 5 options for combining search results.

Operator	:	Combined Result will contain:
 1 INTERSECT 2	:	IDs in common between the two lists
 1 UNION 2	:	IDs from list 1 and list 2
 1 MINUS 2	:	IDs unique to 1
 2 MINUS 1	:	IDs unique to 2
 1 Relative to 2	:	IDs whose features are near each other (collocated) in the genome

If the searches return the same type of genomic feature they can be combined using any of the 5 operators (i.e. search 1 returns genes, search 2 returns genes as in screenshot group A below).

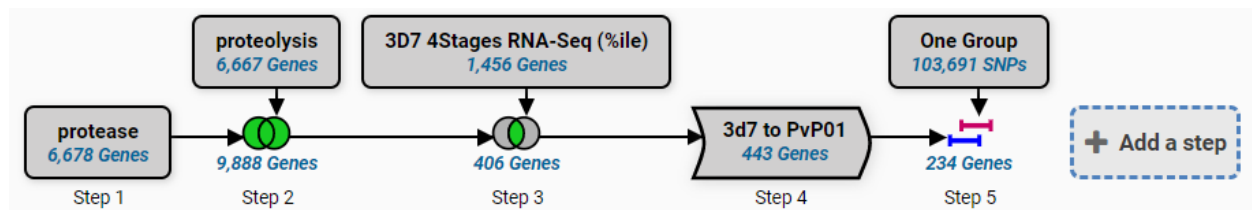


However, searches that return different genomic features will yield no results when combined with intersect, union or minus operators. The Genomic Co-Location tool takes advantage of the genomic location of each gene and each SNP and returns features based on their relative genomic location, i.e. SNPs that are near or within genes.



Building the Strategy:

Find *P. vivax* genes that are possible proteases, likely expressed during the gametocyte stages and contain SNPs in their upstream regions. The final strategy will look like this.



Step by Step Instructions

1. Run a text search using protease as the text term.

Navigation: >PlasmoDB >Search for Genes >Text >Text (product name, notes, etc.)

Identify Genes based on Text (product name, notes, etc.)

[Reset values](#)

Organism

58 selected, out of 58
[select all](#) | [clear all](#) | [expand all](#) | [collapse all](#)

Filter list below...

- ☒ Haemoproteidae
- ☒ Plasmodiidae

Text term (use * as wildcard)

Protease

Fields

- ☒ Alternate product descriptions
- ☒ EC descriptions and numbers
- ☒ Epitopes from IEDB
- ☒ External links
- ☒ Gene ID
- ☒ Gene name or symbol
- ☒ Gene type
- ☒ Genomic sequence ID
- ☒ GO terms
- ☒ InterPro domains
- ☒ Metabolic pathways
- ☒ Names, IDs, and aliases
- ☒ Notes from annotators
- ☒ Organism
- ☒ Ortholog group
- ☒ Orthologs
- ☒ PDB chains
- ☒ Product descriptions
- ☒ PubMed
- ☒ Rodent malaria phenotype
- ☒ Transcripts
- ☒ User comments

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

[Get Answer](#)

Choose all organisms. 59 now in PlasmoDB

Enter protease

Leave all fields checked. We will use the default setting here.

Click Get Answer to initiate the search

You created a one-step strategy by running the text search. The strategy returns 6678 genes that are annotated with the word 'protease'. This annotation could appear in any field that you searched. You can analyze this result by exploring the hits. Look at the data in the columns of the result table. You can add more data with the **Add Columns** button. Clicking a gene ID in the first column will take you to that gene's record page. Please explore your results to see if they make sense. For example, gene product names might contain the word 'protease'.

protease
6,678 Genes

+ Add a step

Step 1

Strategy Box showing your one-step strategy

5,701 Genes (603 ortholog groups) [Revise this search](#)

Gene Results [Genome View](#) [Analyze Results](#)

Genes: 5,701 Transcripts: 5,711 ☐ Show Only One Transcript Per Gene

Rows per page: 100

Download Add to Basket Add Columns

Organism Filter

select all | clear all | expand all | collapse all

☐ Hide zero counts

Search organisms...

Plasmodium adleri 132
Plasmodium berghei 113
Plasmodium bilcollinsi 177
Plasmodium blacklocki 182
Plasmodium chabaudi 100
Plasmodium coatneyi 86
Plasmodium cynomolgi 189
Plasmodium falciparum 2,411
Plasmodium fragile 99
Plasmodium gaboni 234
Plasmodium gallinaceum 101
Plasmodium inui 95
Plasmodium knowlesi 196
Plasmodium malariae 128
Plasmodium ovale curtisi 102
Plasmodium praefalciparum 141
Plasmodium reichenowi 317
Plasmodium relictum 109
Plasmodium vinckei 174
Plasmodium vivax 226

Gene ID	Transcript ID	Organism	Genomic Location (Gene)	Product Description	Score
PADL01_0015600	PADL01_0015600-t36_1	Plasmodium adleri G01	PADLG01_00_20:261..1,218(+)	serine repeat antigen 4, putative	7.613
PADL01_0015700	PADL01_0015700-t36_1	Plasmodium adleri G01	PADLG01_00_20:1,269..2,159(+)	serine repeat antigen 4, putative	8.084
PADL01_0015800	PADL01_0015800-t36_1	Plasmodium adleri G01	PADLG01_00_20:3,838..7,290(+)	serine repeat antigen 5, putative	7.622
PADL01_0015900	PADL01_0015900-t36_1	Plasmodium adleri G01	PADLG01_00_20:7,775..14,786(+)	serine repeat antigen 6, putative	7.622
PADL01_0016000	PADL01_0016000-t36_1	Plasmodium adleri G01	PADLG01_00_20:14,786..21,797(+)	serine repeat antigen 7, putative	7.622
PADL01_0016100	PADL01_0016100-t36_1	Plasmodium adleri G01	PADLG01_00_20:21,797..28,808(+)	serine repeat antigen 8, putative	7.613
PADL01_0004000	PADL01_0004000-t36_1	Plasmodium adleri G01	PADLG01_00_20:28,808..35,819(+)	serine repeat antigen 1, putative	7.622
PADL01_0004100	PADL01_0004100-t36_1	Plasmodium adleri G01	PADLG01_00_20:35,819..42,830(+)	serine repeat antigen 2, putative	7.622

Result List showing all hits from the search

COMMUNITY CHAT

Filter table showing the distribution of hits across the organisms we searched. Apply a filter to see genes from a limited number of species

Add a step choosing to run a search for genes annotated with the biological process gene ontology term – **GO:0006508: proteolysis**. Gene Ontology annotations offer a second line of evidence for finding proteases.

Navigation: Add Step >Combine with other Genes >1 union 2 > A new search >GO Term

protease
6,678 Genes

+ Add a step

Step 1

Add a step to your search strategy

Combine with other Genes

Choose how to combine with other Genes

☐ 1 INTERSECT 2 ☒ 1 UNION 2 ☐ 1 MINUS 2 ☐ 2 MINUS 1

Choose which Genes to combine. From...

☒ A new search ☐ An existing strategy ☐ My basket

GO

Function prediction
Q GO Term
Text
Q Text (product name, notes, etc.)

Search for and choose the GO Term

Which organism is chosen by default for this search? Click 'select all' to run the search on all organisms

Begin typing Proteolysis and then choose the correct GO term from the list

Click Run Step to initiate the search

Add a step to your search strategy

Search for Genes by GO Term

The results will be unioned with the results of Step 1.

Configure Search Learn More View Data Sets Used

Organism

59 selected, out of 59
select all | clear all | expand all | collapse all

Filter list below...

- ☒ Haemoproteidae
- ☒ Plasmodiidae

Evidence

☒ Curated
☒ Computed
select all | clear all

Limit to GO Slim terms

☐ Yes
☒ No

GO Term or GO ID

Select...

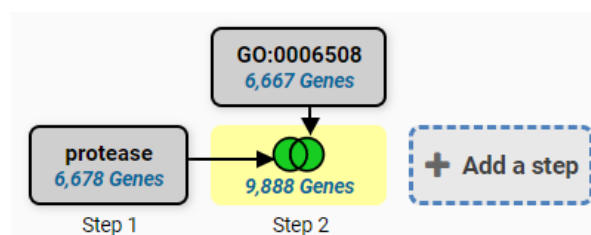
☐ GO Term or GO ID wildcard search

Run Step

Give this search a name (optional)

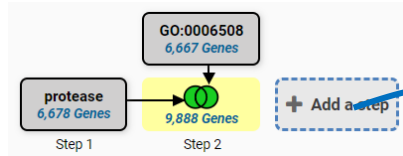
Give this search a weight (optional)

Strategy Result: The GO term search returned 6667 genes annotated with the proteolysis GO term. The union of the text and GO search returns 9888 genes that are suspected to have proteolytic activity.



2. **Add a step choosing to run a search for genes based on Transcript Expression using RNA-Seq Evidence.** Since PlasmoDB has integrated several RNA sequencing data sets you must first choose what data set (experiment) to search before you are taken to the search form to choose parameters. You want the resulting genes to be proteases AND show expression in gametocytes so choose intersect to combine the steps.

Navigation: Add Step >Combine with other Genes >2 intersect 3 >A new search >RNA Seq Evidence



Add a step to your search strategy

Combine with other Genes

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

RNA

Gene models
Gene Model Characteristics
Transcriptomics
Microarray Evidence
RNA-Seq Evidence

Search for and choose

Add a step to your search strategy

Search for Genes by RNA-Seq Evidence

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

Filter Data Set: strand

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

Organism	Data Set	FC	P	SA
Plasmodium falciparum 3D7	Strand specific transcriptomes of 4 life cycle stages (Lopez-Barragan et al.)	FC	P	SA
Plasmodium falciparum 3D7	Intraerythrocytic development cycle transcriptome (2018) (Toenhake et al.)	FC	P	SA
Plasmodium falciparum 3D7	Strand specific transcriptome of the intraerythrocytic developmental cycle (Siegel et al.)	FC	P	SA

Configure Search

Experiment

☒ Strand specific transcriptomes of 4 life cycle stages - Sense
☐ Strand specific transcriptomes of 4 life cycle stages - Antisense

Samples

☐ Late Trophozoite
☐ Schizont
☒ Gametocyte II
☒ Gametocyte V
[select all](#) [clear all](#)

Minimum expression percentile

80

Maximum expression percentile

100

Matches Any or All Selected Samples?

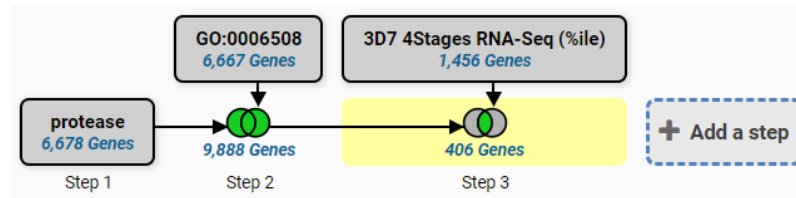
any

Protein Coding Only:

protein coding

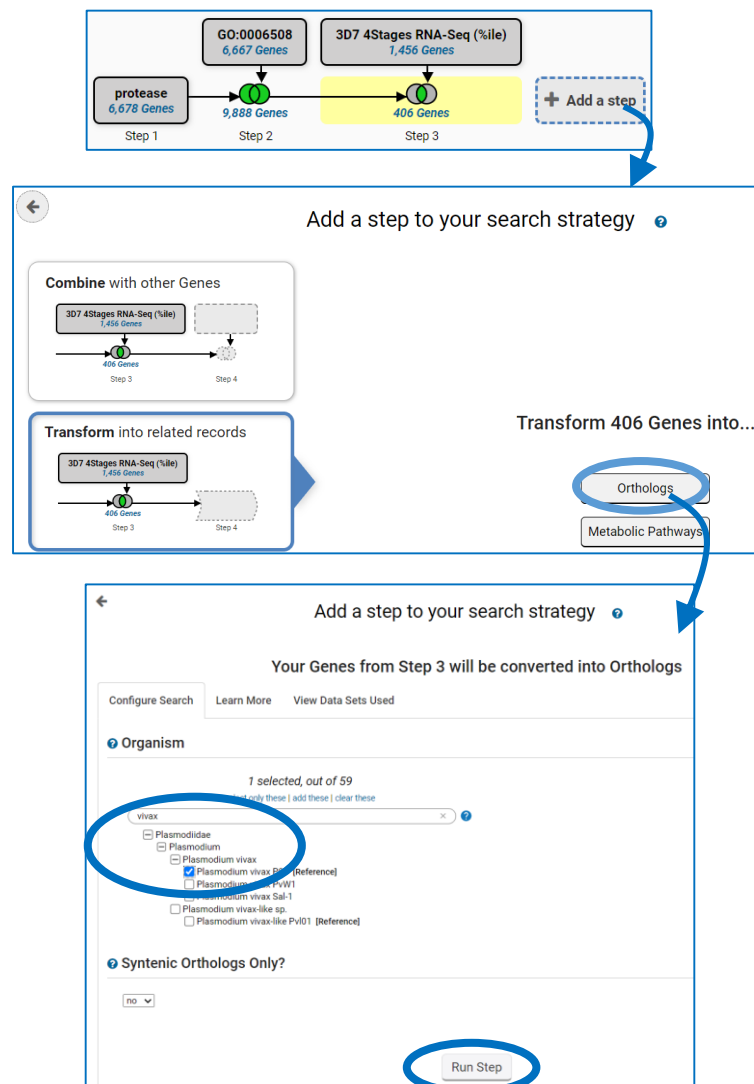
Run Step

Strategy result: We have a three-step strategy that returns 406 *P. falciparum* genes that are suspected proteases with evidence for expression in gametocytes based on RNA Sequencing data. Explore your gene list!!

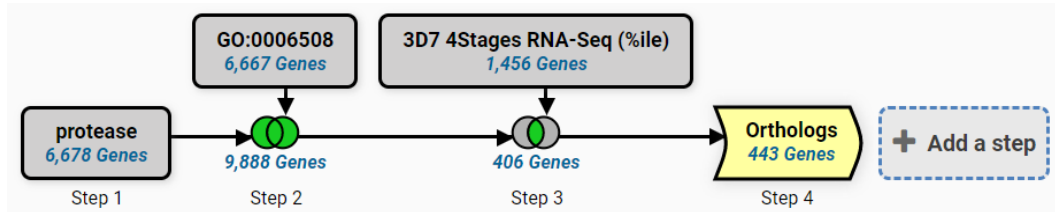


3. **Add a step to the strategy that transforms the 406 *P. falciparum* genes into *P. vivax* genes.** *P. falciparum* is a well-studied organism with active curatorial efforts and large amounts of functional data. A researcher interested in *P. vivax* can take advantage of the *P. falciparum* data by creating a strategy based on *P. falciparum* data then transforming the results to their *P. vivax* orthologs.

Navigation: >Add Step >Transform into related records >Orthologs

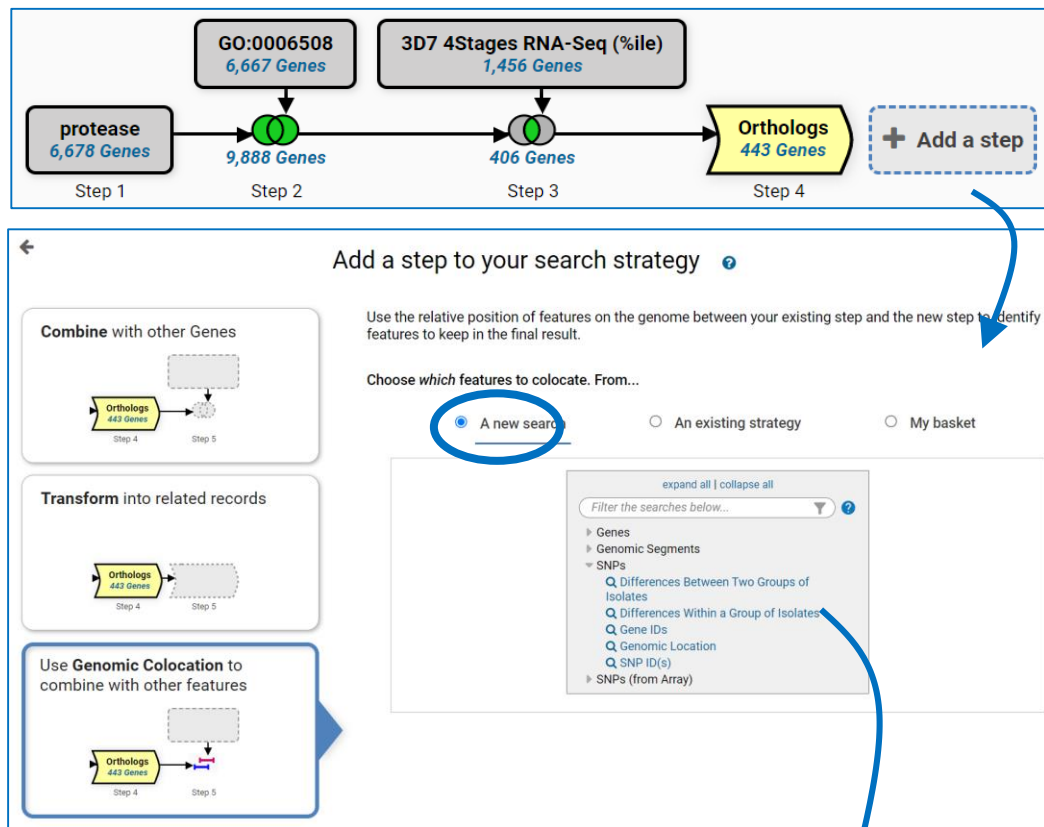


Strategy Result: We have a four-step strategy that returns 443 *P. vivax* genes that are suspected proteases with evidence for expression in gametocytes based on RNA Sequencing data.



4. **Add a step to the strategy that returns *P. vivax* SNPs and collocate those SNPs to the upstream 1000bp of the *P. vivax* genes in step 4.** We can look for variation (SNPs) associated with the genes from Step 4. PlasmoDB integrates whole genome resequencing data from many isolates, and PlasmoDB contains 220 datasets from whole-genome sequencing of *P. vivax* isolates. The search we will use analyzes whole genome resequencing data from all *P. vivax* isolates to find SNPs shared between all isolates. You will notice that initiating the search does not immediately bring up the result, but instead leads you to the collocation tool.

Navigation: >Add Step >Use Genomic Colocation >A new search >Differences Within a Group of Isolates



← Add a step to your search strategy ⓘ

🔍 Organism

The organism you choose will determine the genome to which the SNPs have been mapped. That will also restrict the set of isolates you may choose as SNPs are identified by aligning the reads from those isolates to this genome.

Plasmodium vivax P01

Choose *Plasmodium vivax* P01

🔍 Samples

No filters applied

Use all 220 isolates (Do not filter)

expand all | collapse all

Find a variable

Sample type

Type of sample

Check items below to apply this filter

182 (93%) of 195 Samples have data for this variable

	Sample type	Remaining Samples	Samples	Distribution	%
		182 (100%)	182 (100%)		
<input type="checkbox"/>	Blood	177 (97%)	177 (97%)		(100%)
<input type="checkbox"/>	Specimen from organism	5 (3%)	5 (3%)		(100%)

Read frequency threshold

80%

Minor allele frequency >=

0

Percent isolates with a base call >=

70

Percent isolates with base call = 70

Continue...

Colocation: Because this search returns SNPs and not genes, the only option for combining the two result lists is by relative genomic location. Arrange the statement in the Colocation popup to: **Return Genes from the current step whose upstream 1000bp region overlaps the exact region of a SNP in Step 5 and is on either strand.** Remember to indicate that you want to locate the SNPs in the upstream region of the gene.

← Add a step to your search strategy ⓘ

"Return each **Gene from the current step** whose **upstream region** overlaps the **exact region** of a SNP from the new step and is on **either strand**"

Region

Gene

Region

SNP

Exact

Upstream: 1000 bp

Downstream: 1000 bp

Custom:

begin at: start - 1000 bp

end at: start - 1 bp

Exact

Upstream: 1000 bp

Downstream: 1000 bp

Custom:

begin at: start + 0 bp

end at: stop + 0 bp

Run Step

Strategy: Congratulations! You have completed the strategy and have a list of 234 *P. vivax* genes that are possible proteases, are likely expressed in gametocytes and have upstream SNPs.

This link will retrieve the completed strategy:

<https://plasmodb.org/plasmo/app/workspace/strategies/import/8b35d6c9de221090>

Strategies exercise May 2023 