

Strategies Training Module

In this tutorial you will find genes expressed in gametocytes that are likely proteases and have variation in their upstream regions, possibly affecting promoter regions and other gene features. The strategy you build will combine three different searches that query *P. falciparum* data, then transform the *P. falciparum* genes returned by those searches into their *P. vivax* orthologs and look 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.

Strategies Overview:

The strategy system offers over 100 structured searches that can be combined to produce multi-step strategies. Each search queries a specific data set and **returns a list of IDs** that share the biological characteristic defined by the data.

Searches are accessible from the 'Search For...' menu on the home page and from the 'Searches' dropdown menu. Searches listed under Genes will return a list of gene IDs, while searches listed under 'SNPs' or 'Metabolic Pathways' will return record IDs representing SNPs, or metabolic pathways.

The image shows a screenshot of the 'Search for...' menu and a search results snippet. The menu is on the left, and the search results are on the right. Numbered annotations (1-5) highlight specific features:

- 1** points to the 'Text (product name, notes, etc.)' search option under the 'Text' category.
- 2** points to the 'GO Term' search option under the 'Genes' category.
- 3** points to the 'RNA-Seq Evidence' search option under the 'Transcriptomics' category.
- 4** points to the 'Transform into related records' button in the search results snippet.
- 5** points to the 'Differences Within a Group of Isolates' search option under the 'SNPs' category.






The search results snippet shows a search for 'PF3D7_1133400' or 'reductase or binding protein'. The results are filtered by 'Genes' and 'SNPs'. The 'Transform into related records' button is highlighted with a blue arrow and the number 4. The search results show a list of records, including 'Orthologs' and '107 Genes'.

The 5 searches you will use in this tutorial are:

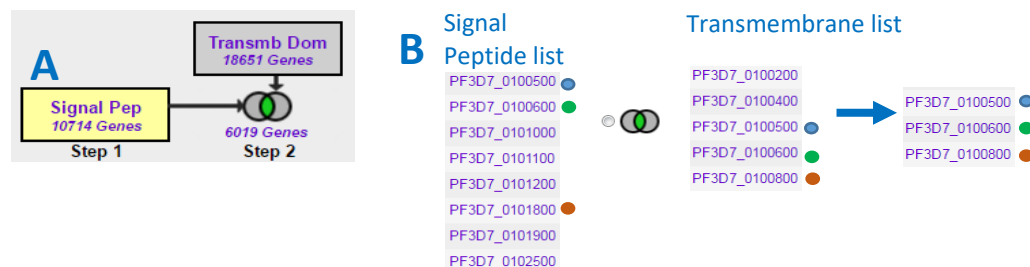
1. Identify Genes by Text (product name, notes, etc.) – The search compares your term against the text in the fields that you specify, returning genes that have a match.
2. Identify Genes by GO Term – Find genes based on the Gene Ontology (GO) Term(s) or ID(s) assigned to them.
3. Identify Genes based on RNA Seq Evidence – PlasmoDB integrates raw RNA sequencing data from many different experiments and analyzes all data according to the same workflow to produce expression values. This search returns genes based on their transcript expression as measure by RNA sequencing.
4. Transform by Orthology – PlasmoDB integrates ortholog profiles from OrthoMCL. The OrthoMCL algorithm clusters proteins into ortholog groups based on BLAST similarity across at 150 genomes that span the tree of life. The transform we perform here will convert a list of genes in one organism to their orthologs in a different organism. In this case, we will transform a list of *P. falciparum* genes into their *P. vivax* orthologs.
5. Identify SNPs based on Differences within a Group of Isolates – PlasmoDB integrates whole genome resequencing of isolates and analyzes each isolate for single nucleotide polymorphisms compared to a reference genome. This search returns SNPs that are shared between all the *P. vivax* isolates that are integrated in PlasmoDB.

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).

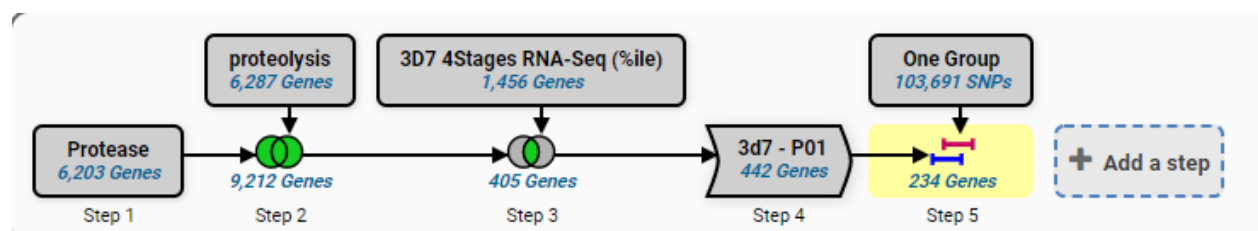


However, searches that return different genomic features will yield no results when combined with intersect, union or minus operators. This is illustrated in screenshot groupings C and D below. Because genes and SNPs are different genomic features, there are no IDs in the list of genes (Step 1) that are present in the list of SNPs (Step 2). To combine a search that returns genes with a search that returns SNPs, you must use the collocation option (1 relative to 2). We know the genomic location of each gene and each SNP and the collocation option is designed to return 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. This search strategy employs 4 searches, an ortholog transform and the collocation tool to integrate SNP information. Steps 1 and 2 return *P. falciparum* proteases using two different lines of evidence – a text search in step 1 and a Gene Ontology (GO) term search in step 2. These searches are combined with a union to obtain a more comprehensive list of possible proteases. Step 3 returns genes with evidence for expression during the gametocyte stages based on RNA sequencing data collected in *P. falciparum*. Steps 2 and 3 are combined using the intersect operator to produce a list of genes that have BOTH biological properties: these genes are likely proteases with evidence for expression during gametocyte stages. In the next step, the *P. falciparum* genes returned in the step 3 result are transformed into their *P. vivax* orthologs. This results in a set of 442 *P. vivax* genes with suspected protease activity and expression in gametocytes based on annotation and experimental evidence from *P. falciparum*, an organism for which more complete annotation and functional genomics data is available. In Step 5 we look for single nucleotide polymorphisms (SNPs) among isolates of *P. vivax* and collocate these SNPs to the upstream regions of the *P. vivax* genes. The final result is a set of 218 *P. vivax* genes that are likely proteases expressed in the gametocyte stage and that have SNPs in their upstream regions. Your strategy should look like this when you are done:



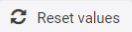
Step by Step Instructions

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

Identify Genes by Text (product name, notes, etc.): Using the Text Search, find genes whose records contain the term 'protease'. To reach the text search, click on the link in the home page 'Search For...' menu. The page opens showing a list of parameters that are needed to query the data. Every search is loaded with default parameters so that you can click Get Answer and run the search. Change the Text term to 'protease' and click Get Answer to initiate the search. The search results are displayed in the My Strategies section which consists of a strategy panel followed by a filter table and a result table.

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

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



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

☒ Aconoidasida

☒ Haemosporida

select all | clear all | expand all | collapse all

Text term (use * as wildcard)

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

Choose all 56 organisms

Enter protease

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

Click Get Answer to initiate the search.

Parameters:

Organism	:	Default - all
Text term (use * as wildcard)	:	protease
Fields	:	Default - all

Results and strategy: You created a one-step strategy by running the text search. The strategy returns 6203 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'.

Strategy Box showing your one-step strategy

Text 5,701 Genes Step 1 Add a step

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

Organism Filter
select all | clear all | expand all | collapse all
Hide zero counts
Search organisms...

Organism	Count
Plasmodium adleri	132
Plasmodium berghei	113
Plasmodium billcollinsi	177
Plasmodium blacklocki	192
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

Result List showing all hits

Gene ID	Transcript ID	Organism	Genomic Location (Gene)	Product Description	Score
PADL01_0015600	PADL01_0015600-136_1	Plasmodium adleri G01	PADLG01_00_20:261..1,218(+)	serine repeat antigen 4, putative	7.613
PADL01_0015700	PADL01_0015700-136_1	Plasmodium adleri G01	PADLG01_00_20:1,269..2,159(+)	serine repeat antigen 4, putative	8.084
PADL01_0015800	PADL01_0015800-136_1	Plasmodium adleri G01	PADLG01_00_20:3,838..7,290(+)	serine repeat antigen 5, putative	7.622
PADL01_0015900	PADL01_0015900-136_1	Plasmodium adleri G01	PADLG01_00_20:7,775..11,227(+)	serine repeat antigen 6, putative	7.622
PADL01_0016000	PADL01_0016000-136_1	Plasmodium adleri G01	PADLG01_00_20:11,700..15,152(+)	serine repeat antigen 7, putative	7.622
PADL01_0016100	PADL01_0016100-136_1	Plasmodium adleri G01	PADLG01_00_20:15,625..19,077(+)	serine repeat antigen 8, putative	7.613
PADL01_0004000	PADL01_0004000-136_1	Plasmodium adleri G01	PADLG01_00_20:19,600..23,052(+)	serine repeat antigen 1	7.622
PADL01_0004100	PADL01_0004100-136_1	Plasmodium adleri G01	PADLG01_00_5:10,775..14,786(+)	serine repeat antigen 2	7.622

Filter table showing the distribution of hits across the organisms we searched. Click a # to show only that 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. The ontologies are a controlled vocabulary for describing the molecular function, biological process and subcellular location of a gene product. GO annotations in PlasmoDB were either provided by the sequencing and annotation centers or inferred based on a gene's similarity to protein domains from the InterPro databases. The GO Term search returns a gene if it is annotated with the GO term that you are looking for. Let's use that search to look for genes annotated with GO:0006508: proteolysis. We will union the text search results with our GO term results when we combine the results of the two searches.

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

Protease
6,203 Genes
Step 1

+ Add a step

Add a step to your search strategy

Combine with other Genes

1 Choose *how* to combine with other Genes

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

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

☒ A new search ☐ An existing strategy ☐ My basket

GO

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

Search for and choose the GO Term search.

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

Add Step 2 : GO Term

Organism

0 selected, out of 45

Filter list below...

Plasmodium
select all | clear all | expand all | collapse all

Evidence

☒ Curated
☒ Computed

Limit to GO Slim terms

☐ Yes
☒ No

GO Term or GO ID

Begin typing to see suggestions...

Begin typing to see suggestions to choose from (CTRL or CMD click to select multiple)

GO Term or GO ID wildcard search

N/A

Run Step

Click Run Step to initiate the search

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

Give this search a name (optional)

Give this search a weight (optional)

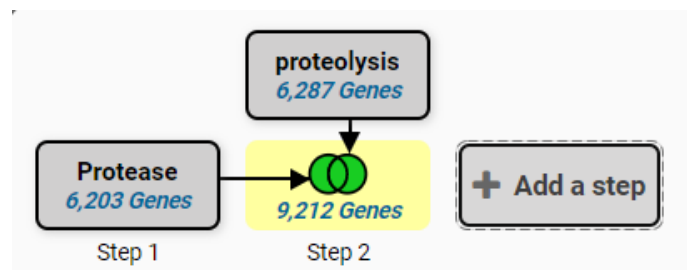
Parameters:

Organism	:	Choose All
Evidence	:	Default
Limit to GO Slim Terms?		Default
GO Term or GO ID	:	GO:0006508 : proteolysis
Free Text (use '*' for wildcard)	:	N/A

Combine:

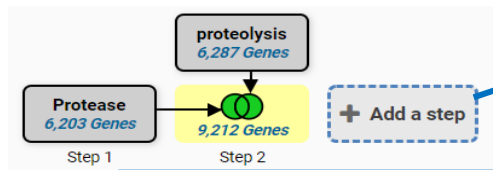


Strategy Result: The GO term search returned 6287 genes annotated with the proteolysis GO term. The union of the text and GO search returns 9212 genes that are suspected to have proteolytic activity.



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. Use the Filter Data set tool to choose the Percentile search (P) for 'Strand specific Transcriptomes of 4 life cycle stages (Lopez-Barragan et al)'. This data set contains the RNA sequencing analysis of two gametocyte samples. Running the percentile search using the default parameters will return the genes whose expression levels are in the top 20% for those samples.

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 the RNA-Seq evidence.

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:

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

Organism	Data Set	FC	P	SA
<i>Plasmodium falciparum</i> 3D7	Strand specific transcriptomes of 4 life cycle stages (Lopez-Barragan et al.)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<i>Plasmodium falciparum</i> 3D7	Intraerythrocytic development cycle transcriptome (2018) (Toenhake et al.)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<i>Plasmodium falciparum</i> 3D7	Strand specific transcriptome of the intraerythrocytic developmental cycle (Siegel et al.)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Add a step to your search strategy

Experiment

Strand specific transcriptomes of 4 life cycle stages - Sense

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:

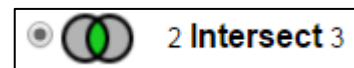
protein coding

Run Step

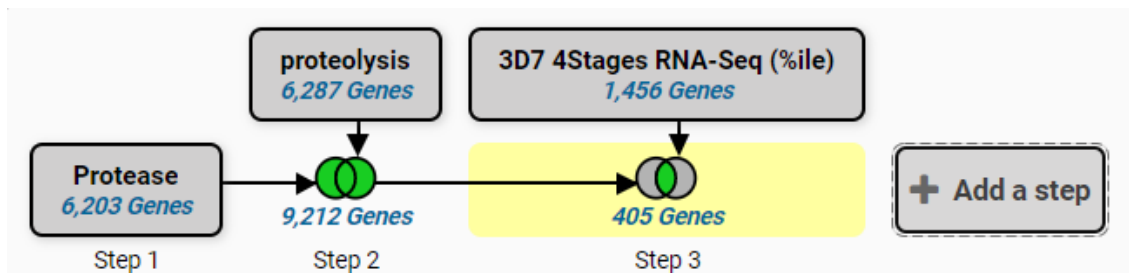
Parameters:

Experiment	:	Strand specific transcriptomes of 4 life cycle stages sense strand
Samples	:	Gametocyte II, Gametocyte V
Minimum expression percentile	:	default
Maximum expression percentile	:	default
Matches Any or All Selected Samples?	:	default
Protein Coding Only:	:	default

Combine: Intersecting this search with the previous result will produce a list of genes that are common to both result lists.



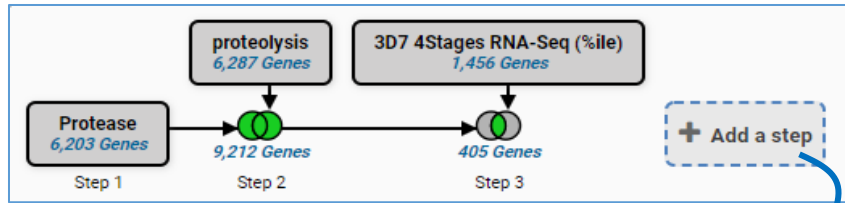
Strategy result: We have a three-step strategy that returns 405 *P. falciparum* genes that are suspected proteases with evidence for expression in gametocytes based on RNA Sequencing data. Explore your gene list!!



3. Add a step to the strategy that transforms the 405 *P. falciparum* genes into *P. vivax* genes.

P. falciparum is a well-studied organism with active curatorial efforts and large amounts of functional data. For example, PlasmoDB has 22 RNA sequencing and 11 microarray data sets integrated for *P. falciparum*, but only 5 RNA-Seq and 2 microarray for *P. vivax*. A researcher interested in *P. vivax* can take advantage of the *P. falciparum* data by creating a strategy based on *P. falciparum* data to retrieve genes with the biological properties they are interested in, and then transforming the results to their *P. vivax* orthologs.

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



← Add a step to your search strategy ?

Combine with other Genes

3D7 4Stages RNA-Seq (%ile) (1,456 Genes)

91 Genes (Step 3) → Step 4

Transform into related records

3D7 4Stages RNA-Seq (%ile) (1,456 Genes)

91 Genes (Step 3) → Step 4

Use Genomic Colocation to combine with other features

3D7 4Stages RNA-Seq (%ile) (1,456 Genes)

91 Genes (Step 3) → Step 4

Transform 91 Genes into...

- Orthologs
- Metabolic Pathways
- Compounds

← Add a step to your search strategy ?

Your Genes from Step 3 will be converted into Orthologs

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

viva

☐ Plasmodium vivax

☒ Plasmodium vivax P01

☐ Plasmodium vivax Sal-1

☐ Plasmodium vivax-like sp.

☐ Plasmodium vivax-like Pvl01

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

Syntenic Orthologs Only?

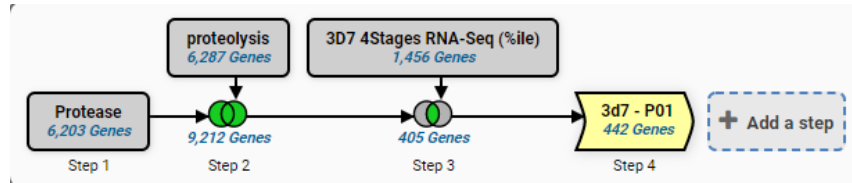
no

Run Step

Parameters: Choose only *P. vivax* P01 in the Organism parameter of the Add Step Popup.

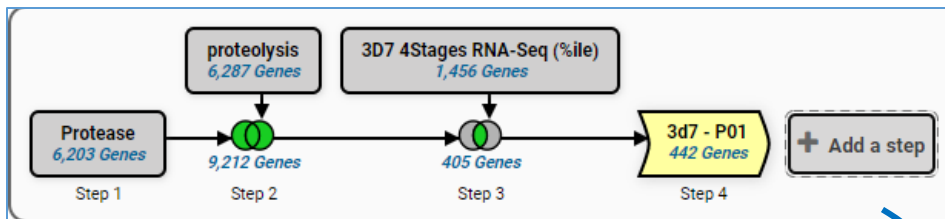
Combine: The ortholog transform function does not combine lists, but instead transforms the results into orthologs from a different species.

Strategy Result: We have a four-step strategy that returns 442 *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. PlasmoDB analyzes the whole genome sequencing reads by aligning them to the reference genome and then walking down the genome one base at a time looking for bases in the isolate that do not match the reference sequence. The SNPs are loaded in the database along with other information such as how many sequencing reads supported the SNP call and the genomic location of the SNP. 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

Use the relative position of features on the genome between your existing step and the new step to identify features to keep in the final result.

Choose which features to colocate. From...

☒ A new search ☐ An existing strategy ☐ My basket

expand all | collapse all

filter the searches below...

- Genes
- Genomic Segments
- SNPs
 - Differences Between Two Groups of Isolates
 - Differences Within a Group of Isolates
 - Gene IDs
 - Genomic Location
 - SNP ID(s)
 - SNPs (from Array)

Combine with other Genes

Transform into related records

Use Genomic Colocation to combine with other features

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

Samples

195 Samples Total

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	%
Blood	177 (97%)	177 (97%)		(100%)
Specimen from organism	5 (3%)	5 (3%)		(100%)

Read frequency threshold

80%

Minor allele frequency >=

0

Percent isolates with a base call >=

70

Run Step

Parameters:

Organism	:	<i>P. vivax</i> P01
Isolates	:	Default = All Isolates (195)
Read frequency threshold	:	Default - 80%
Minor allele frequency >=	:	Default - 0
Percent isolates with a base call >=	:	Default - 70

Colocation: Our SNP search returned 103691 SNPs (number in red in Colocation tool below). 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/085f18c1ec57312d>

