

# Search Strategies in VEuPathDB

## 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

## Introduction

**Search strategies** in VEuPathDB are a unique tool for mining our vast data resources. They enable genome-wide queries as part of *in silico* experiments.

In this tutorial you will find *P. vivax* genes<sup>1</sup> 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 to 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.

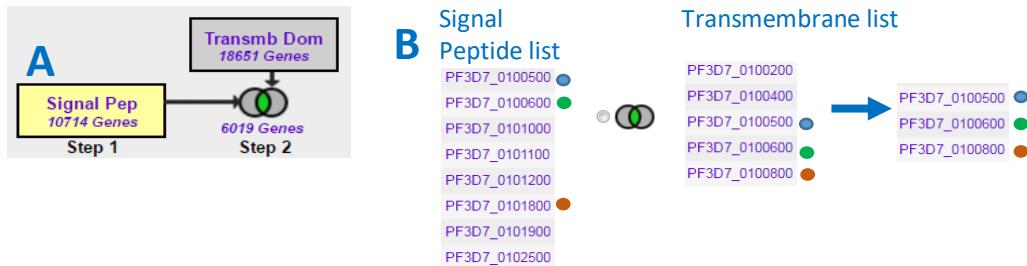
<sup>1</sup> Note: This exercise uses PlasmoDB.org as an example, but the same functionality is available on all VEuPathDB genomics resources

## 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	1 INTERSECT 2	:	IDs in common between the two lists
2	1 UNION 2	:	IDs from list 1 and list 2
3	1 MINUS 2	:	IDs unique to 1
4	2 MINUS 1	:	IDs unique to 2
5	1 Relative to 2	:	IDs whose features are near each other (collocated) in the genome
6	IGNORE 2	:	Ignore the next step
7	IGNORE 1	:	Ignore previous step

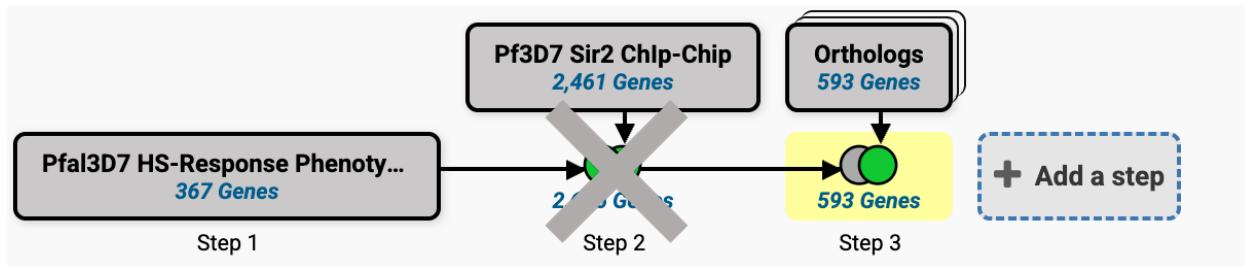
If the searches return the same type of genomic feature they can be combined using any of the 5 operators (i.e. search 1 returns genes, search 2 returns genes as in screenshot group A below).



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.



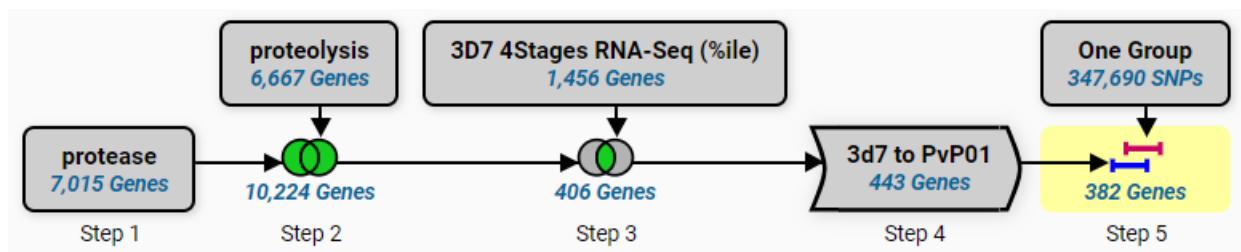
In multistep search strategies, you can also use the **Ignore** operator (options 6 and 7 in the table above) to mask off steps before or after certain search results. This also allows you to apply different search criteria without duplicating search strategies or deleting individual steps.



## 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<sup>2</sup>.



## 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.)

<sup>2</sup> VEuPathDB is updated regularly, so the numbers you see while completing the exercise may differ slightly from those shown in the screenshots.

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

 Reset values

## Organism

62 selected, out of 62

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

Filter list below...



 ?

Reference only

▶  Haemoproteidae

▶  Plasmodiidae



Choose all organisms

## Text term (use \* as wildcard)

Protease



Enter 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](#)

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

Protease  
7,015 Genes

Step 1

+ Add a step

Click Get Answer to  
initiate the search



You created a one-step strategy by running the text search. The strategy returns 7017 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'.

Unnamed Search Strategy \*

**Text**  
7,017 Genes

+ Add a step

Step 1

Strategy Box showing your one-step strategy

6,454 Genes (944 ortholog groups) Revise this search

Gene Results Genome View Analyze Results

Genes: 6,454 Transcripts: 6,474 (hiding 20)  Show Only One Transcript Per Gene

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

Rows per page: 50

Download Send to... Add Columns

Gene ID Transcript ID Organism Genomic Location Prod

Htart\_000017900 Htart\_000017900.1 Haemoproteus tartakovskii strain SISKIN1 LSRZ01000007.31,041..31,490(+) hypothetical protein  
Htart\_000021300 Htart\_000021300.1 Haemoproteus tartakovskii strain SISKIN1 LSRZ01000009.63,972..65,153(+) 26S protease subunit 1  
Htart\_000033100 Htart\_000033100.1 Haemoproteus tartakovskii strain SISKIN1 LSRZ01000018.51,994..53,506(+) rhomboid protein  
Htart\_000035200 Htart\_000035200.1 Haemoproteus tartakovskii strain SISKIN1 LSRZ01000020.43,196..46,273(+) ATP-dependent protease  
Htart\_000035500 Htart\_000035500.1 Haemoproteus tartakovskii strain SISKIN1 LSRZ01000020.50,926..54,873(+) ubiquitin  
Htart\_000050500 Htart\_000050500.1 Haemoproteus tartakovskii strain SISKIN1 LSRZ01000034.11,611..13,812(-) ubiquitin  
Htart\_000094500 Htart\_000094500.1 Haemoproteus tartakovskii strain SISKIN1 LSRZ01000080.485..1,801(+) 26S protease subunit 2

Result List showing all hits from the search

Filter table showing the distribution of hits across the organisms we searched. Click a # to show only that species

Organism Filter

select all | clear all | expand all | collapse all  Hide zero counts

Search organisms...

Rows per page: 50

Download Send to... Add Columns

Gene ID Transcript ID Organism Genomic Location Prod

Htart\_000017900 Htart\_000017900.1 Haemoproteus tartakovskii strain SISKIN1 LSRZ01000007.31,041..31,490(+) hypothetical protein  
Htart\_000021300 Htart\_000021300.1 Haemoproteus tartakovskii strain SISKIN1 LSRZ01000009.63,972..65,153(+) 26S protease subunit 1  
Htart\_000033100 Htart\_000033100.1 Haemoproteus tartakovskii strain SISKIN1 LSRZ01000018.51,994..53,506(+) rhomboid protein  
Htart\_000035200 Htart\_000035200.1 Haemoproteus tartakovskii strain SISKIN1 LSRZ01000020.43,196..46,273(+) ATP-dependent protease  
Htart\_000035500 Htart\_000035500.1 Haemoproteus tartakovskii strain SISKIN1 LSRZ01000020.50,926..54,873(+) ubiquitin  
Htart\_000050500 Htart\_000050500.1 Haemoproteus tartakovskii strain SISKIN1 LSRZ01000034.11,611..13,812(-) ubiquitin  
Htart\_000094500 Htart\_000094500.1 Haemoproteus tartakovskii strain SISKIN1 LSRZ01000080.485..1,801(+) 26S protease subunit 2

Result List showing all hits from the search

Filter table showing the distribution of hits across the organisms we searched. Click a # to show only that species

2. 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

**Step 1:** Protease 7,015 Genes

**Add a step**

**Step 2: Add a step to your search strategy**

- 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

Search for and choose the GO Term search.

**Step 3: Add Step 2 : GO Term**

Organism: 0 selected, out of 45

Evidence: Curated, Computed

Limit to GO Slim terms: 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

Give this search a name (optional)

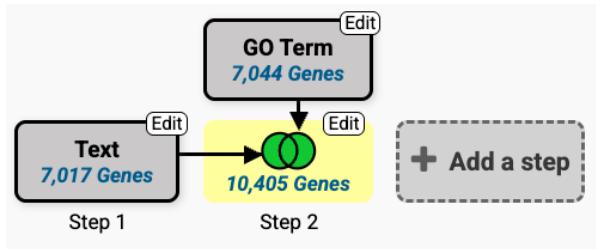
Give this search a weight (optional)

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

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

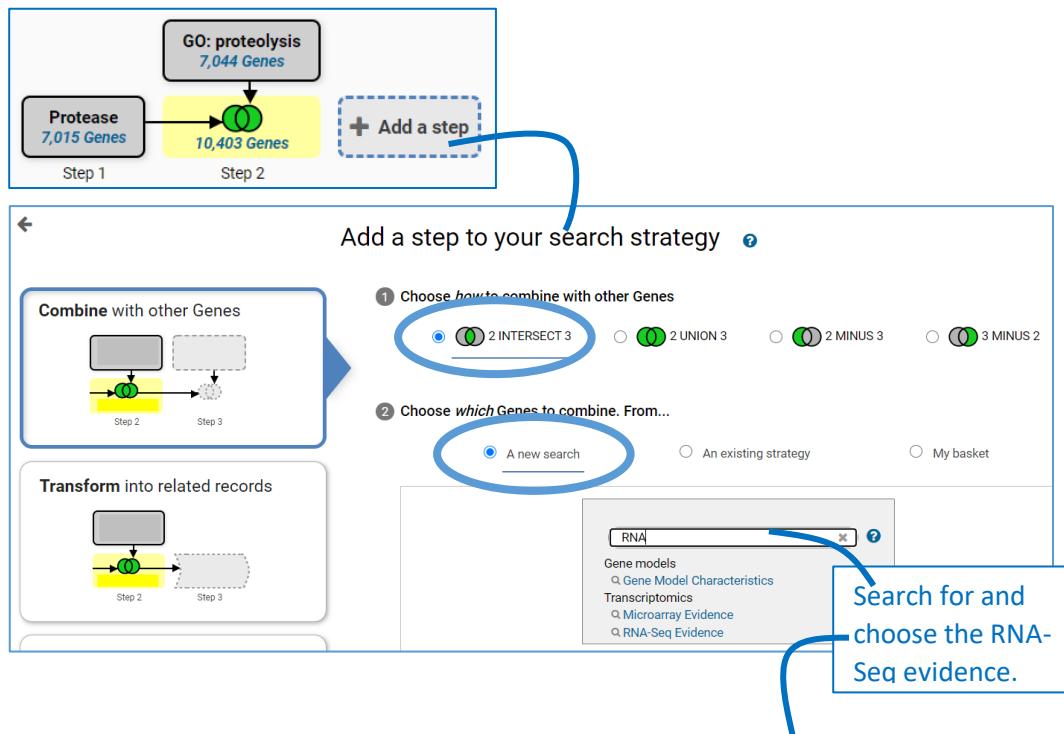
Click Run Step to initiate the search

**Strategy Result:** The GO term search returned 7044 genes annotated with the proteolysis GO term. The union of the text and GO search returns 10,405 genes that are suspected to have proteolytic activity.



3. **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. Choose the experiment “Strand specific transcriptomes of 4 life cycle stages (Lopez-Barragan et al.)”. This data contains RNA-Seq transcriptomes for trophozoites, schizonts and gametocytes. Since you want the resulting genes to be proteases AND show expression in gametocytes, 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 ?

### Search for Genes by RNA-Seq Evidence

The results will be (+) intersected with | (−) subtracted from | (×) multiplied by | (÷) divided by the results of Step 2.

Filter Data Set:  ?

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

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

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

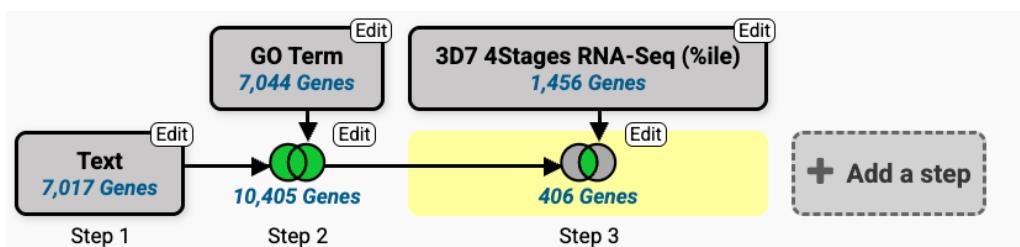
**Maximum expression percentile**  
100

**Matches Any or All Selected Samples?**  
any

**Protein Coding Only:**  
protein coding

Run Step

**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!!



#### 4. Add a step to the strategy that transforms the 406 *P. falciparum* genes into *P. vivax* genes.

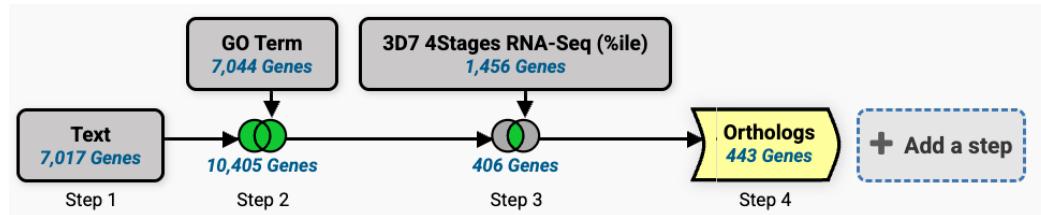
*P. falciparum* is a well-studied organism with active curatorial efforts and large amounts of functional data. A researcher interested in *P. vivax* can take advantage of the *P. falciparum* data by creating a strategy based on *P. falciparum* data then transforming the results to their *P. vivax* orthologs.

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

The screenshot shows a bioinformatics search interface with the following components:

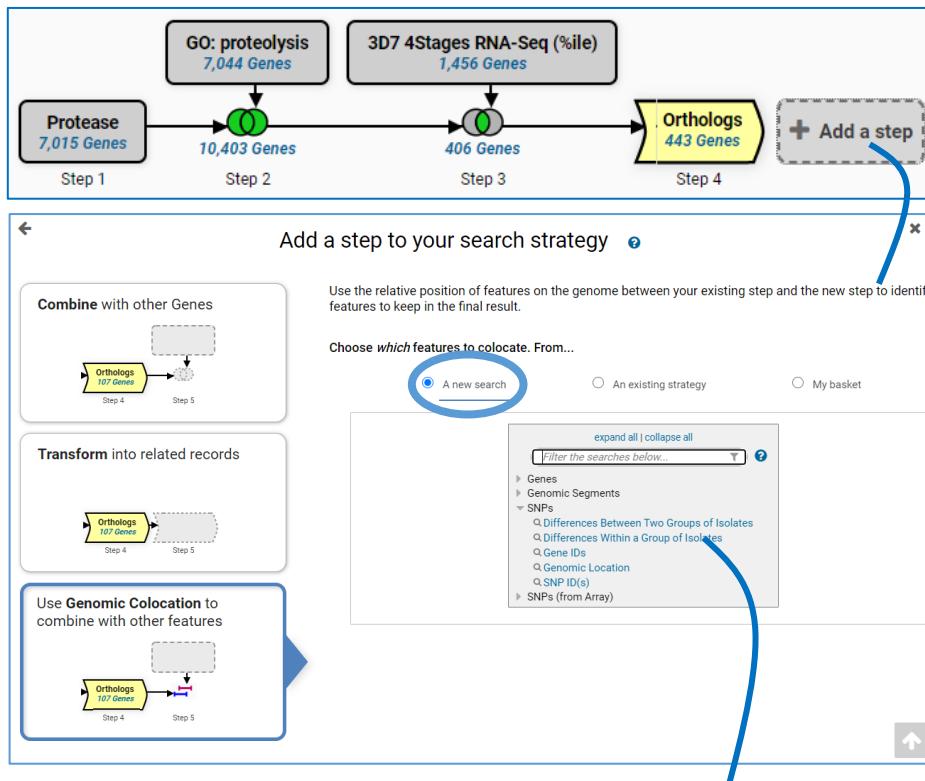
- Top Navigation Bar:**
  - GO: proteolysis 7,044 Genes
  - 3D7 4Stages RNA-Seq (%ile) 1,456 Genes
  - Step 1: Protease 7,015 Genes
  - Step 2: 10,403 Genes
  - Step 3: 406 Genes
  - Add a step (button)
- Central Search Strategy Editor:**
  - Combine with other Genes:** Step 3 (406 Genes) → Step 4 (dashed box)
  - Transform into related records:** Step 3 (406 Genes) → Step 4 (dashed box)
    - 3D7 4Stages RNA-Seq (%ile) 1,456 Genes
    - 406 Genes
  - Use Genomic Colocation to combine with other features:** Step 3 (406 Genes) → Step 4 (dashed box)
  - Transform 406 Genes into...** (dropdown menu)
    - Orthologs (selected, circled)
    - Metabolic Pathways
    - Compounds
- Detailed Configuration Panel:**
  - Add a step to your search strategy**
  - Your Genes from Step 3 will be converted into Orthologs
  - Configure Search (button)
  - Learn More (link)
  - View Data Sets Used (link)
  - Reset values to default (button)
  - Organism:** vivax (selected, circled)
    - 1 selected, out of 62
    - select only these | add these | clear these
    - vivax
      - Plasmodiidae
      - Plasmodium
        - Plasmodium vivax
          - Plasmodium vivax P01 [Reference] (selected, checked)
          - Plasmodium vivax PAM
          - Plasmodium vivax PvV1
          - Plasmodium vivax Sal-1
          - Plasmodium vivax-like sp.
          - Plasmodium vivax-like PvI01 [Reference]
    - Syntenic Orthologs Only?**
      - no (radio button)
    - Run Step** (button, circled)

**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.



- 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 236 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 colocation 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.

Choose Plasmodium vivax P01

**Samples**

No filters applied

Sample type Type of sample

Check items below to apply this filter 182 (93%) of 195 Samples have data for this variable

<input type="checkbox"/>	Sample type	Remaining Samples <span style="color: blue;">?</span>	Samples <span style="color: blue;">?</span>	Distribution <span style="color: blue;">?</span>	% <span style="color: blue;">?</span>
<input type="checkbox"/>	Blood	177 (97%)	177 (97%)	<div style="width: 97%; background-color: #e6f2ff;"></div>	(100%)
<input type="checkbox"/>	Specimen from organism	5 (3%)	5 (3%)	<div style="width: 3%; background-color: #ff9999;"></div>	(100%)

Use all 236 isolates (Do not filter)

Read frequency threshold 80%

Minor allele frequency >= 0

Percent isolates with base call = 70

Percent isolates with a 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

Exact  
 Upstream: 1000 bp  
 Downstream: 1000 bp  
 Custom:  
begin at: start - 1000 bp  
end at: stop + 1 bp

Region SNP

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 382 *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/85844dc5fe1aa1f9>

