

## Strategies Training Module

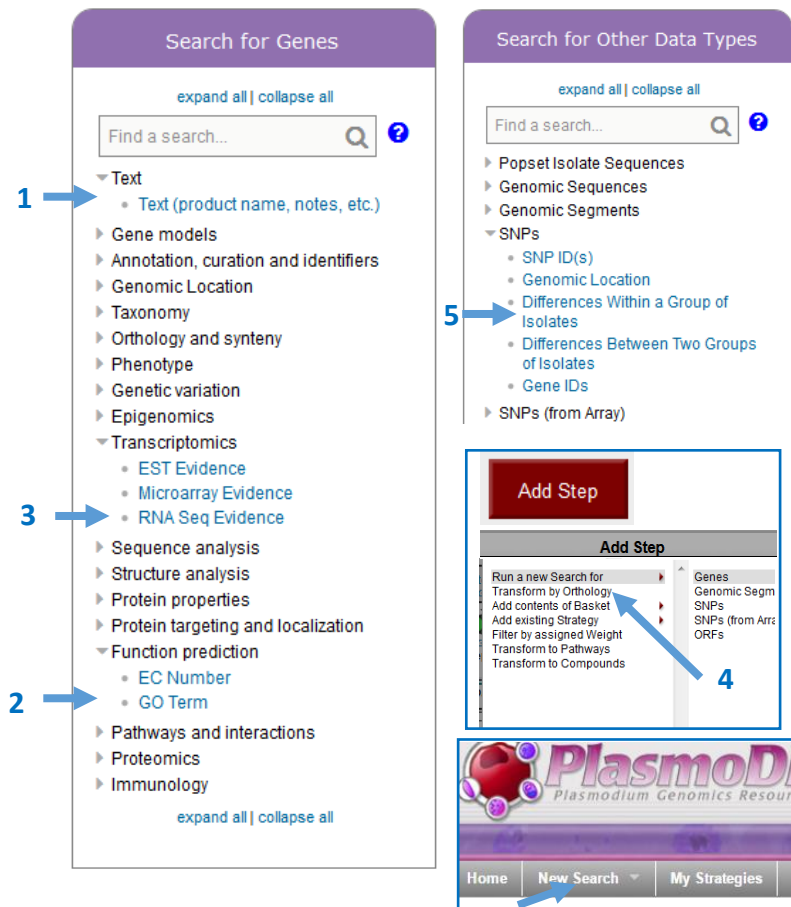
In this tutorial you will find genes expressed in gametocytes that are likely proteases and have variation in their upstream regions, possibly affecting promoter regions and other gene features. The strategy you build will combine three different searches that query *P. falciparum* data, then transform 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 home page and the New Search dropdown menu (screenshots left). Searches listed under 'Search for Genes' will return a list of genes, while those listed under 'Search for Other Data Types' will return other entities such as SNPs, ORFs, ESTs, isolates, compounds, etc.

**\*\*Click the links on the home page or navigate the New Search dropdown menu to visit the text search page.**








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 the data to produce expression values. This search returns genes based on their transcript expression as measure by RNA sequencing.

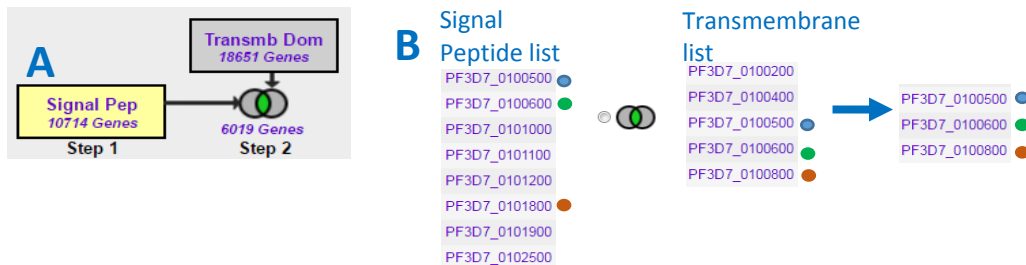
4. Transform by Orthology – PlasmoDB integrates ortholog profiles from OrthoMCL. The OrthoMCL algorithm clusters proteins into ortholog groups based on BLAST similarity across at least 150 genomes that span the tree of life. The transform we perform here will convert a list of genes in one organism to 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 SNPs 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).



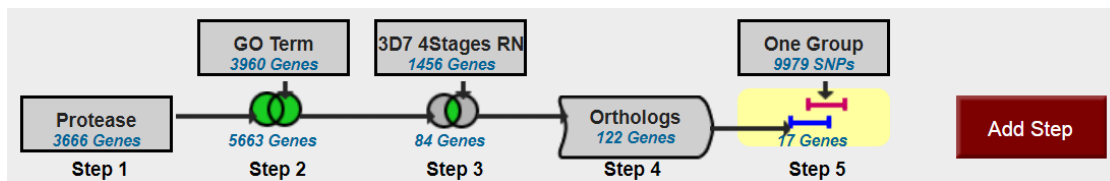
However, searches that return different genomic features will yield no results when combined with intersect, union or minus operators. This is illustrated in screenshot groupings C and D below. There are no IDs in the list of genes (Search 1, Step 1 result) that are present in the list of SNPs (Search 2, Step 1 result). To combine a search that returns genes with a search that returns SNPs, you must use the collocation option (1 relative to 2). Since we know the genomic location of each gene and each SNP the

colocation option will 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 uses the colocation 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 suspected proteases with evidence for expression during gametocyte stages. The *P. falciparum* genes returned in the step 3 result are transformed into their *P. vivax* orthologs. This results in a set of 84 *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 37 *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.): Find genes whose records contain the term 'protease' using the Text Search. To reach the text search, click on the link in the home page menu (screenshot group A below). 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 (see Parameter table and screenshot group A). 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 (see screenshot group B).

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

**A**

Search for Genes

expand all | collapse all

Find a search...

Text

- Text (product name, notes, etc.)
- Gene models
- Annotation, curation and identifiers
- Genomic Location

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

Organism

30 selected, out of 30

Plasmodium

Text term (use \* as wildcard)

protease

Fields

- Altus
- EC descriptions
- Gene ID
- Gene notes
- Gene product
- Gene name
- GO terms and definitions
- Metabolic pathway names and descriptions
- Protein domain names and descriptions
- PubMed
- Human Malaria Phenotype
- Similar proteins (BLAST hits v. NCBI/Protein)
- User comments

Advanced Parameters

Get Answer

Type protease in the text box but leave the other parameters as default. Click Get Answer to initiate the search.

**Parameters:**

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

**Results and strategy:** You created a one-step strategy by running the text search. The strategy returns 3666 genes that are annotated with the word Protease. You can analyze any 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.

**B**

My Strategies: [New](#) [Opened \(3\)](#) [All \(141\)](#) [Basket](#) [Public Strategies \(34\)](#) [Help](#)

Hide search strategy panel

(Genes)

**Protease**  
3666 Genes  
Step 1

Add Step

Strategy: Text \*  
Rename  
Duplicate  
Save As  
Share  
Delete

3666 Genes from Step 1 [Revise](#)

Strategy: Text

Click on a number in this table to limit/filter your results

All Results	Ortholog Groups	<i>P. adleri</i>	<i>P. berghii</i>	<i>P. bilcollinsi</i>	<i>P. blacklocki</i>	<i>P. chabaudi</i>	<i>P. coatneyi</i>	<i>P. cynomolgi</i> (nr Genes: 159)			
	G01	ANKA	G01	G01	chabaudi	Hackeri	strain B	strain M	307	798	
3666	133	99	92	84	83	87	61	75	84	110	85

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

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

Genes: 3666 Transcripts: 3676 (hiding 10) ☒ Show Only One Transcript Per Gene

First 1 2 3 4 5 Next Last [Advanced Paging](#) [Download](#) [Add to Basket](#) [Add Columns](#)

Gene ID	Transcript ID	Organism	Paralog count	Gene (Gen)	Score	Gene Name or Symbol
PKNH_1242000	PKNH_1242000.1	<i>P. knowlesi</i> strain H	0	PKNH_1242000	7	NA
PY17X_1308000	PY17X_1308000.1	<i>P. yoelii</i> yoelii 17X	1	PY17X_1308000	7	NA

Result List showing all hits from the search

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. 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 two steps to combine the two lists.

Navigation: Add Step >Run a new search for >>Genes >>>Function Prediction >>>>GO Term

**Step 1**

Protease  
3666 Genes

Add Step

**Add Step**

Run a new Search for  
 Transform by Ontology  
 Add contents of Basket  
 Add existing Strategy  
 Filter by assigned Weight  
 Transform to Pathways  
 Transform to Compounds

Genes  
 Genomic Segments  
 SNPs  
 SNPs (from Array)  
 ORFs

Text  
 Gene models  
 Annotation, curation and identifiers  
 Genomic Location  
 Taxonomy  
 Orthology and synteny  
 Phenotype  
 Genetic variation  
 Epigenomics  
 Transcriptomics  
 Sequence analysis  
 Structure analysis  
 Protein properties  
 Protein targeting and localization  
 Function prediction  
 Pathways and interactions  
 Proteomics  
 Immunology

EC Number  
 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

Choose union

**Add Step 2 : GO Term**

**Organism**  
 1 selected, out of 30  
 Filter list below...  
 Plasmidium  
 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

**Combine Genes in Step 1 with Genes in Step 2:**

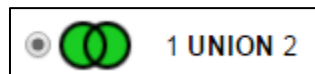
☐ 1 Intersect 2  
☒ 1 Union 2  
☐ 2 Minus 1  
☐ 1 Relative to 2, using genomic colocation

Run Step

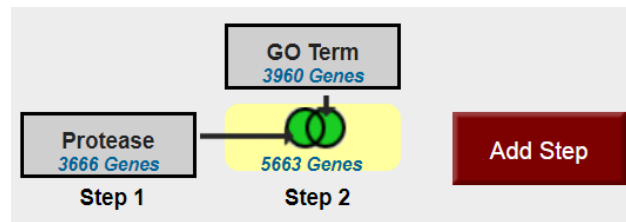
Parameters:

<b>Organism</b>	:	Default = all
<b>GO Term or GO ID</b>	:	GO:0006508 : proteolysis
<b>Free Text (use '*' for wildcard)</b>	:	N/A

Combine:



Strategy Result: The GO term search returned 3960 genes annotated with the proteolysis GO term. The union of the text and GO search returns 5663 genes 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 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 'expression percentile parameters will return the genes whose expression levels are in the top 20 % for those samples.

Navigation: >Add Step >Run a new search for >>Genes >>>Transcriptomics >>>>RNA Seq Evidence

**Add Step**

Run a new Search for

- Transform by Orthology
- Add contents of Basket
- Add existing Strategy
- Filter by assigned Weight
- Transform to Pathways
- Transform to Compounds

**Genes**

- Genomic Segments
- SNPs
- SNPs (from Array)
- ORFs

**Transcriptomics**

- Text
- Gene models
- Annotation, curation and identifiers
- Genomic Location
- Taxonomy
- Orthology and synteny
- Phenotype
- Genetic variation
- Epigenomics
- Sequence analysis
- Structure analysis

**EST Evidence**

- Microarray Evidence
- RNA Seq Evidence

**Add Step 3 : RNA Seq Evidence**

Filter Data Sets: Type keyword(s) to filter

Legend: DE Differ... FC Fold... P Perce...

Organism	Data Set	Choose a search
<i>P. berghei</i> ANKA	5 asexual and sexual stage transcriptomes (Hoeijmakers et al.)	FC P
<i>P. chabaudi</i> chabaudi	Trophozoite transcriptomes after mosquito transmission or direct injection into mice (Spence et al.)	DE FC P
<i>P. falciparum</i> 3D7	NSR-seq Transcript Profiling of malaria-infected pregnant women and children (Vignali et al.)	FC P
<i>P. falciparum</i> 3D7	Polysomal and steady-state asexual stage transcriptomes (Bunnik et al.)	FC P
<i>P. falciparum</i> 3D7	Blood stage transcriptome (3D7) (Otto et al.)	FC P
<i>P. falciparum</i> 3D7	Ribosome and steady state mRNA sequencing of asexual cell cycle stages (Caro et al.)	FC P
<i>P. falciparum</i> 3D7	Transcriptomes of 7 sexual and asexual life stages (Lopez-Barragan et al.)	FC P
<i>P. falciparum</i> 3D7	Intraerythrocytic cycle transcriptome (3D7) (Hoeijmakers et al.)	FC P
<i>P. falciparum</i> 3D7	Strand specific transcriptomes of 4 life cycle stages (Lopez-Barragan et al.)	FC P
<i>P. falciparum</i> 3D7	Transcriptome during intraerythrocytic development (Bartfai et al.)	FC P
<i>P. falciparum</i> 3D7	Mosquito or cultured sporozoites and blood stage transcriptome (NF54) (Hoffmann et al.)	FC P
<i>P. yoelii</i> yoelii 17X	Salivary gland sporozoite transcriptomes: WT vs. Puf2-KO (Lindner et al.)	FC P

Add Step

Add Step 3 : P.falciparum Strand specific transcriptomes of 4 life cycle stages RNASeq (percentile)

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

Combine Genes in Step 2 with Genes in Step 3:

☒ 2 Intersect 3

☐ 2 Union 3

☐ 2 Minus 3

☐ 3 Minus 2

☐ 2 Relative to 3, using genomic colocation

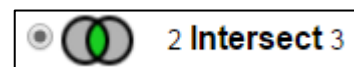
Run Step

Choose  
2 intersect 3

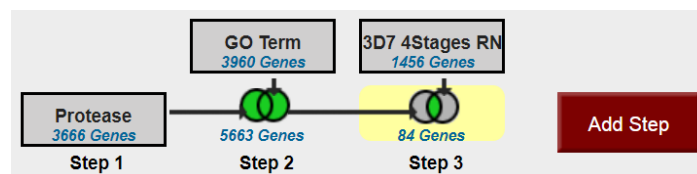
Parameters:

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 84 *P. falciparum* genes that are suspected proteases with evidence for expression in gametocytes based on RNA Sequencing data. Explore your gene list!!



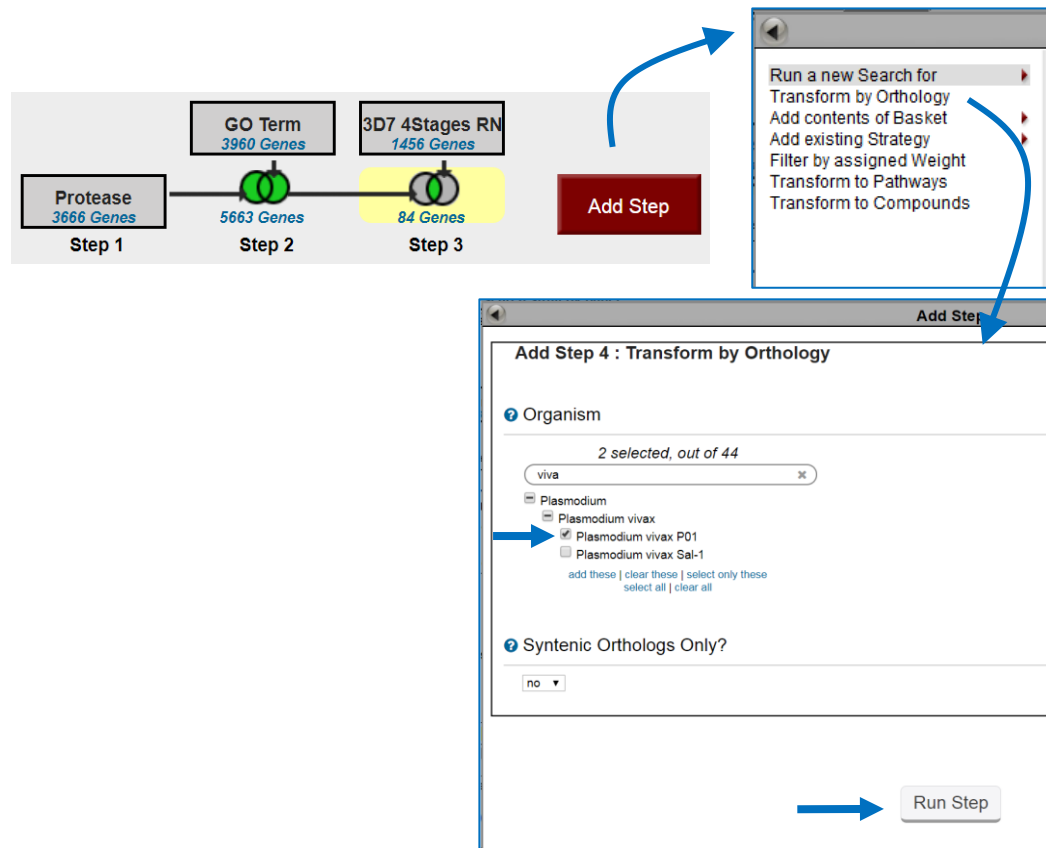


4. Add a step to the strategy that transforms the 84 *P. falciparum* genes into *P. vivax* genes.

*P. falciparum* is a well-studied organism with active curatorial efforts and large amounts of functional data. For example, PlasmoDB has 10 RNA sequencing and 9 microarray data sets integrated for *P. falciparum*, but only 2 each for *P. vivax*. A *P. vivax* researcher 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.

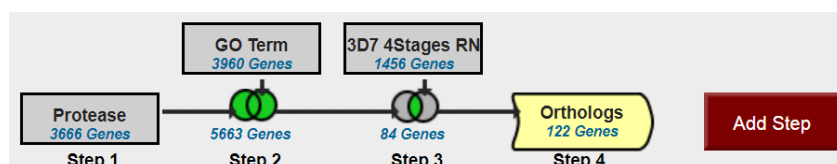
Navigation: >Add Step >Transform by Orthology

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 122 *P. vivax* genes that are suspected proteases with evidence for expression in gametocytes based on RNA Sequencing data. Explore the result table.



5. **Add a step to the strategy that returns SNP and collocate those SNPs to the upstream 1000bp of the *P. vivax* genes.** We can look for variation (SNPs) associated with the genes from Step 4. PlasmoDB integrates whole genome resequencing data from many isolates, and there are 178 datasets from whole-genome sequencing of *P. vivax* isolates in PlasmoDB. We analyze 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 >Run a new search for >>SNPs >>> Differences Within a Group of Isolates

**Workflow Steps:**

- Step 1: Protease (366 Genes)
- Step 2: GO Term (3960 Genes)
- Step 3: 3D7 4Stages RN (1456 Genes)
- Step 4: Orthologs (122 Genes)

**Add Step Dialog:**

- Run a new Search for
  - Transform by Orthology
  - Add contents of Basket
  - Add existing Strategy
  - Filter by assigned Weight
  - Transform to Pathways
  - Transform to Compounds
- Genes
- Genomic Segments
- SNPs
- SNPs (from Array)
- ORFs
- SNP ID(s)
- Genomic Location
- Differences Within a Group of Isolates
- Differences Between Two Groups of Isolates
- Gene IDs

**Add Step 5: Differences Within a Group of Isolates**

**Organism:** Plasmodium vivax P01

**Samples:** 178 Samples Total

**data set:** A data item that is an aggregate of other data items of the same type that have something in common. Averages are calculated.

**Check items below to apply this filter:**

data set	Remaining Samples	Samples	Distribution
Aligned genomic sequence reads - Field and monkey adapted isolates	178 (100%)	9 (5%)	9 (5%)
Aligned genomic sequence reads - Field isolates	178 (100%)	4 (2%)	4 (2%)
Aligned genomic sequence reads - Hybrid Selection Project	178 (100%)	162 (91%)	162 (91%)
Aligned genomic sequence reads - strain IQ07	178 (100%)	1 (1%)	1 (1%)
Aligned genomic sequence reads - strain P01	178 (100%)	1 (1%)	1 (1%)
Plasmodium vivax P01 Genome Sequence and Annotation	178 (100%)	1 (1%)	1 (1%)

**Read frequency threshold:** 80%

**Minor allele frequency >=:** 0

**Percent isolates with a base call >=:** 70

**Combine Genes in Step 4 with SNPs in Step 5:**

- 4 Intersect
- 4 Union
- 4 Minus
- 4 Relative to 5, using genomic collocation

**Set the Percent isolates with a base call >= 70**

**Continue...**

**Clicking continue will initiate the search and bring up the collocation tool.**

Parameters:

Organism	:	Pvivax P01
Isolates	:	Default = All Isolates (178)
Read frequency threshold	:	Default - 80%
Minor allele frequency >=	:	Default - 0
Percent isolates with a base call >=	:	Default - 70

Colocation: Our SNP search returned 9979 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 each Gene from step 4 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 Step**

**Genomic Colocation**

Combine Step 4 and Step 5 using relative locations in the genome

You had 122 Genes in your Strategy (Step 4). Your new SNPs search (Step 5) returned 9979 SNPs.

"Return each **Gene from Step 4** whose **upstream region** overlaps the **exact region** of a SNP in Step 5 and is on **either strand**"

(122 Genes in Step)

Region

Gene

Exact

Upstream: 1000 bp

Downstream: 1000 bp

Custom:

begin at: start - 1000 bp

end at: start 1 bp

(9979 SNPs in Step)

Region

SNP

Exact

Upstream: 1000 bp

Downstream: 1000 bp

Custom:

begin at: start + 0 bp

end at: stop + 0 bp

Submit

Close

**Strategy: Congratulations!** You have completed the strategy and have a list of 17 *P. vivax* genes that are possible proteases, are likely expressed in gametocytes and have upstream SNPs.

<http://plasmodb.org/plasmo/im.do?s=922b4adf5a70d44c>

