Strategies Training Module

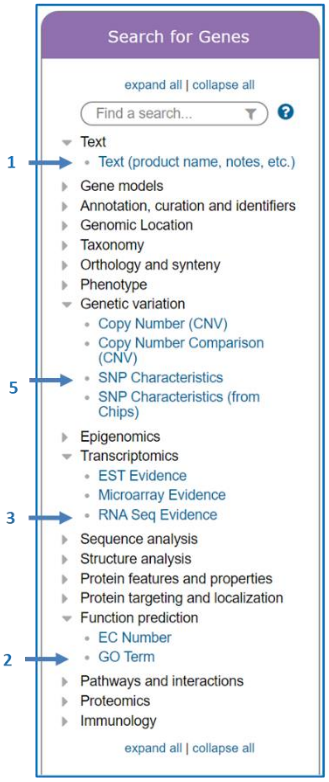
For this tutorial let’s imagine that we are *Plasmodium vivax* researchers interested in genes involved in the egress of gametocytes from the infected red blood cells and that might be under selective pressure. Turning to PlasmoDB, we will use the strategy system to find *P. vivax* genes that are likely proteases expressed in gametocytes and that contain non-synonymous SNPs. The strategy you build will combine three different searches that query *P. falciparum* data, then transform the *P. falciparum* genes into their *P. vivax* orthologs and determine which of the *P.vivax* genes have non-synonymous SNPs. 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 final strategy share two biological properties, proteolytic activity and expression in gametocytes, suggesting they may be involved in egress. They also are likely to contain non-synonymous SNPs, an indication of selective pressure.

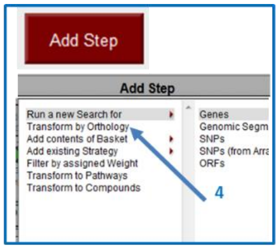
Strategies Overview:

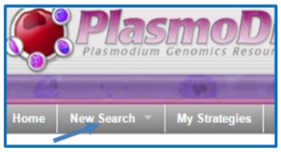
The strategy system offers over 100 structured searches that can be combined to produce multi-step strategies. Each search queries a specific data set and **returns a list of IDs** that share the biological characteristic defined by the data.

Searches are accessible from the home page and the New Search dropdown menu (screenshots on right). 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.

**Searches are available from the center panels on the home page or from the dropdown menu called ‘New Search’.







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. The ontologies are a controlled vocabulary of terms for describing the molecular function, biological process and subcellular location of a gene product. The Gene Ontology standardizes the representation of gene and gene product attributes across species and databases. This search returns genes with the GO Term or ID that you specify.
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.

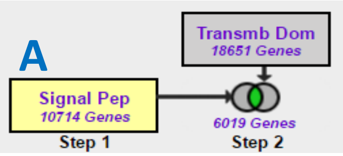
4. Transform by Orthology – PlasmoDB integrates ortholog profiles from OrthoMCL. The OrthoMCL algorithm clusters proteins into ortholog groups based on BLAST similarity across at least 150 genomes that span the tree of life. The transform we perform here will convert a list of genes in one organism to 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 Genes based on SNP Characteristics – PlasmoDB integrates whole genome resequencing of isolates and analyzes each isolate for SNPs compared to a reference genome. The SNPs are then analyzed for their effect on the gene product. This search returns genes based on their effect on the gene product (synonymous, non-synonymous, etc).

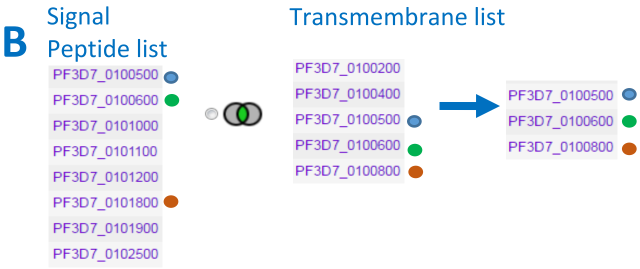
Before we get started… a few words about combining search results:

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).





Signal Peptide list

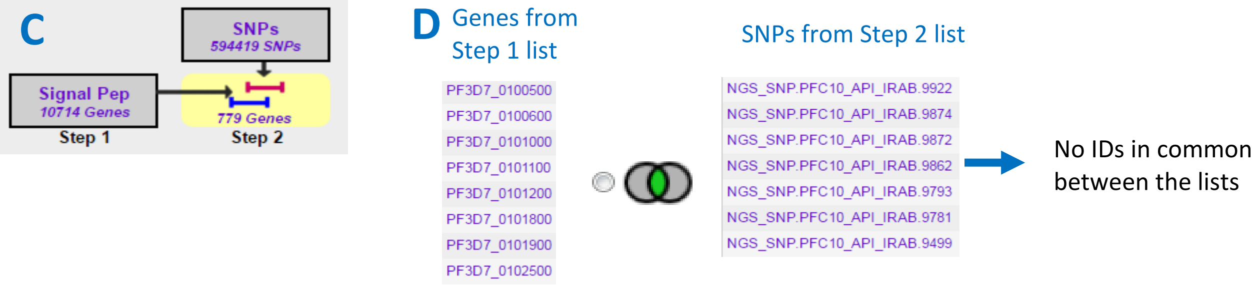
PF3D7_0100500
PF3D7_0100600
PF3D7_0101000
PF3D7_0101100
PF3D7_0101200
PF3D7_0101800
PF3D7_0101900
PF3D7_0102500

Transmembrane list

PF3D7_0100200
PF3D7_0100400
PF3D7_0100500
PF3D7_0100600
PF3D7_0100800

PF3D7_0100500
PF3D7_0100600
PF3D7_0100800

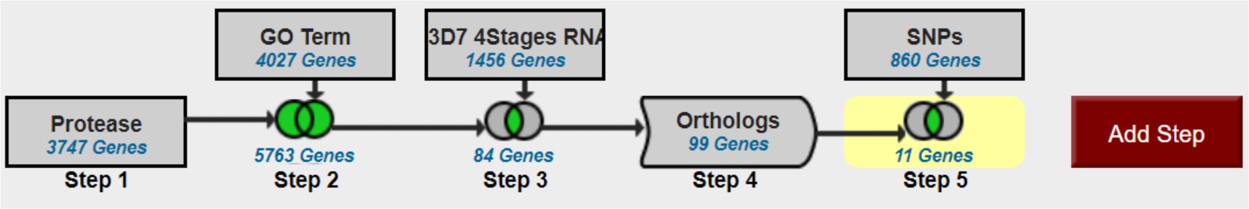
However, searches that return different genomic features will yield no results when combined with intersect, union or minus operators. This is illustrated in screenshot groupings C and D below. Because genes and SNPs are different genomic features, there are no IDs in the list of genes (Search 1, Step 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.


C
SNPs
594419 SNPs
Signal Pep
10714 Genes
779 Genes
Step 1
Step 2
D
Genes from Step 1 list
PF3D7_0100500
PF3D7_0100600
PF3D7_0101000
PF3D7_0101100
PF3D7_0101200
PF3D7_0101800
PF3D7_0101900
PF3D7_0102500
SNPs from Step 2 list
NGS_SNP.PFC10_API_IRAB.9922
NGS_SNP.PFC10_API_IRAB.9874
NGS_SNP.PFC10_API_IRAB.9872
NGS_SNP.PFC10_API_IRAB.9862
NGS_SNP.PFC10_API_IRAB.9793
NGS_SNP.PFC10_API_IRAB.9781
NGS_SNP.PFC10_API_IRAB.9499
No IDs in common between the lists

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 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 99 *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 11 *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:

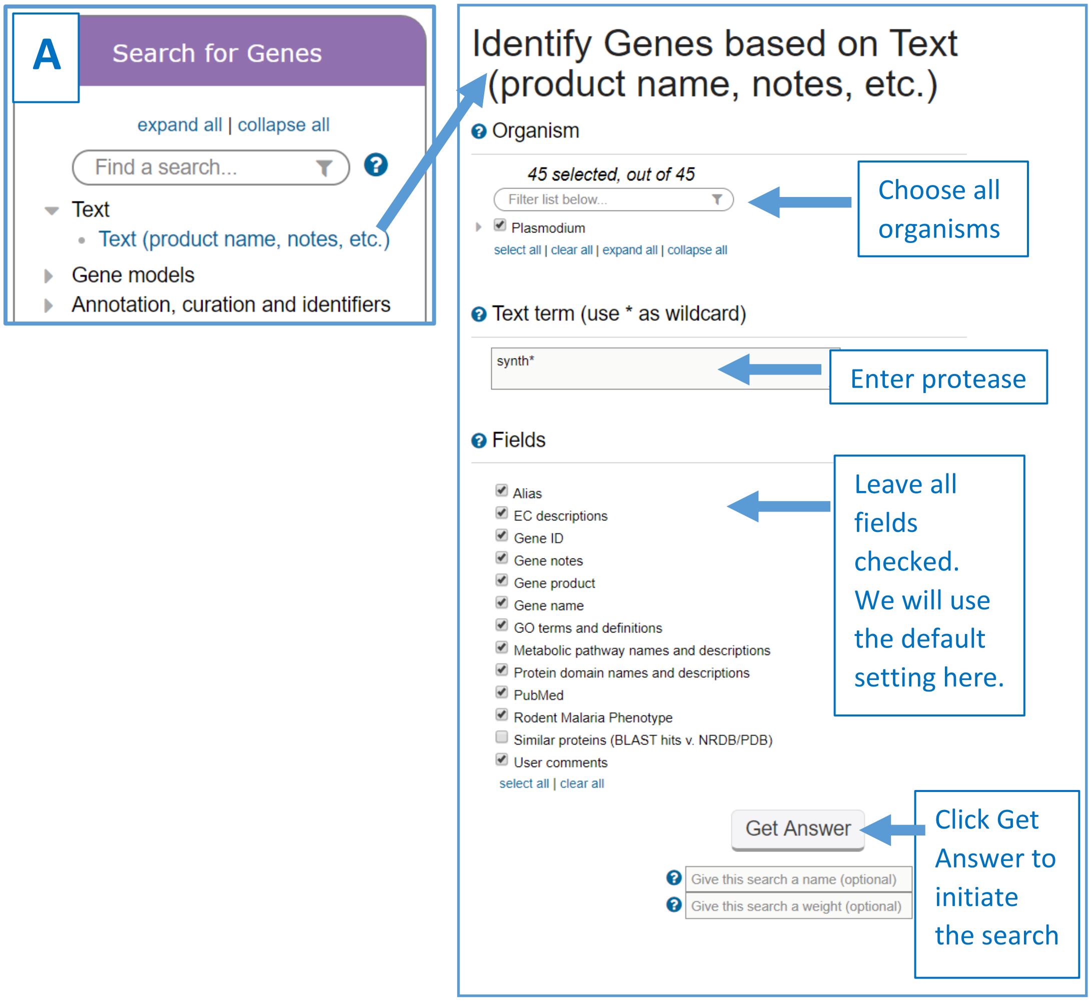
https://plasmodb.org/plasmo/im.do?s=e51e31931fb515b4



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 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

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

Organism

45 selected, out of 45

Filter list below...

☑ Plasmodium

select all | clear all | expand all | collapse all

Choose all organisms

Text term (use * as wildcard)

synth*

Enter protease

Fields

☑ Alias
☑ 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
☑ Rodent Malaria Phenotype
☐ Similar proteins (BLAST hits v. NRDB/PDB)
☑ User comments

select all | clear all

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

Get Answer

Give this search a name (optional)

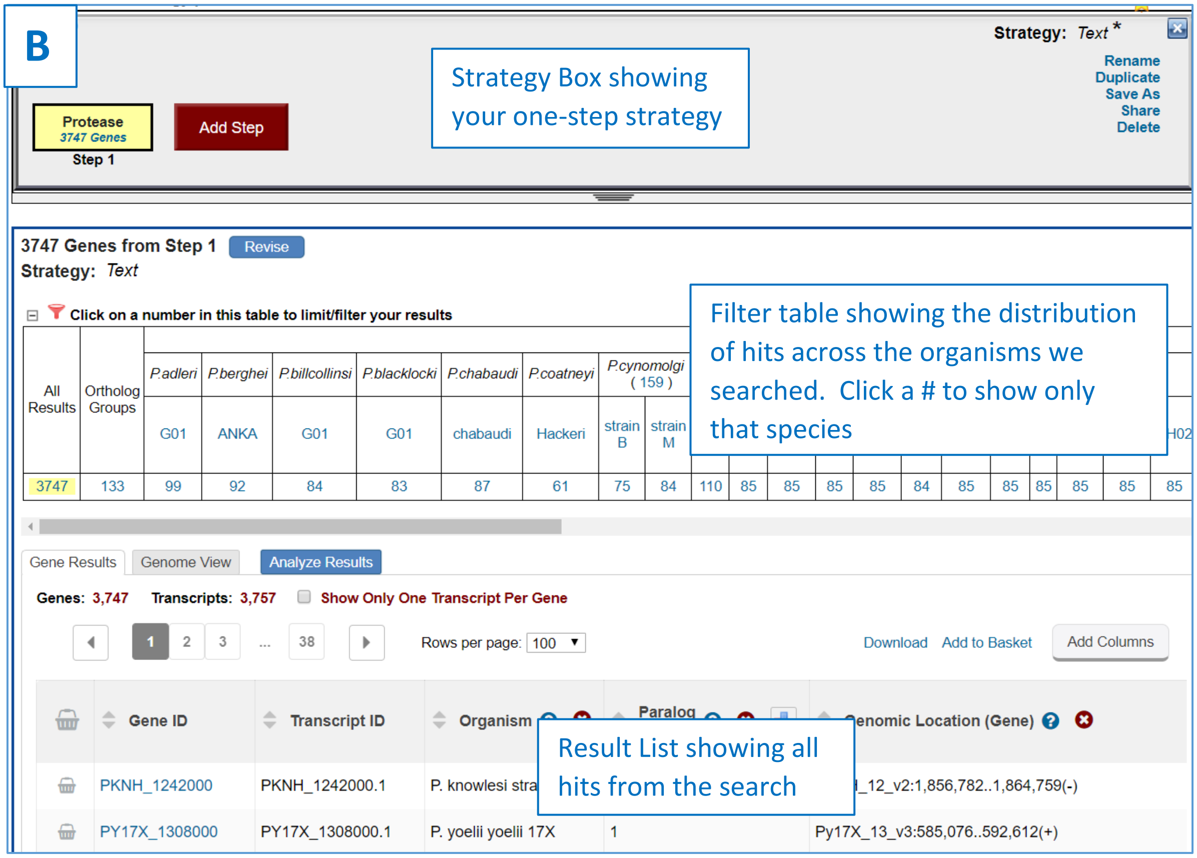
Give this search a weight (optional)

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 3747 genes that are annotated with the word 'protease'. 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

Strategy: *Text* *

3747 Genes from Step 1

Strategy: *Text*

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

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

All Results	Ortholog Groups	P.adleri	P.berghei	P.billcollinsi	P.blacklocki	P.chabaudi	P.coatneyi	P.cynomolgi (159)													
		G01	ANKA	G01	G01	chabaudi	Hackeri	strain B	strain M												H02
3747	133	99	92	84	83	87	61	75	84	110	85	85	85	85	84	85	85	85	85	85	85

Genes: 3,747 Transcripts: 3,757 □ Show Only One Transcript Per Gene

◀ 1 2 3 ... 38 ▶ Rows per page: 100 ▼

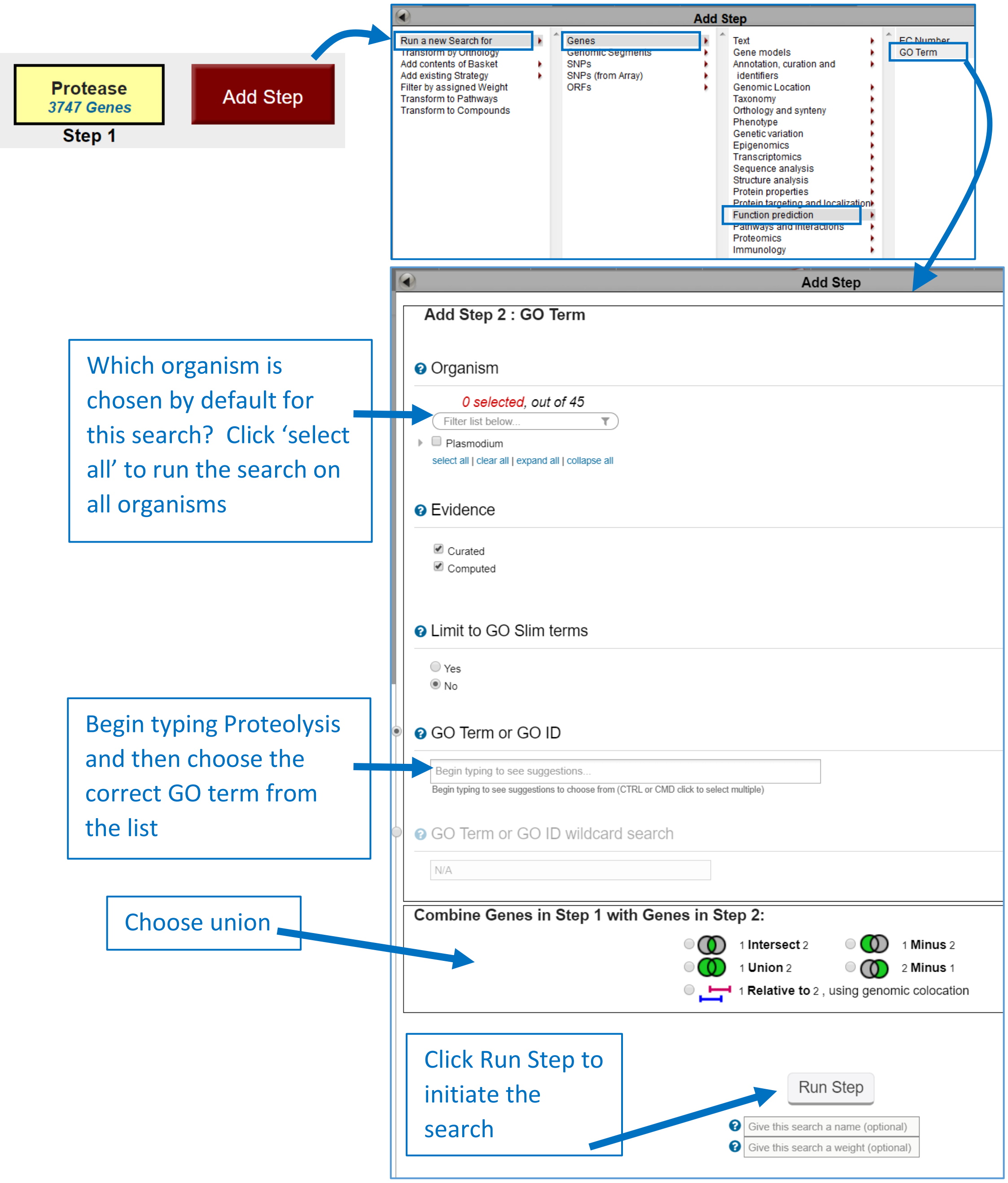
Result List showing all hits from the search

	Gene ID	Transcript ID	Organism	Paralog	enomic Location (Gene)
	PKNH_1242000	PKNH_1242000.1	P. knowlesi stra		_12_v2:1,856,782..1,864,759(-)
	PY17X_1308000	PY17X_1308000.1	P. yoelii yoelii 17X	1	Py17X_13_v3:585,076..592,612(+)

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 text search results with our GO term results when we combine the results of the two searches.

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



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

Combine Genes in Step 1 with Genes in Step 2:

○ 1 **Intersect** 2 ○ 1 **Minus** 2

○ 1 **Union** 2 ○ 2 **Minus** 1

○ 1 **Relative to** 2 , using genomic colocation

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 organisms

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

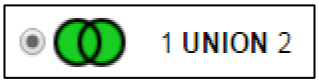
Choose union

Click Run Step to initiate the search

Parameters:

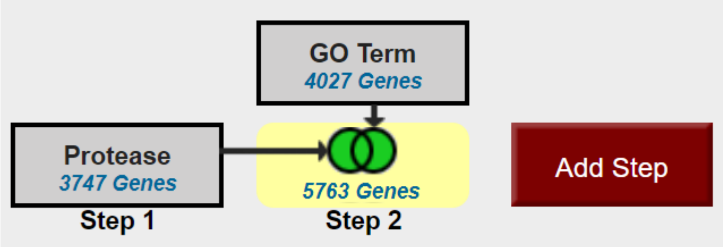
Organism	:	Default = all
GO Term or GO ID	:	GO:0006508 : proteolysis
Free Text (use '*' for wildcard)	:	N/A

Combine:



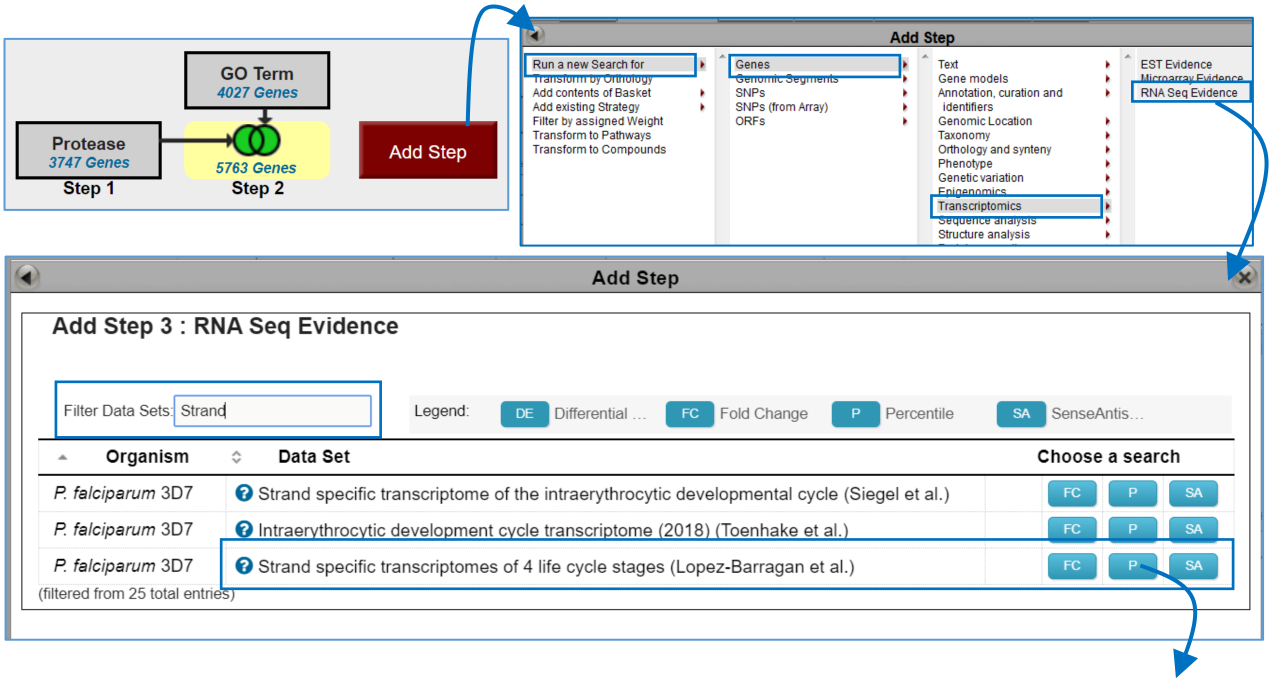
● 1 UNION 2

Strategy Result: The GO term search returned 4027 genes annotated with the proteolysis GO term. The union of the text and GO search returns 5763 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. 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 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



Protease
3747 Genes
Step 1

GO Term
4027 Genes

5763 Genes
Step 2

Add Step

Add Step

Run a new Search for	Genes	Text	EST Evidence
Transform by Orthology	Genomic Segments	Gene models	Microarray Evidence
Add contents of Basket	SNPs	Annotation, curation and identifiers	RNA Seq Evidence
Add existing Strategy	SNPs (from Array)	Genomic Location	
Filter by assigned Weight	ORFs	Taxonomy	
Transform to Pathways		Orthology and synteny	
Transform to Compounds		Phenotype	
		Genetic variation	
		Epigenomics	
		Transcriptomics	
		Sequence analysis	
		Structure analysis	

Add Step

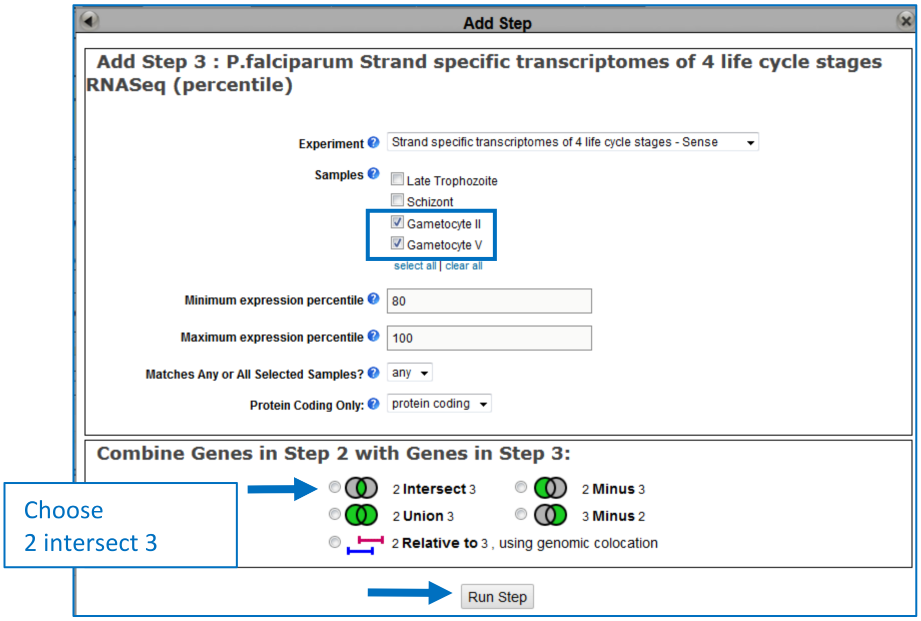
Add Step 3 : RNA Seq Evidence

Filter Data Sets: Strand

Legend: DE Differential … FC Fold Change P Percentile SA SenseAntis…

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

(filtered from 25 total entries)



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 **Minus** 3
○ 2 **Union** 3 ○ 3 **Minus** 2
○ 2 **Relative to** 3 , using genomic colocation

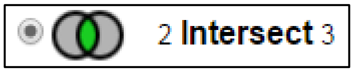
Run Step

Choose
2 intersect 3

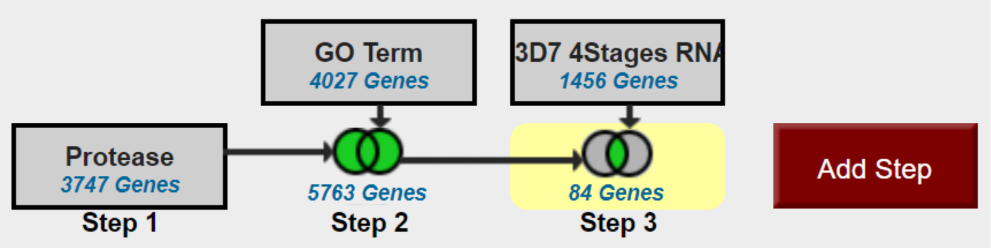
Parameters:

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.

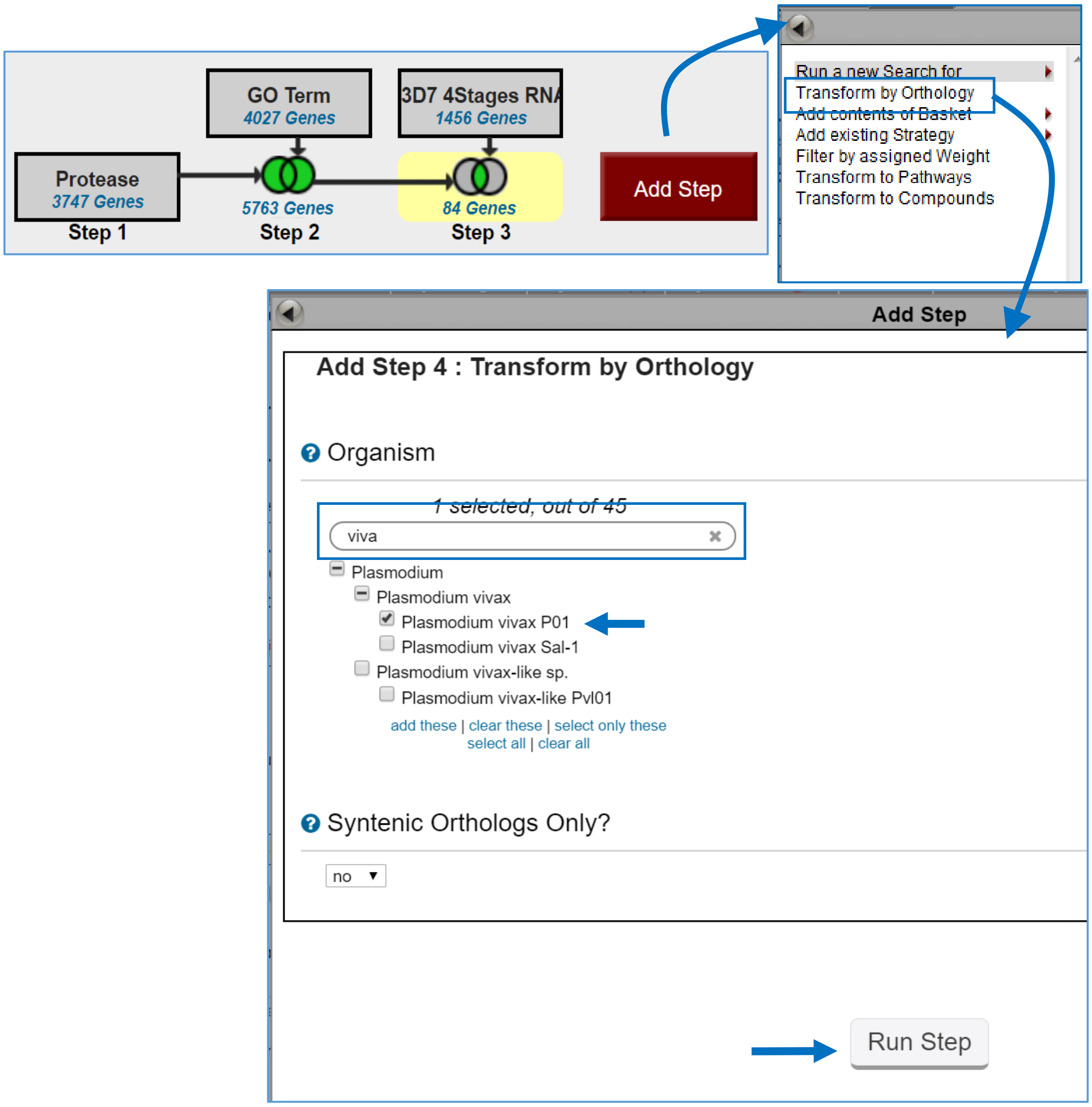


Strategy result: We have a three-step strategy that returns 84 *P. falciparum* genes that are suspected proteases with evidence for expression in gametocytes based on RNA Sequencing data. Explore your gene list!!



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 13 RNA sequencing and 10 microarray data sets integrated for *P. falciparum*, but only 4 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 by Orthology



Add Step 4 : Transform by Orthology

Organism

1 selected, out of 45

viva

- ⊟ Plasmodium
  - ⊟ Plasmodium vivax
    - ☑ Plasmodium vivax P01
    - □ Plasmodium vivax Sal-1
  - □ Plasmodium vivax-like sp.
    - □ Plasmodium vivax-like PvI01

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

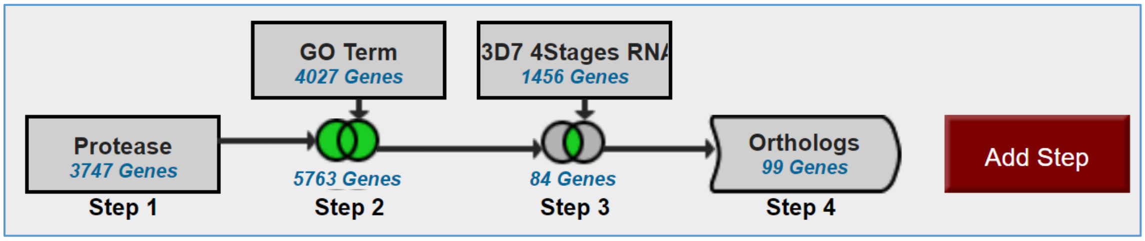
Syntenic Orthologs Only?

no ▼

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

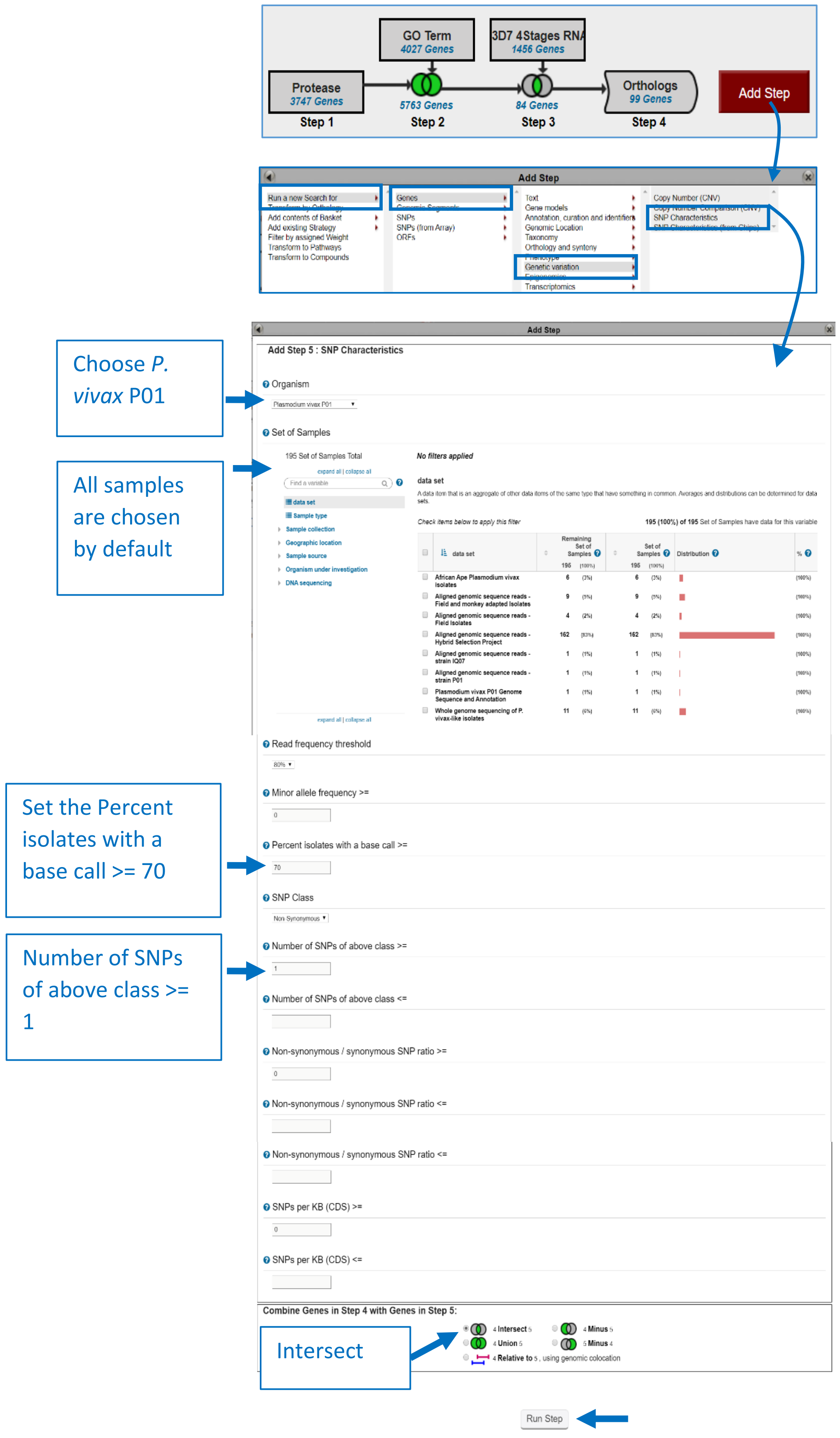


5. **Add a step to the strategy that returns *P. vivax* genes that contain non-synonymous SNPs.**
   PlasmoDB integrates whole genome resequencing data from many isolates, and there are 195 datasets from whole-genome sequencing of *P. vivax* isolates in PlasmoDB. We analyze the whole genome sequencing reads by aligning them to the reference genome (*P. vivax* P01)and then walking down the genome one base at a time looking for bases in the isolate that do not match the reference sequence. Each SNP is loaded in the database along with other characteristics such as how many sequencing reads supported the SNP call, it's genomic location of the SNP, its effect on the gene it overlaps with (in any). The search we will use returns genes based on the characteristics of the SNPs that they contain.

   Navigation: >Add Step >Run a new search for >>Genes >>>Genetic Variation >>>>SNP characteristics

   Parameters:

Organism	:	*P. vivax* P01
Set of Samples	:	Default = All Isolates (195)
Read frequency threshold	:	Default - 80%
Minor allele frequency >=	:	Default - 0
Percent isolates with a base call >=	:	Default - 70
SNP Class	:	Non-Synonymous
Number of SNPs of above class >=	:	1



Add Step 5 : SNP Characteristics

Choose *P. vivax* P01

Organism

Plasmodium vivax P01

Set of Samples

All samples are chosen by default

195 Set of Samples Total

expand all | collapse all

Find a variable

- data set
- Sample type
- Sample collection
- Geographic location
- Sample source
- Organism under investigation
- DNA sequencing

expand all | collapse all

No filters applied

data set

A data item that is an aggregate of other data items of the same type that have something in common. Averages and distributions can be determined for data sets.

Check items below to apply this filter

195 (100%) of 195 Set of Samples have data for this variable

□	data set	Remaining Set of Samples		Set of Samples		Distribution	%
		195	(100%)	195	(100%)		
□	African Ape Plasmodium vivax isolates	6	(3%)	6	(3%)		(100%)
□	Aligned genomic sequence reads - Field and monkey adapted Isolates	9	(5%)	9	(5%)		(100%)
□	Aligned genomic sequence reads - Field Isolates	4	(2%)	4	(2%)		(100%)
□	Aligned genomic sequence reads - Hybrid Selection Project	162	(83%)	162	(83%)		(100%)
□	Aligned genomic sequence reads - strain IQ07	1	(1%)	1	(1%)		(100%)
□	Aligned genomic sequence reads - strain P01	1	(1%)	1	(1%)		(100%)
□	Plasmodium vivax P01 Genome Sequence and Annotation	1	(1%)	1	(1%)		(100%)
□	Whole genome sequencing of P. vivax-like isolates	11	(6%)	11	(6%)		(100%)

Read frequency threshold

80%

Minor allele frequency >=

0

Set the Percent isolates with a base call >= 70

Percent isolates with a base call >=

70

SNP Class

Non Synonymous

Number of SNPs of above class >= 1

Number of SNPs of above class >=

1

Number of SNPs of above class <=

Non-synonymous / synonymous SNP ratio >=

0

Non-synonymous / synonymous SNP ratio <=

Non-synonymous / synonymous SNP ratio <=

SNPs per KB (CDS) >=

0

SNPs per KB (CDS) <=

Combine Genes in Step 4 with Genes in Step 5:

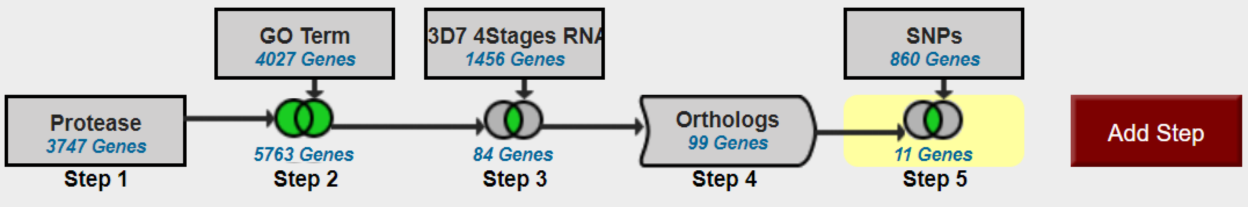
Intersect

● 4 **Intersect** 5
○ 4 **Union** 5
○ 4 **Relative to** 5 , using genomic colocation
○ 4 **Minus** 5
○ 5 **Minus** 4

Run Step

Strategy: Congratulations! You have completed the strategy and have a list of 11 *P. vivax* genes that are possible proteases, are likely expressed in gametocytes and contain non-synonymous SNPs.

This link will retrieve the completed strategy:
https://plasmodb.org/plasmo/im.do?s=e51e31931fb515b4


Protease
3747 Genes
Step 1
GO Term
4027 Genes
5763 Genes
Step 2
3D7 4Stages RNA
1456 Genes
84 Genes
Step 3
Orthologs
99 Genes
Step 4
SNPs
860 Genes
11 Genes
Step 5
Add Step