Strategies Tutorial

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

Learning objectives:

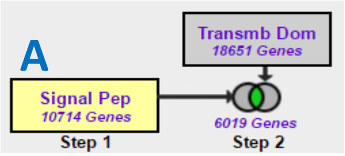
- Build a multistep strategy
- Use the Text, GO Term, RNA-Seq, and SNP searches
- Transform genes of one organism into their orthologs in another organism
- Combine search results using Boolean operators
- Co-locate two different record types – genes and SNPs
- Infer expression timing from a well-studied organism onto an organism with less data.
- Use the nested strategy function to add data to the strategy and increase the stringency of evidence used to find genes.

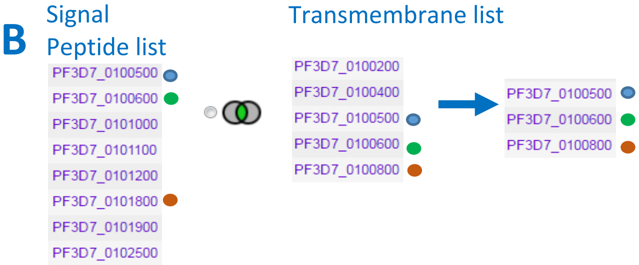
In this tutorial you will find *P. vivax* genes that are likely expressed in gametocytes, act as proteases and have variation in their upstream regions, possibly affecting promoter regions and other gene features. The strategy you build will take advantage of the data rich organism of *P. falciparum*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

Before we get started… a few words about combining search results:
Each search returns a list of IDs. When two searches are combined, the two result sets (list of IDs) are merged. The table shows the 5 options for combining search results.

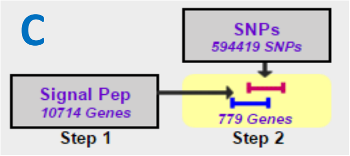
Operator	:	Combined Result will contain:
○ 1 INTERSECT 2	:	IDs in common between the two lists
● 1 UNION 2	:	IDs from list 1 and list 2
○ 1 MINUS 2	:	IDs unique to 1
○ 2 MINUS 1	:	IDs unique to 2
○ 1 Relative to 2	:	IDs whose features are near each other (co-located) in the genome

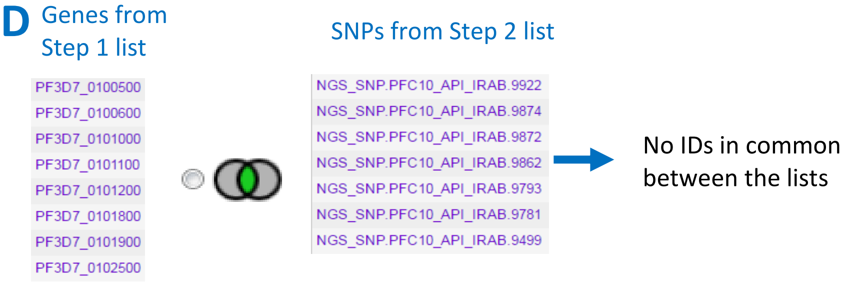
If the searches return the same type of genomic feature they can be combined using any of the 5 operators (i.e. search 1 returns genes, search 2 returns genes as in screenshot group A below).





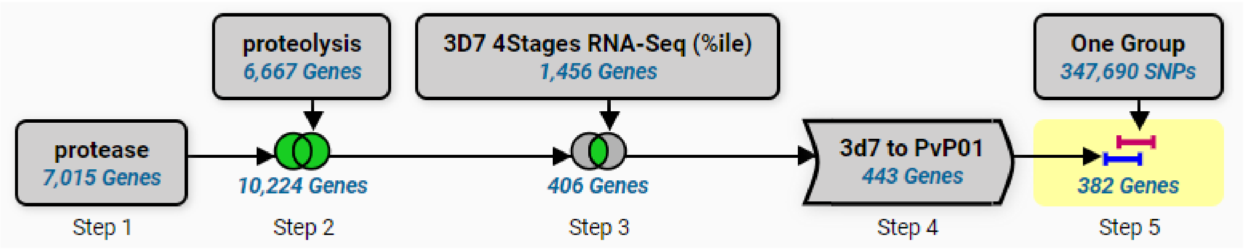
However, searches that return different genomic features will yield no results when combined with intersect, union or minus operators. The Genomic Co-Location tool takes advantage of the genomic location of each gene and each SNP and returns features based on their relative genomic location, i.e. SNPs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




D
Genes from Step 1 list
SNPs from Step 2 list
PF3D7_0100500
PF3D7_0100600
PF3D7_0101000
PF3D7_0101100
PF3D7_0101200
PF3D7_0101800
PF3D7_0101900
PF3D7_0102500
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

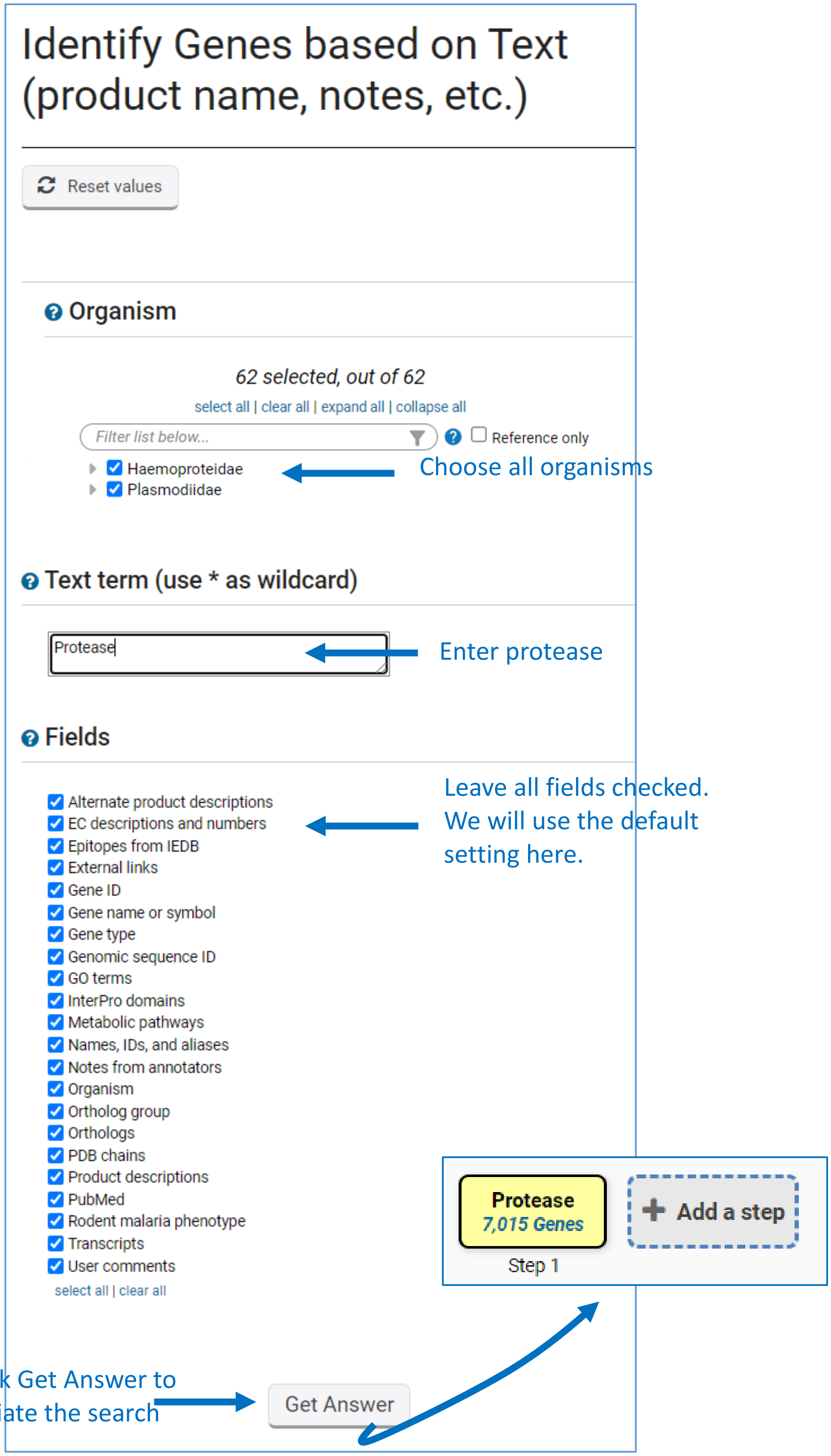
Building the Strategy:

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



Step by Step Instructions

1. **Run a text search using protease as the text term.**
   Navigation: >PlasmoDB >Search for Genes >Text > Text (product name, notes, etc.)



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

Organism

62 selected, out of 62

☑ 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

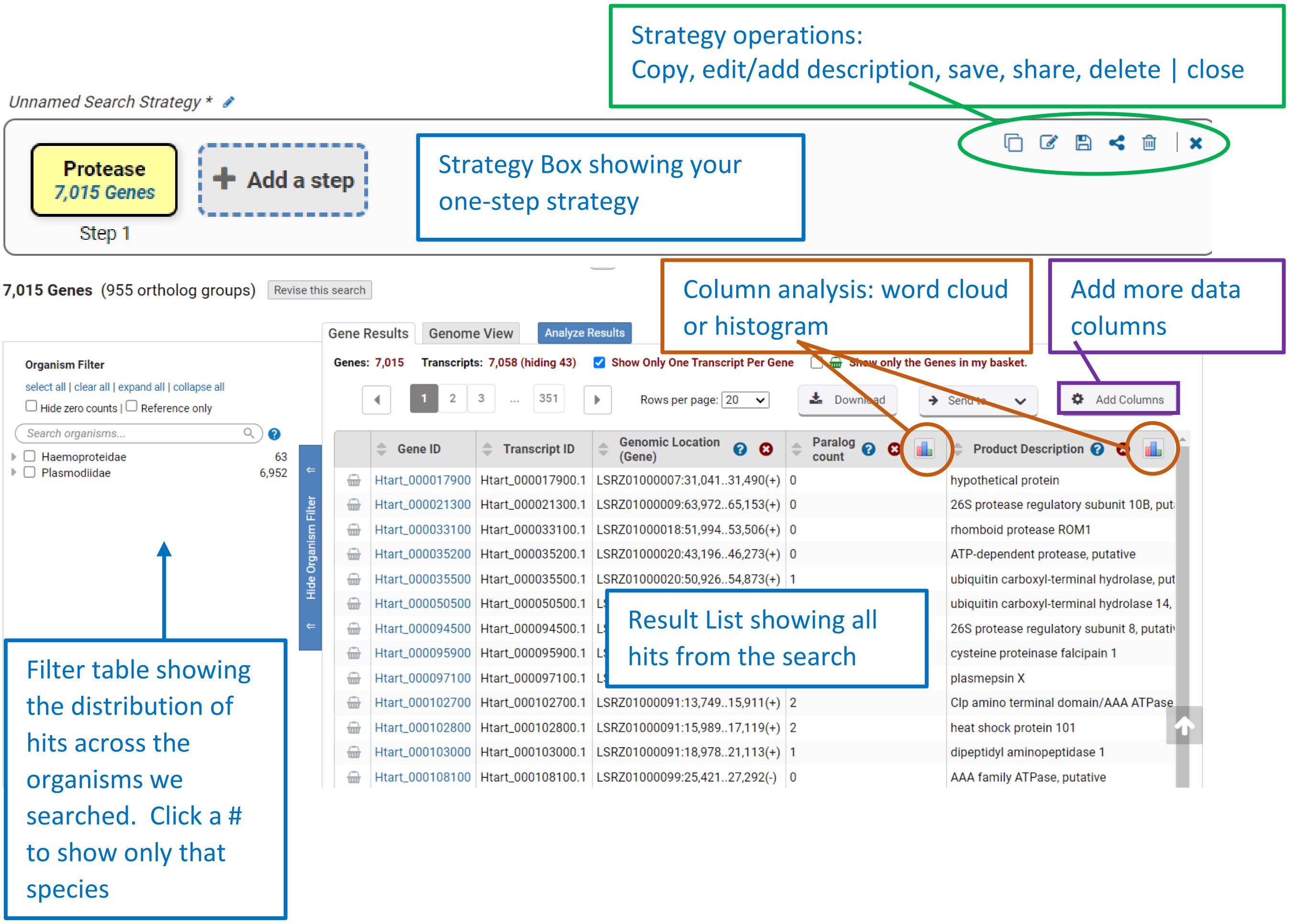
Leave all fields checked. We will use the default setting here.

Protease
7,015 Genes
Step 1

Click Get Answer to initiate the search

You created a one-step strategy by running the text search. The strategy returns 7015 genes that are annotated with the word ‘protease’. This annotation could appear in any field that you searched. Please explore your results to see if they make sense:

- Look at the data in the columns of the result table. For example, gene product names might contain the word ‘protease’.
- Add more data columns to investigate other data types
- Run a column analysis.



Strategy operations:
Copy, edit/add description, save, share, delete | close

Strategy Box showing your one-step strategy

Column analysis: word cloud or histogram

Add more data columns

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

Result List showing all hits from the search

7,015 Genes (955 ortholog groups)

Genes: 7,015 Transcripts: 7,058 (hiding 43) ☑ Show Only One Transcript Per Gene ☐ Show only the Genes in my basket.

◀ 1 2 3 ... 351 ▶ Rows per page: 20

☐ Hide zero counts | ☐ Reference only

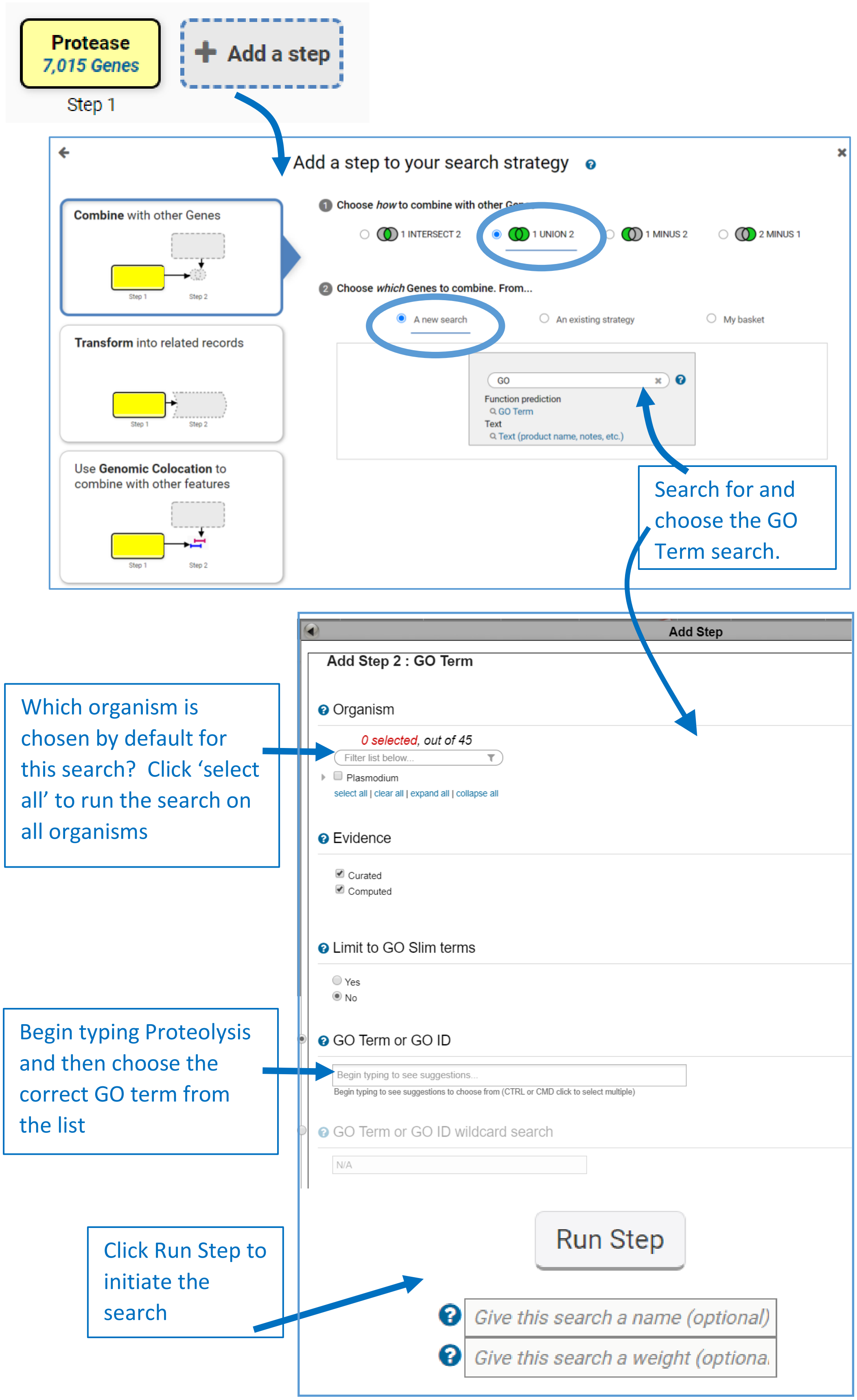
Search organisms...

☐ Haemoproteidae	63
☐ Plasmodiidae	6,952

Gene ID	Transcript ID	Genomic Location (Gene)	Paralog count	Product Description
Htart_000017900	Htart_000017900.1	LSRZ01000007:31,041..31,490(+)	0	hypothetical protein
Htart_000021300	Htart_000021300.1	LSRZ01000009:63,972..65,153(+)	0	26S protease regulatory subunit 10B, put
Htart_000033100	Htart_000033100.1	LSRZ01000018:51,994..53,506(+)	0	rhomboid protease ROM1
Htart_000035200	Htart_000035200.1	LSRZ01000020:43,196..46,273(+)	0	ATP-dependent protease, putative
Htart_000035500	Htart_000035500.1	LSRZ01000020:50,926..54,873(+)	1	ubiquitin carboxyl-terminal hydrolase, put
Htart_000050500	Htart_000050500.1	LS		ubiquitin carboxyl-terminal hydrolase 14,
Htart_000094500	Htart_000094500.1	LS		26S protease regulatory subunit 8, putati
Htart_000095900	Htart_000095900.1	LS		cysteine proteinase falcipain 1
Htart_000097100	Htart_000097100.1	LS		plasmepsin X
Htart_000102700	Htart_000102700.1	LSRZ01000091:13,749..15,911(+)	2	Clp amino terminal domain/AAA ATPase
Htart_000102800	Htart_000102800.1	LSRZ01000091:15,989..17,119(+)	2	heat shock protein 101
Htart_000103000	Htart_000103000.1	LSRZ01000091:18,978..21,113(+)	1	dipeptidyl aminopeptidase 1
Htart_000108100	Htart_000108100.1	LSRZ01000099:25,421..27,292(-)	0	AAA family ATPase, putative

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

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



Protease
7,015 Genes
Step 1

➕ Add a step

Add a step to your search strategy

Combine with other Genes
Step 1
Step 2
Transform into related records
Step 1
Step 2
Use Genomic Colocation to combine with other features
Step 1
Step 2

1 Choose *how* to combine with other Gen…

○ 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.

Add Step

Add Step 2 : GO Term

Organism

0 selected, out of 45

Filter list below...

□ Plasmodium

select all | clear all | expand all | collapse all

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

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)

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

○ GO Term or GO ID wildcard search

N/A

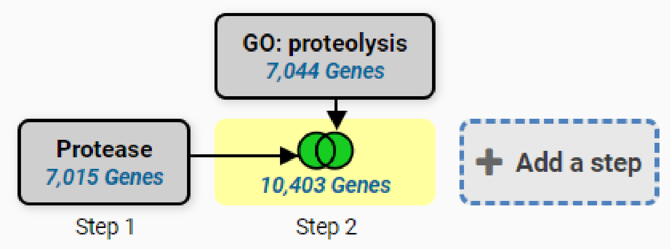
Run Step

Give this search a name (optional)

Give this search a weight (optiona

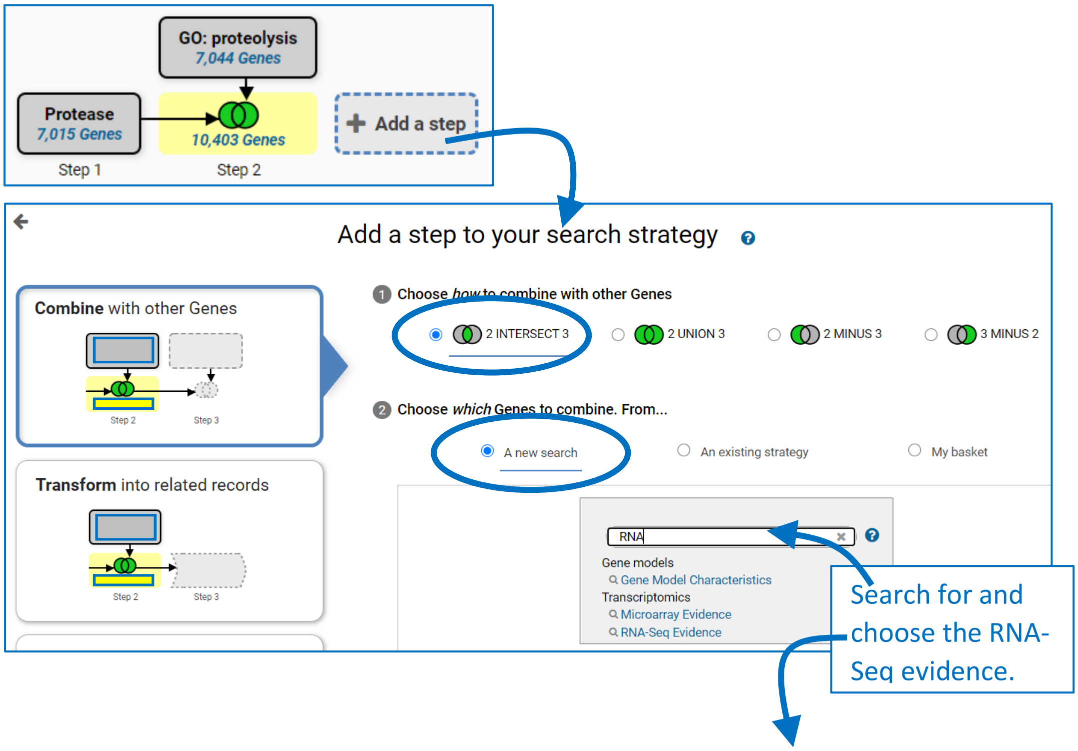
Click Run Step to initiate the search

Strategy Result: The GO term search returned 7044 genes annotated with the proteolysis GO term. The union of the text and GO search returns 10,403 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 many RNA sequencing data sets you must first choose what data set (experiment) to search before you can 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



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...

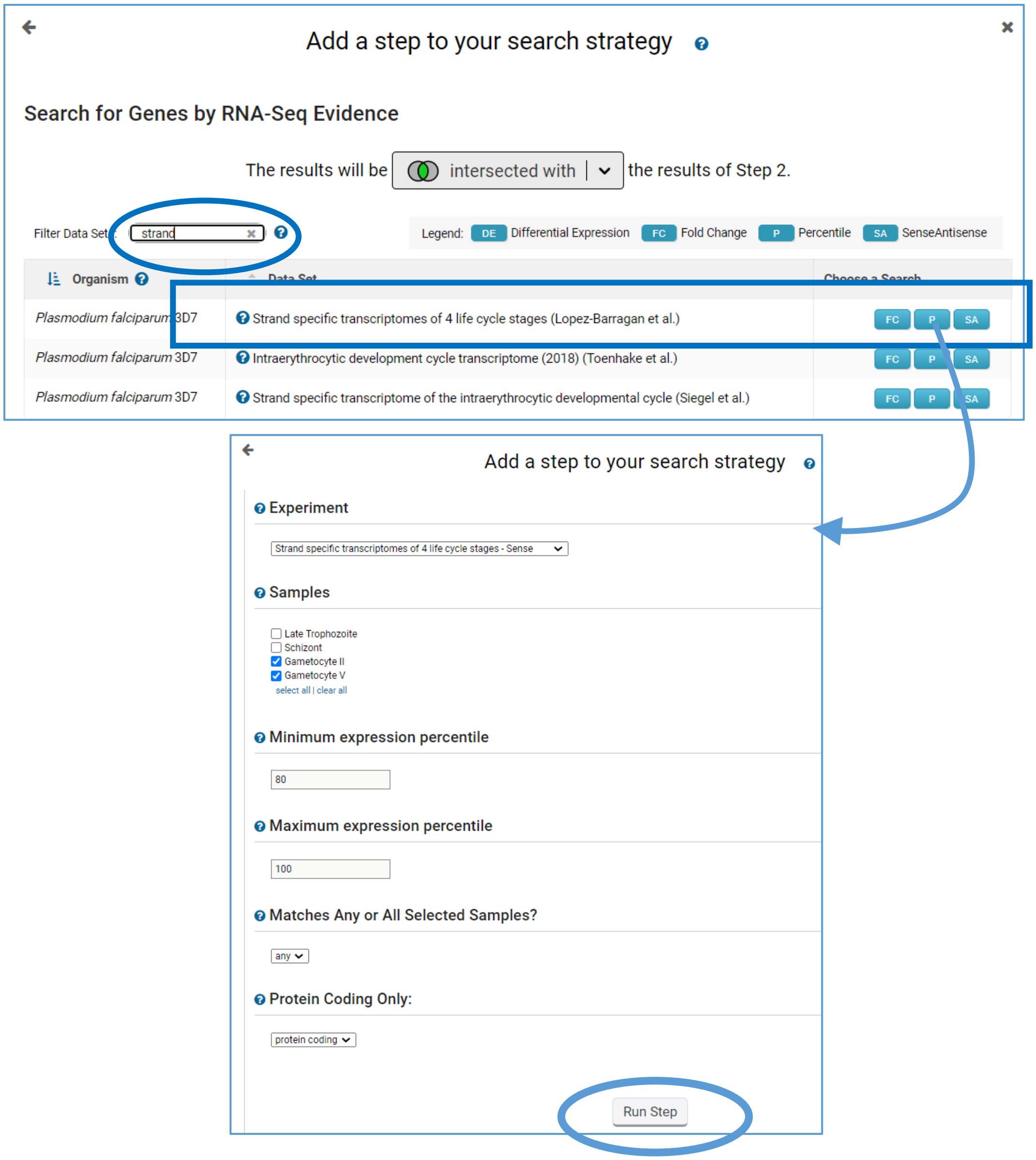
● A new search ○ An existing strategy ○ My basket

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: strand

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

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

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

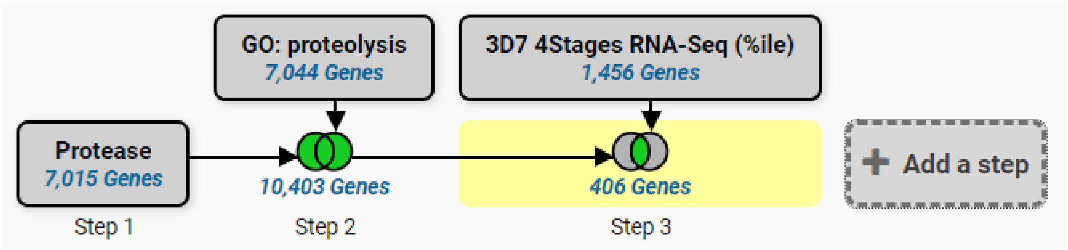
Matches Any or All Selected Samples?

any

Protein Coding Only:

protein coding

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!!



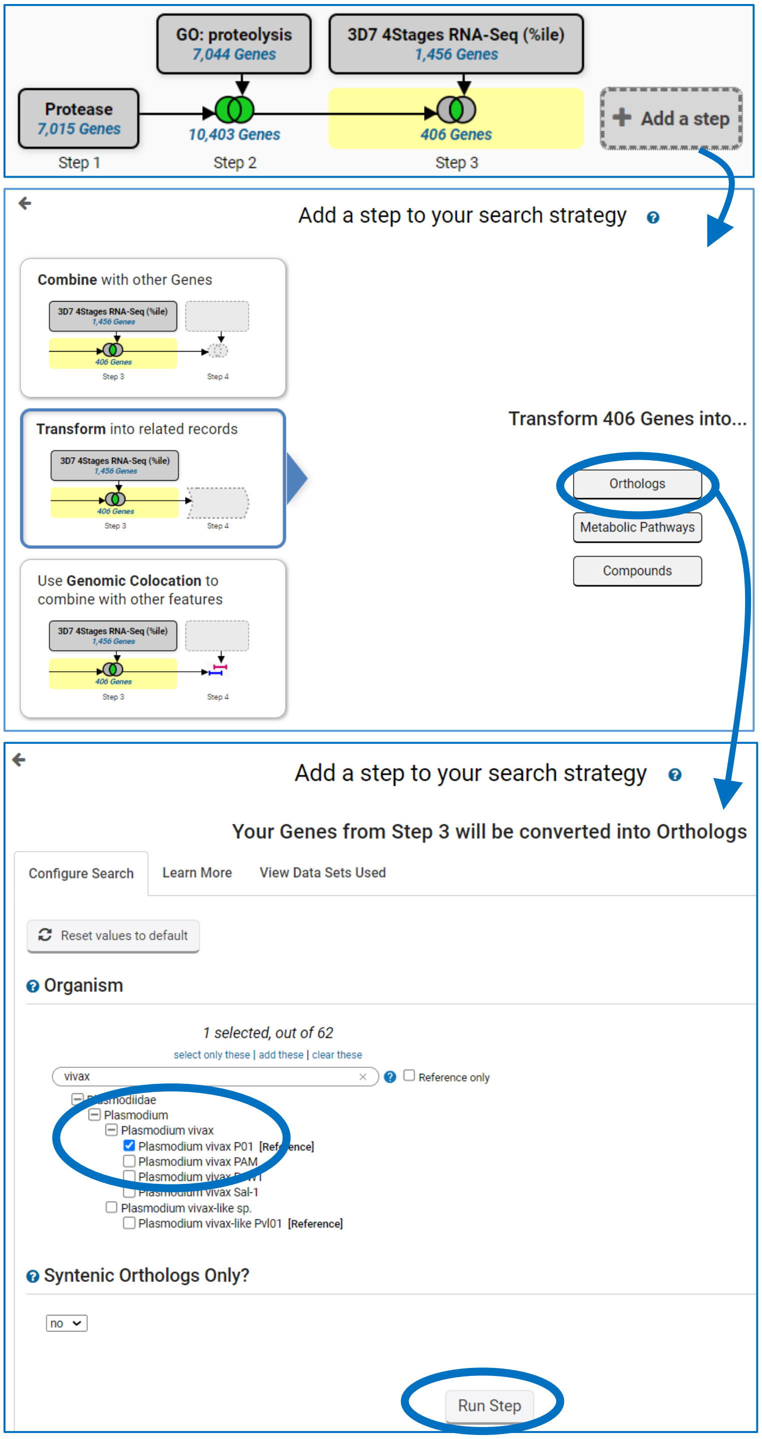
Step 1: Protease — 7,015 Genes

Step 2: GO: proteolysis — 7,044 Genes; 10,403 Genes

Step 3: 3D7 4Stages RNA-Seq (%ile) — 1,456 Genes; 406 Genes

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

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



Add a step to your search strategy

Transform 406 Genes into...

Add a step to your search strategy

Your Genes from Step 3 will be converted into Orthologs

Organism

1 selected, out of 62

vivax

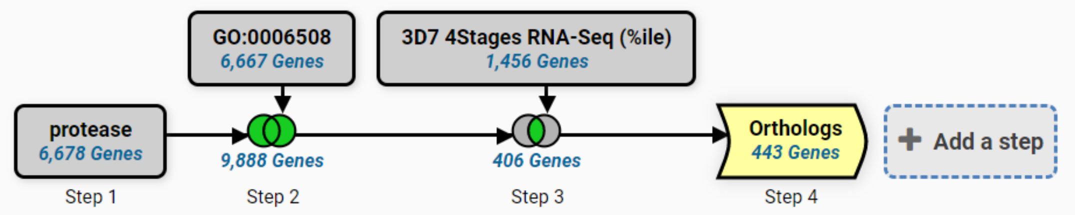
□ Reference only

⊟ Plasmodiidae
⊟ Plasmodium
⊟ Plasmodium vivax
☑ Plasmodium vivax P01 [Reference]
□ Plasmodium vivax PAM
□ Plasmodium vivax PvW1
□ Plasmodium vivax Sal-1
□ Plasmodium vivax-like sp.
□ Plasmodium vivax-like Pvl01 [Reference]

Syntenic Orthologs Only?

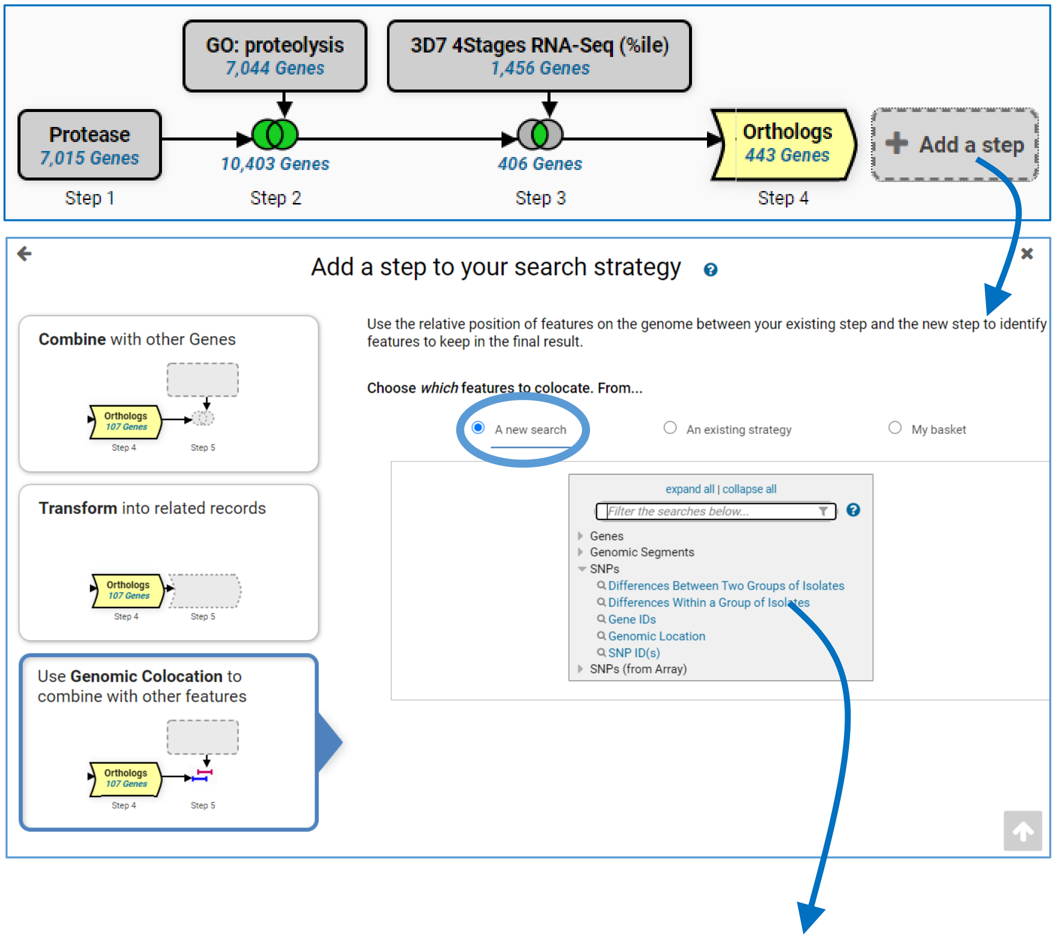
no

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



4. **Add a step to the strategy that returns *P vivax* SNPs and co-locate 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 whole-genome sequences 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 co-location tool.

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

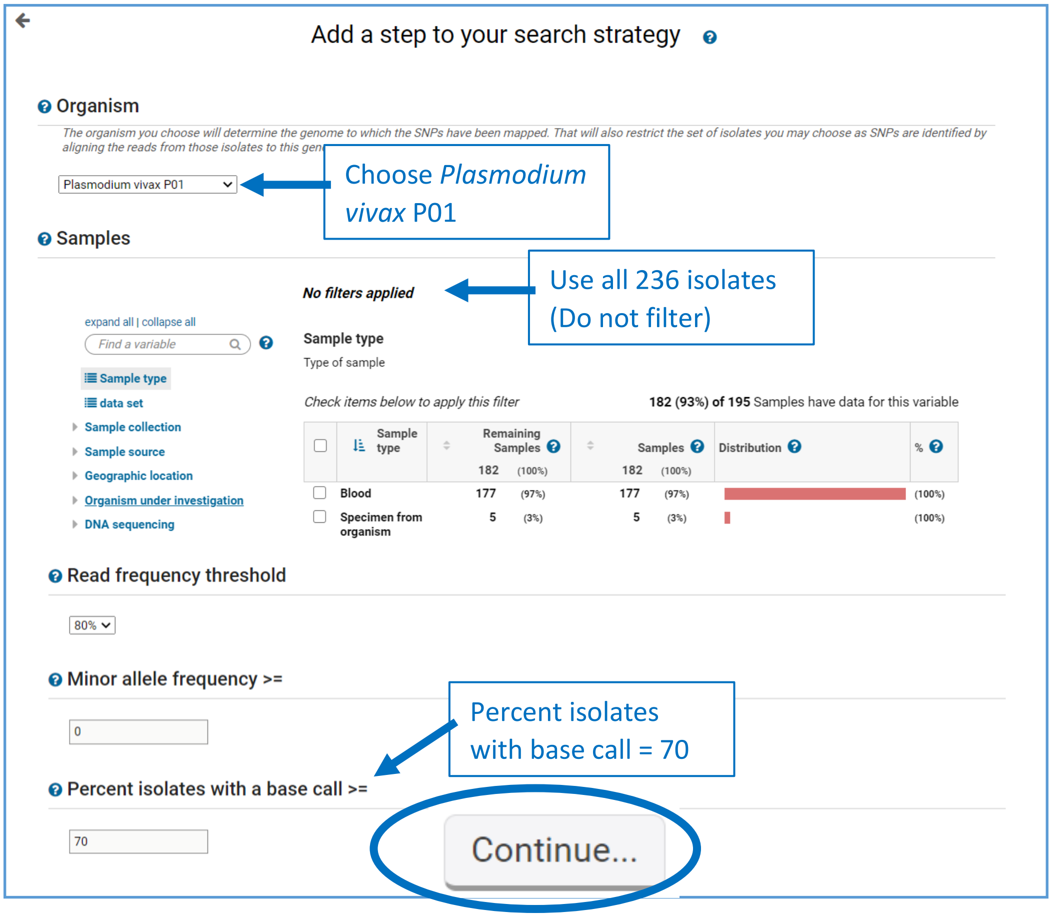


Add a step to your search strategy

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



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 gen

Plasmodium vivax P01

Choose *Plasmodium vivax* P01

Samples

No filters applied

Use all 236 isolates (Do not filter)

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 182 (100%)	Samples 182 (100%)	Distribution	%
□	Blood	177 (97%)	177 (97%)	🟥	(100%)
□	Specimen from organism	5 (3%)	5 (3%)	🟥	(100%)

Read frequency threshold

80%

Minor allele frequency >=

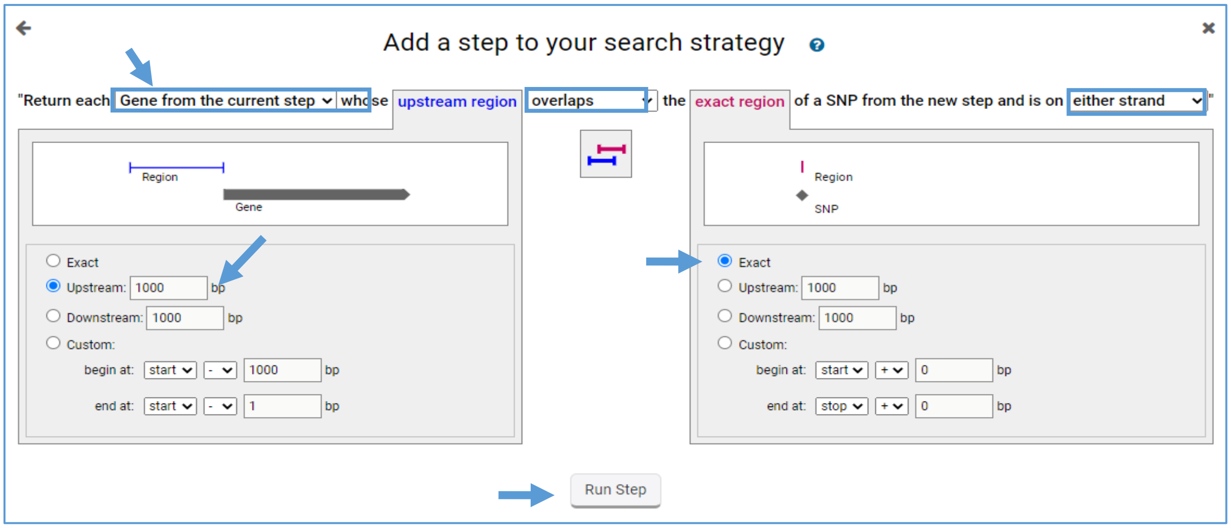
0

Percent isolates with a base call >=

70

Percent isolates with base call = 70

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.



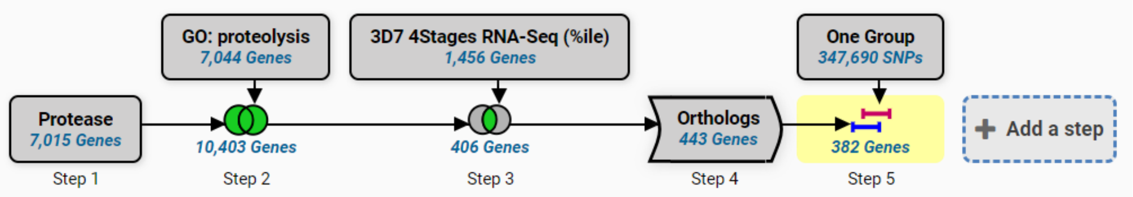
Add a step to your search strategy

"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"

○ 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

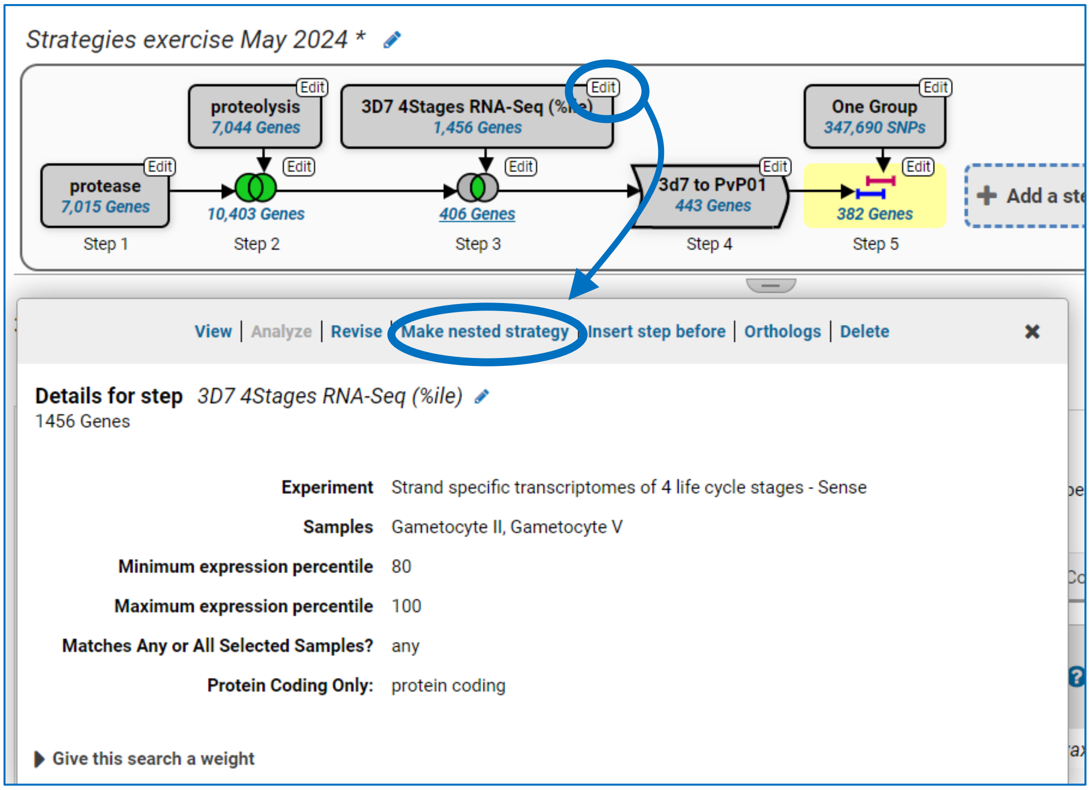
Strategy Result: You have completed a 5-step strategy and have a list of 382 *P. vivax* genes that are possible proteases, are likely expressed in gametocytes and have upstream SNPs. This link will retrieve the strategy so far: https://plasmodb.org/plasmo/app/workspace/strategies/import/d67d74edca408d0b



5. **Increase the specificity of the gametocyte calls at Step 3. Use a nested strategy at Step 3 to remove genes that also showed high expression in the asexual stages.** Nested strategies allow for controlling the logic of the strategy and help organize long strategies. The searches and operations in a nested strategy are performed first (like a parenthesis in mathematical equations) and the nested strategy result is sent to the parent strategy.

Open a nested strategy

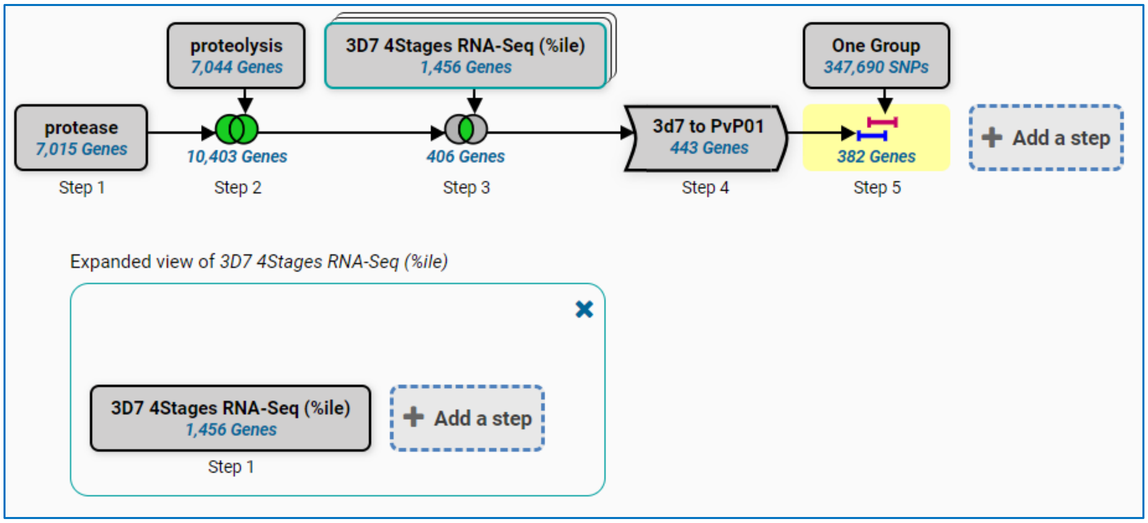
Navigation: >hover over the strategy >click Edit at Step 3 >Make nested strategy



Details for step *3D7 4Stages RNA-Seq (%ile)*
1456 Genes

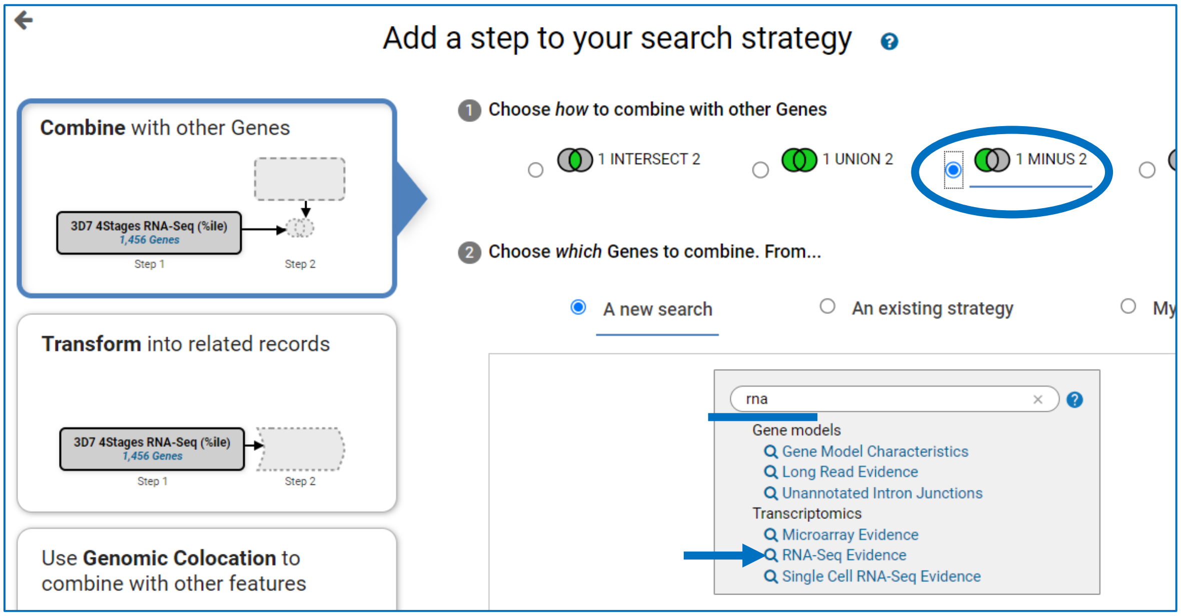
Experiment Strand specific transcriptomes of 4 life cycle stages - Sense
Samples Gametocyte II, Gametocyte V
Minimum expression percentile 80
Maximum expression percentile 100
Matches Any or All Selected Samples? any
Protein Coding Only: protein coding

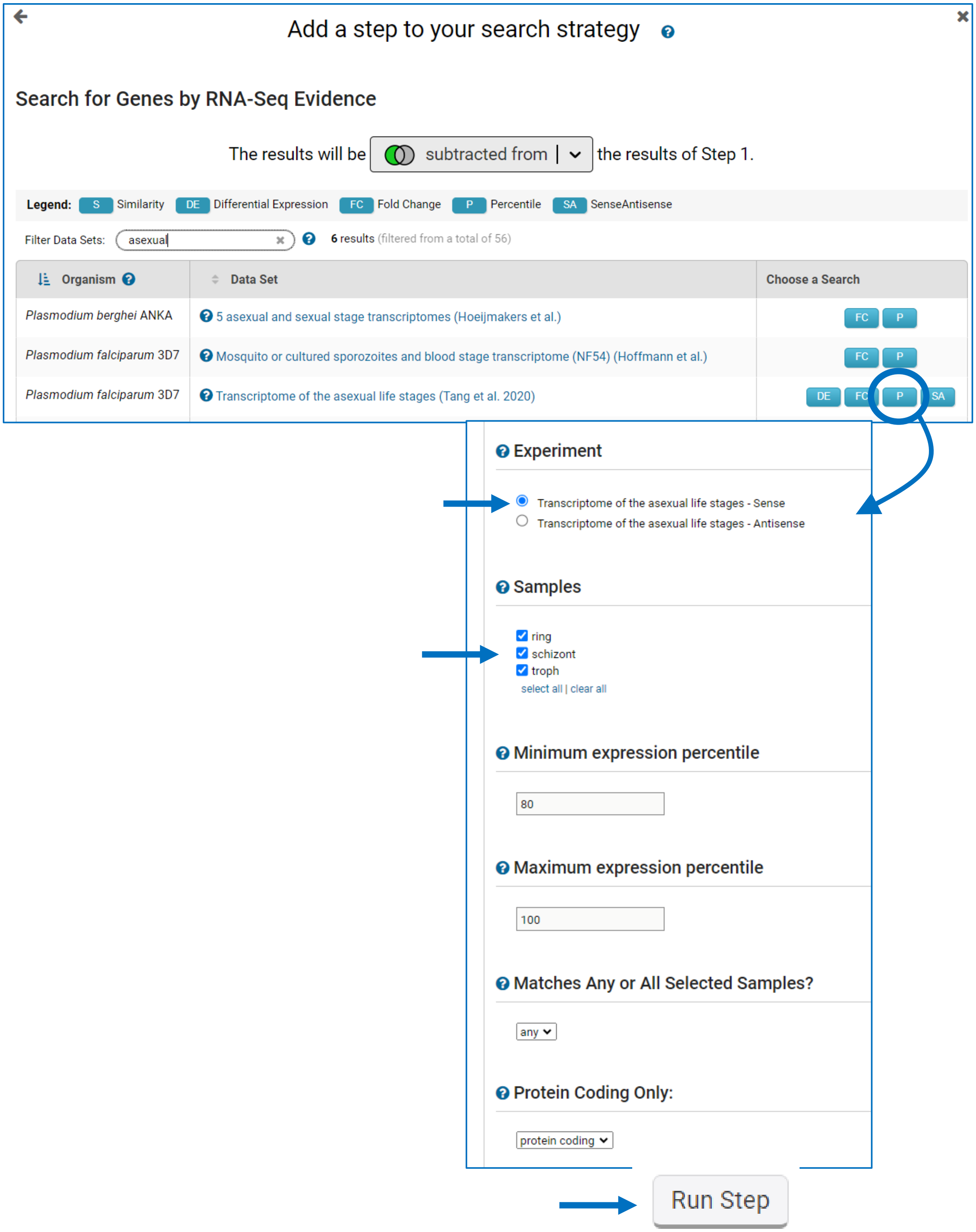
▶ Give this search a weight



Add a step in the nested strategy that subtracts genes with high expression (80-100 percentile) in the asexual stages. The RNA-sequence data set called **Transcriptome of the asexual life stages (Tang et al. 2020)** This data set contains transcriptomes for Schizont, ring and trophozoite stages. We can subtract genes with high expression at these stages to make our 'gametocyte gene' list more stage specific.

Navigation: >Add Step >1 minus 2 >RNA-Seq Evidence >Percentile search for Transcriptome of Asexual Stages


Add a step to your search strategy
Combine with other Genes
3D7 4Stages RNA-Seq (%ile)
1,456 Genes
Step 1
Step 2
Transform into related records
3D7 4Stages RNA-Seq (%ile)
1,456 Genes
Step 1
Step 2
Use Genomic Colocation to combine with other features
1 Choose how to combine with other Genes
○ 1 INTERSECT 2
○ 1 UNION 2
● 1 MINUS 2
2 Choose which Genes to combine. From...
● A new search
○ An existing strategy
○ My
rna
Gene models
Gene Model Characteristics
Long Read Evidence
Unannotated Intron Junctions
Transcriptomics
Microarray Evidence
RNA-Seq Evidence
Single Cell RNA-Seq Evidence



Add a step to your search strategy

Search for Genes by RNA-Seq Evidence

The results will be subtracted from the results of Step 1.

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

Filter Data Sets: asexual

6 results (filtered from a total of 56)

Organism	Data Set	Choose a Search
Plasmodium berghei ANKA	5 asexual and sexual stage transcriptomes (Hoeijmakers et al.)	FC P
Plasmodium falciparum 3D7	Mosquito or cultured sporozoites and blood stage transcriptome (NF54) (Hoffmann et al.)	FC P
Plasmodium falciparum 3D7	Transcriptome of the asexual life stages (Tang et al. 2020)	DE FC P SA

Experiment

● Transcriptome of the asexual life stages - Sense
○ Transcriptome of the asexual life stages - Antisense

Samples

☑ ring
☑ schizont
☑ troph
select all | clear all

Minimum expression percentile

80

Maximum expression percentile

100

Matches Any or All Selected Samples?

any

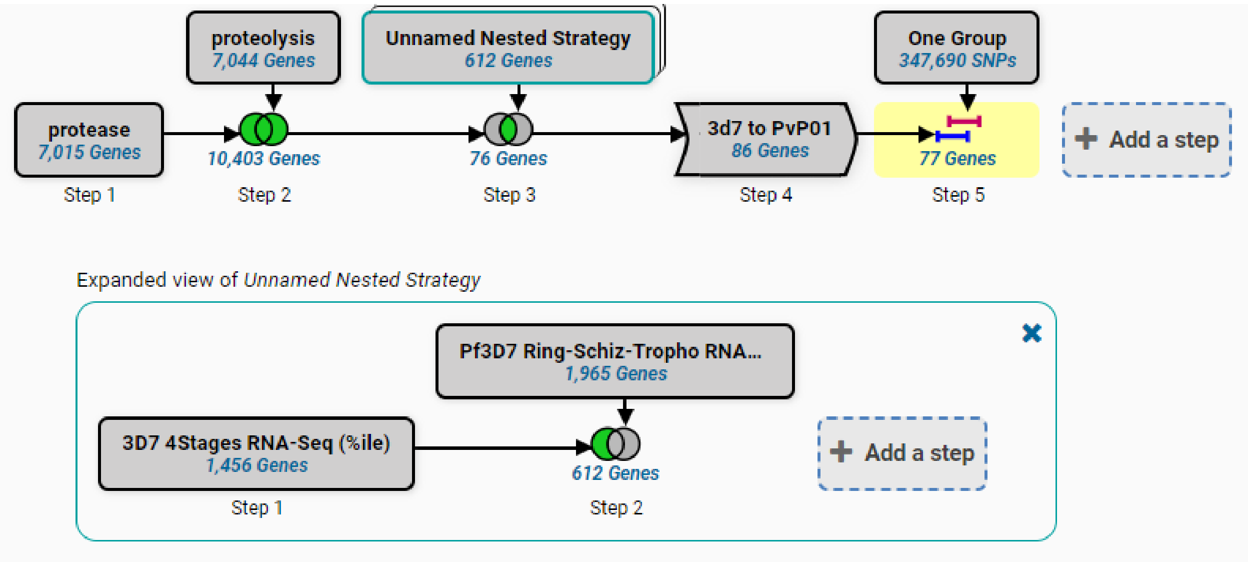
Protein Coding Only:

protein coding

Run Step

Strategy Result: Subtracting genes with high expression in the asexual stages reduces the number of genes in the final result from 382 to 77.

Here is a link to the final strategy
https://plasmodb.org/plasmo/app/workspace/strategies/import/65e11c1ac70478b1


proteolysis
7,044 Genes
Unnamed Nested Strategy
612 Genes
One Group
347,690 SNPs
protease
7,015 Genes
10,403 Genes
76 Genes
3d7 to PvP01
86 Genes
77 Genes
Add a step
Step 1
Step 2
Step 3
Step 4
Step 5
Expanded view of *Unnamed Nested Strategy*
Pf3D7 Ring-Schiz-Tropho RNA...
1,965 Genes
3D7 4Stages RNA-Seq (%ile)
1,456 Genes
612 Genes
Add a step
Step 1
Step 2