VEuPathDB
Eukaryotic Pathogen, Vector & Host Informatics Resources

# 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[1] 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 search 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.

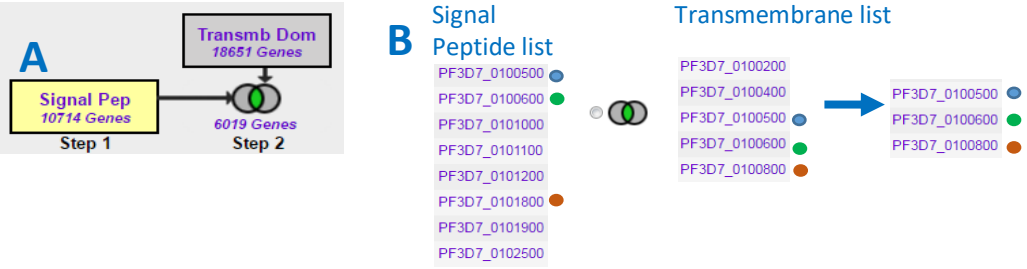
[1] Note: This exercise uses PlasmoDB.org as an example, but the same functionality is available on all VEuPathDB genomics resources

## 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.

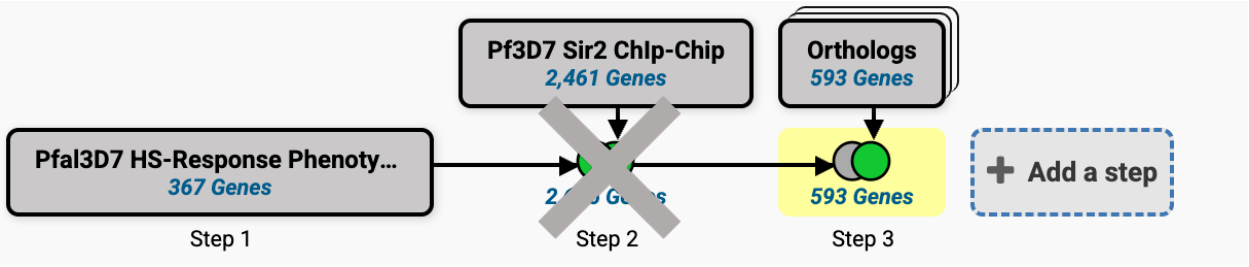
| | Operator | : | Combined Result will contain: |
|---|---|---|---|
| 1 | 1 INTERSECT 2 | : | IDs in common between the two lists |
| 2 | 1 UNION 2 | : | IDs from list 1 and list 2 |
| 3 | 1 MINUS 2 | : | IDs unique to 1 |
| 4 | 2 MINUS 1 | : | IDs unique to 2 |
| 5 | 1 Relative to 2 | : | IDs whose 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.

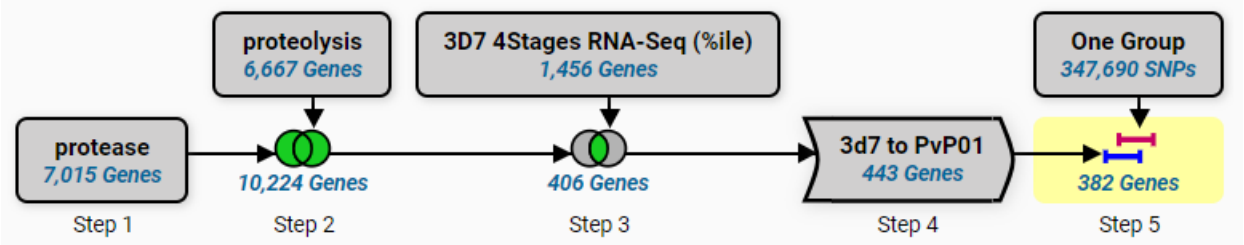
In multistep search strategies, you can also use the **Ignore** operator (options 6 and 7 in the table above) to mask off steps before or after certain search results. This also allows you to apply different search criteria without duplicating search strategies or deleting individual steps.




## Building the Strategy

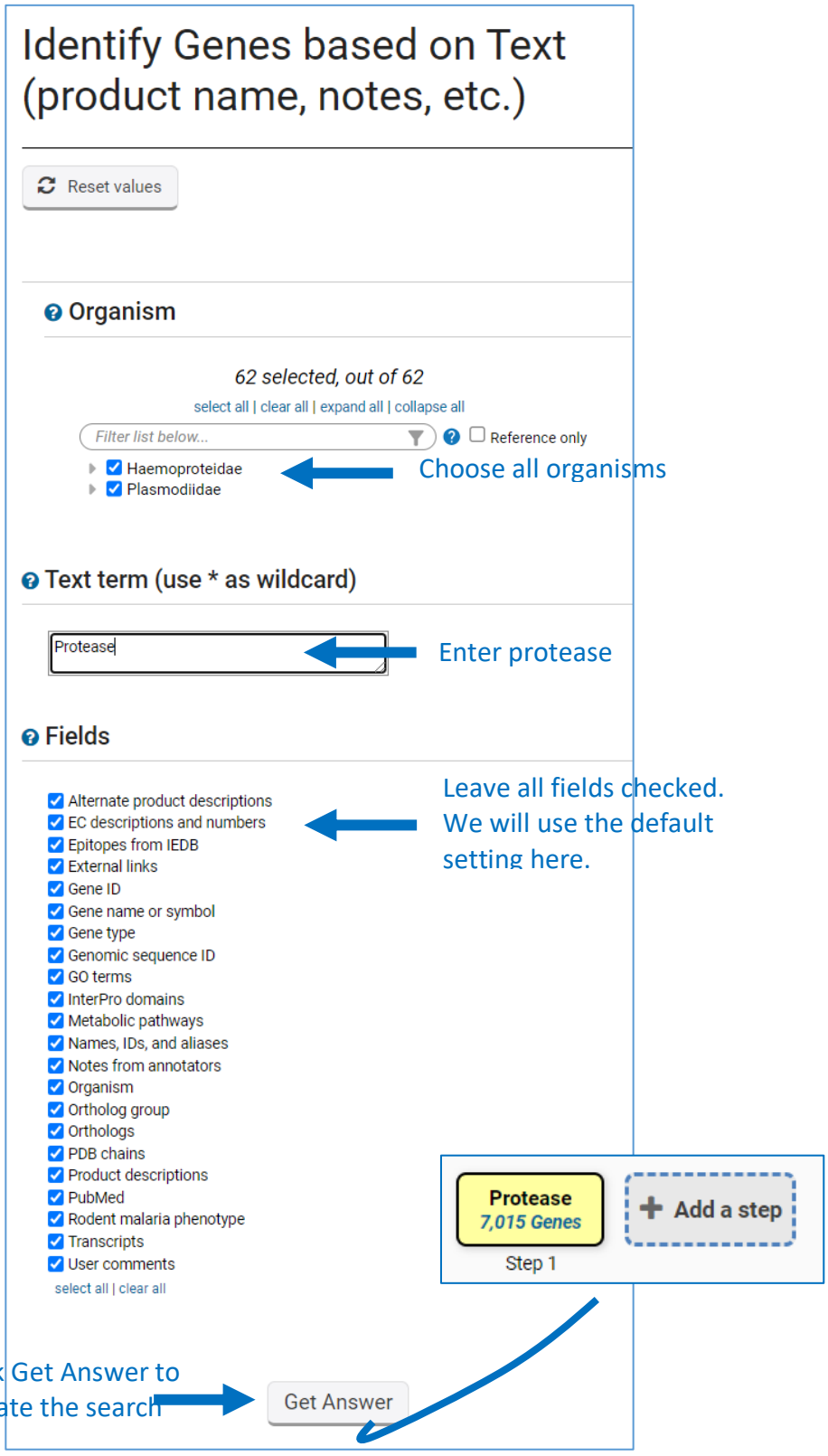
**Find *P. vivax* genes that are possible proteases, likely expressed during the gametocyte stages, and contain SNPs in their upstream regions.**

The final strategy will look like this.



## Step by Step Instructions

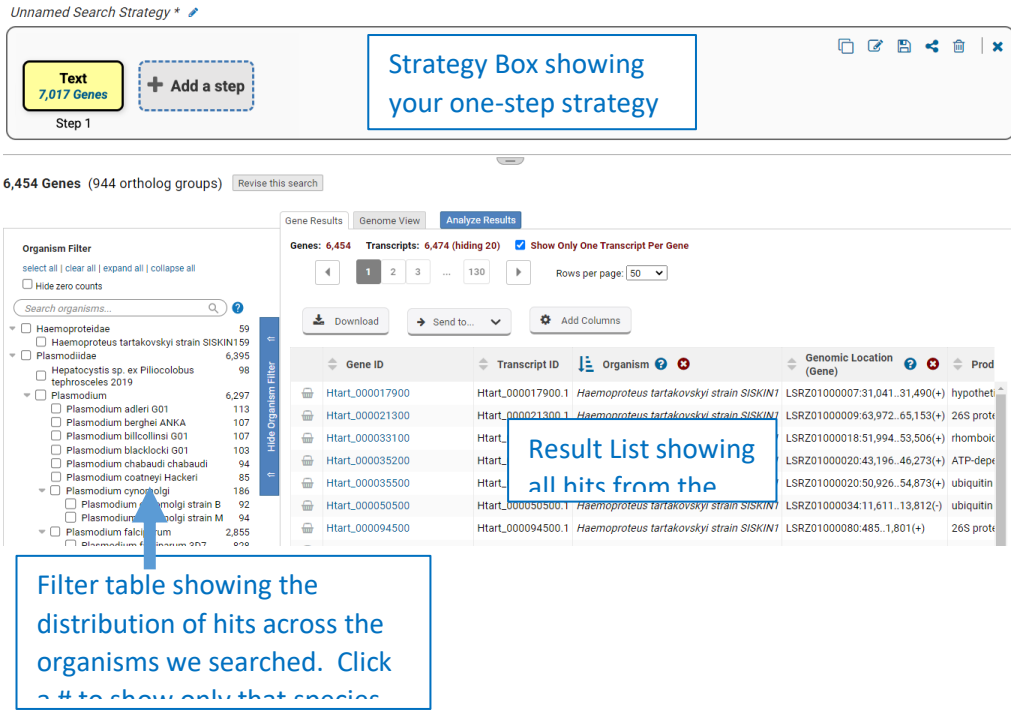
1. **Run a text search using protease as the text term.**
   **Navigation**: >PlasmoDB >Search for Genes >Text > Text (product name, notes, etc.)



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. You can analyze this result by exploring the hits.

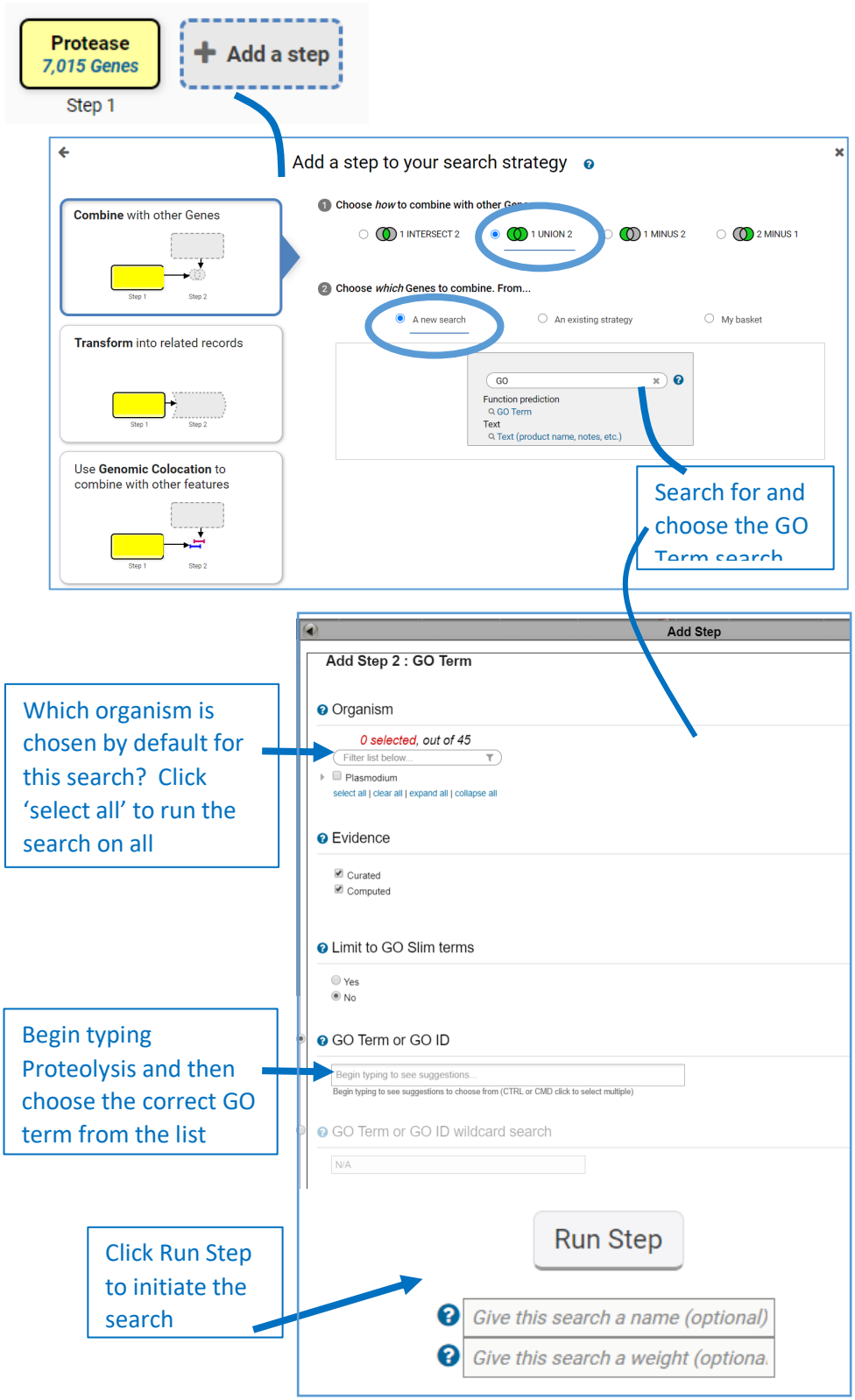
Look at the data in the columns of the result table.  You can add more data with the **Add Columns** button.

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’.

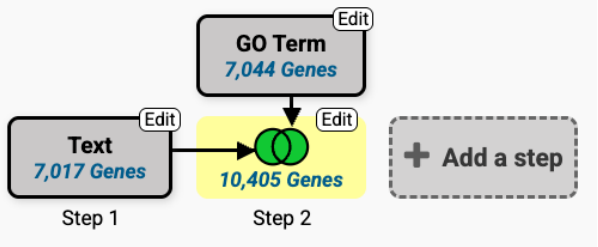


2. **Add a step choosing to run a search for genes annotated with the biological process gene ontology term – GO:0006508: proteolysis.** Gene Ontology annotations offer a second line of evidence for finding proteases.

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

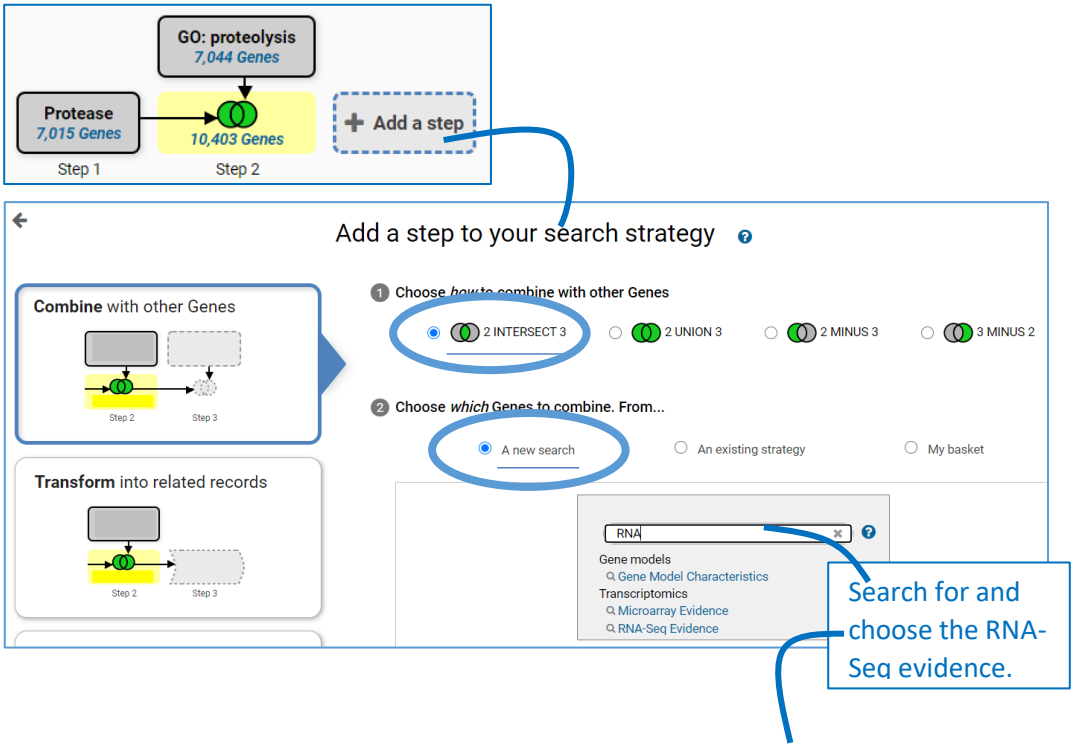


**Strategy Result**: The GO term search returned 7044 genes annotated with the proteolysis GO term. The union of the text and GO search returns 10,403 genes that are suspected to have proteolytic activity.

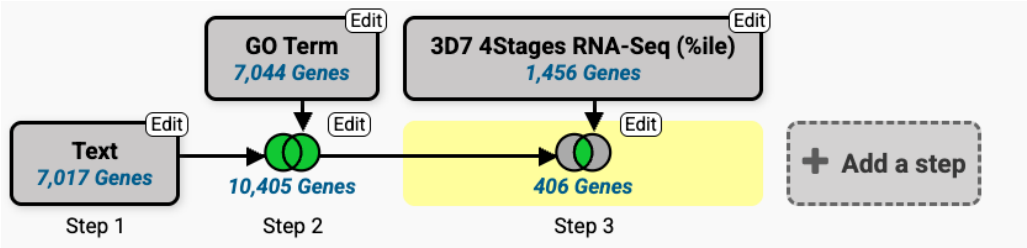



3. **Add a step choosing to run a search for genes based on Transcript Expression using RNA-Seq Evidence.** Since PlasmoDB has integrated several RNA sequencing data sets you must first choose what data set (experiment) to search before you are taken to the search form to choose parameters. Choose the 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

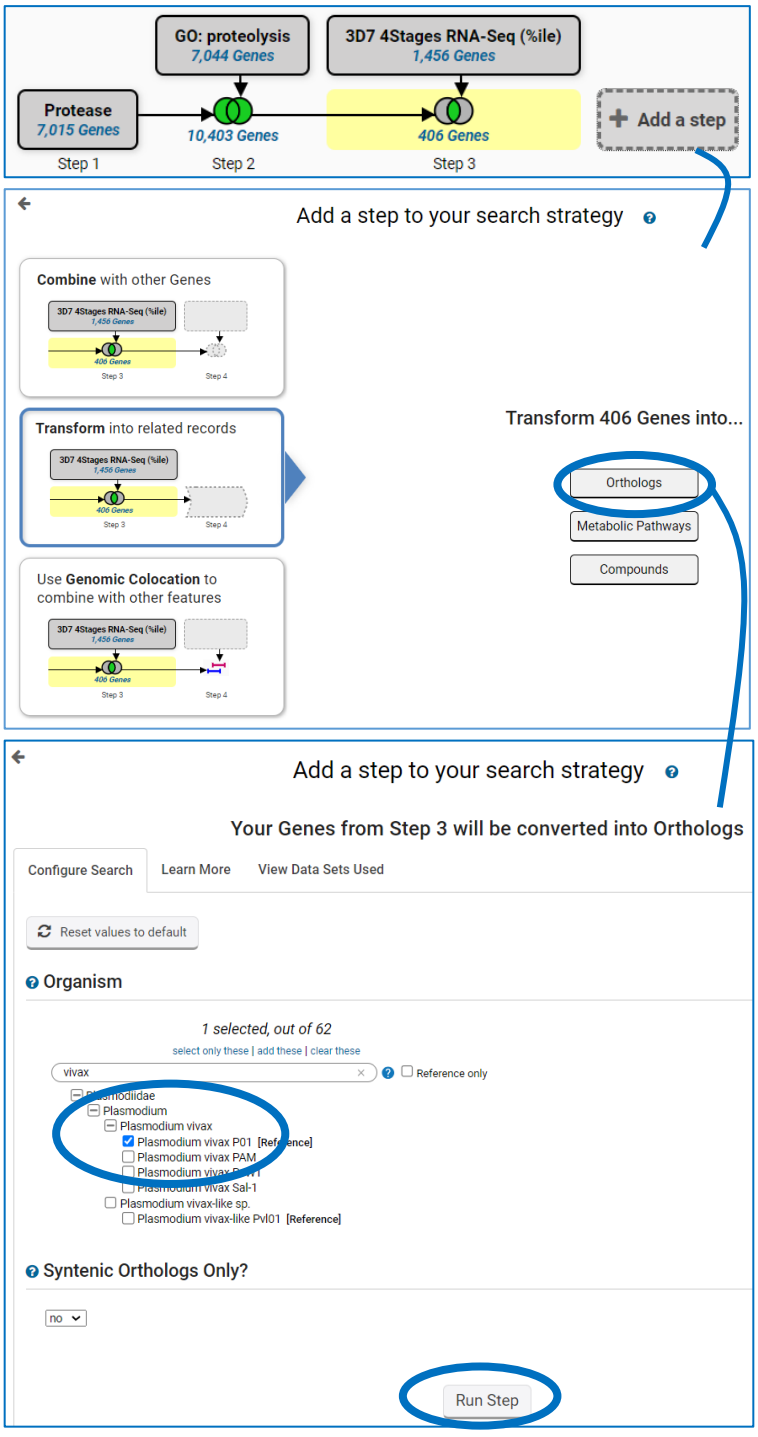



**Strategy result**: We have a three-step strategy that returns 406 *P. falciparum* genes that are suspected proteases with evidence for expression in gametocytes based on RNA Sequencing data. Explore your gene list!!

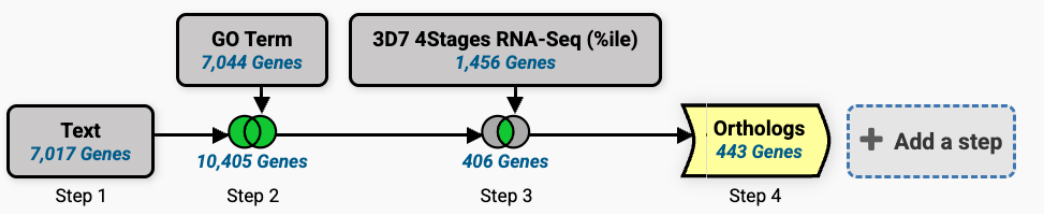


4. **Add a step to the strategy that transforms the 406 *P. falciparum* genes into *P. vivax* genes.**
   *P. falciparum* is a well-studied organism with active curatorial efforts and large amounts of functional data. A researcher interested in *P. vivax* can take advantage of the *P. falciparum* data 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

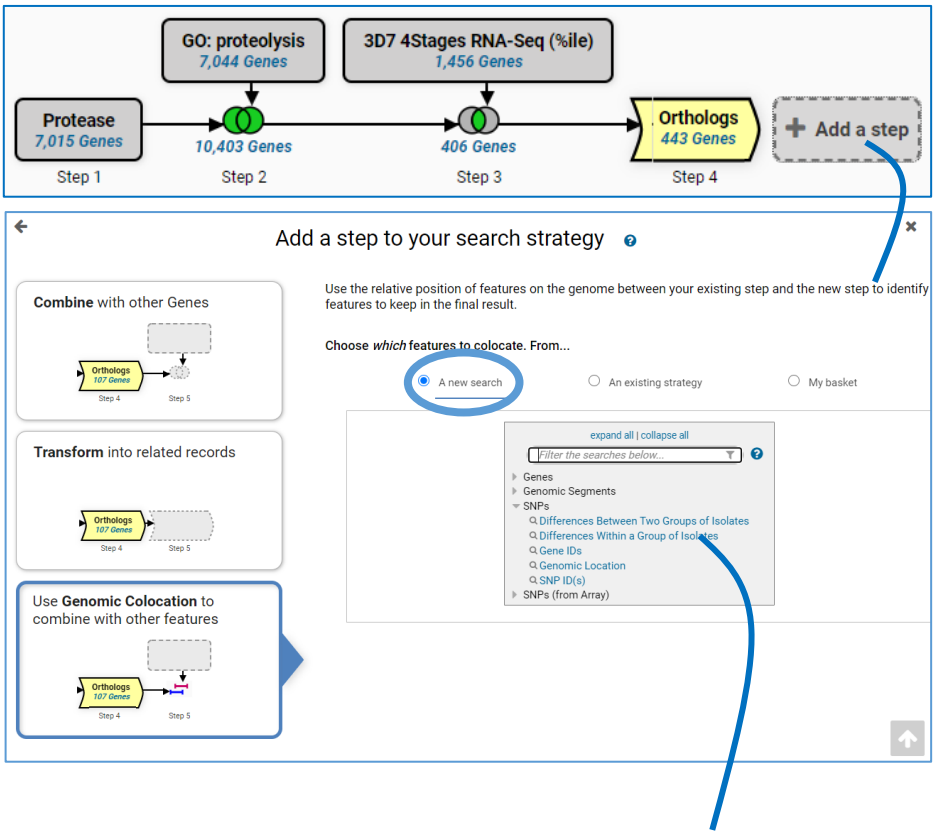


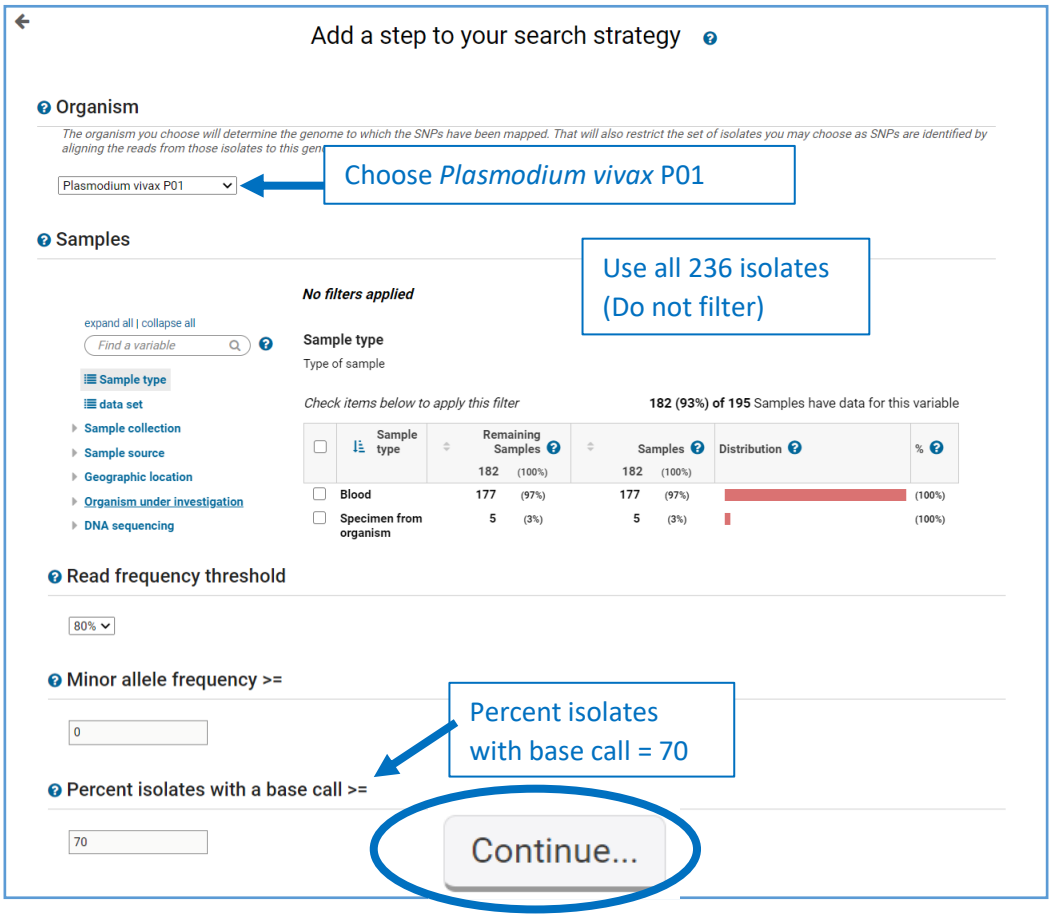
**Strategy Result**: We have a four-step strategy that returns 443 *P. vivax* genes that are suspected proteases with evidence for expression in gametocytes based on RNA Sequencing data.



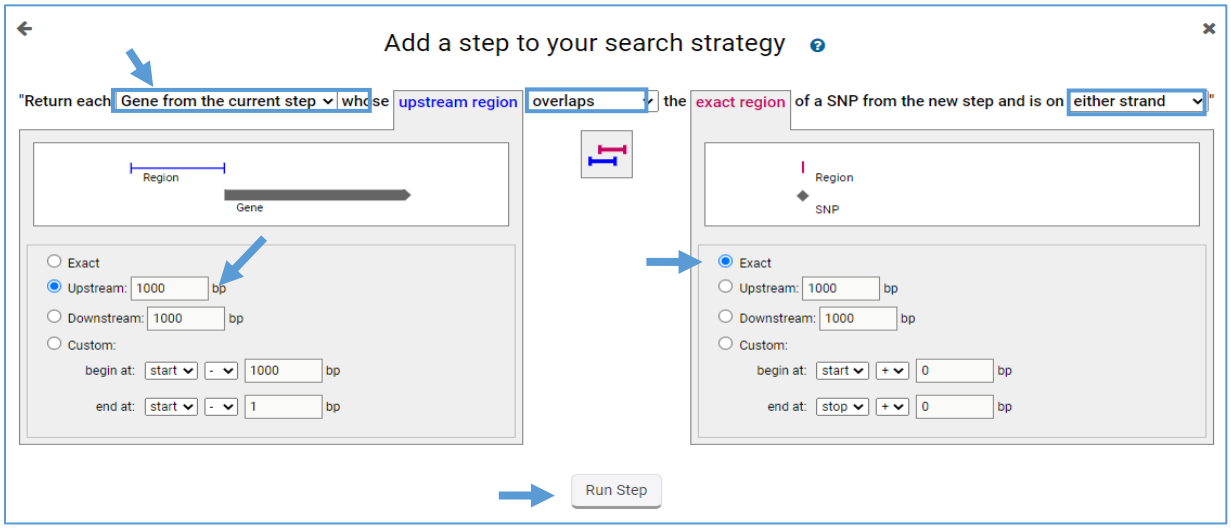

5. **Add a step to the strategy that returns *P vivax* SNPs and collocate those SNPs to the upstream 1000bp of the *P. vivax* genes in step 4.** We can look for variation (SNPs) associated with the genes from Step 4. PlasmoDB integrates whole genome resequencing data from many isolates, and PlasmoDB 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






**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.



**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