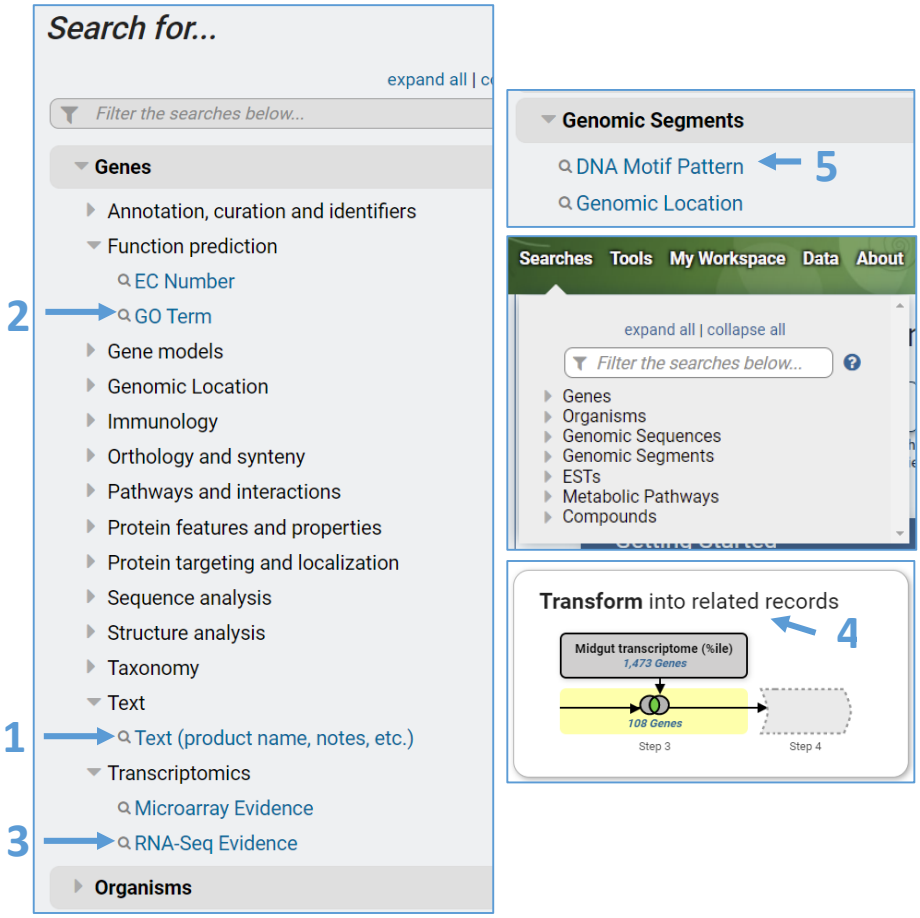
## Strategies Training Module

In this tutorial you will find genes expressed in the midgut that are likely proteases and have a DNA motif in their upstream regions, possibly acting as a binding site for transcription factors. The strategy you build will combine three different searches to retrieve a set of *Anopheles culicifacies* A-37 genes that are likely proteases expressed in the midgut. Then you will transform those genes into their *Anopheles gambiae* PEST orthologs and use a fourth search to look for a DNA motif in the upstream regions of the *Anopheles gambiae* PEST genes. The ortholog transform enables you to use expression data collected in *Anopheles culicifacies* to make inferences about genes in *Anopheles gambiae.* The *Anopheles gambiae* genes returned by the completed strategy are likely to share two biological properties, proteolytic activity and expression in the midgut, and have a DNA motif in their upstream regions, possibly acting as a binding site for regulatory proteins.

## 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 'Search For…' menu on the home page and from the 'Searches' dropdown menu. Searches listed under Genes will return a list of gene IDs, while searches listed under 'SNPs' or 'Genomic Segments' will return record IDs representing those features. Genomic Segments are distinct regions of the genome that are not associated with genes, such as DNA motifs.



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 – VectorBase 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 – VectorBase integrates ortholog profiles from OrthoMCL. The OrthoMCL algorithm clusters proteins into ortholog groups based on BLAST similarity across at 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 *Anopheles culicifacies* A-37 genes into their *Anopheles gambiae* PEST orthologs.
5. Identify Genomic Segments based on DNA Motif Pattern – A nucleotide or amino-acid sequence pattern that is widespread and can have a biological significance. The chromosomes and contigs of whole genome sequence can be considered to be text strings, and it is easy to search text strings for specific patters. This search will find genomic segments (pieces of the chromosome or contig) whose DNA contains a motif pattern (or short text string) that you specify.

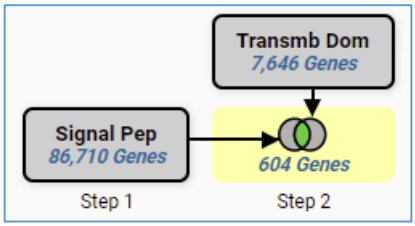
## Before we get started… a few words about combining search results:

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.

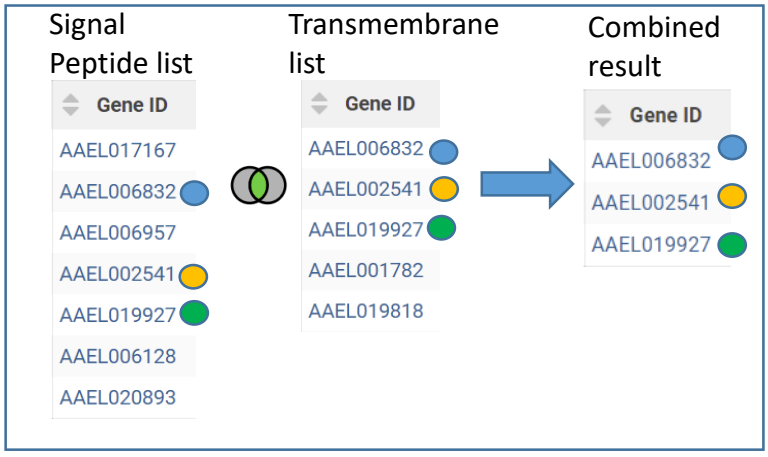
| Operator | : | Combined Result will contain: |
|---|---|---|
| 1 INTERSECT 2 | : | IDs in common between the two lists |
| 1 UNION 2 | : | IDs from list 1 and list 2 |
| 1 MINUS 2 | : | IDs unique to 1 |
| 2 MINUS 1 | : | IDs unique to 2 |
| 1 Relative to 2 | : | IDs whose features are near each other (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 and B below).

A

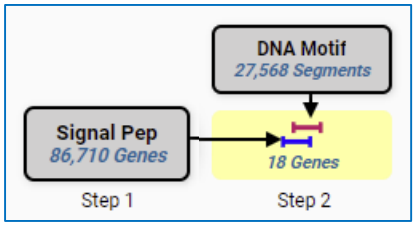



B

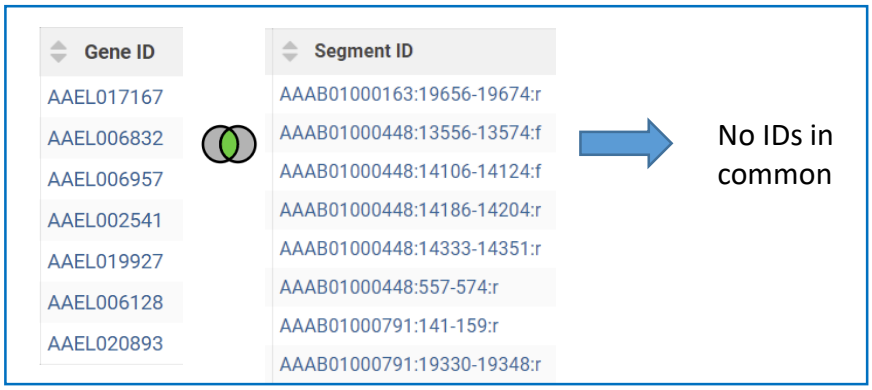



However, searches that return different genomic features will yield no results when combined with intersect, union or minus operators. This is illustrated in screenshot groupings C and D below. Because genes and motifs are different genomic features, there are no IDs in the list of genes (Step 1) that are present in the list of SNPs (Step 2). To combine a search that returns genes with a search that returns motifs, you must use the collocation option (1 relative to 2). We know the genomic location of each gene and each motif and the colocation option is designed to return features based on their relative genomic location, i.e. motifs that are near or within genes.

C

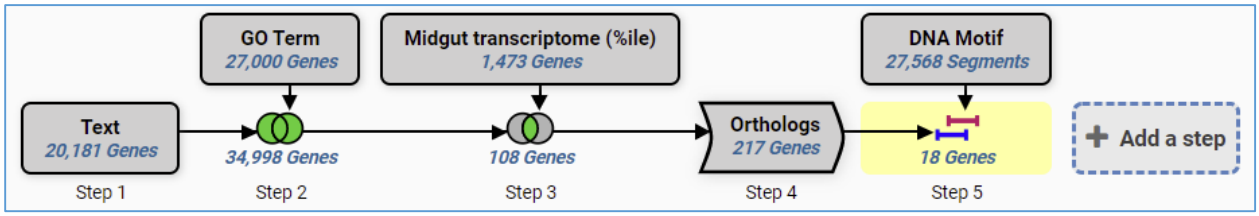



D



## Building the Strategy:

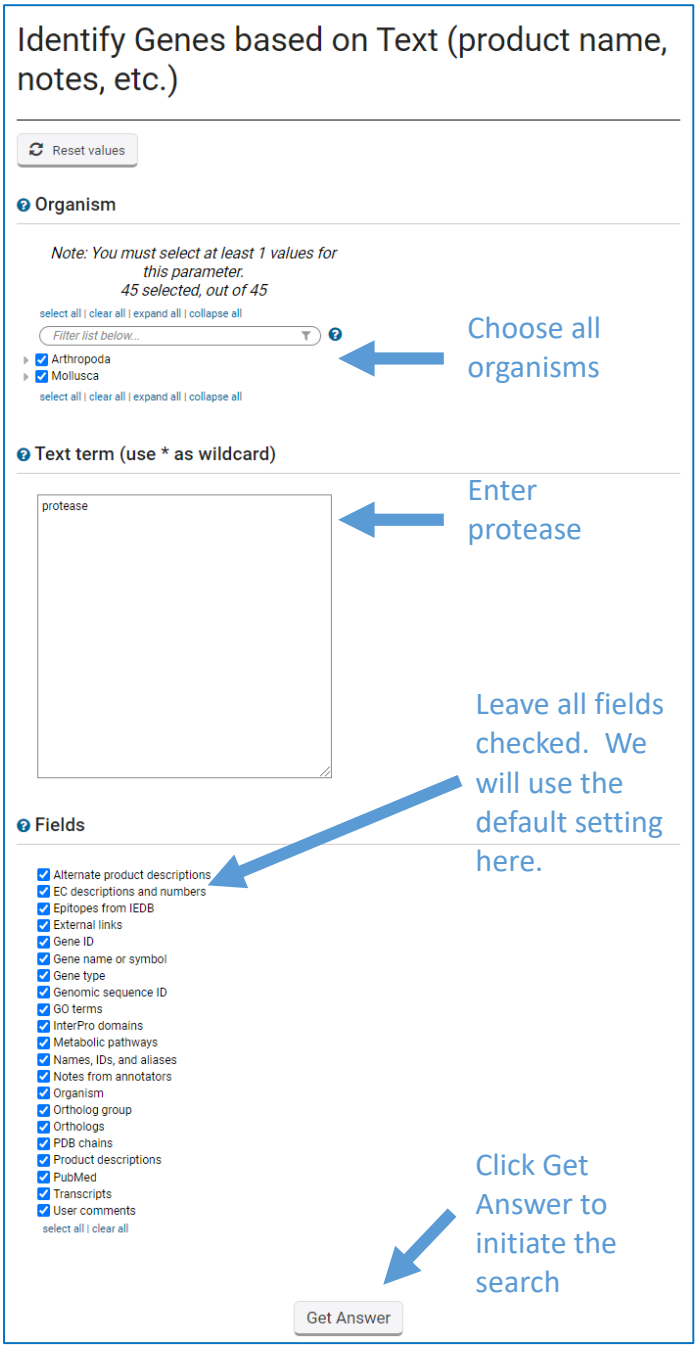
**Find *Anopheles gambiae* genes that are possible proteases, likely expressed in the midgut and contain a DNA motif in their upstream regions.** This search strategy employs 4 searches, an ortholog transform and the colocation tool to integrate the motif information. Steps 1 and 2 return proteases using 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 in the midgut based on RNA sequencing data collected in *Anopheles culicifacies* A-37. Steps 2 and 3 are combined using the intersect operator to produce a list of genes that have BOTH biological properties: these genes are likely proteases with evidence for expression in the midgut. In the next step, the *Anopheles culicifacies* genes returned in the step 3 result are transformed into their *Anopheles gambiae* orthologs. This results in a set of 217 *Anopheles gambiae* genes with suspected protease activity and expression in midguts based on annotation and experimental evidence from *Anopheles culicifacies*. In Step 5 we look for a DNA motif that is known to be a transcription factor binding site in other organisms, and collocate these DNA motifs to the upstream regions of the *Anopheles gambiae* genes. The final result is a set of 18 *Anopheles gambiae* genes that are likely proteases expressed in the midgut stage and that have DNA motifs in their upstream regions that may attract regulatory proteins. 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.): 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 'Search For…' menu. 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. 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.

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

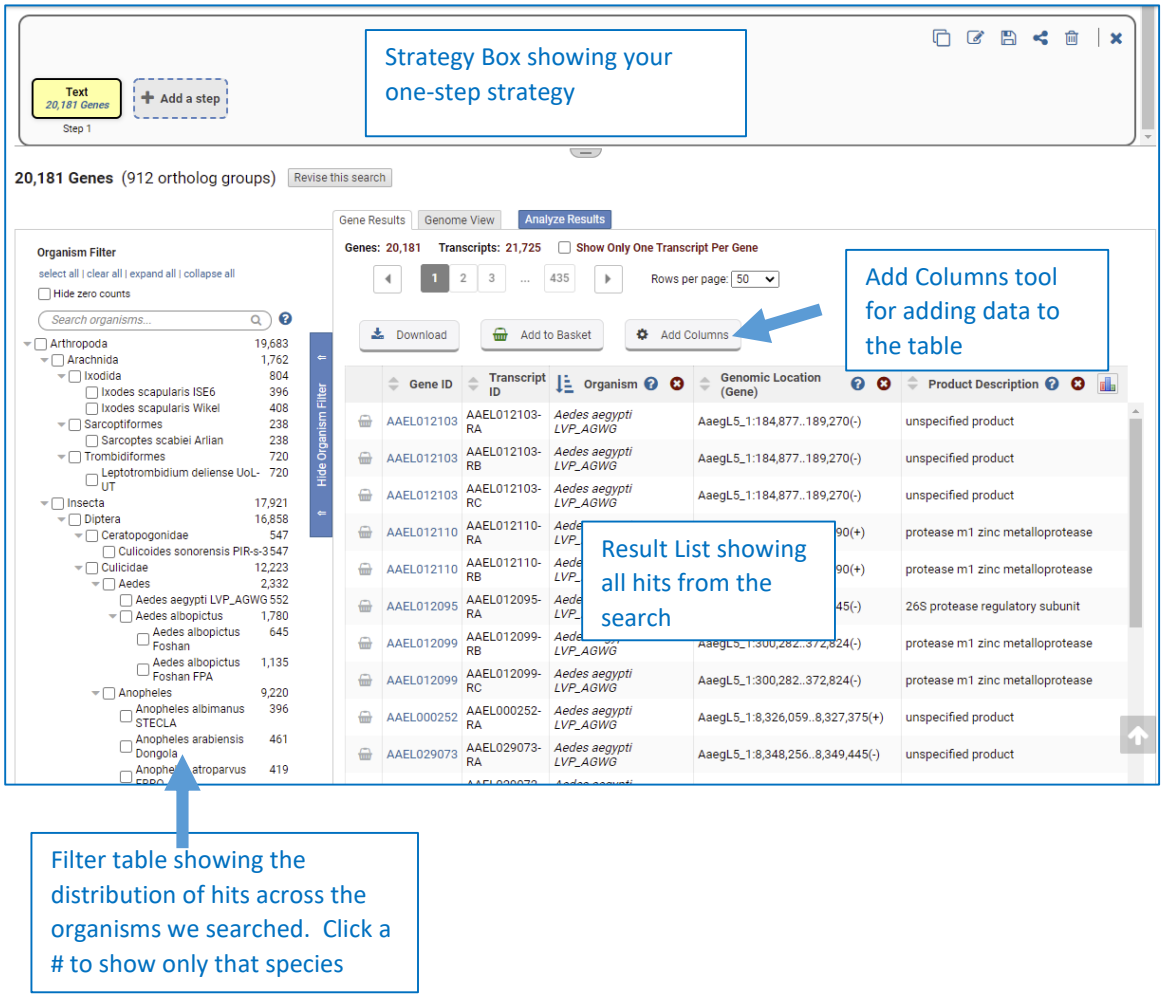



**Parameters:**

| | | |
|---|---|---|
| **Organism** | **:** | Default - all |
| **Text term (use * as wildcard)** | **:** | protease |
| **Fields** | **:** | Default - all |

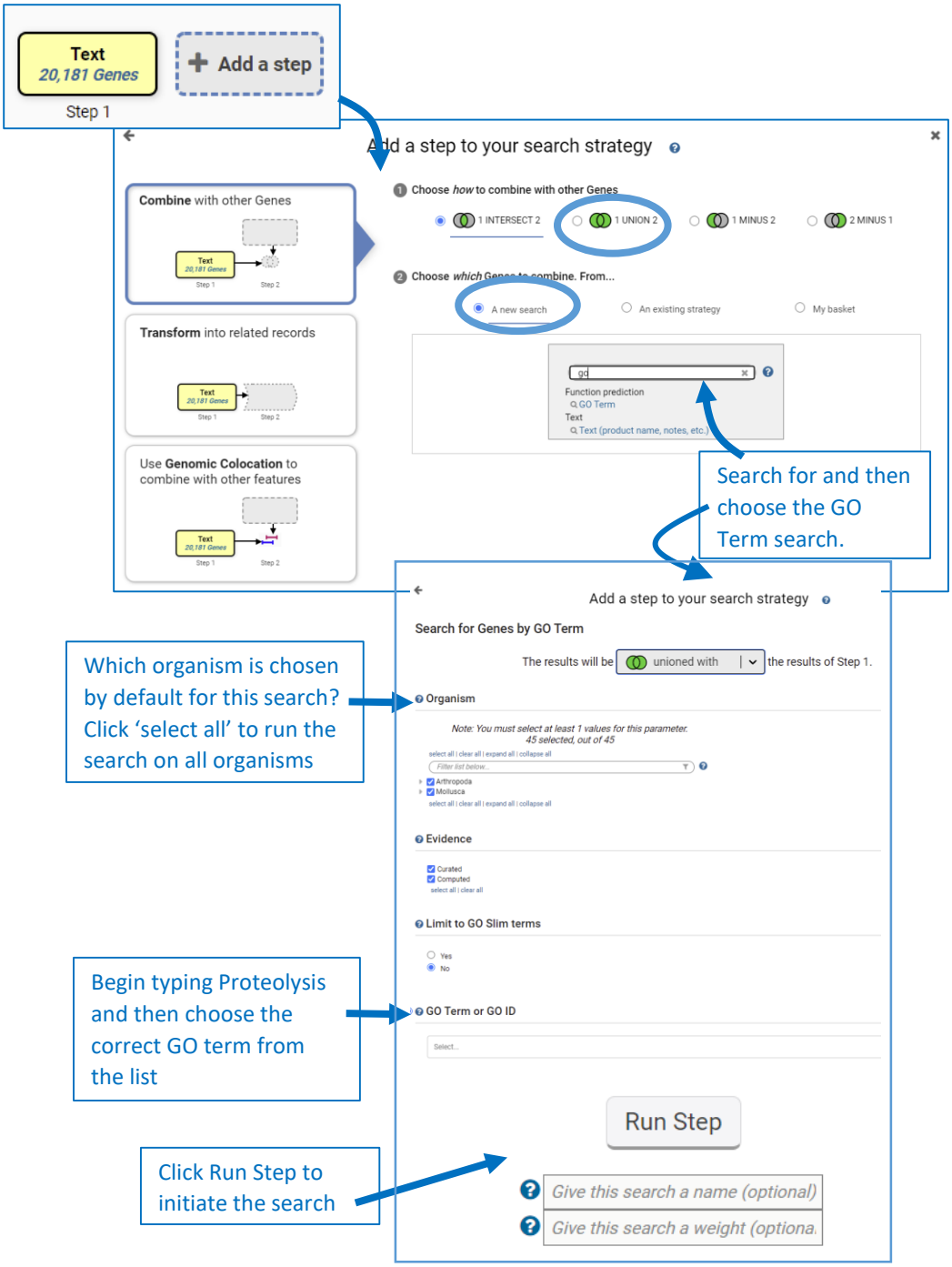
**Results and strategy:** You created a one-step strategy by running the text search. The strategy returns 20,181 genes that are annotated with the word 'protease'. This annotation could appear in any field that you searched. You can analyze this result by exploring the hits. Look at the data in the 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'.




**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 VectorBase 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 >Combine with other Genes >1 union 2 > A new search >GO Term

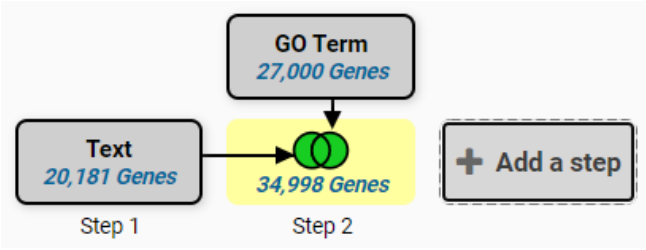



**Parameters:**

| | | |
|---|---|---|
| **Organism** | : | Choose All |
| **Evidence** | : | Default |
| **Limit to GO Slim Terms?** | | Default |
| **GO Term or GO ID** | : | GO:0006508 : proteolysis |
| **Free Text (use '*' for wildcard)** | : | N/A |

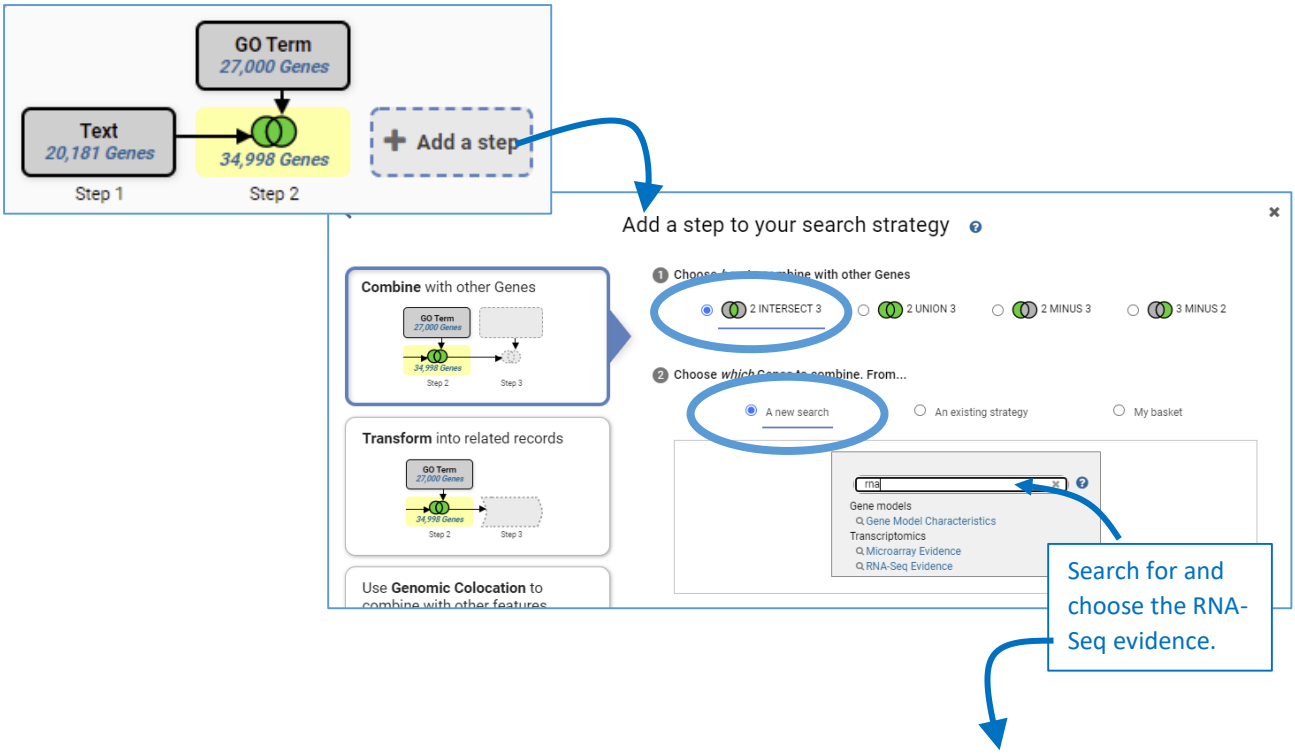
**Combine:**

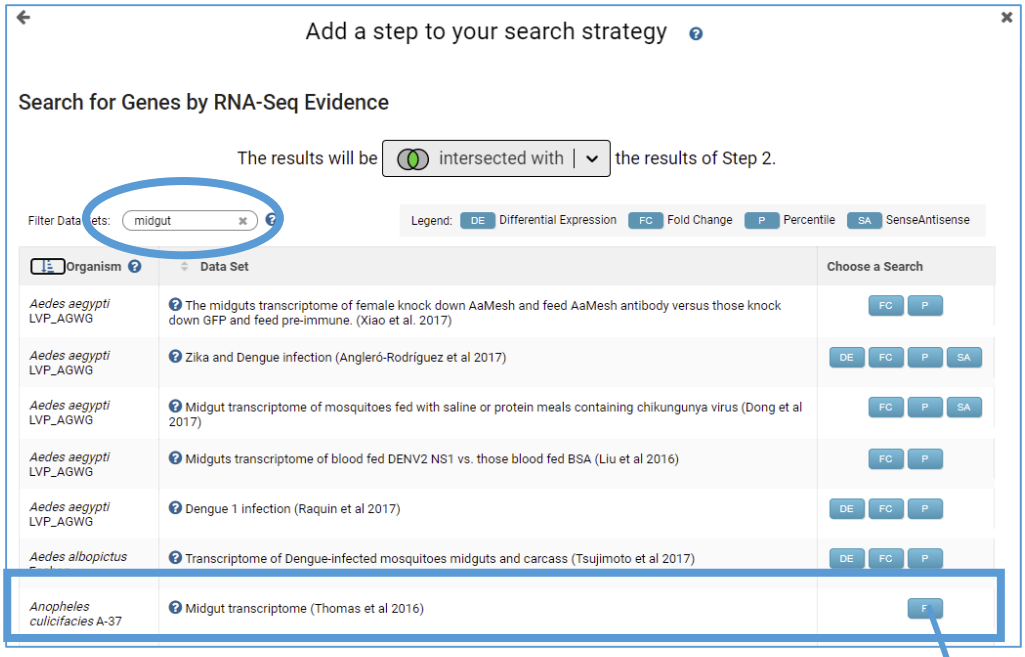

**Strategy Result:** The GO term search returned 20,000 genes annotated with the proteolysis GO term. The union of the text and GO search returns 34,998 genes that are suspected to have proteolytic activity.

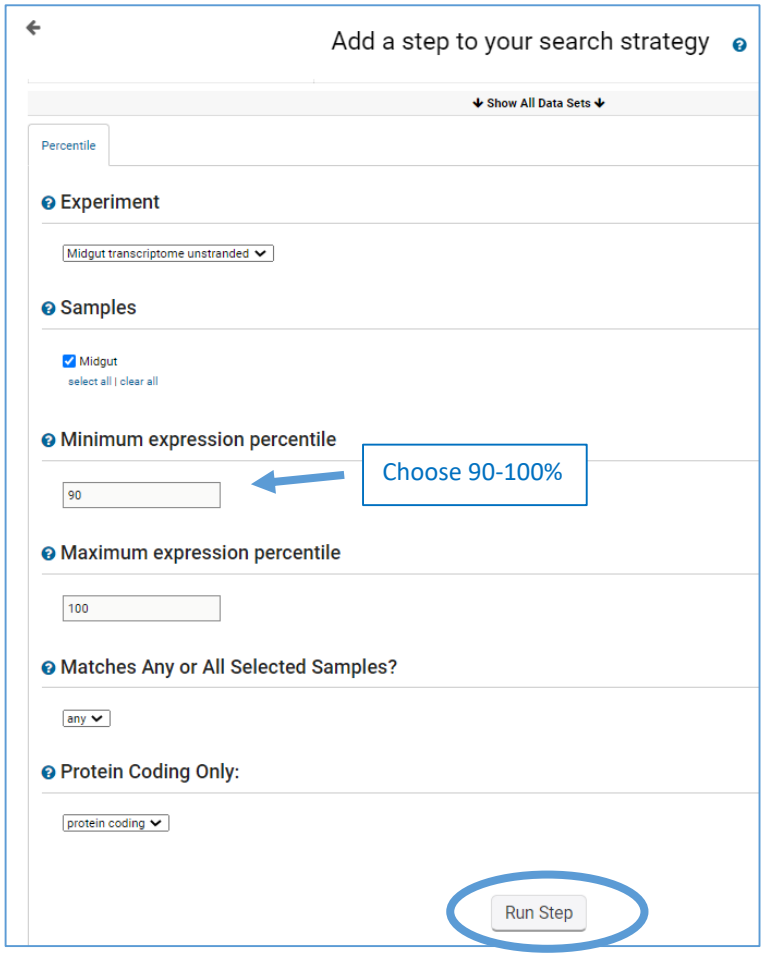



2. **Add a step choosing to run a search for genes based on Transcript Expression using RNA-Seq Evidence.** Since VectorBase has integrated many RNA sequencing data sets you must first choose what data set (experiment) to search before you are taken to the search form to choose parameters. Use the Filter Data set tool to choose the Percentile search (P) for 'Midgut transcriptome (Thomas et al 2016)'. This data set contains the RNA sequencing analysis of two midgut samples. Running the percentile search using the default parameters will return the genes whose expression levels are in the top 10% for those samples.

**Navigation**: Add Step >Combine with other Genes >2 intersect 3 >A new search >RNA Seq Evidence



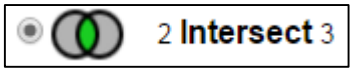

Add a step to your search strategy
Search for Genes by RNA-Seq Evidence
The results will be intersected with the results of Step 2.
Filter Data Sets: midgut
Legend: DE Differential Expression FC Fold Change P Percentile SA SenseAntisense
Organism
Data Set
Choose a Search
Aedes aegypti LVP_AGWG
The midguts transcriptome of female knock down AaMesh and feed AaMesh antibody versus those knock down GFP and feed pre-immune. (Xiao et al. 2017)
FC P
Aedes aegypti LVP_AGWG
Zika and Dengue infection (Angleró-Rodríguez et al 2017)
DE FC P SA
Aedes aegypti LVP_AGWG
Midgut transcriptome of mosquitoes fed with saline or protein meals containing chikungunya virus (Dong et al 2017)
FC P SA
Aedes aegypti LVP_AGWG
Midguts transcriptome of blood fed DENV2 NS1 vs. those blood fed BSA (Liu et al 2016)
FC P
Aedes aegypti LVP_AGWG
Dengue 1 infection (Raquin et al 2017)
DE FC P
Aedes albopictus
Transcriptome of Dengue-infected mosquitoes midguts and carcass (Tsujimoto et al 2017)
DE FC P
Anopheles culicifacies A-37
Midgut transcriptome (Thomas et al 2016)
P


Add a step to your search strategy
Show All Data Sets
Percentile
Experiment
Midgut transcriptome unstranded
Samples
Midgut
select all | clear all
Minimum expression percentile
90
Choose 90-100%
Maximum expression percentile
100
Matches Any or All Selected Samples?
any
Protein Coding Only:
protein coding
Run Step

**Parameters:**

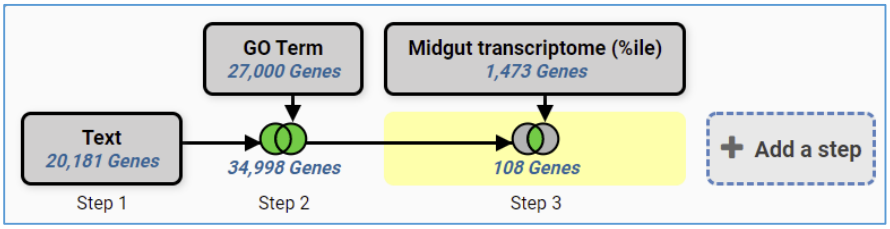
| | | |
|---|---|---|
| **Experiment** | : | Midgut Transcriptome |
| **Samples** | : | midgut |
| **Minimum expression percentile** | : | 90 |
| **Maximum expression percentile** | : | default |
| **Matches Any or All Selected Samples?** | : | default |
| **Protein Coding Only:** | : | default |

**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 108 *Anopheles culicifacies* A-37 genes that are suspected proteases with evidence for expression in the midgut based on RNA Sequencing data. Explore your gene list!!

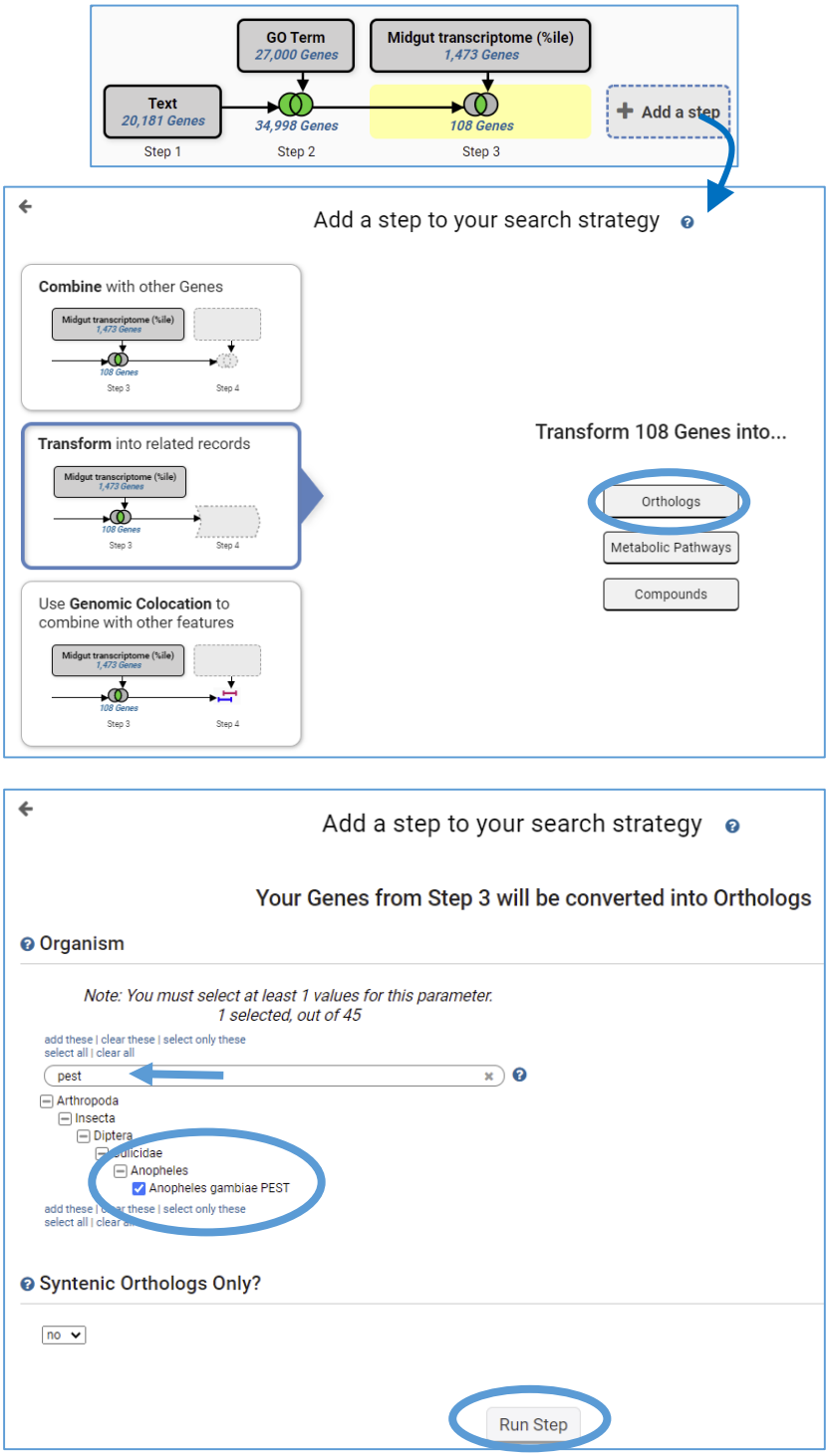
Notice that the list only contains *Anopheles culicifacies* A-37 genes now. This is because the RNA seq experiment was run in that organism and the data were aligned to the *Anopheles culicifacies* A-37 genome. When intersecting a list of genes from multiple organisms, with a list from only one organism, the only genes in common between the lists is the organism in the second list.




3. **Add a step to the strategy that transforms the 108 *Anopheles culicifacies* A-37 genes into *Anopheles gambiae PEST* genes.**

*Anopheles gambiae PEST* is a well-studied organism with large amounts of functional data. For example, VectorBase has 13 RNA sequencing and 26 microarray data sets integrated for *Anopheles gambiae PEST*, but only 1 RNA-Seq and no microarray for *Anopheles culicifacies*. A researcher interested in Anopheles *culicifacies* can take advantage of the *Anopheles gambiae* data by creating a strategy based on *Anopheles gambiae* data to retrieve genes with the biological properties they are interested in, and then transforming the results to their *Anopheles culicifacies* orthologs. In this case, there was the perfect experiment performed in *Anopheles culicifacies* but not in *A gambiae*, so we took advantage of the *A. culicifacies* data and transformed the results to *A. gamiae*. Because we use determine orthology for all organisms in VectorBase, you can transform between organisms such as mosquito and sand fly or tick if you so desire.

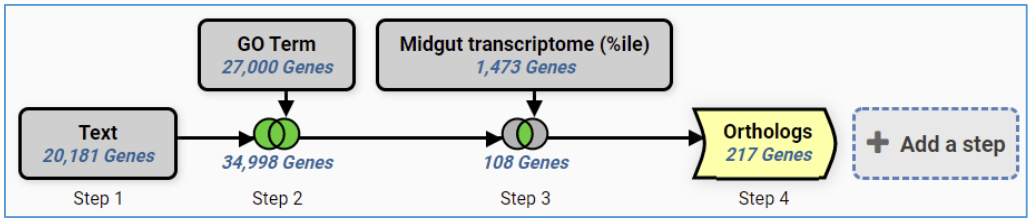
**Navigation:** >Add Step >Transform into related records >Orthologs




**Parameters:** Choose only *Anopheles gambiae PEST* in the Organism parameter of the Add Step Popup.

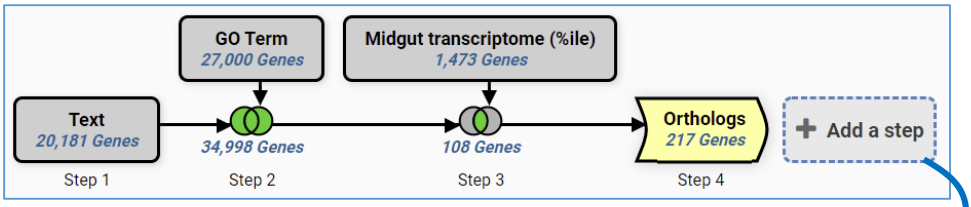
**Combine:** The ortholog transform function does not combine lists, but instead transforms the results into orthologs from a different species.

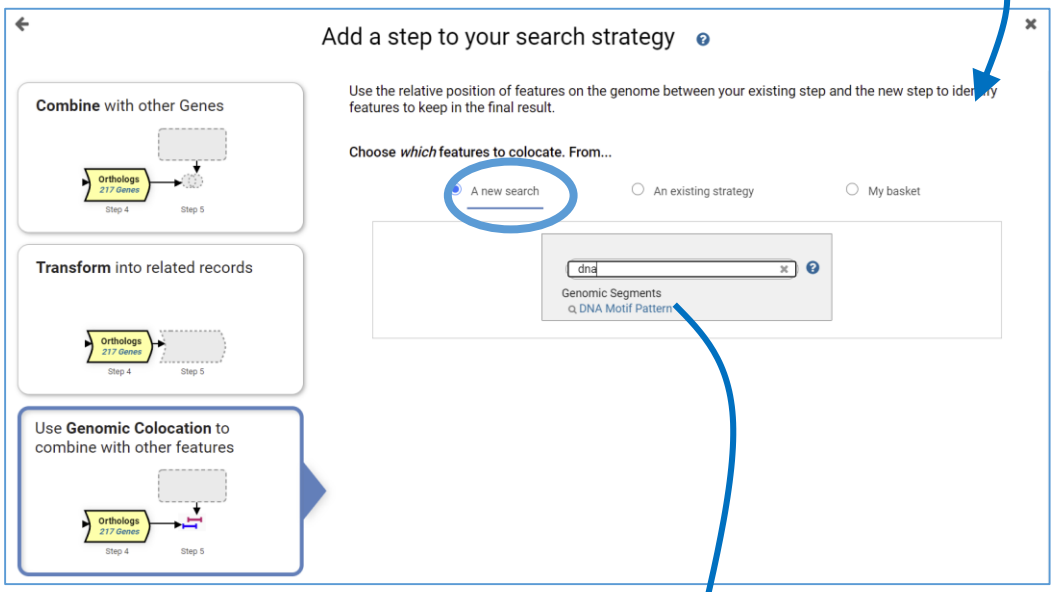
**Strategy Result:** We have a four-step strategy that returns 217 *Anopheles gambiae PEST* genes that are suspected proteases with evidence for expression in midgut based on RNA Sequencing data.

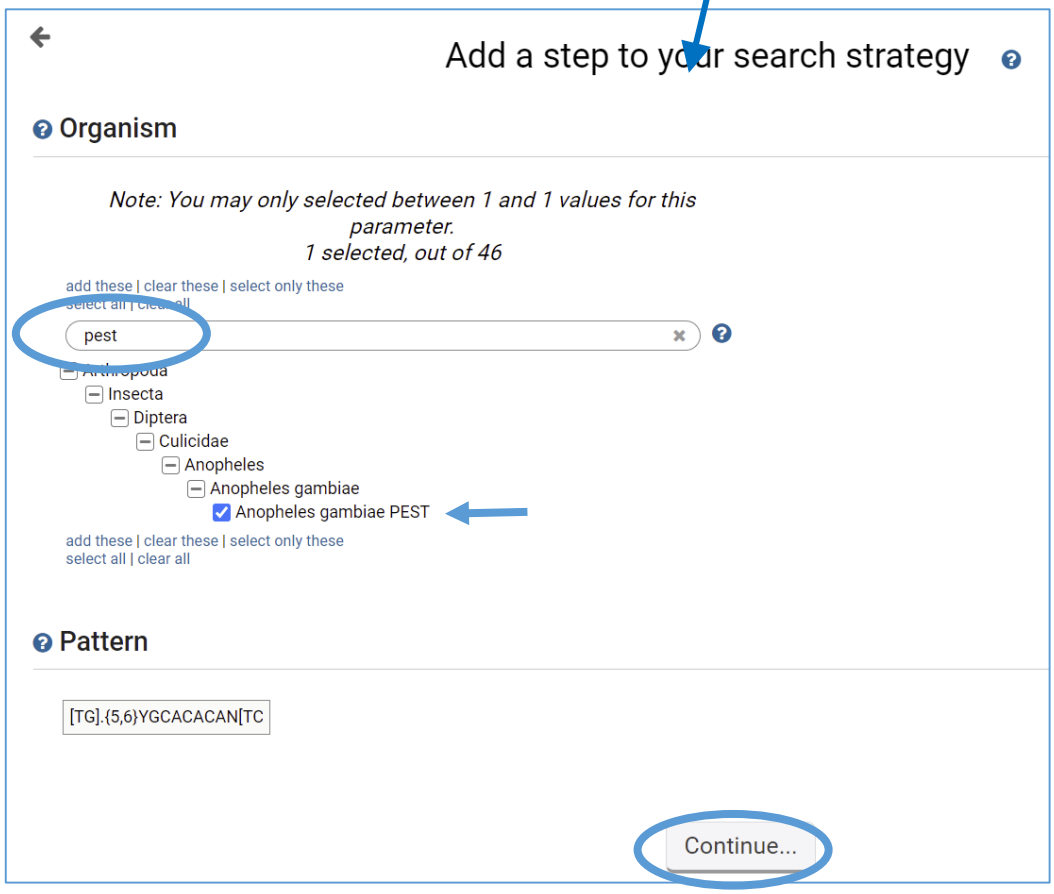



4. **Add a step to the strategy that returns transcription factor binding sites in *Anopheles gambiae PEST* and collocate those 'genomic segments' to the upstream 1000bp of the *Anopheles gambiae PEST* genes in step 4.** DNA motifs are nucleotide sequence patterns that occur often in the genome and can have a biological significance. Regulatory protein binding sites, for example, can occur upstream of the first exon and function provide a binding site for proteins that initiate transcription, such as a transcription factor. We will use the DNA Motif Pattern search to find the genomic locations of a particular transcription factor binding site. The search records the locations and creates 'genomic segment' records for each occurrence. Since these records are not genes, we cannot use the Boolean operations to combine the search results. However, since we do know the locations of the DNA motifs, we can use the collocation tool to identify which genes contain the DNA motifs in their upstream regions. The collocation tool is a powerful way to associate genes with other genomic features such as DNA motifs or single nucleotide polymorphisms. You will notice that initiating the DNA Motif Pattern 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 >DNA Motif Pattern


GO Term
27,000 Genes
Midgut transcriptome (%ile)
1,473 Genes
Text
20,181 Genes
34,998 Genes
108 Genes
Orthologs
217 Genes
Add a step
Step 1
Step 2
Step 3
Step 4

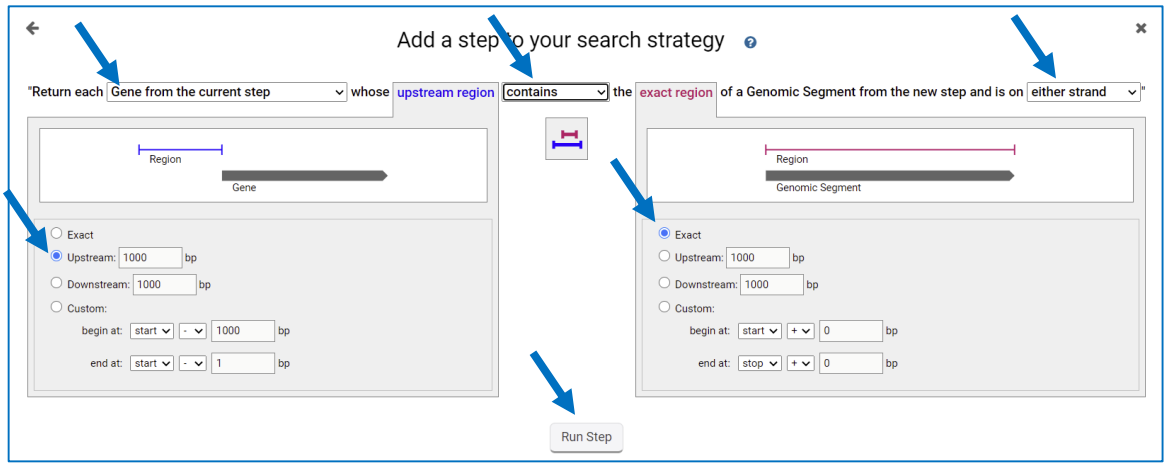

Add a step to your search strategy
Combine with other Genes
Transform into related records
Use Genomic Colocation to combine with other features
Use the relative position of features on the genome between your existing step and the new step to identify features to keep in the final result.
Choose which features to colocate. From...
A new search
An existing strategy
My basket
dna
Genomic Segments
DNA Motif Pattern


Add a step to your search strategy
Organism
Note: You may only selected between 1 and 1 values for this parameter.
1 selected, out of 46
add these | clear these | select only these
pest
Insecta
Diptera
Culicidae
Anopheles
Anopheles gambiae
Anopheles gambiae PEST
add these | clear these | select only these
select all | clear all
Pattern
[TG].{5,6}YGCACACAN[TC
Continue...

**Parameters:**

| | | |
|---|---|---|
| **Organism** | : | *Anopheles gambiae PEST* |
| **Pattern** | : | Default = [TG].{5,6}YGCACACAN[TCA]H |

**Collocation:** Because the DNA motif pattern search returns genomic segments and not genes, the only option for combining the two result lists is by relative genomic location. Arrange the statement in the Collocation popup to: **Return each Gene from the current step whose upstream 1000bp region contains the exact region of a genomic segment from the new step and is on either strand**. Remember to indicate that you want to locate the genomic segments in the upstream region of the gene.




**Strategy: Congratulations!** You have completed the strategy and have a list of 18 *Anopheles gambiae PEST* genes that are possible proteases, are likely expressed in the midgut and have upstream transcription factor binding sites.

This link will retrieve the completed strategy:
https://vectorbase.org/vectorbase/app/workspace/strategies/import/fef6c85026b3360d