

Strategies Tutorial

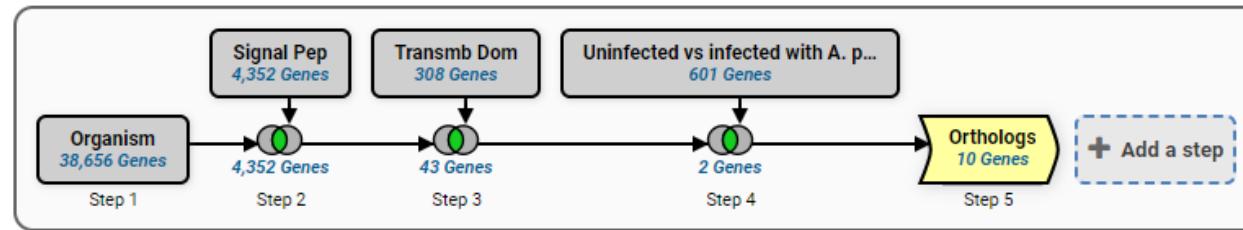
Note: This exercise uses **VectorBase.org** as an example, but the same functionality is available on all VEuPathDB resources.

Learning objectives

- Build a multistep strategy
- Use the organism genes, RNA-Seq, transmembrane domains, and signal peptide searches
- Combine search results using Boolean operators
- Transform genes of one genome into their orthologs in another species' genome
- 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

This is an example of one VectorBase search strategies: Find genes expressed in the tick *Ixodes scapularis*, when infected with the bacteria *Anaplasma phagocytophilum*, that are probably secreted GPCRs (G protein-coupled receptors). GPCRs are characterized by having 7TMDs (seven transmembrane domains). The genes will be transformed into their *Ix. persulcatus* orthologs. The ortholog transform enables you to make inferences about genes in *Ix. persulcatus*, an organism with less functional data, based on existing data in the closely related and better-studied *Ix. scapularis*. This strategy has five steps and will look like this:

AGS 2024 wks 



Strategies Overview

The strategy system offers approximately 100 searches. As shown in the example above, these 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 in the header of every page.

The screenshot shows the VectorBase website interface. At the top, there's a navigation bar with links for 'My Strategies', 'Searches' (which is highlighted with a pink oval), 'Tools', 'My Workspace', 'Data', and 'About'. Below the navigation is a search bar with placeholder text 'Site search, e.g. AGAP004730 or *reductase or "binding protein"'. The main content area has a sidebar on the left titled 'Search for...' containing a list of search categories: Genes, Organisms, Genomic Sequences, Genomic Segments, ESTs, Metabolic Pathways, and Compounds. To the right of the sidebar is an 'Overview' section with a 'Getting Started' guide. A dropdown menu titled 'Filter the searches below...' is open, showing expanded categories: Genes, Organisms, Genomic Sequences, Genomic Segments, ESTs, Metabolic Pathways, and Compounds.

Searches listed under **Genes** will return a list of **gene IDs**, while searches listed under **Metabolic Pathways** will return **metabolic pathways**, respectively.

The image contains three side-by-side screenshots of the VectorBase search interface. The left screenshot shows the 'Genes' section of the search sidebar, with the 'Genes' category highlighted by a pink box. The middle screenshot shows the expanded 'Genes' category, which includes 'Annotation, curation and identifiers', 'Function prediction', 'Gene models', 'Genomic Location', 'Immunology', 'Orthology and synteny', 'Pathways and interactions', 'Protein features and properties', 'Protein targeting and localization', 'Proteomics', 'Sequence analysis', 'Structure analysis', 'Taxonomy', 'Text', and 'Transcriptomics'. Two arrows point from the 'Function prediction' and 'Gene models' items to the right screenshot. The right screenshot shows the 'Metabolic Pathways' section of the search sidebar, with the 'Metabolic Pathways' category highlighted by a pink box. This section includes 'Compounds', 'Genes', 'Identifier (pathway, gene, compound, etc.)', and 'Pathway Name/ID'.

The five searches you will use in this tutorial are:

1. Identify Genes based on Organism – Find all genes from one or more species/organisms.
2. Identify genes based on predicted signal peptide - Signal P is used to identify signal peptides and their likely cleavage sites. A signal peptide is a short peptide present at the N-terminus of most newly synthesized proteins that are destined towards the secretory pathway.

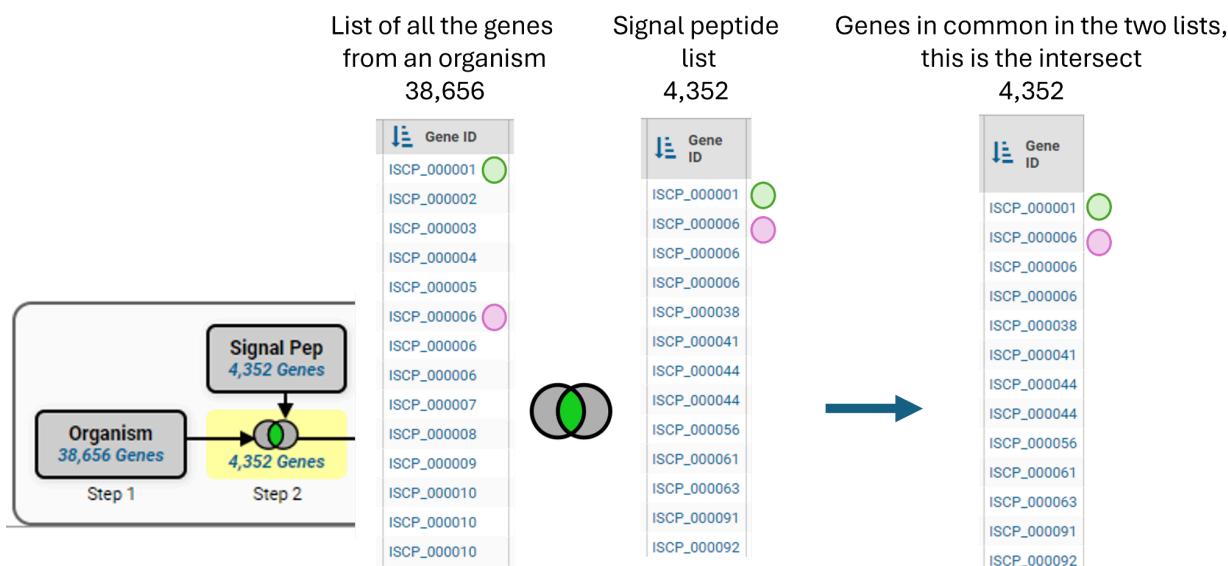
3. Identify genes based on predicted transmembrane domainis - TMHMM is used to predict transmembrane domain presence and topology from protein sequences.
 4. Identify Genes based on RNA Seq Evidence – VectorBase integrates raw RNA sequencing data from many different experiments and analyzes all data according to the same workflow to produce expression values. This search returns genes based on their transcript expression as measured by RNA sequencing. Pipeline details here
<https://vectorbase.org/vectorbase/app/static-content/methods.html#RNA-Sequence> .
 5. Transform by Orthology – VectorBase integrates ortholog profiles from OrthoMCL (<https://orthomcl.org/orthomcl/app>). The OrthoMCL algorithm clusters proteins into ortholog groups based on BLAST similarity across hundreds of genomes (<https://orthomcl.org/orthomcl/app/release-summary>), that span the tree of life. The transform converts a list of genes in one organism to their orthologs in a different organism.

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

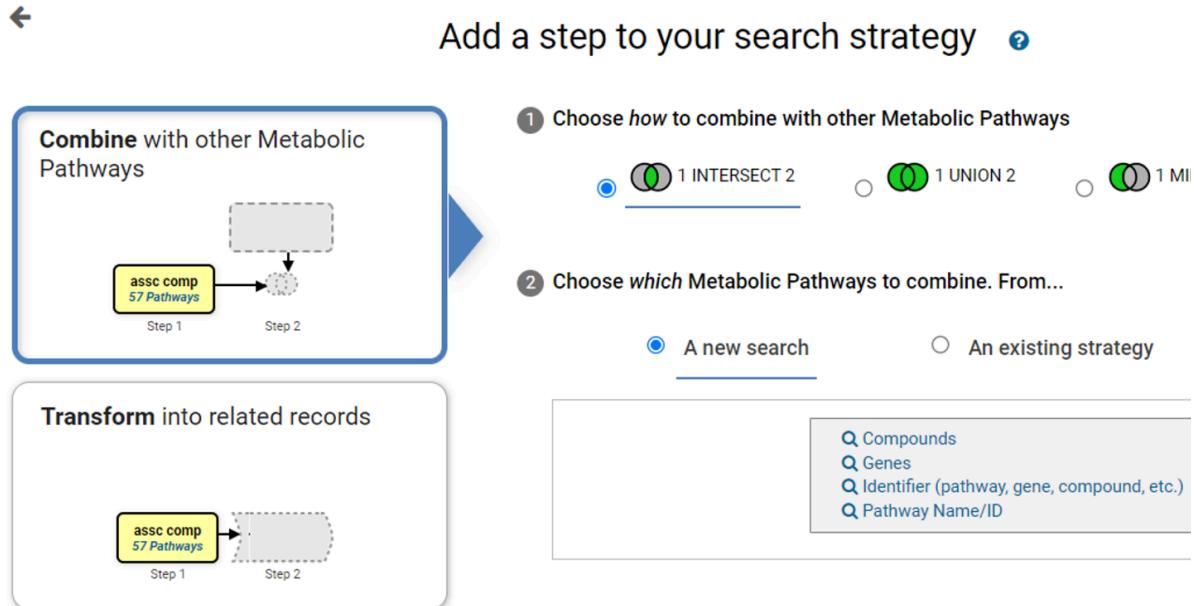
Each search returns a list of IDs. When two searches are combined, the two result sets (list of IDs) are merged. The table shows the 5 options for combining search results.

Operator	:	Combined Result will contain
<input checked="" type="radio"/> 	:	IDs in common between the two lists
<input type="radio"/> 	:	IDs from list 1 and list 2
<input type="radio"/> 	:	IDs unique to 1
<input type="radio"/> 	:	IDs unique to 2
<input type="radio"/> 	:	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 the image below:



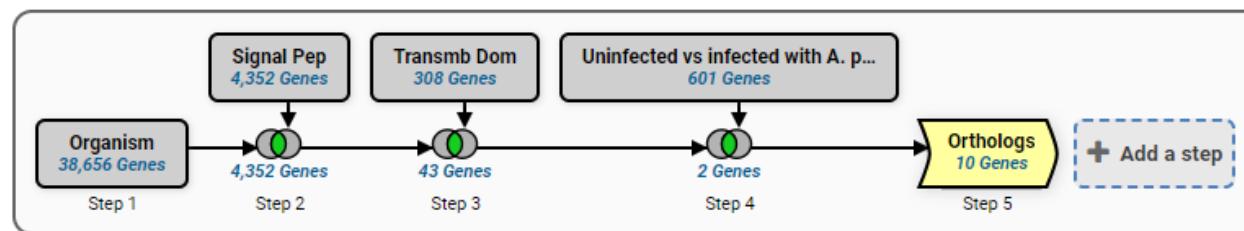
For this reason, the site interface will only show you compatible searches, which can return the same type of genomic features when combined with intersect, union, or minus operators. This is illustrated in the image below. Using a Metabolic Pathways step, the only other four options available are compounds, genes, identifiers, and pathway name/ID.



Build the Strategy

Find genes expressed in the tick *Ixodes scapularis*, when infected with the bacteria *Anaplasma phagocytophilum*, that are probably secreted and have seven transmembrane domains; convert them to orthologs in *Ix. persulcatus*. The strategy combines five steps (or Searches).

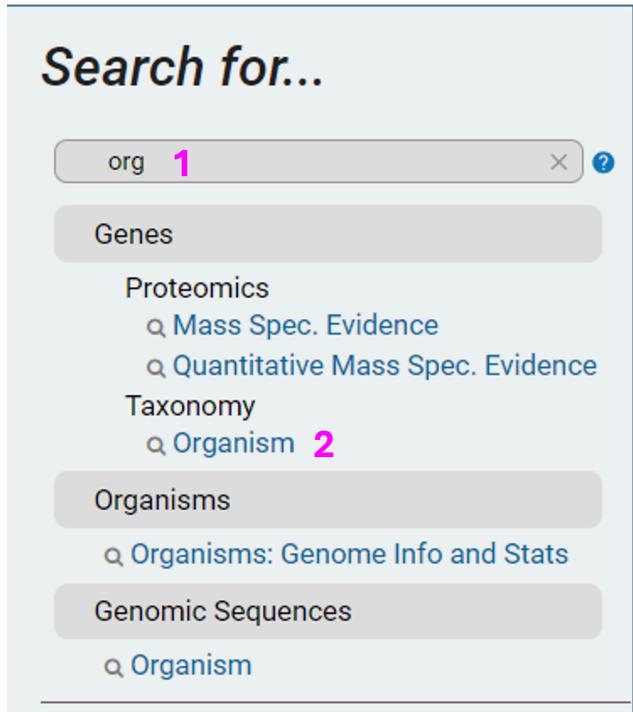
AGS 2024 wks



- Step 1: returns all *Ix. scapularis* PalLabHiFi genes
- Step 2: intersection with genes that have a signal peptide
- Step 3: intersection with genes that have 7 TMDs
- Step 4: intersection with the genes that have RNAseq differential expression (tick un- vs infected w/*Anaplasma*, up- or down-regulated, 2x fold change, p <= 0.1)
- Step 5: transforms the *Ix. scapularis* genes into *Ix. persulcatus* orthologs

Step-by-Step Instructions

- Run an organism search to obtain all the genes from *Ix. scapularis* PalLabHiFi



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. Reset the values if you have run a search before, and set the parameters as shown below.

Identify Genes based on Organism

Configure Search Learn More View Data Sets Used

[Reset values to default](#) **3**

[Organism](#)

1 selected, out of 81
select only these | add these | clear these

ixodes **4** [Reference only](#)

- └ Arthropoda
- └ Arachnida
- └ Ixodida
- └ Ixodes
 - └ Ixodes persulcatus Iper-2018 [Reference]
 - └ Ixodes scapularis
 - └ Ixodes scapularis ISE6
 - 5** Ixodes scapularis PalLabHiFi [Reference]
 - └ Ixodes scapularis Wikel

6 [Get Answer](#)

The search results are displayed in the My Strategies menu

The screenshot shows the VectorBase search interface. At the top, there is a navigation bar with links for 'My Strategies', 'Searches', 'Tools', 'My Workspace', 'Data', 'About', 'Help', and 'Contact Us'. A pink arrow points upwards from the search bar area towards the navigation bar.

The main search area is titled 'Unnamed Search Strategy *' with a pencil icon. It contains a 'Strategy box showing your one-step strategy' which includes a yellow box labeled 'Organism 38,656 Genes' and a blue dashed box labeled '+ Add a step'. A pink arrow points from the text 'You can add a step/search' to the '+ Add a step' button. Below this is a table with the header '38,656 Genes (10,724 ortholog groups)' and a 'Revise this search' link.

To the left, there is a 'Filter table showing the distribution of hits across the VectorBase genomes'. It includes an 'Organism Filter' section with checkboxes for 'Arthropoda' and 'Mollusca', both of which are selected. A pink arrow points from the text 'Filter table showing the distribution of hits across the VectorBase genomes' to the 'Arthropoda' checkbox.

The right side of the interface shows 'Gene Results' with tabs for 'Gene Results', 'Genome View', and 'Analyze Results'. It displays statistics: 'Genes: 38,656', 'Transcripts: 47,545', and two unchecked checkboxes for 'Show Only One Transcript Per Gene' and 'Show only the Genes in my basket'. A pink arrow points from the text 'You can add columns/attributes to the list of result genes' to the 'Add Columns' button. Below this is a table with columns: 'Gene ID', 'Transcript ID', 'Previous ID(s)', and 'Genomic Sequence ID'. The first row of the table is highlighted with a red box, and a pink arrow points from the text 'Link to gene page' to the 'Gene ID' column of this row. The table also includes a 'Download' button and a 'Send to...' dropdown.

Result: You created a one-step strategy by running the organism search. The strategy returns 38,656 genes. You can analyze this result by exploring the hits. Look at the data in the columns of the result table. Explore your results.

2. Add a step to the strategy that intersects the 38,656 genes with genes with signal peptides.

Add a step and follow the numbers shown below



Add a step to your search strategy [?](#)

1 Combine with other Genes

2 Choose how to combine with other Genes

1 INTERSECT 2 1 UNION 2 1 MINUS 2

3 Choose which Genes to combine. From...

A new search An existing strategy

4 signal

Protein targeting and localization
Q Predicted Signal Peptide **5**
Q Transmembrane Domain Count

6 ixodes

select only these | add these | clear these

- Arthropoda
- Arachnida
- Ixodida
- 7** Ixodes
 - Ixodes persulcatus Iper-2018 [Reference]
 - Ixodes scapularis
 - Ixodes scapularis ISE6
 - Ixodes scapularis PalLabHiFi [Reference] **8**
 - Ixodes scapularis Wikel

8 Run Step

3. Add a step to the strategy that intersects the 4,352 genes with genes with 7 TMDs.

Add a step and follow the numbers shown below

Add a step to your search strategy [?](#)

1 Combine with other Genes

2 Choose how to combine with other Genes

2 INTERSECT 3 2 UNION 3 2 MINUS 3

3 Choose which Genes to combine. From...

A new search An existing strategy

4 trans [4](#) [?](#)

Protein targeting and localization
 Predicted Signal Peptide
 Transmembrane Domain Count **5**
Transcriptomics
 Microarray Evidence
 RNA-Seq Evidence

5 selected, out of 81
[select only these](#) | [add these](#) | [clear these](#)

6 ixodes [6](#) [?](#) Reference only

- Arthropoda
- Arachnida
- Ixodida
- Ixodes
 - Ixodes persulcatus Iper-2018 [Reference]
 - Ixodes scapularis
 - Ixodes scapularis ISE6
 - Ixodes scapularis PaillabHiFi [Reference] **7**
 - Ixodes scapularis Wikle

8 Minimum Number of Transmembrane Domains [?](#)

7 **8**

9 Maximum Number of Transmembrane Domains [?](#)

7 **9**

10 Run Step

4. Add a step choosing to run a search for genes based on Transcript Expression using RNA-Seq Evidence. Since VectorBase has integrated more than a couple hundred RNA sequencing data sets (231 total in the last release, 7 May 2024), 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 “*I. scapularis* ISE6 cells uninfected (control) vs infected with *Anaplasma phagocytophilum* human NY18 isolate (aligned to Wikle strain) (Villar et al 2015)”. This data set

contains the RNA sequencing analysis of a tick cell line infected with the bacteria *Anaplasma phagocytophilum* human NY18 isolate.

Add a step and follow the numbers shown below

Add a step to your search strategy

1 Combine with other Genes

2 Choose how to combine with other Genes

3 INTERSECT 4 3 UNION 4 3 MINUS 4

3 Choose which Genes to combine. From...

A new search An existing strategy

4 RNA

5 RNA 4

Gene models
Gene Model Characteristics
Unannotated Intron Junctions
Transcriptomics
Microarray Evidence
RNA-Seq Evidence

Search for Genes by RNA-Seq Evidence

6

The results will be intersected with | the results of Step 3.

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

Filter Data Sets: Anap **7** 1 result (filtered from a total of 231)

Organism	Data Set	Choose a Search
<i>Ixodes scapularis</i> PalLabHiFi	Uninfected (control) vs infected with <i>Anaplasma phagocytophilum</i> human NY18 isolate (Villar et al 2015)	<input checked="" type="checkbox"/> DE <input type="checkbox"/> FC <input type="checkbox"/> P <input type="checkbox"/> SA

8

RNA-seq parameters > Run Step

(SRP058346 Integrated metabolomics transcriptomics and proteomics identifies metabolic pathways in *Leishmania* and *Leishmania*-infected macrophages. [View details](#))

iscaPalLabHiFi - Antisense

(Reference Sample

- Uninfected ISE6 cells
 Infected ISE6 cells with *A. phagocytophilum*

(Comparator Sample

- Uninfected ISE6 cells
 Infected ISE6 cells with *A. phagocytophilum*

(Direction**(fold difference >=****(adjusted P value less than or equal to****5. Add a step to transform your list of genes into orthologs.**

Add a step and follow the numbers shown below

Combine with other Genes

Step 4 Step 5

1 Transform into related records

Step 4 Step 5

Use Genomic Colocation to

Transform 2 Genes into...

2 Orthologs
Metabolic Pathways
Compounds

1 selected, out of 81
select only these | add these | clear these

ixodes 3 Reference only

- Arthropoda
- Arachnida
- Ixodida
- Ixodes
 - Ixodes persulcatus Iper-2018 [Reference]
 - Ixodes scapularis
 - Ixodes scapularis ISE6
 - Ixodes scapularis PallLabHiFi [Reference]
 - Ixodes scapularis Wikel

4 Syntenic Orthologs Only?

no

5

Congratulations, you have completed the strategy!

Here is a link to the final strategy

<https://vectorbase.org/vectorbase/app/workspace/strategies/import/af0bf3cf6a1e949f>