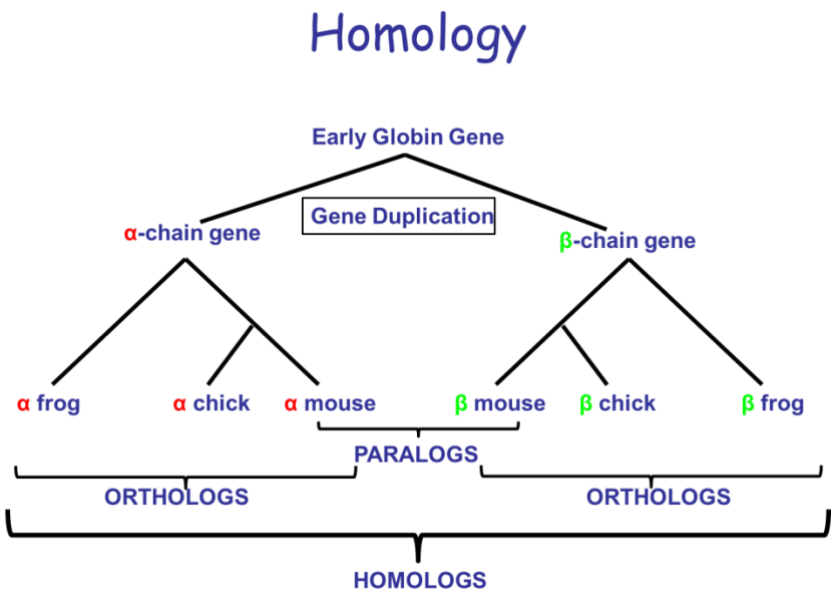
Homology gene relationships via OrthoMCL DB



Learning objectives:

- Explore the orthology table on VEuPathDB gene pages
- Getting to OrthoMCL from VEuPathDB gene pages
- Run searches in OrthoMCL
- Explore the cluster graphs in OrthoMCL
- Leverage the phyletic pattern search
- Leverage the orthology transform tool

OrthoMCL is a genome-scale algorithm for grouping homologous protein sequences. Such homologous sequences share evolutionary history and might also share function. Thus, homology predictions are important in predicting the function of newly identified genes. Indeed, detection of homologs has become more widespread with the rapid progress in genome sequencing and the discovery of protein sequences. Importantly, proteins in OrthoMCL groups have been shown to display a high degree of functional conservation (e.g., a group's proteins have consistent EC numbers), highlighting that OrthoMCL is useful for functional annotation of newly sequenced genomes.

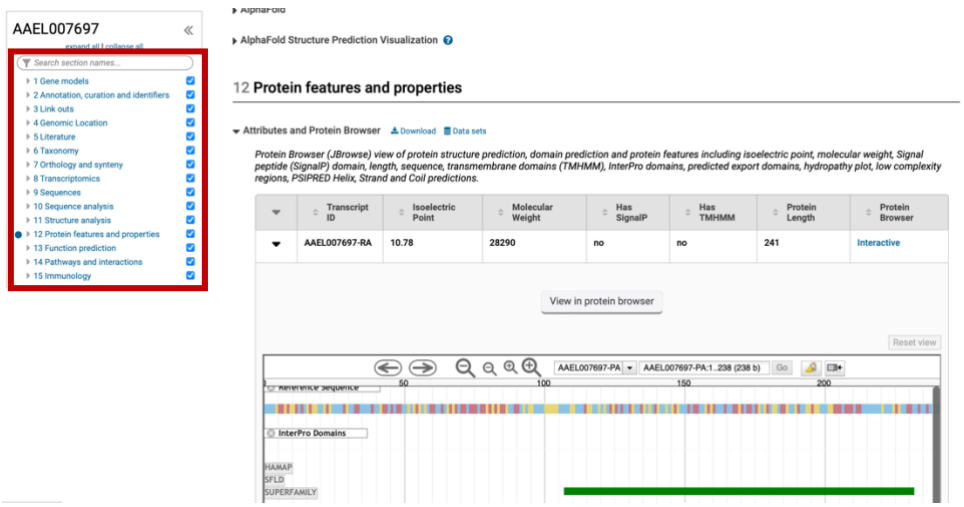
OrthoMCL not only identifies groups shared by proteins from two or more species, but also groups representing species-specific gene expansion families. To achieve this, the OrthoMCL algorithm starts with reciprocal best BLAST hits within each proteome as potential in-paralog/recent paralog pairs and reciprocal best hits across any two proteomes as potential ortholog pairs. Related proteins are interlinked in a similarity graph. Then, MCL (Markov Clustering algorithm; www.micans.org/mcl) is invoked to split mega-clusters. This process is analogous to the manual review in COG construction. MCL clustering is based on weights between each pair of proteins. Thus, to account for differences in evolutionary distance between any two organisms, the weights are normalized before running MCL.

The organism specific orthology information garnered from our OrthoMCL analysis of VEuPathDB organisms is presented on gene pages and integrated into an Orthology Phylogenetic Profile search. The OrthoMCL.org site offers a deep look into all data associated with the OrthoMCL results for orthology groups and proteins.

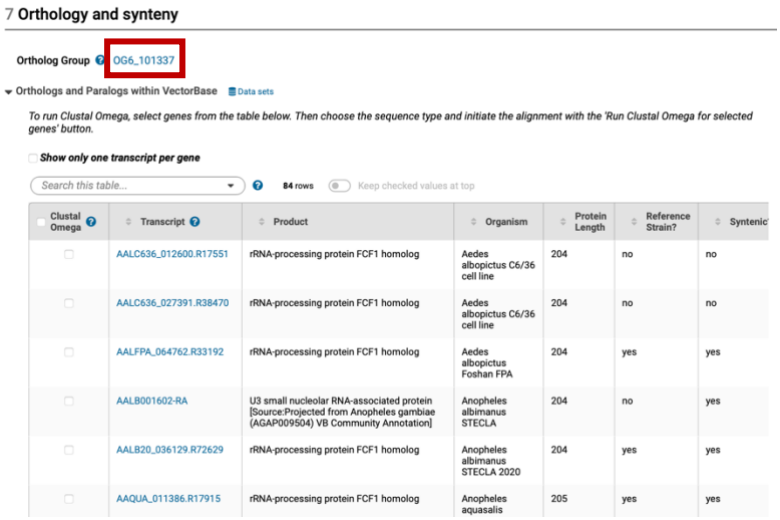
1. **Getting to OrthoMCL from VEuPathDB databases**
   Note: For this exercise use https://vectorbase.org and http://orthomcl.org

a. Use the VectorBase Site Search to visit AAEL007697 gene page for *Aedes aegypti* LVP_AGWG.

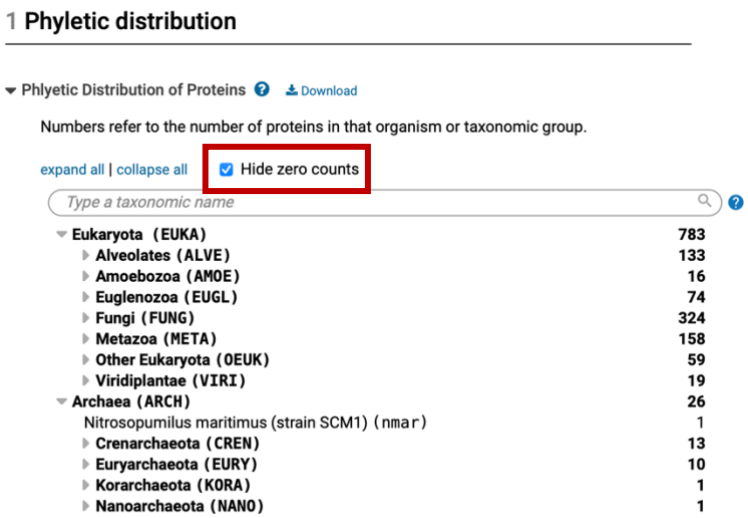
b. What information on the gene page can you use to guess a function for this gene? It is annotated as an unspecified product! Hint: look at the orthologs table and the domains in the protein features graph. You can use the navigation panel on the left to get to different gene page sections.



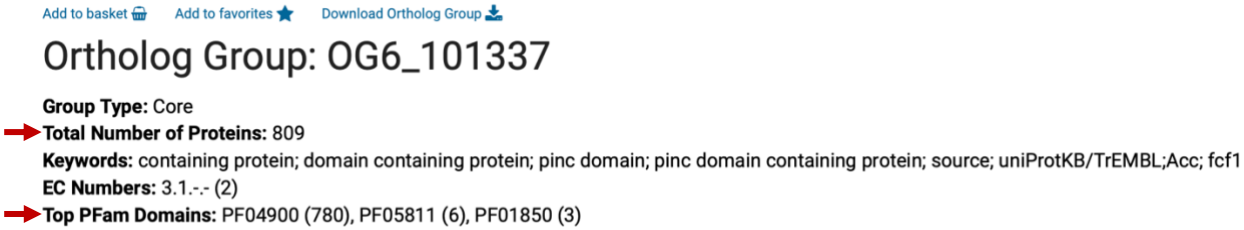
c. Go to the Orthology and Synteny section and look at the table labeled "Orthologs and Paralogs within VectorBase". Does this gene have orthologs in other mosquitoes? What about other organisms? (hint: scan the organism column in the table)



d. What about orthologs in organisms not in VEuPathDB? (hint: click on the Ortholog Group link above the table to examine the orthology information for the group at OrthoMCL.org). Does it have any orthologs in bacteria or archaea? (hint: click on Hide zero counts).



e. Scroll down to the PFam domains section. Domain architectures are found under the PFam Architecture of Each Protein table and are described in the PFam Legend table. Do all the proteins in this group have similar domain architecture? What is the distribution of the PF04900 domain across the 809 proteins in this ortholog group? PF00149? (see summary at the top of the page)



f. Based on the orthologs and the PFam domains shared by the group, what do you think this protein might be doing? If you had to give this gene a name, what would you call it?

PFam Legend Download

Search this table... 7 rows

Accession	Symbol	Description	Count	Legend
PF04900	Fcf1	Fcf1	780	
PF05811	DUF842	Eukaryotic protein of unknown function (DUF842)	6	
PF01850	PIN	PIN domain	3	
PF00160	Pro_isomerase	Cyclophilin type peptidyl-prolyl cis-trans isomerase/CLD	1	
PF13638	PIN_4	PIN domain	1	
PF00227	Proteasome	Proteasome subunit	1	
PF00149	Metallophos	Calcineurin-like phosphoesterase	1	

PFam Architecture of Each Protein Download

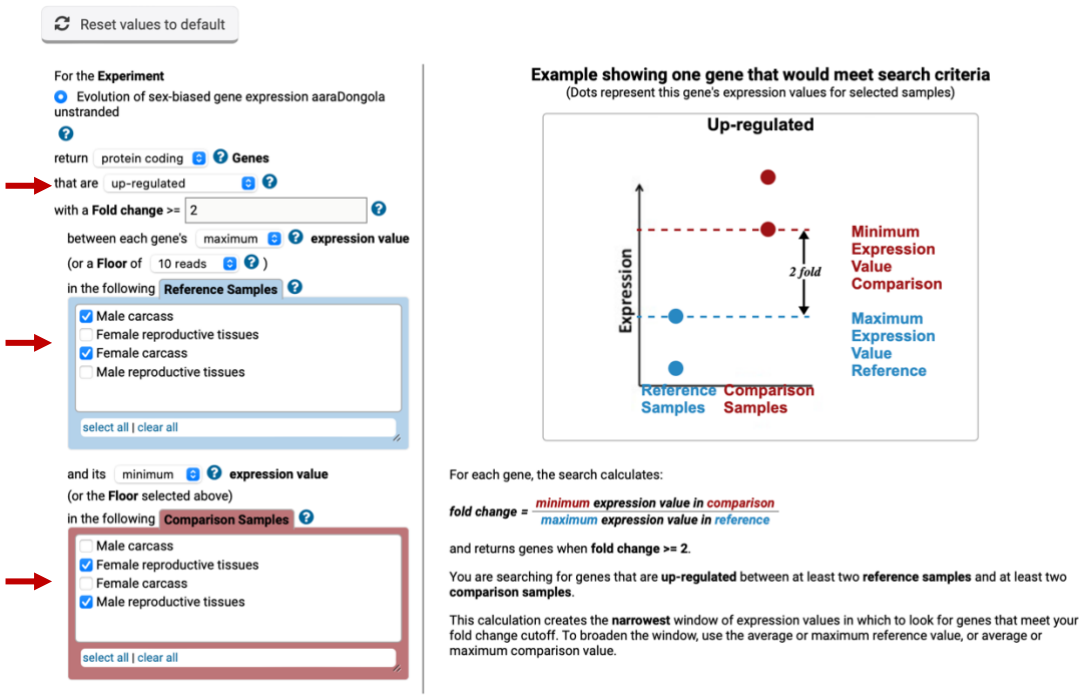
Search this table... 809 rows

Accession	Taxon	Core/Peripheral	Protein Length	
aacu|ASPACDRAFT_77294	Aspergillus aculeatus ATCC 16872	Peripheral	189	
aaeg-old|AAEL007697	Aedes aegypti LVP_AGWG (old build 2019-12-20)	Core	241	
aaeg|AAEL007697	Aedes aegypti LVP_AGWG	Peripheral	241	
aalb|AALFPA_064762	Aedes albopictus Foshan FPA	Peripheral	204	
aalc|AALC636_012600	Aedes albopictus C6/36 cell line	Peripheral	204	
aalc|AALC636_027391	Aedes albopictus C6/36 cell line	Peripheral	204	

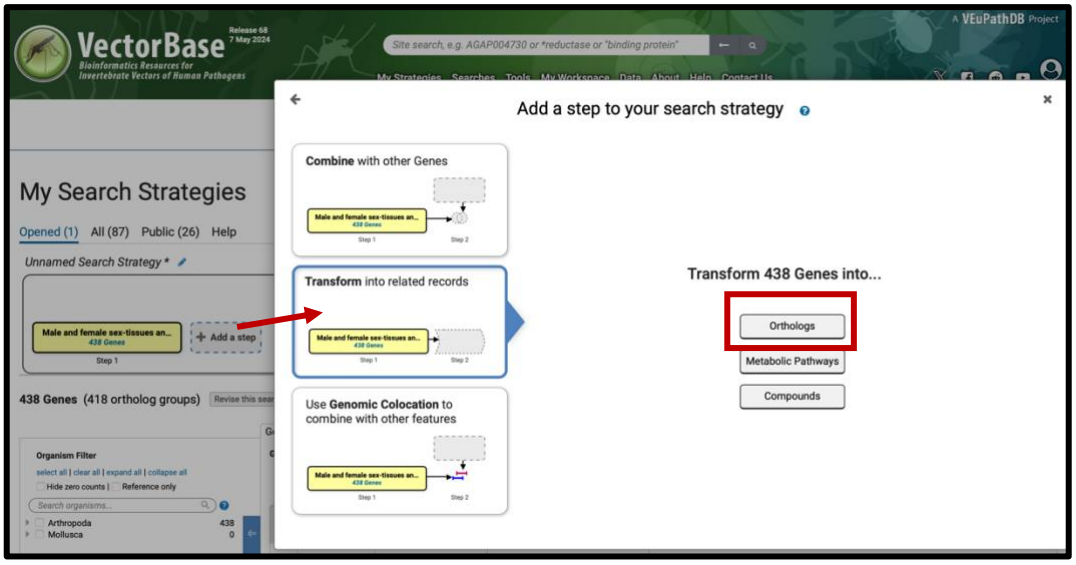
2. Using the orthology transform and phyletic pattern search in VectorBase.

The goal of this exercise is to identify all vector genes that are upregulated in male and female reproductive tissue identified from an experiment performed in *A. arabiensis* that are conserved in nematodes but absent in mammals.

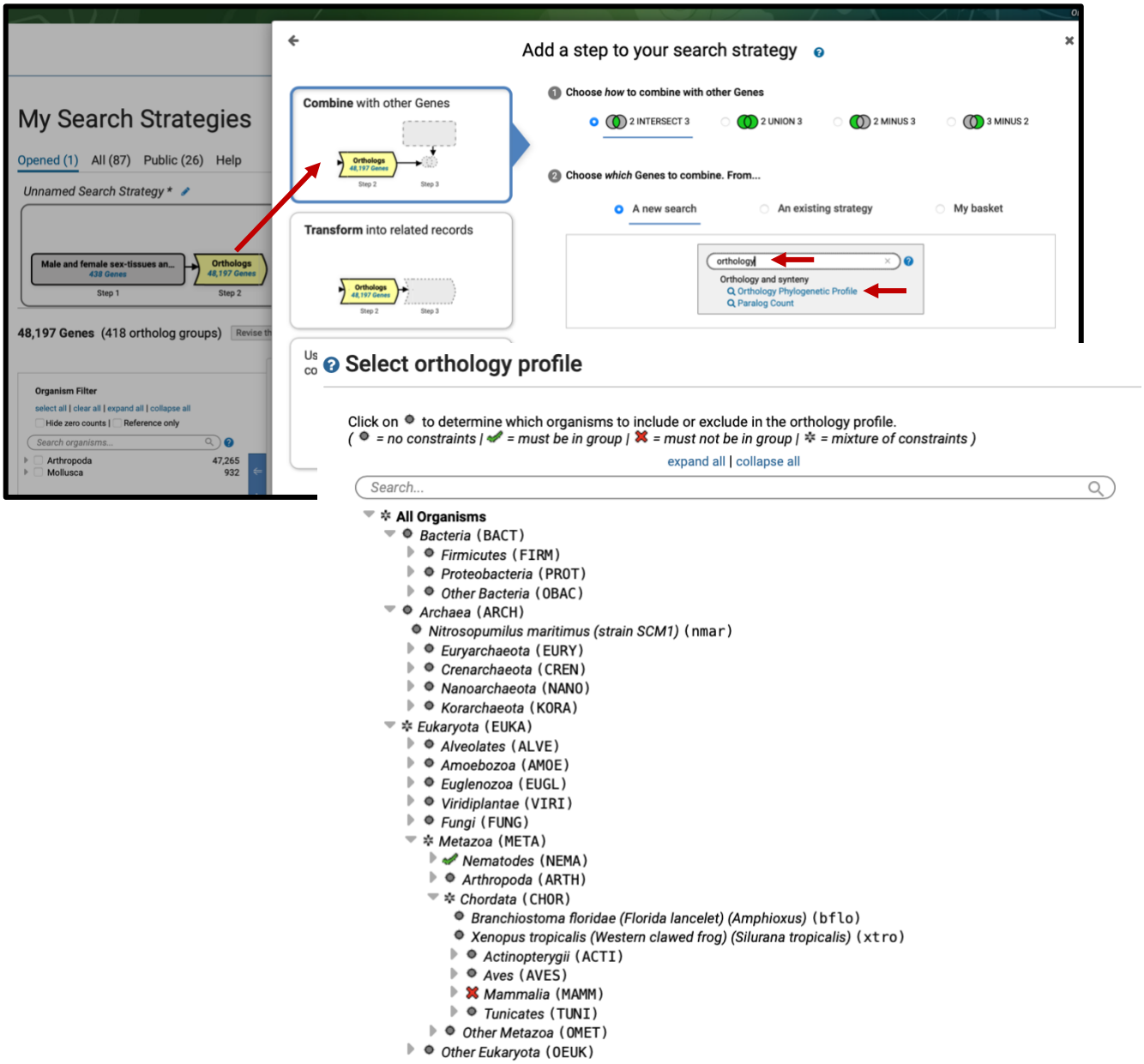
a. Start by identifying all genes that are upregulated by 2-fold in male and female reproductive tissue compared to the rest of the insect carcass. To do this find the RNAseq experiment "A. arabiensis DONGOLA 2021 Evolution of sex-biased gene expression aaraDongola" and run a fold change search.

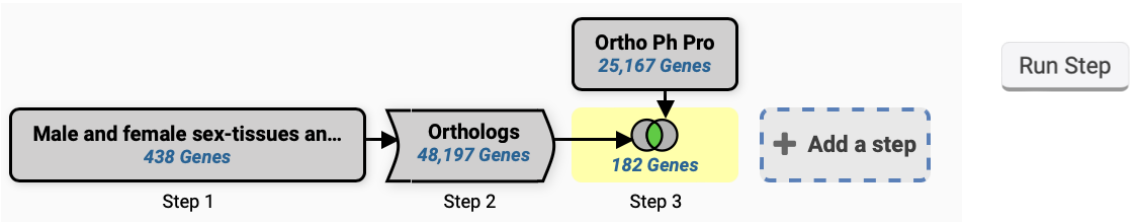


b. Once you find these genes, transform the results to all organisms in vectorbase.



c. Next add a step and use the orthology phylogenetic profile search to identify all vector genes that are conserved in nematodes but absent from mammals. Remember you can filter the searches by typing a word like “orthology”.

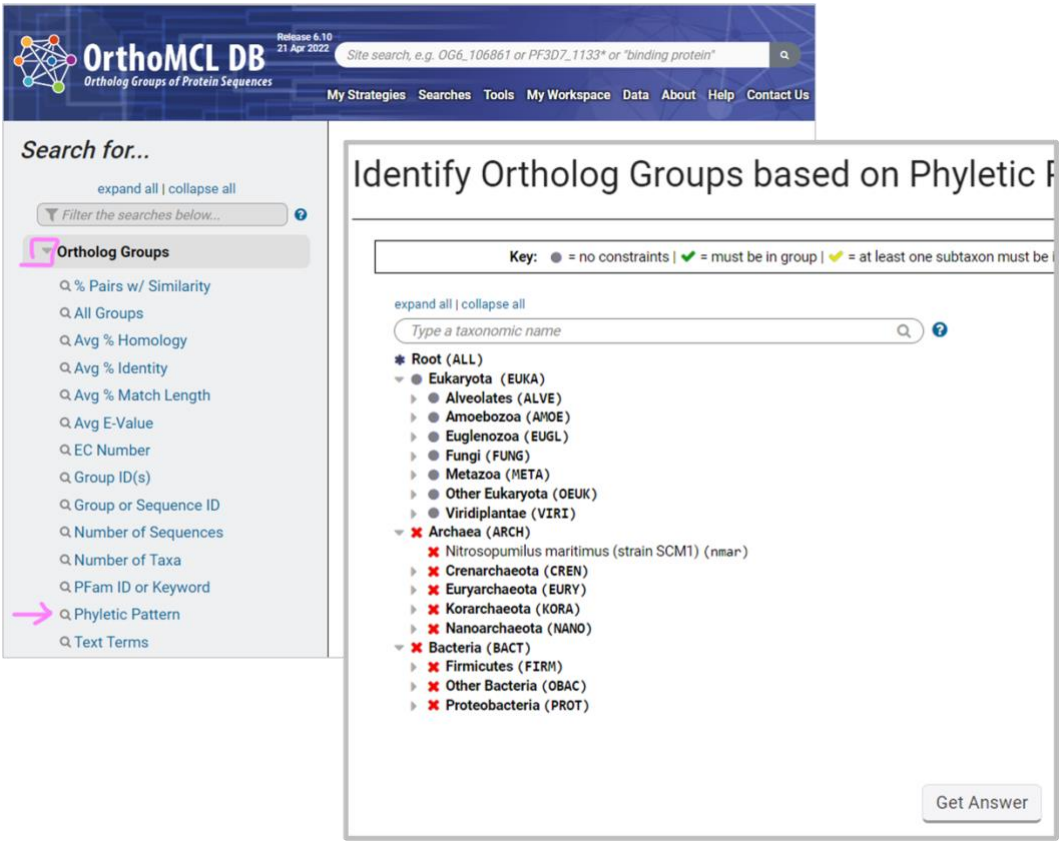


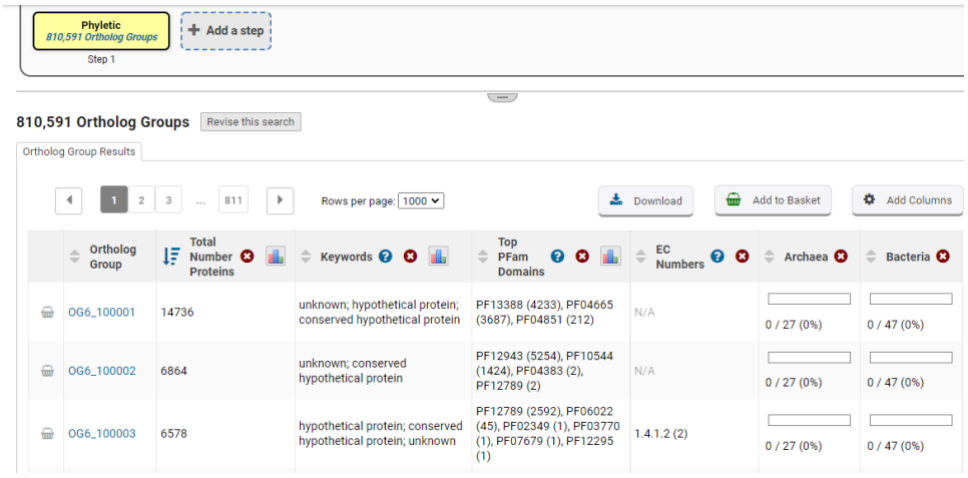


*********Below are optional exercises*********

3. Using the phyletic pattern tool in OrthoMCL
Note: For this exercise use http://orthomcl.org/

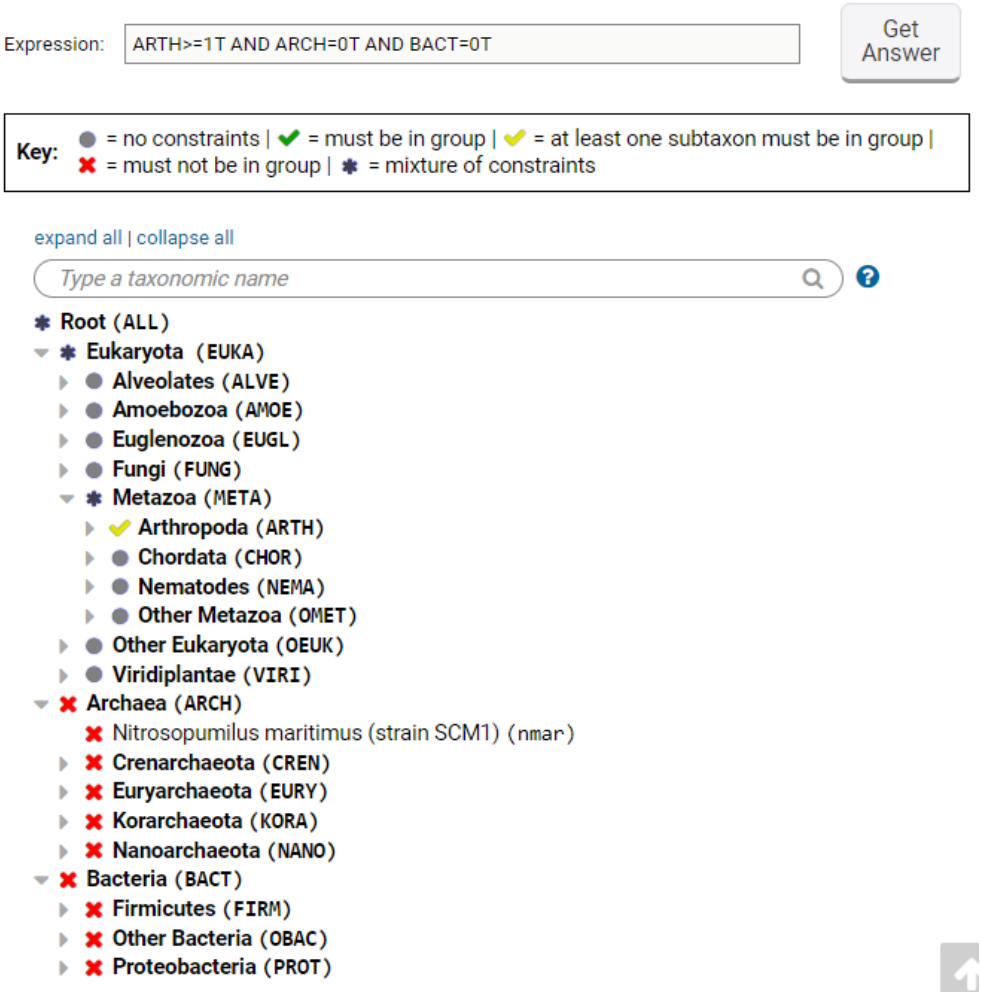
a. How many orthology groups OrthoMCL do not have any orthologs in bacteria or archaea? How many protein groups do not contain orthologs from bacteria and archaea?





b. Find all groups that contain orthologs from at least one species of Arthropoda but not from bacteria or archaea.

Before looking at the answer below, try this on your own or with your neighbor classmate.

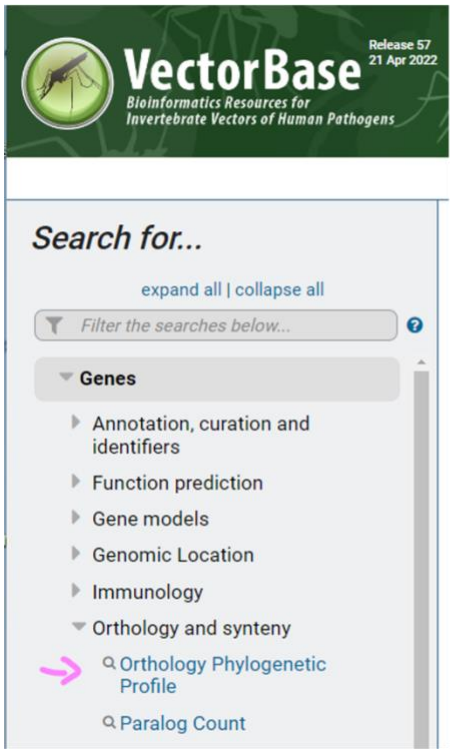


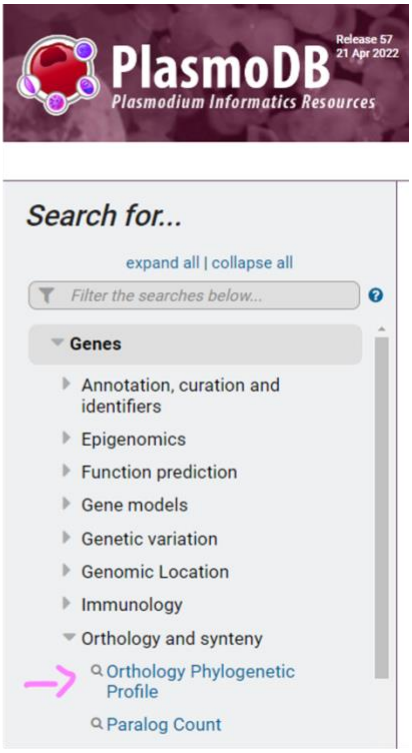
Before you click on Get Answer, scroll down to the bottom of the page to find additional information about expression parameters.

The expression parameter of the query we just did is:

Expression ARTH>=1T AND ARCH=0T AND BACT=0T

If just with clicks you cannot construct a query, try typing an expression.





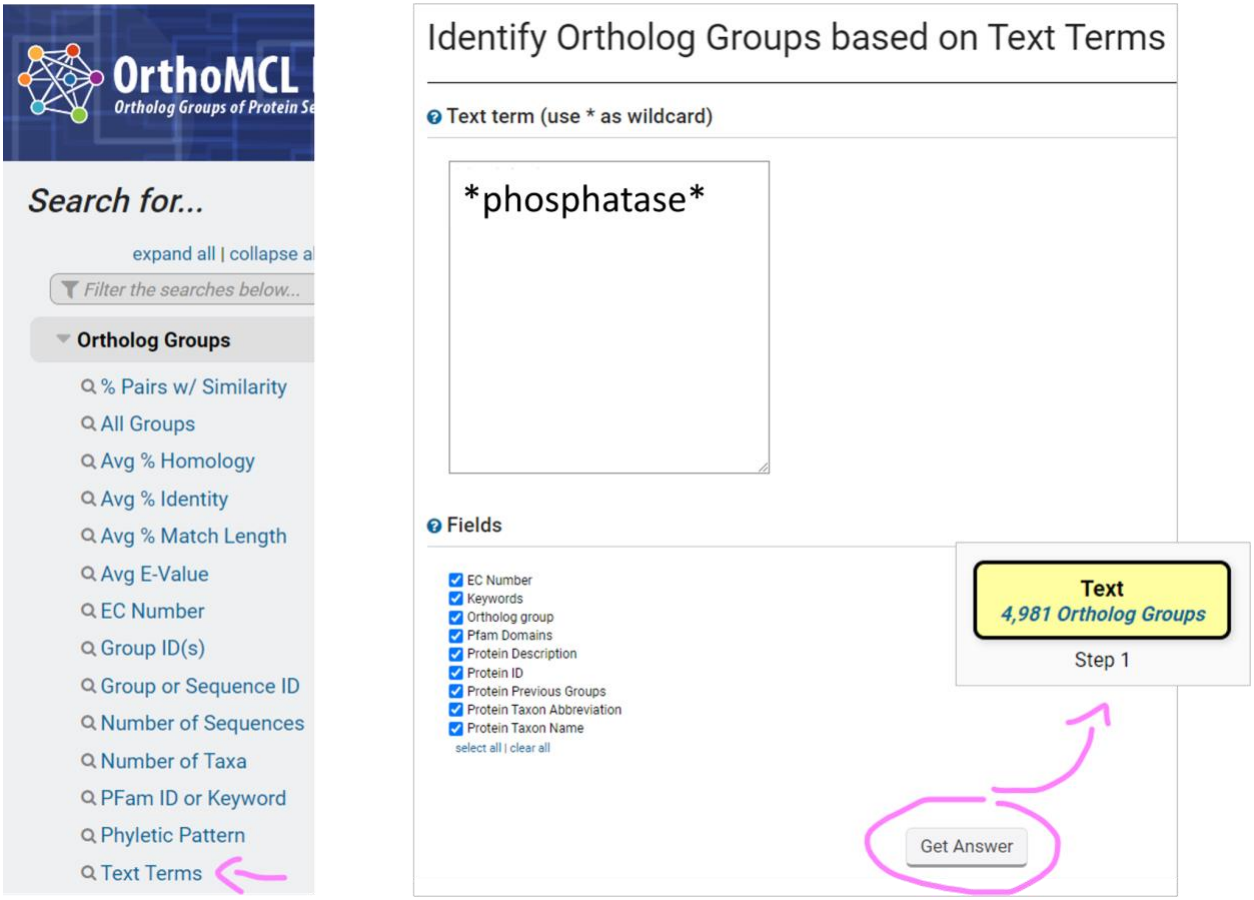
c. All VEuPathDB sites also have a phyletic pattern search that uses OrthoMCL data under Genes -> Orthology and synteny -> Orthology Phylogenetic Profile.

This search is very useful to identify genes in your organism of interest that are restricted in their profile. For example, if you are working with a parasite species, you might want to identify genes that are conserved among organisms in your genus (e.g., *Plasmodium*) but not present in the host (e.g., human) as these genes may make good drug targets or vaccine candidates. Optional: go to your favorite VEuPathDB site and run this search to identify all genes that are not present in human or mouse.

4. Combining searches in OrthoMCL (Use http://orthomcl.org for this exercise).

Find all plant proteins that are likely phosphatases that do not have orthologs outside of plants.

a. Use the text search **to find OrthoMCL groups** that contain the word *phosphatase* (note that the search should be run with the asterisks).



b. Add a step and run a phyletic pattern search for groups that contain any plant protein but do not contain any other organism outside plants. (hint: make sure everything has a red x on it except for plants (Viridiplantae (VIRI)), which should be a grey circle).

- Root (ALL)
  - Eukaryota (EUKA)
    - Alveolates (ALVE)
    - Amoebozoa (AMOE)
    - Euglenozoa (EUGL)
    - Fungi (FUNG)
    - Metazoa (META)
    - Other Eukaryota (OEUK)
    - Viridiplantae (VIRI)
  - Archaea (ARCH)
    - Nitrosopumilus maritimus (strain SCM1) (nmar)
    - Crenarchaeota (CREN)
    - Euryarchaeota (EURY)
    - Korarchaeota (KORA)
    - Nanoarchaeota (NANO)
  - Bacteria (BACT)
    - Firmicutes (FIRM)
    - Other Bacteria (OBAC)
    - Proteobacteria (PROT)

c. Examine your results. How many groups were returned by the search? What is the distribution of plant proteins in each orthology group?

404 Ortholog Groups

Ortholog Group Results

Rows per page: 1000 | Download | Add to Basket | Add Columns

Ortholog Group	Total Number Proteins	Keywords	Top PFam Domains	EC Numbers	Viridiplantae	Archaea	Alveolata
OG6_134309	58	containing protein; domain containing protein; leucine rich repeat containing protein; nb-arc dom...	PF00931 (47), PF13855 (9), PF07985 (1)	3.1.3.16 (31)	5 / 14 (36%)	0 / 27 (0%)	0 / 125 (0%)
OG6_108065	37	ppm-type phosphatase domain containing protein; uncharacterized protein	PF00481 (25), PF00227 (5)	N/A	1 / 14 (7%)	0 / 27 (0%)	0 / 125 (0%)
OG6_112109	26	phosphatase; ppm-type phosphatase domain containing protein	PF00481 (26), PF02148 (1), PF07576 (1), PF13639 (1)	3.1.3.16 (6)	10 / 14 (71%)	0 / 27 (0%)	0 / 125 (0%)
OG6_112423	24	ippc domain containing protein	PF03372 (22)	3.1.3.36 (2), 3.1.3.56 (2), 3.1.3.86 (2), 3.1.3.- (1)	7 / 14 (50%)	0 / 27 (0%)	0 / 125 (0%)

d. Run a multiple sequence alignment for OG6_112109. Click on the group ID in your result table and navigate to the <u>List of Proteins</u> section of the group page. The Clustal Omega tool is integrated into the table. There are several formats available for the Clustal output, making it easy to take these results to other visualization programs.

OrthoMCL DB

Site search, e.g. OG6_106861 or PF3D7_1133* or "binding protein"

My Strategies Searches Tools My Workspace Data About Help Contact Us

OG6_112109

expand all | collapse all

Search section names...

1 Phyletic distribution
2 Group summary
3 List of proteins
4 PFam domains
5 Cluster graph

expand all | collapse all

3 List of proteins

List of All Proteins Download

To align sequences, select proteins from the table below. Then choose the 'Output format' and click the 'Run Clustal Omega for selected

Search this table...

Clustal Omega	Accession	Description	Organism	Taxon
☑	vcar|D8UBL1	PPM-type phosphatase domain-containing protein	Volvox carteri f. nagariensis	Viridiplantae
☑	crei|A0A2K3DZC7	PPM-type phosphatase domain-containing protein	Chlamydomonas reinhardtii (Chlamydomonas smithii)	Viridiplantae
☑	vcar|D8TYP9	Uncharacterized protein	Volvox carteri f. nagariensis	Viridiplantae
☑	apro|A0A087SRW5	PPM-type phosphatase domain-containing protein	Auxenochlorella protothecoides (Green microalga) (Chlorella protothecoides)	Viridiplantae
☑	cbra|A0A388JMB4	PPM-type phosphatase domain-containing protein	Chara braunii (Braun's stonewort)	Viridiplantae
☑	apro|A0A087SJZ6	PPM-type phosphatase domain-containing protein	Auxenochlorella protothecoides (Green microalga) (Chlorella protothecoides)	Viridiplantae
☑	crei|A0A2K3DBF3	PPM-type phosphatase domain-containing protein	Chlamydomonas reinhardtii (Chlamydomonas smithii)	Viridiplantae
☑	osat|Q0JMD4	Probable protein phosphatase 2C 3	Oryza sativa subsp. japonica (Rice)	Viridiplantae
☑	zmay|A0A1D6PCB8	PPM-type phosphatase domain-containing protein	Zea mays (Maize)	Viridiplantae
☑	ppat|A0A2K1L7H1	PPM-type phosphatase domain-containing protein	Physcomitrium patens (Spreading-leaved earth moss) (Physcomitrella patens)	Viridiplantae
☑	knit|A0A1Y1ILB8	PPM-type phosphatase domain-containing protein	Klebsormidium nitens (Green alga) (Ulothrix nitens)	Viridiplantae
☑	zmay|A0A1D6MTG2	PPM-type phosphatase domain-containing protein	Zea mays (Maize)	Viridiplantae
☑	ppat|A0A2K1K2S8	PPM-type phosphatase domain-containing protein	Physcomitrium patens (Spreading-leaved earth moss) (Physcomitrella patens)	Viridiplantae
☑	ppat|A0A2K1JUE3	PPM-type phosphatase	Physcomitrium patens (Spreading-	Viridiplantae

Check All | Uncheck All

Please note: selecting a large number of proteins will take several minutes to align.

Output format: Mismatches highlighted

Run Clustal Omega for selected proteins