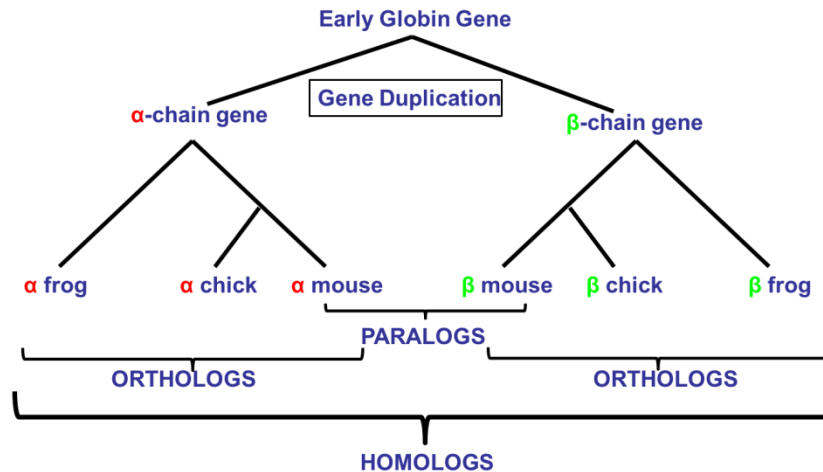# Homology

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 orthologous protein sequences which not only share evolutionary history, but also share function. Thus, ortholog prediction is important in predicting the function of newly identified proteins. Detection of orthologs has become more widespread with the rapid progress in genome sequencing and the discovery of protein sequences (Glover et al. 2019). 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) (Li et al. 2003), 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; Dongen 2000; 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 cryptodb.org and orthomcl.org

a. Use the CryptoDB site search to visit the gene page for the *Cryptosporidium muris* gene, CMU_034340, hypothetical protein, conserved.

b. What information on the gene page can you use to guess a function for this gene? It is annotated as a hypothetical protein! Hint: look at the orthologs table and the domains in the protein features graph. You may also want to visit some of the external links or take a look at InterPro domains.

c. Go to the Orthology and Synteny section and look at the table labeled "Orthologs and Paralogs within CryptoDB". Does this gene have orthologs in other *Cryptosporidium* species? What about other organisms? (hint: scan the organism column in the table) The orthologs may have defined functions (apparent from the product description) which can be used to infer the same function on CMU_034340.

7 Orthology and synteny

Ortholog Group OG6_101337

▾ Orthologs and Paralogs within CryptoDB ≡ Data sets

To run Clustal Omega, select genes from the table below. Then choose the sequence type and initiate the alignment with the 'Run Clustal Omega for selected genes' button.

Search this table... 🔍

↕ Clustal Omega ❓	↕ Gene	↕ Organism	↕ Product	↕ is syntenic	↕ has comments ❓
☐	Cvel_467	Chromera velia CCMP2878	rRNA-processing protein FCF1 homolog, putative	no	no
☐	cand_030400	Cryptosporidium andersoni isolate 30847	hypothetical protein	yes	no
☐	Chro.70261	Cryptosporidium hominis TU502	hypothetical protein	yes	no
☐	CHUDEA7_2290	Cryptosporidium hominis UdeA01	unspecified product	yes	no
☐	GY17_00002025	Cryptosporidium hominis isolate 30976	rRNA-processing protein Fcf1/Utp23	yes	no
☐	ChTU502y2012_407g1140	Cryptosporidium	Fcf1	yes	no

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?

1 Phyletic distribution

▾ Phlyetic Distribution of Proteins ❓ ⬇ Download

Numbers refer to the number of proteins in that organism or taxonomic group.

expand all | collapse all ☑ Hide zero counts

Type a taxonomic name 🔍 ❓
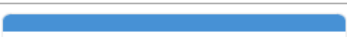
▾ Eukaryota (EUKA)	734
▸ Alveolates (ALVE)	130
▸ Amoebozoa (AMOE)	14
▸ Euglenozoa (EUGL)	73
▸ Fungi (FUNG)	299
▸ Metazoa (META)	143
▸ Other Eukaryota (OEUK)	56
▸ Viridiplantae (VIRI)	19
▾ Archaea (ARCH)	26
Nitrosopumilus maritimus (strain SCM1) (nmar)	1
▸ Crenarchaeota (CREN)	13
▸ Euryarchaeota (EURY)	10
▸ Korarchaeota (KORA)	1
▸ Nanoarchaeota (NANO)	1

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 760 proteins in this ortholog group? PF00149?

4 PFam domains

PFam Legend ⬇ Download

Search this table... ▼ ❓

⇅ Accession	⇅ Symbol	⇅ Description	⇅ Count ❓	Legend
PF04900	Fcf1	Fcf1	731	
PF05811	DUF842	Eukaryotic protein of unknown function (DUF842)	5	
PF01850	PIN	PIN domain	3	
PF00227	Proteasome	Proteasome subunit	1	
PF00160	Pro_isomerase	Cyclophilin type peptidyl-prolyl cis-trans isomerase/CLD	1	
PF13638	PIN_4	PIN domain	1	
PF00149	Metallophos	Calcineurin-like phosphoesterase	1	

PFam Architecture of Each Protein ⬇ Download

Search this table... ▼ ❓

⇅ 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	

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?

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

a. How many orthology groups in OrthoMCL do not have any orthologs in bacteria or archaea?
   (Use the Groups by Phyletic Pattern search)

b. How many protein sequences do not contain orthologs from bacteria and archaea?

c. Find all groups that contain orthologs from at least one species of *Cryptosporidium* and *Giardia* but not from bacteria or archaea. If you are getting frustrated trying to figure this one out, you have a right to be! You cannot answer this question by using the check boxes in the Groups by Phyletic Pattern search. However, OrthoMCL has an added feature that allows you to enter an expression to define the phyletic pattern. This option provides additional flexibility. For example, the phyletic patterns search we just ran can be expressed as ARCH=0T AND BACT=0T. Can you figure out what expression to use to answer this question? (hint: scroll down to the bottom of the Phyletic Pattern search page to find additional information about expression parameters.

Identify Ortholog Groups based on Phyletic Pattern

In the graphical tree display:
- Click on the ▶ icons to show or hide subtaxa and species.
- Click on the ● icons to specify which taxa or species to include or exclude in the profile.
- Refer to the legend below to understand other icons.

Expression: `ARCH=0T AND BACT=0T`

Key: ● = no constraints | ✔ = must be in group | ✔ = at least one subtaxon must be in group | ✖ =

expand all | collapse all

`Type a taxonomic name` 🔍 ❓

✳ **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)**

[Get Answer]

Build a Web Services URL from this Search >>

❓ `Give this search a name (optional)`
❓ `Give this search a weight (optiona.`

Description

Find Ortholog Groups by phyletic pattern.

Phyletic Pattern Expression is a flexible and powerful way to identify ortholog groups with a certain conservation pattern. The pattern is used to identify gr finds groups with a certain copy number (e.g., duplications) within specified taxa.

Examples

These expressions find ortholog groups in which...

Before looking at the answer below, try this on your own or with the people in your breakout room.

Identify Ortholog Groups based on Phyletic Pattern

Find Ortholog Groups that have a particular phyletic pattern, i.e., that include or exclude taxa or species that you specify.

The search is controlled by the Phyletic Pattern Expression (PPE) shown in the text box. Use either the text box or the graphical tree display, or both, to specify your pattern. The graphical tree display is a friendly way to generate expression. You can always edit the expression directly. For PPE help see the instructions at the bottom of this page.

In the graphical tree display:
- Click on the ▶ icons to show or hide subtaxa and species.
- Click on the ● icons to specify which taxa or species to include or exclude in the profile.
- Refer to the legend below to understand other icons.

Expression: `ARCH=0T AND BACT=0T AND cand+chom+chod+choi+chot+cmei+cmur+cpia+cpar+cpar-old+ctyz+cubi>=1T AND gass+gass-old+gadh+gasb+gabb+gase+gmur>=1T` [Get Answer]

Key: ● = no constraints | ✔ = must be in group | ✔ = at least one subtaxon must be in group | ✖ = must not be in group | ✦ = mixture of constraints

`giard` ✖ ❓

✳ **Root (ALL)**
✳ **Eukaryota (EUKA)**
✳ **Other Eukaryota (OEUK)**
- ✔ Giardia Assemblage A isolate WB (gass)
- ✔ Giardia Assemblage A isolate WB (old build 2013-02-08) (gass-old)
- ✔ Giardia Assemblage A2 isolate DH (gadh)
- ✔ Giardia Assemblage B isolate GS (gasb)
- ✔ Giardia Assemblage B isolate GS_B (gabb)
- ✔ Giardia Assemblage E isolate P15 (gase)
- ✔ Giardia muris strain Roberts-Thomson (gmur)

ARCH=0T AND BACT=0T AND cand+chom+chod+choi+chot+cmel+cmur+cpia+cpar+cpar-old+ctyz+cubi>=1T AND gass+gass-old+gadh+gasb+gabb+gase+gmur>=1T
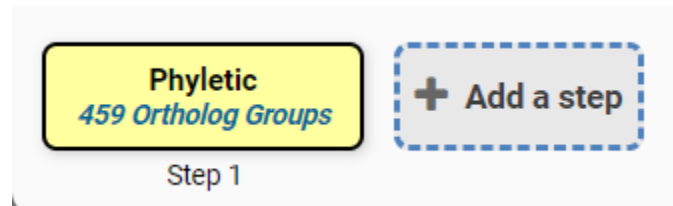
d. All VEuPathDB sites also have a phyletic pattern search that uses OrthoMCL data (PlasmoDB, Genes -> Orthology and synteny -> Orthology Phylogenetic Profile). This search is very useful to identify genes in your organism of interest that have a restricted orthology profile. For example, genes may make good drug targets or vaccine candidate may be conserved among organisms in your genus but not present in the host as these s. Optional: go to your favorite VEuPathDB site and run this search to identify all genes that are not present in human or mouse.
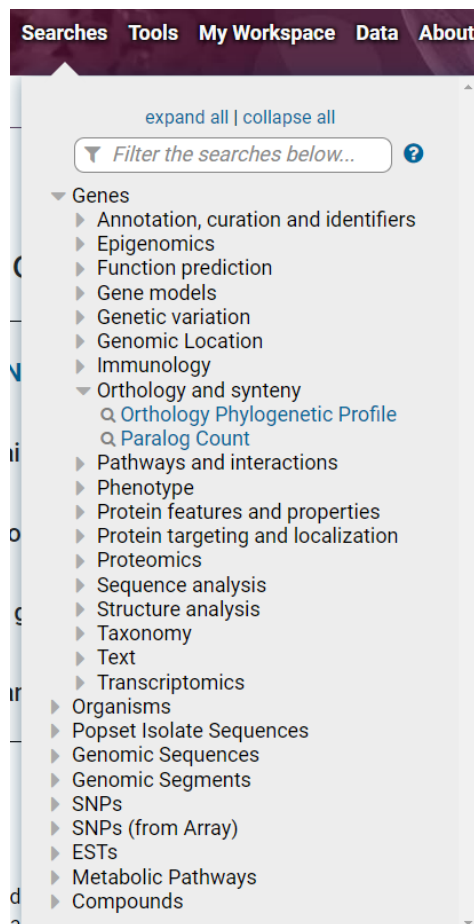
3. Combining searches in OrthoMCL (Use http://orthomcl.org for this exercise).

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 without the quotation marks but 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).
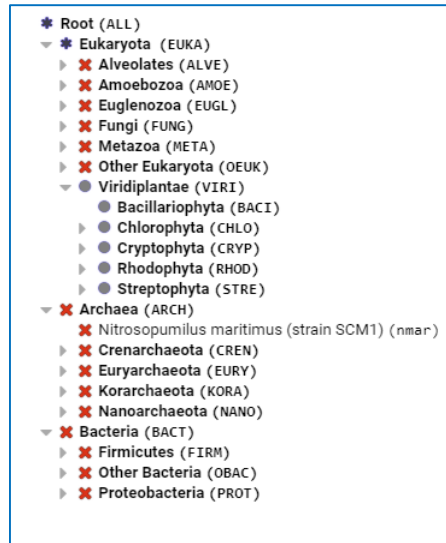
c. Examine your results. How many groups were returned by the search? What is the distribution of plant proteins in each orthology group? (use the Add Columns tool to turn on the Viridiplantae column if it is not already on)

Unnamed Search Strategy * ✏

399 Ortholog Groups

Ortholog Group Results

| ◀ | **1** | 2 | 3 | ... | 20 | ▶ | Rows per page: 20 ▼ | | ⬇ Download | 🧺 Add to Basket |

	Ortholog Group ↓≡	Total Number Proteins ✖ 📊	⇕ Viridiplantae ✖	⇕ Top PFam Domains ❓ ✖ 📊	⇕ Keywords ❓ ✖ 📊
🛒	OG6_112109	26	10 / 14 (71%)	PF00481 (26), PF02148 (1), PF07576 (1), PF13639 (1)	phosphatase; ppm-type phosphatase domain containing protei
🛒	OG6_112423	24	7 / 14 (50%)	PF03372 (22)	ippc domain containing protein
🛒	OG6_130528	21	4 / 14 (29%)	PF00931 (15), PF13855 (1)	disease resistance; disease resistance protein; containing prot
🛒	OG6_153369	20	3 / 14 (21%)	PF13855 (14), PF00560 (13), PF08263 (13)	lrrnt_2 domain containing protein; plant-type; serine-threonine p
🛒	OG6_114630	18	1 / 14 (7%)	PF00481 (11)	ppm-type phosphatase domain containing protein; uncharacter

d. Run a multiple sequence alignment for OG6_112109. Click on the group ID in your result table and navigate to the List of Proteins 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.

4. **Explore a specific OrthoMCL group - examining the cluster graph.** Use http://orthomcl.org

a. Visit the OrthoMCL group OG6_131670. Use the site search to navigate to OG6_131670.

b. Examine the Phyletic Distribution.   What is the phylogenetic distribution of the members of this group? The distribution is presented as a tree. Expand the tree to view the distribution.
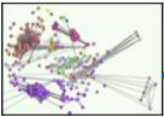
c. Navigate to the Cluster graph tab. Modify the E-value cutoff slider. What happens when you increase or decrease the E-value? Can you identify subclusters of orthologs? The view of the graph can be changed using the Edge type options and the Node options.

5. **Using the orthology transform tool to identify apicoplast targeted genes in *Toxoplasma* and *Neospora*.   Note: For this exercise use http://veupathdb.org**
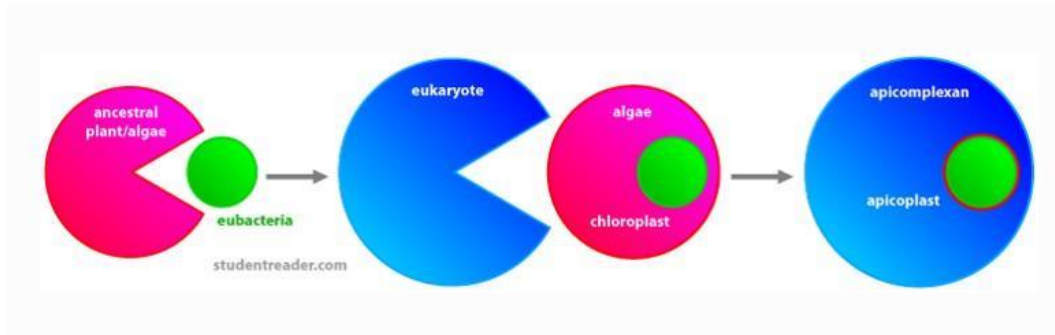
The apicoplast likely became encased in four membranes via a double endosymbiotic event. The chloroplast arose by engulfment of a cyanobacteria by a plant/algae ancestor. An algae was then engulfed by the ancestor of all apicomplexans. Thus, an apicoplast organelle arose with four membranes.

a. Start by finding genes in *Plasmodium* that are predicted to target to the apicoplast. Hint: Navigate to the P.f. Subcellular Localization search. You can further expand your list of potentially Apicoplast targeted proteins by running a GO terms search for the term "apicoplast" or the GO ID: GO:0020011 in *P falciparum* 3D7 (hint, click on add step the go to the function prediction category and select the GO term search). Which Boolean operation did you use? Union or intersect?
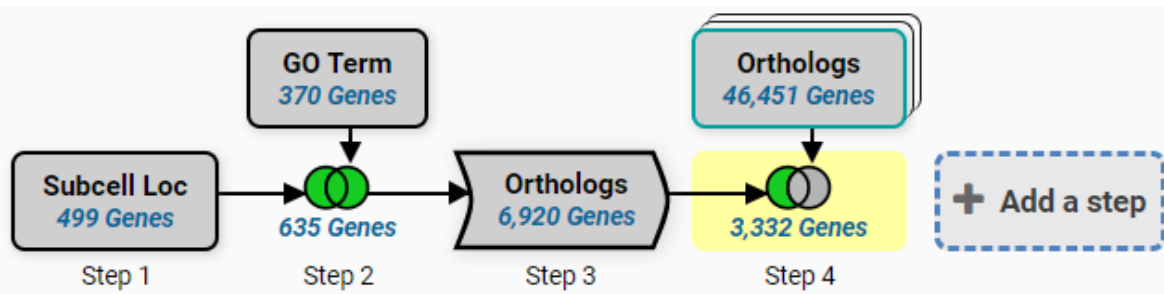
b. Transform the results into their *Toxoplasma* and *Neospora* orthologs. Add a step to your strategy that transforms the results into *Toxoplasma* and *Neospora*.
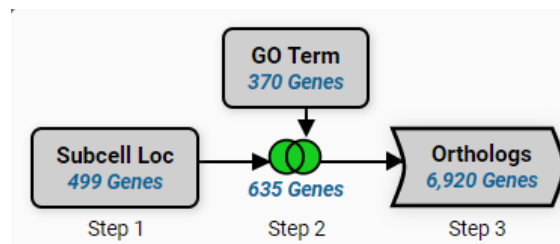
c. Although *Cryptosporidium* is an apicomplexan parasite it has lost its apicoplast! Can you use this fact to refine your results from the above search? Hint: try subtracting out any orthologs present in *Cryptosporidium*. You will need to use a nested strategy. First pull up all Cryptosporidium genes with the Genes by Taxonomy search and then transform these back to their *Toxoplasma* and *Neospora* orthologs for the subtraction to complete.

GO Term
370 Genes

Subcell Loc
499 Genes

635 Genes

Step 1

Step 2

Orthologs
6,920 Genes

Step 3

Orthologs
46,451 Genes

3,332 Genes

Step 4

+ Add a step

Expanded view of *Orthologs*

✖

Organism
54,377 Genes

Orthologs
46,451 Genes

+ Add a step

Step 1

Step 2