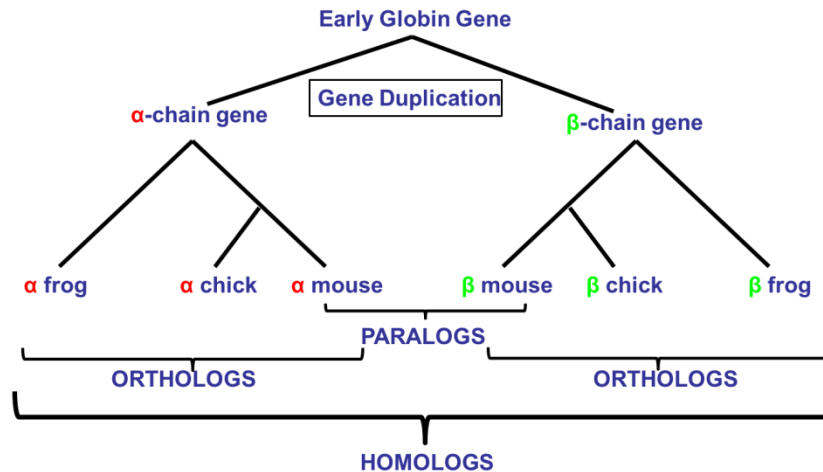


## Orthology and Phyletic Patterns

### 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](http://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](http://cryptodb.org) and [orthomcl.org](http://orthomcl.org)

- Use the CryptoDB site search to visit the gene page for the *Cryptosporidium muris* gene, CMU\_034340, hypothetical protein, conserved.
- 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.

CMU\_034340

expand all | collapse all

Search section names...

1 Gene models

2 Annotation, curation and identifiers

3 Link outs

4 Genomic Location

5 Literature

6 Taxonomy

7 Orthology and synteny

8 Genetic variation

9 Transcriptomics

10 Sequence analysis

11 Sequences

12 Structure analysis

13 Protein features and properties

14 Function prediction

15 Pathways and interactions

16 Immunology

expand all | collapse all

Proteins Properties and Features

Download

Data sets

Transcript ID	Isoelectric Point	Molecular Weight	Has SignalP	Has TMHMM	Protein Length	Pro Bro
CMU_034340-t26_1	10.5	23784	no	no	206	Interac

View in protein browser

Reference Sequence

InterPro Domains

Transmembrane Domains (TMHMM)

Signal Peptide

- 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 syteny

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
<input type="checkbox"/>	<a href="#">Cvel_467</a>	Chromera velia CCMP2878	rRNA-processing protein FCF1 homolog, putative	no	<a href="#">no</a>
<input type="checkbox"/>	<a href="#">cand_030400</a>	Cryptosporidium andersoni isolate 30847	hypothetical protein	yes	<a href="#">no</a>
<input type="checkbox"/>	<a href="#">Chro.70261</a>	Cryptosporidium hominis TU502	hypothetical protein	yes	<a href="#">no</a>
<input type="checkbox"/>	<a href="#">CHUDEA7_2290</a>	Cryptosporidium hominis UdeA01	unspecified product	yes	<a href="#">no</a>
<input type="checkbox"/>	<a href="#">GY17_00002025</a>	Cryptosporidium hominis isolate 30976	rRNA-processing protein Fcf1/Utp23	yes	<a href="#">no</a>
<input type="checkbox"/>	<a href="#">ChTU502y2012_407q1140</a>	Cryptosporidium	Fcf1	yes	<a href="#">no</a>

- 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

▼ Phyletic 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)** 555
  - ▶ **Alveolates (ALVE)** 107
  - ▶ **Amoebozoa (AMOE)** 13
  - ▶ **Euglenozoa (EUGL)** 60
  - ▶ **Fungi (FUNG)** 198
  - ▶ **Metazoa (META)** 113
  - ▶ **Other Eukaryota (OEUK)** 45
  - ▶ **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 697 proteins in this ortholog group? PF00149?

## 4 Pfam domains

▼ Pfam Legend [Download](#)

Search this table...



Accession	Symbol	Description	Count	Legend
PF04900	Fcf1	Fcf1	668	
PF01850	PIN	PIN domain	3	
PF00149	Metallophos	Calcineurin-like phosphoesterase	1	
PF13638	PIN_4	PIN domain	1	
PF05811	DUF842	Eukaryotic protein of unknown function (DUF842)	1	
PF00227	Proteasome	Proteasome subunit	1	
PF00160	Pro_isomerase	Cyclophilin type peptidyl-prolyl cis-trans isomerase/CLD	1	

▼ Pfam Architecture of Each Protein [Download](#)

Search this table...



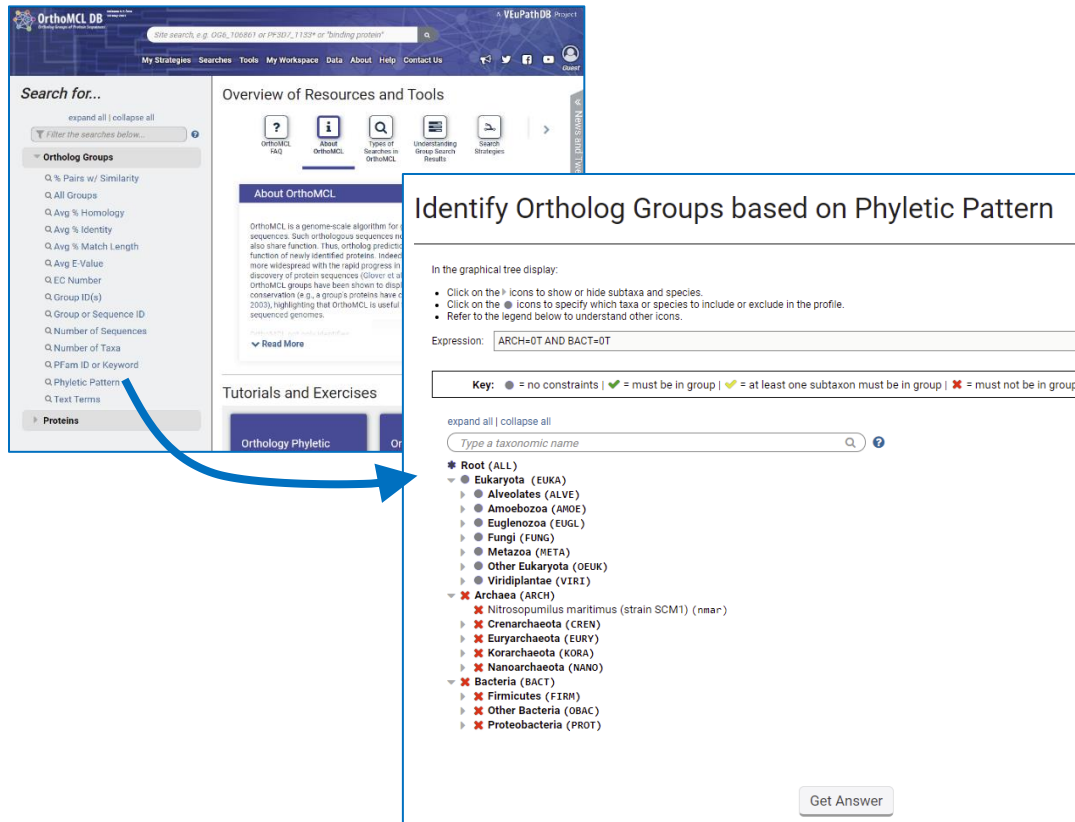
Accession	Taxon	Core/Peripheral	Protein Length	
aacu ASPACDRAFT_77294	Aspergillus aculeatus ATCC 16872	Peripheral	189	
aaeg-old AAEL007697	Aedes aegypti LVP_AGWG (old build 2019-12-20)	Core	241	
aaeg AAEL007697	Aedes aegypti LVP_AGWG	Peripheral	241	
aalb AALFPA_064762	Aedes albopictus Foshan FPA	Peripheral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

- How many orthology groups in OrthoMCL do not have any orthologs in bacteria or archaea?



The screenshot shows the OrthoMCL website interface. On the left, the 'Search for...' section has a dropdown menu for 'Orthology Groups' with various filters. A blue arrow points from the 'Phyletic Pattern' option in this menu to the 'Identify Ortholog Groups based on Phyletic Pattern' tool window on the right.

**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=OT AND BACT=OT

Key: = no constraints | = must be in group | = at least one subtaxon must be in group | = must not 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

Phyletic  
834,492 Ortholog Groups

Step 1

834,492 Ortholog Groups [Revise this search](#)

Ortholog Group Results

1 2 3 ... 41,725 Rows per page: 20

[Download](#) [Add to Basket](#) [Add Columns](#)

Ortholog Group	Total Number Proteins	Keywords	Top Pfam Domains	EC Numbers	Archaea	Bacteria	Alveolates
OG6_100001	14741	unknown; hypothetical protein; conserved hypothetical protein	PF13388 (4233), PF04665 (3687), PF04851 (212)	N/A	0 / 27 (0%)	0 / 47 (0%)	8 / 129 (6%)
OG6_100002	6864	unknown; conserved hypothetical protein	PF12943 (5254), PF10544 (1424), PF04383 (2), PF12789 (2)	N/A	0 / 27 (0%)	0 / 47 (0%)	2 / 129 (2%)
OG6_100003	6580	hypothetical protein; conserved hypothetical protein; unknown	PF12789 (2592), PF06022 (45), PF02349 (1), PF03770 (1), PF07679 (1), PF12295 (1)	1.4.1.2 (2)	0 / 27 (0%)	0 / 47 (0%)	10 / 129 (8%)

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

The screenshot illustrates the workflow in the OrthoMCL web interface to convert ortholog groups into protein sequences. It starts with a 'Phyletic' search strategy (Step 1) containing 834,492 Ortholog Groups. An 'Add a step' button leads to a configuration panel where a second step, 'To Proteins', is added. This step is configured to 'Transform 834,492 Ortholog' groups into 'Proteins'. The final workflow shows 'Phyletic' (Step 1) leading to 'To Proteins' (Step 2), resulting in 5,714,402 Proteins. Below the workflow, the 'Protein Results' section displays a table of the first few results.

Accession	Taxon Name	Description	Length	EC Number
aacuASPACDRAFT_10177	Aspergillus aculeatus ATCC 16872	unknown	1634	N/A
aacuASPACDRAFT_10224	Aspergillus aculeatus ATCC 16872	Protein FYV10 [Source:UniProtKB/TrEMBL;Acc:A0A1L9WJ01]	1427	N/A
aacuASPACDRAFT_10238	Aspergillus aculeatus ATCC 16872	Spc7 domain-containing protein [Source:UniProtKB/TrEMBL;Acc:A0A1L9X3H8]	1572	N/A
aacuASPACDRAFT_10268	Aspergillus aculeatus ATCC 16872	KIX_2 domain-containing protein [Source:UniProtKB/TrEMBL;Acc:A0A1L9X308]	1497	N/A

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=OT AND BACT=OT. 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:

Key: = no constraints | = must be in group | = at least one subtaxon must be in group | = must not be in group

[expand all](#) | [collapse all](#)

- \* 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 (optional)

---

### 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 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:

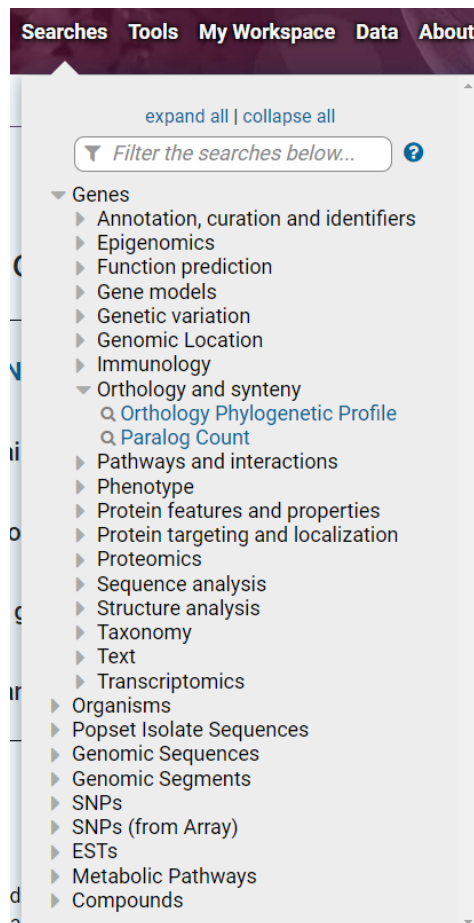
Key: = no constraints | = must be in group | = at least one subtaxon must be in group | = must not be in group | = mixture of constraints

- \* Root (ALL)
  - \* Eukaryota (EUKA)
    - \* Other Eukaryota (OEUK)
      - Giardia Assemblage A isolate WB (gass)
      - Giardia Assemblage A2 isolate DH (gadn)
      - Giardia Assemblage B isolate GS (gabn)
      - Giardia Assemblage B isolate GS\_B (gabn)
      - Giardia Assemblage E isolate P15 (gase)
      - Giardia muris strain Roberts-Thomson (gmur)

[Get Answer](#)

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 under 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).

expand all | collapse all  
Filter the searches below...

Ortholog Groups  
Q % Pairs w/ Similarity  
Q All Groups  
Q Avg % Homology  
Q Avg % Identity  
Q Avg % Match Length  
Q Avg E-Value  
Q EC Number  
Q Group ID(s)  
Q Group or Sequence ID  
Q Number of Sequences  
Q Number of Taxa  
Q PFam ID or Keyword  
Q Phyletic Pattern  
Q Text Terms  
Proteins

Identify Ortholog Groups based on Text Terms

Text term (use \* as wildcard)

Fields

- ☒ EC Number
- ☒ Keywords
- ☒ Ortholog group
- ☒ PFam Domains
- ☒ Protein Description
- ☒ Protein ID
- ☒ Protein Previous Groups
- ☒ Protein Taxon Abbreviation
- ☒ Protein Taxon Name

select all | clear all

Get Answer

Text  
5,028 Ortholog Groups

Add a step

- 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)
  - x Alveolates (ALVE)
  - x Amoebozoa (AMOE)
  - x Euglenozoa (EUGL)
  - x Fungi (FUNG)
  - x Metazoa (META)
  - x Other Eukaryota (OEUK)
  - Viridiplantae (VIRI)
    - Bacillariophyta (BACI)
    - Chlorophyta (CHLO)
    - Cryptophyta (CRYP)
    - Rhodophyta (RHOD)
    - Streptophyta (STRE)
- x Archaea (ARCH)
  - x Nitrosopumilus maritimus (strain SCM1) (nmar)
  - x Crenarchaeota (CREN)
  - x Euryarchaeota (EURY)
  - x Korarchaeota (KORA)
  - x Nanoarchaeota (NANO)
- x Bacteria (BACT)
  - x Firmicutes (FIRM)
  - x Other Bacteria (OBAC)
  - x 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?



#### 403 Ortholog Groups

Ortholog Group Results							
<div> <div>1 2 3 ... 21</div> <div>Rows per page: 20</div> <div>Download</div> <div>Add to Basket</div> <div>Add C</div> </div>							
Ortholog Group	Total Number Proteins	Keywords	Top Pfam Domains	EC Numbers	Viridiplantae	Archaea	
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%)	
OG6_108065	37	ppm-type phosphatase domain containing protein; uncharacterized protein	PF00481 (25), PF00227 (5)	N/A	1 / 14 (7%)	0 / 27 (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%)	
OG6_112423	24	lppc 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%)	
OG6_130528	21	disease resistance; disease resistance protein; containing protein; disease resistance protein rp...	PF00931 (15), PF13855 (1)	3.1.3.16 (4)	4 / 14 (29%)	0 / 27 (0%)	

- 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.

OrthoMCL DB

Site search, e.g. OG6\_108065 or PF007\_11331 or "binding protein"

My Strategies Searches Tools My Workspace Data About Help Contact Us

OG6\_112109

Search random names

1 Phyletic distribution

2 Group summary

3 List of proteins

4 Pfam domains

5 Cluster graph

List of proteins

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

Search this table...

Clustal Omega	Accession	Description	Organism	Taxon	Core/Peripheral	Length
<input checked="" type="checkbox"/>	vcariDSUBL1	PPM-type phosphatase domain-containing protein	Volvox carterii f. nageiensis	Viridiplantae	Peripheral	1309
<input checked="" type="checkbox"/>	creiAQA2K3DZC7	PPM-type phosphatase domain-containing protein	Chlamydomonas reinhardtii (Chlamydomonas smithii)	Viridiplantae	Core	1237
<input checked="" type="checkbox"/>	vcariD8TYP9	Uncharacterized protein	Volvox carterii f. nageiensis	Viridiplantae	Peripheral	998
<input checked="" type="checkbox"/>	aproAQA087SRW5	PPM-type phosphatase domain-containing protein	Auxenochlorella protothecoides (Green microalgae) (Chlorella protothecoides)	Viridiplantae	Core	708
<input checked="" type="checkbox"/>	cbraiAQA388JMB4	PPM-type phosphatase domain-containing protein	Chara braunii (Braun's stonewort)	Viridiplantae	Core	704
<input checked="" type="checkbox"/>	aproAQA087SJZ6	PPM-type phosphatase domain-containing protein	Auxenochlorella protothecoides (Green microalgae) (Chlorella protothecoides)	Viridiplantae	Core	543
<input checked="" type="checkbox"/>	creiAQA2K3DBF3	PPM-type phosphatase domain-containing protein	Chlamydomonas reinhardtii (Chlamydomonas smithii)	Viridiplantae	Core	491
<input checked="" type="checkbox"/>	osatiQQJMD4	Probable protein phosphatase 2C 3	Oryza sativa subsp. japonica (Rice)	Viridiplantae	Core	485

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

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

- Visit the OrthoMCL group OG6\_131670. Use the site search to navigate to OG6\_131670.
- 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.

### 1 Phyletic distribution

▼ Phyletic 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 🔍 ⓘ

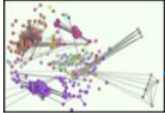
- ▼ Eukaryota (EUKA) 113
  - ▶ Alveolates (ALVE) 110
    - ▶ Metazoa (META) 3

- 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 Cluster graph

Click to open the Cluster graph in a new tab

Cluster graph of all proteins ⓘ



Cluster Graph: OG6\_131670 (97 proteins) ⓘ

Back to Group page

▼ Edge Options ⓘ

Edge Type

- ☒ Ortholog
- ☒ Coortholog
- ☒ Inparalog
- ☒ Peripheral-Core
- ☒ Peripheral-Peripheral
- ☐ Other Similarities

E-Value Cutoff

Max E-Value: 1E-22

▼ Node Options ⓘ

Show Nodes By

- ☒ Taxa
- ☐ EC Numbers
- ☐ Pfam Domains

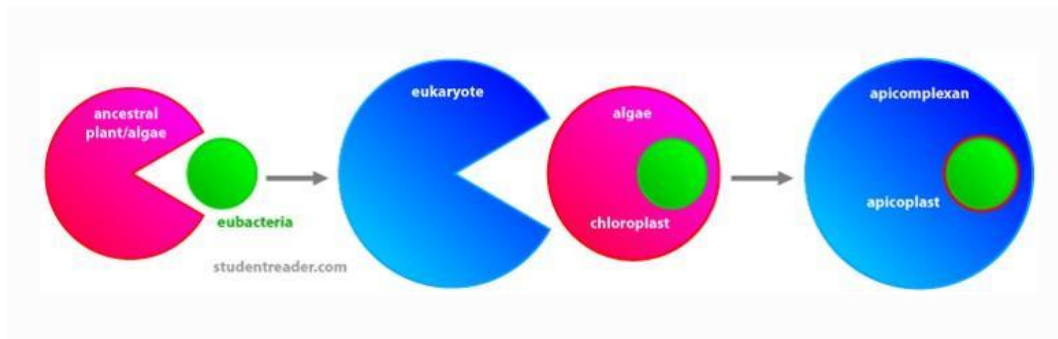
Mouse over a taxon legend to highlight sequences of that taxon.

gnip (1) gnip-old (1) hpl (1)  
gad (1) gber (1) pbl (1)  
pog (1) pche (1) pcca (1)  
pym (1) pzym (1) pfal (1)  
pfal-old (1) pfag (1) pfac (1)  
pfad (1) pfga (1) pfub (1)  
pfgn (1) pfho (1) pfai (1)  
pfka (1) pfkn (1) pfkt (1)  
pfml (1) pfed (1) pfan (1)

Sequence List ⓘ Node Details ⓘ

Accession	Taxon	Length
dbesBESB_010830	dbes	536
canldicand_009760	cand	318
ccayloc_07054	ccay	553
ccnfLOC34623090	ccnf	553
chodCHUDEA8_5030	chod	325
chodGY17_0000954	chod	325
chomChvo_80575	chom	325
chodCHTUS2cy2012_408fg0200	chod	325
cmelCmelJ03MEL1_07710	cmel	325
cmurCMUL011520	cmur	318
cpar-oldcpdl_5030	cpar-old	324
cparcpdl_5030	cpar	324
cpaiCPATCC_004380	cpai	324
csuCSU_001537	csu	577
ctyzCTYZ_00000545	ctyz	325
cubiCubi_03730	cubi	325
eeaceEAH_00023650	eeace	550
eburEBH_0061790	ebur	551
efalEfaB_MINUS_13398.g1198	efal	547
emaxEMWEY_00025950	emax	652
emrEMR_0005590	emr	551
enecENH_00050470	enec	553
epreEPH_0014470	epre	551
eteneETH_00031620	etene	553
gnip-oldGN_098890	gnip-old	387

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?

**Search for...**  
expand all | collapse all  
Filter the searches below...  
Pathways and interactions  
Phenotype  
Protein features and properties  
Protein targeting and localization  
Exported Protein  
P.f. Subcellular Localization  
Predicted Signal Peptide  
Transmembrane Domain Count  
Proteomics

**Identify Genes based on P.f. Subcellular Localization**  
Localization  
Apicoplast  
Get Answer

**Subcell Loc**  
513 Genes

+ Add a step

Step 1

**Organism**

1 selected, out of 439

add these | clear these | select only these  
select all | clear all

3d7

- ☐ Apicomplexa
  - ☐ Aconoidasida
    - ☐ Haemosporida
      - ☐ Plasmodium
        - ☐ Plasmodium falciparum
        - ☒ Plasmodium falciparum 3D7 [Reference]

add these | clear these | select only these  
select all | clear all

**Evidence**

☒ Curated  
☒ Computed  
select all | clear all

**Limit to GO Slim terms**

☐ Yes  
☒ No

**GO Term or GO ID**


GO:0020011 | apicoplast:7

**GO Term or GO ID wildcard search**

N/A

Run Step

**Subcell Loc**  
513 Genes

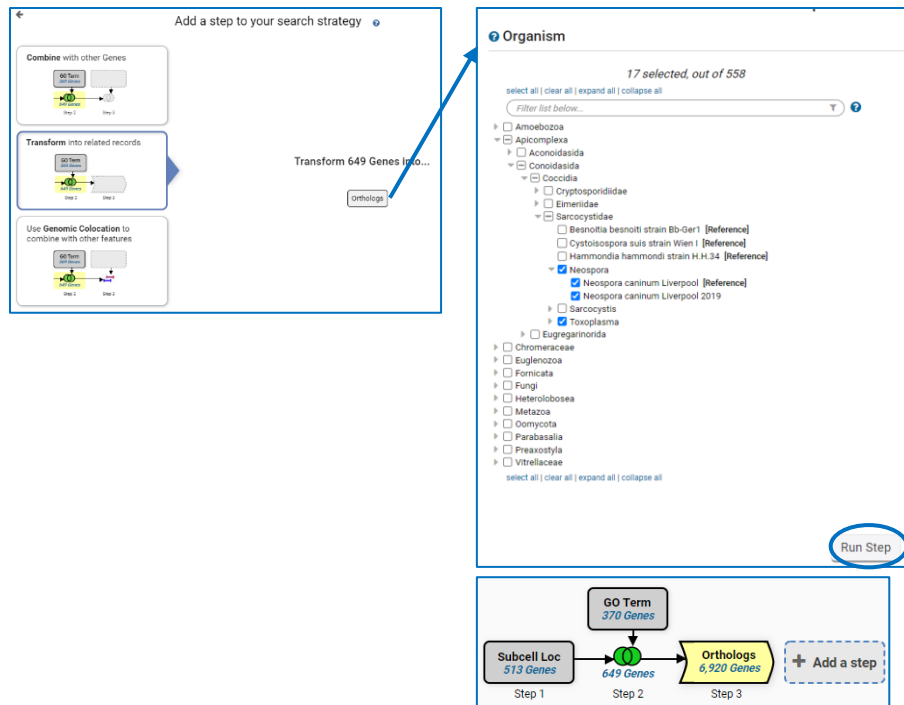


649 Genes

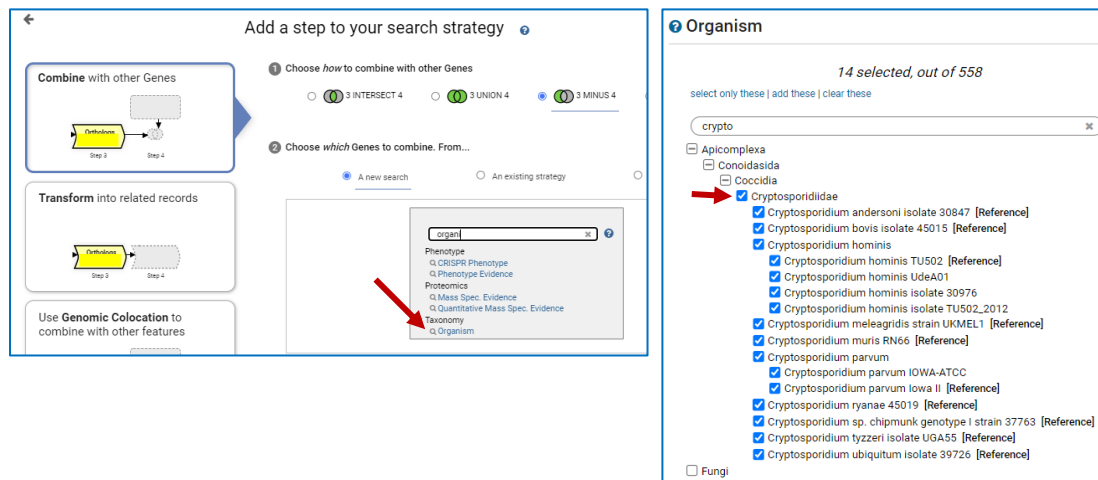
+ Add a step

Step 1
Step 2

- 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 and use the ortholog transform back to *Toxoplasma* and *Neospora* genes for the subtraction to complete.



Opened (1) All (1) Public (13) Help

Unnamed Search Strategy \*

Step 1 Step 2 Step 3 Step 4

6,920 Genes (288 ortholog groups)

View | Analyze | Revis | **Make nested strategy** | Insert step before | Orthologs | Delete

**Details for step** Organism

54766 Genes

**Organism** Cryptosporidium andersoni isolate 30847, Cryptosporidium bovis isolate 45015, Cryptosporidium hominis TU502, Cryptosporidium hominis UdeA01, Cryptosporidium hominis isolate 30976, Cryptosporidium hominis isolate TU502\_2012, Cryptosporidium meleagridis strain UKMEL1, Cryptosporidium muris RN66, Cryptosporidium parvum IOWA-ATCC, Cryptosporidium parvum Iowa II, Cryptosporidium ryanae 45019, Cryptosporidium sp. chipmunk genotype I strain 37763, Cryptosporidium tyzzeri isolate UGA55, Cryptosporidium ubiquitum isolate 39726

Give this search a weight

Some genes in your combined result have transcripts that were not returned by one or both of the two input searches

