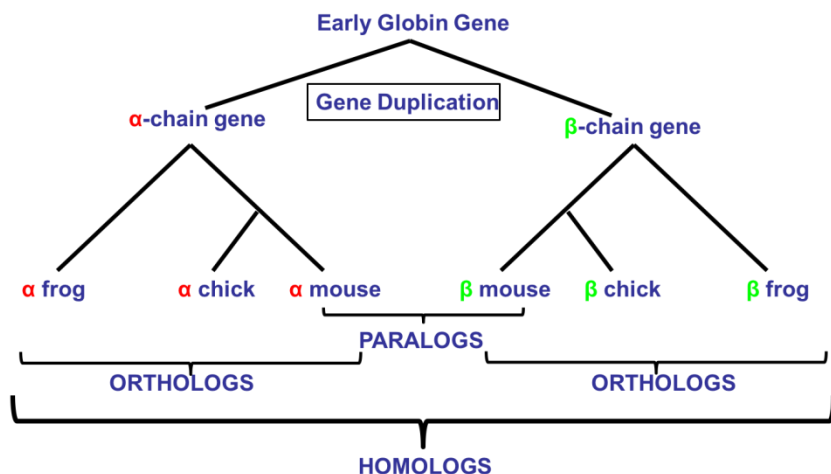


## Orthology and Phyletic Patterns

### Homology



#### 1. Getting to OrthoMCL from EuPathDB databases

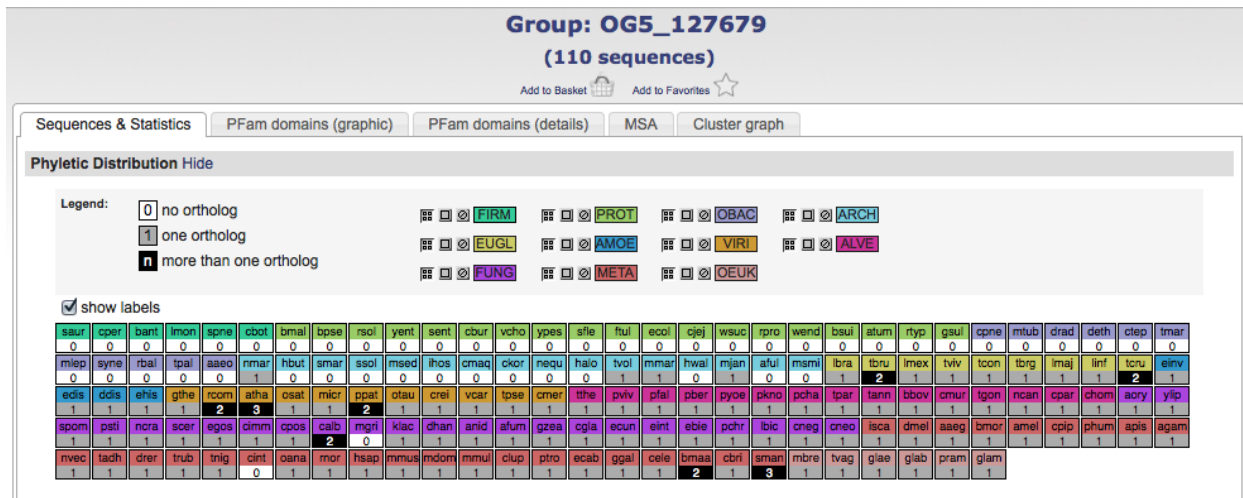
Note: For this exercise use <http://cryptodb.org> and <http://orthomcl.org/>

- Go to the gene page for the *Cryptosporidium parvum* gene with the ID: cgd7\_2290
- 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.
- Scroll down to the table labeled “Orthologs and Paralogs within CryptoDB”. Does this gene have orthologs in other *Cryptosporidium* species? What about other organisms? (hint: click on the link below the table that takes you to OrthoMCL).

Gene	Organism	Product	is syntenic	has comments
Cvel_467	Chromera velia CCMP2878	rRNA-processing protein FCF1 homolog, putative	no	no
Chro.70261	Cryptosporidium hominis TU502	hypothetical protein	yes	no
CMU_034340	Cryptosporidium muris RN66	hypothetical protein, conserved	yes	no
GNI_088410	Gregarina niphandrodes Unknown strain	rRNA-processing Fcf1-like protein	no	no
Vbra_6876	Vitrella brassicaformis CCMP3155	rRNA-processing protein FCF1 homolog, putative	no	no

View the group (OG5\_127679) containing this gene (cgd7\_2290) in the OrthoMCL database

- Does this protein have orthologs in other organisms? Does it have any orthologs in bacteria or archaea? (Hint: mouse over the colorful boxes in the table to reveal the full species and phylum names – see image below).



- e. Take a look at the PFAM domain architectures found under the PFam domains (graphic) tab. Do all the proteins in this group have similar domain architecture?
- f. Based on the orthologs, 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 protein groups in OrthoMCL do not have any orthologs in bacteria or archaea? (Hint: go to the “Phyletic Pattern” search in the Evolution section of the “Identify Ortholog Groups”

**OrthoMCL DB** Version 5 10 May 13 A EuPathDB Project

Groups Quick Search:  synth\*   Sequences Quick Search:  synth\*

About OrthoMCL   Help   Login   Register   Contact Us

Home   New Search   My Strategies   My Basket (0)   Tools   Data Summary   Downloads   Community   My Favorites

**Data Summary**

- Genomes: 150
- Protein Sequences: 1,398,546
- Ortholog Groups: 124,740

**News and Tweets**

**Community Resources**

**Education and Tutorials**

**About OrthoMCL**

**Identify Ortholog Groups**

Text, IDs  
Group ID(s)  
Text Terms

Evolution  
Phyletic Pattern

Function  
PFam ID or Keyword  
Enzyme Commission Assignment

Group Statistics  
Number  
Avg % C  
% Pairs  
Avg % I  
Avg % M  
Avg E-V

**Identify Protein Sequences**

Text, IDs  
Sequence ID(s)  
Group ID(s)  
Text Terms

Function  
PFam ID or Keyword  
Enzyme Commission Assignment

Similarity/Pattern

**Identify 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 handy way to generate a pattern 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 +/- to show or hide subtaxa and species.
- Click on the icon to specify which taxa or species to include or exclude in the profile.
- Refer to the legend below to understand other icons.

Expression: BACT=OT AND ARCH=OT

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

- Root (ALL):
- ✳ Bacteria (BACT):
- ✳ Archaea (ARCH):
- Eukaryota (EUKA):

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

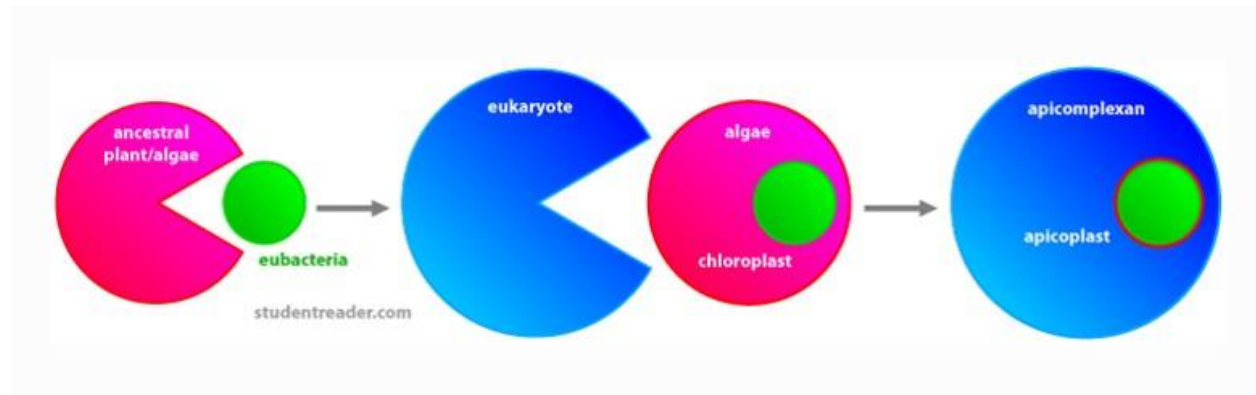
category). To specify a phyletic pattern click on the icon next to the taxonomic group or species to include or exclude it.

- a. How many protein groups do not contain orthologs from eukaryotes?
- b. Find all groups that contain orthologs from at least one species of *Cryptosporidium* and *Giardia* but not from bacteria or archaea.

All EuPathDB sites also have a phyletic pattern search that uses OrthoMCL data under Genes -> Evolution -> Orthology Phylogenetic Profile. This search is very useful to identify genes in your organism of interest that are restricted in their profile. For example, you frequently want to identify genes that are conserved among organisms in your genus but not present in the host as these genes may make good drug targets or vaccine candidates. Optional: go to your favorite EuPathDB site and run this search to identify all genes that are not present in human or mouse.

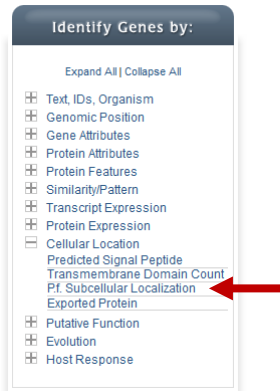
### 3. Using the orthology transform tool to identify apicoplast targeted genes in *Toxoplasma* and *Neospora*.

Note: For this exercise use <http://eupathdb.org>



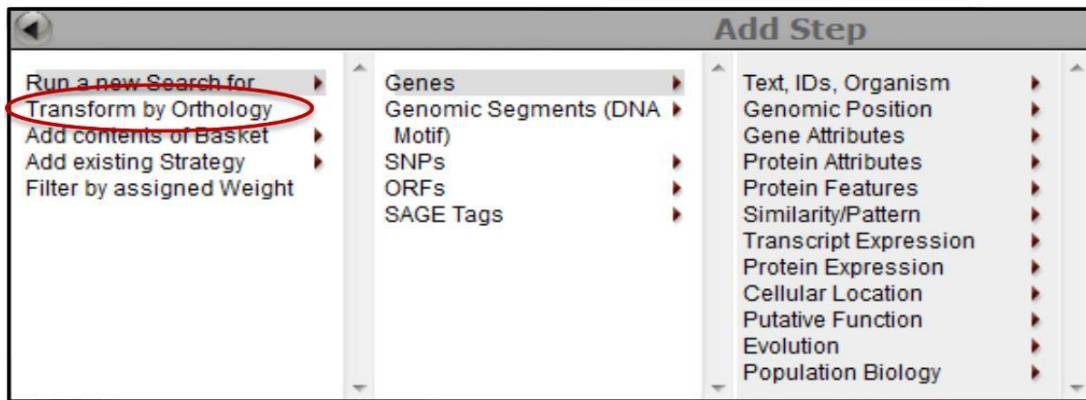
The apicoplast likely became encased in four membranes via a double endosymbiotic event. The chloroplast arose by engulfment of a cyanobacteria by a plant/algae ancestor. An algae was then engulfed by the ancestor of all apicomplexans. Thus an apicoplast organelle arose with four membranes.

- a. Start by finding genes in *Plasmodium* that are predicted to target to the apicoplast. Hint: click on "Cellular Location" then on "P.f. Subcellular Localization".



b. Transform the results of the above search to their *Toxoplasma* orthologs.

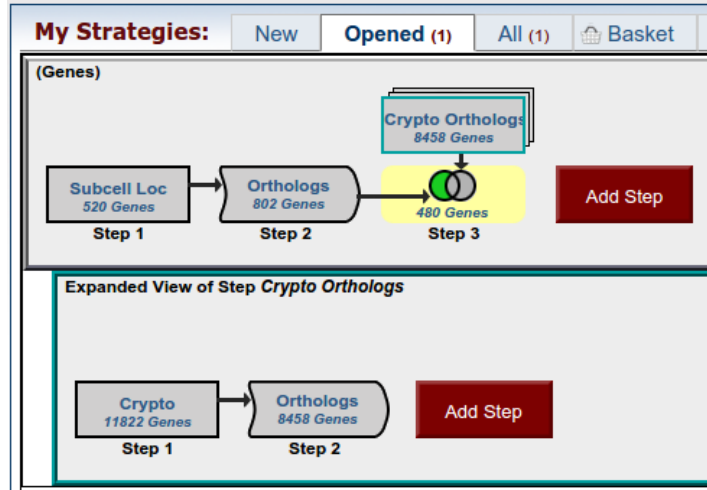
Hint: add a step, then select “Transform by Orthology”. On the search page, select all



*Toxoplasma* and *Neospora*.

c. Although *Cryptosporidium* is an apicomplexan parasite it has actually 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.



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 without the quotation marks but with the asterisks).

The screenshot shows the OrthoMCL DB website interface. The search bar contains the text '\*phosphatase\*'. Below the search bar, there are two search boxes: 'Groups Quick Search:' and 'Sequences Quick Search:'. The search results are displayed in a list format, organized by taxonomic groups. The groups are: Aconoidasida (ACON), Haemosporida (HAEM), Piropiasmida (PIRO), Amoebozoa (AMOE), Euglenozoa (EUGL), Viridiplantae (VIRI), Streptophyta (STRE), Chlorophyta (CHLO), Rhodophyta (RHOD), Cryptophyta (CRYP), Bacillariophyta (BACI), Fungi (FUNG), Microsporidia (MICR), and Basidiomycota (BASI). The search results for each group are listed in a grid format. The search results for the plant groups (Viridiplantae) are highlighted in grey, while the search results for other groups are in red. The search results for the plant groups are: atha, osat, ppat, rcom, micr, crei, otau, vcar, cmer, gthe, tpse, ecun, ebie, eint, cneo, cneg, lbic, pchr.

- 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).

