

# Orthology and Phyletic Patterns

## 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 does this gene do? It is annotated as a hypothetical protein!
- Scroll down to the table labeled “Orthologs and Paralogs within CryptoDB”. Does this gene have orthologs in other *Cryptosporidium* species? What about other

### Orthologs and Paralogs within CryptoDB Hide

| Gene                       | Organism                             | Product                         | is syntenic | has comments |
|----------------------------|--------------------------------------|---------------------------------|-------------|--------------|
| <a href="#">Chro.70261</a> | <i>Cryptosporidium hominis</i> TU502 | hypothetical protein            | yes         | no           |
| <a href="#">CMU_034340</a> | <i>Cryptosporidium muris</i> RN66    | hypothetical protein, conserved | yes         | no           |

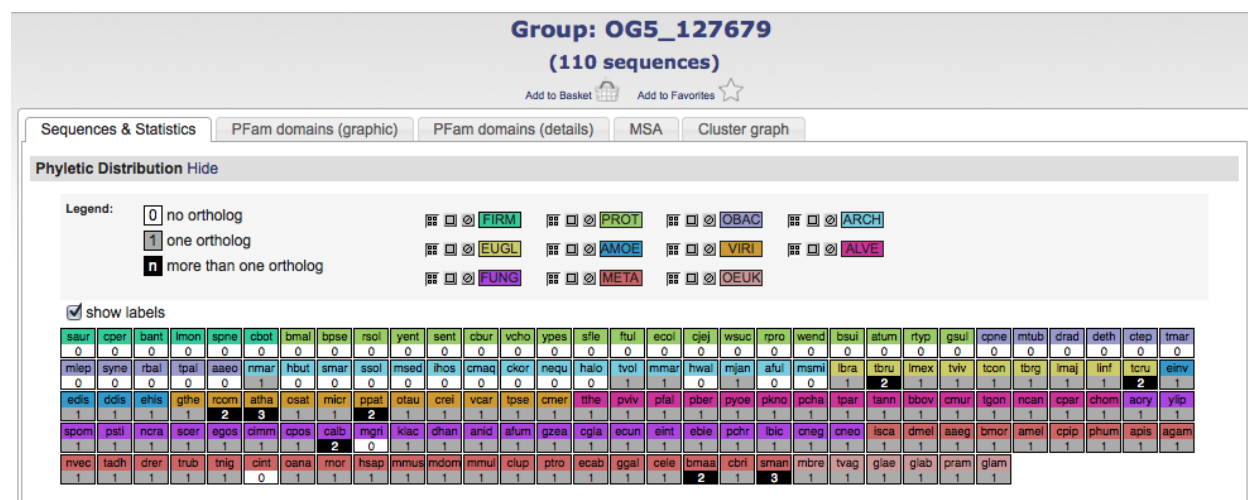
[View the group \(OG5\\_127679\) containing this gene \(cgd7\\_2290\) in the OrthoMCL database](#)



organisms? (hint: click on the link below the table that takes you to OrthoMCL).

- 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).



- 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?
- 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/>

- a. 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” category). To specify a phyletic pattern click on

OrthoMCL DB  
Ortholog Groups of Protein Sequences  
Version 5  
10 May 13  
A EuPathDB Project

Groups Quick Search:  Sequences Quick Search:

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

**Tools:**

- BLAST
- Assign your proteins to groups
- Download OrthoMCL software
- Web Services
- Publications mentioning OrthoMCL

**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 friendly 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

the icon next to the taxonomic group or species to include or exclude it.

- b. How many protein groups do not contain orthologs from eukaryotes?

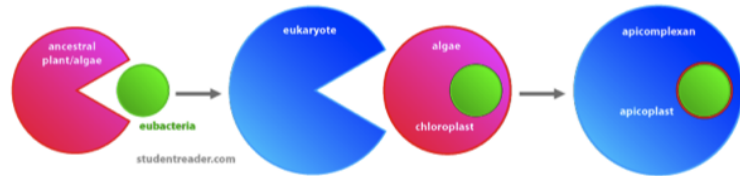
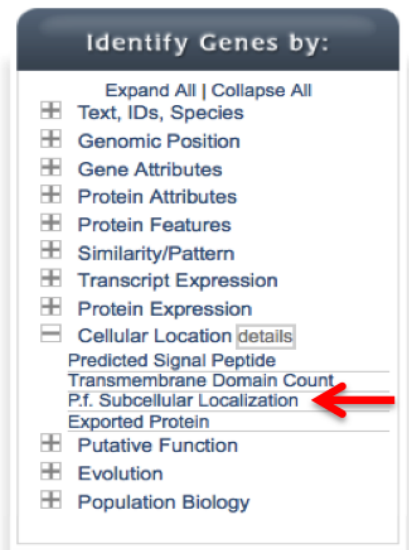
- c. Find all groups that contain orthologs from at least one species of *Cryptosporidium* and *Giardia* but not from bacteria or archaea.

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*.

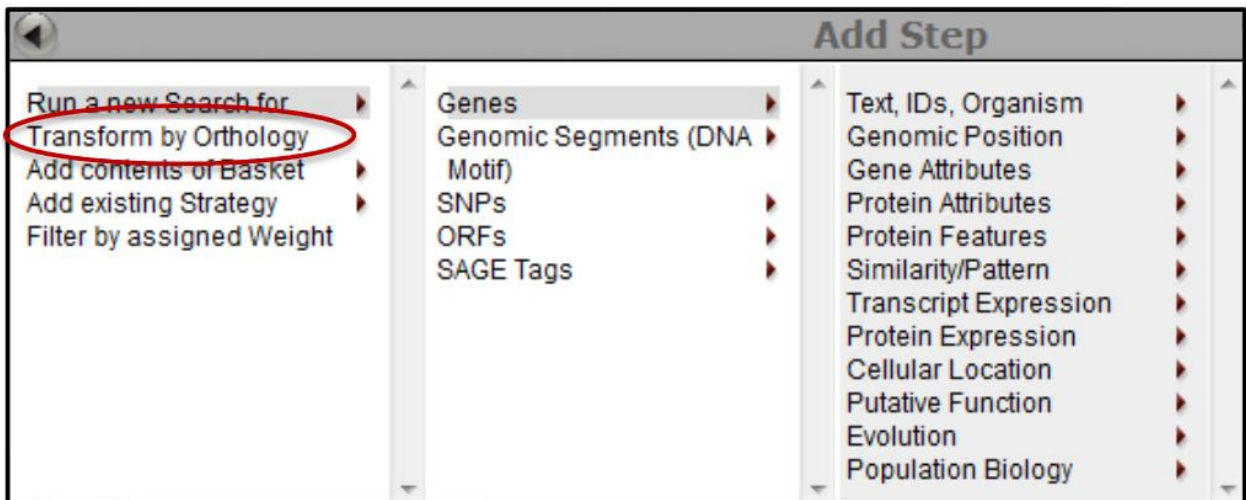
Note: For this exercise use <http://eupathdb.org>

- 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”; see image



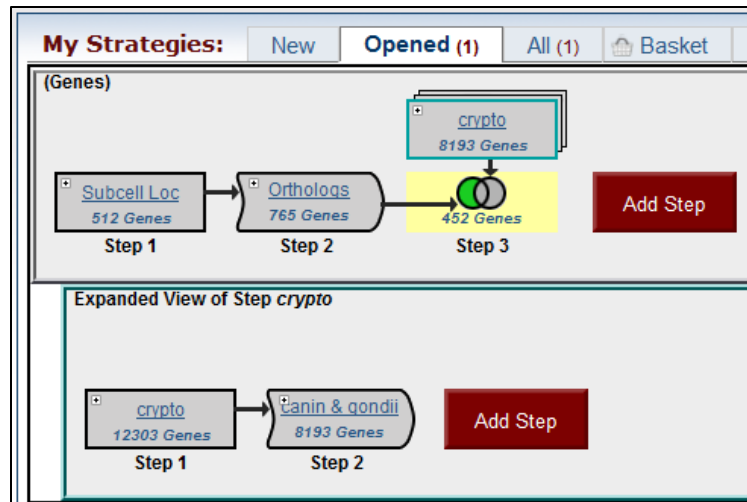
below.

- 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?

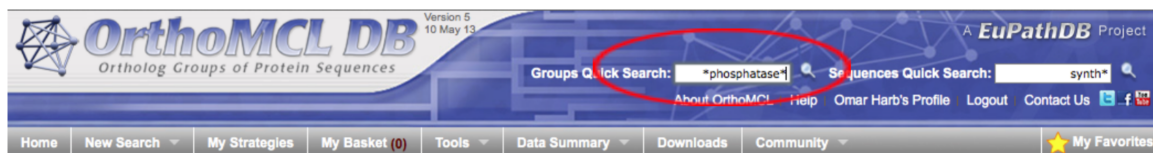
Hint: try subtracting out any orthologs present in *Cryptosporidium*. You will need to use a nested strategy.



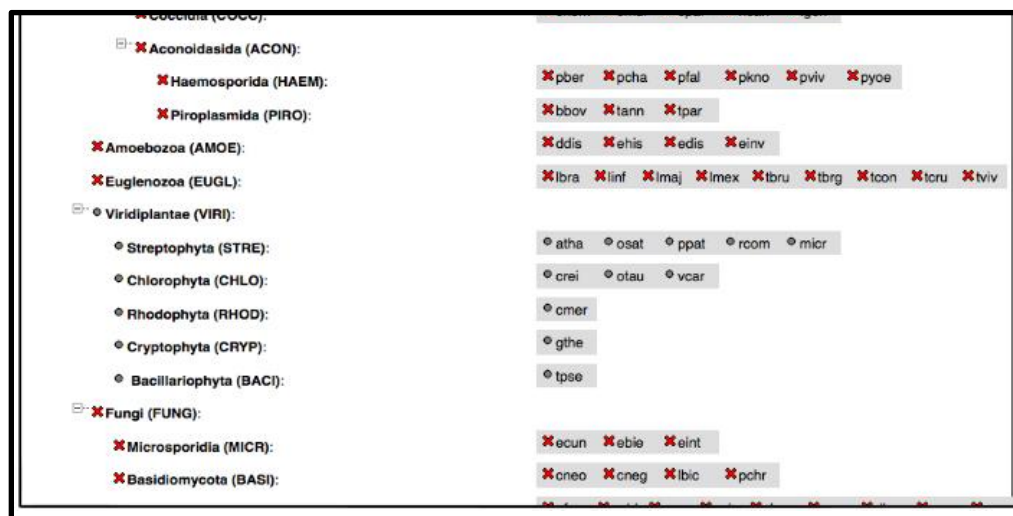
#### 4. Combining searches in OrthoMCL (Use <http://orthomcl.org> for this exercise).

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

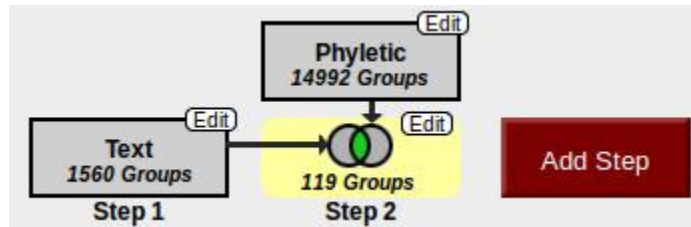
- Use the text search to find OrthoMCL groups that contain the word “\*phosphatase\*” (note that the search should be run without the quotation marks but with the asterisks).



- 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. How many groups did you return? Explore the multiple sequence alignments from some of these groups. (Hint: click on a group ID and open the MSA tab).



Group: OG5\_150204

(10 sequences)

Add to Basket

Add to Favorites

Sequences & Statistics

PFam domains (graphic)

PFam domains (details)

MSA

Cluster graph

Phyletic Distribution Hide

Legend:

0 no o

Sequences & Statistics

PFam domains (graphic)

PFam domains (details)

MSA

Cluster graph

MUSCLE (3.7) multiple sequence alignment

```

otau|estExt_fggeneshl_pg.C_Chrr_06|-----MSRP-----HSDGCLA-----RTVDDDDLD-----
osai|NP_001052931|MAAAAAAATVEAVGVAGGRRRR-----SGSVAGDLLRREASAEASASASAGAGGRERE
osai|NP_001047178|-----MSTRKSKVVPAGGGAATVPLAVLLRREVVSSEKTAEE-----
osai|NP_001050291|-----MEVQGEVD-----GSGVPLAVLLKRELNCQKVE-----
rcou|30170.m013899|-----MMAARE-----HQTVPVSVLLKRELNERVE-----
atha|NP_564504|-----MSTKGE-----HHTVPLSVLLKRESANEKID-----
ppat|e.gwl.29.62.1|-----MVPLASLLAYELKNEIVE-----
atha|NP_001031252|-----MASREGKRRNNHDDKLVPLAALISRETAKAKME-----
atha|NP_177008|-----MASREGKRRNNHDDKLVPLAALISRETAKAKME-----
rcou|30066.m000733|-----MASGEGRCRR-----DNLVPLAALISREMKNEKME-----
                                     .*.

otau|estExt_fggeneshl_pg.C_Chrr_06|---ASIVARSRKVKQGEDFDLSVPALTTPRRATGEDAPAV-----MFLGYLGLFDGHGG
osai|NP_001052931|RRPSVAAGQACRAKKGEDFALLKFCERLPAGGA-----PFSAFALFDGHNG
osai|NP_001047178|---RPELVGLFSGAKKGEDYTLFKPCDCELRPGVPS-----SFSAFGLFDGHNG
osai|NP_001050291|---RPFMLFGEASKKKGEDFTLLPKSRRPGQAQAGEDAGGAGDDDISVFAIFDGHNG
rcou|30170.m013899|---KPEILHGQASQSKKGEDFTLLKTECQVRVLGDGVS-----TYSVFGLFDGHNG
atha|NP_564504|---KPEILHGQASQSKKGEDFTLLKTECQVRVLGDGVS-----TYSVFGLFDGHNG
ppat|e.gwl.29.62.1|---NPLRLCGLALQPRKGEDFALVKTDCQIRPGDGSS-----TPAVFGIFDGHNG
atha|NP_001031252|---KPIVRFQGAASQSKKGEDYVLITKDSLVRVPSNST-----AFSVFVAFDGHNG
atha|NP_177008|---KPIVRFQGAASQSKKGEDYVLITKDSLVRVPSNST-----AFSVFVAFDGHNG
rcou|30066.m000733|---KPTVRCGSAASQSKKGEDYVLITKDCQVRVPSNST-----TFSVFVAFDGHNG
                                     . . . . .
                                     . . . . .

otau|estExt_fggeneshl_pg.C_Chrr_06|KACASHCAETTFAGEVTRGLDGDALREDEGADAFERRIFEALRRAFKVVDFTFVAMDVH
osai|NP_001052931|SGAAVYAKENILSNVMCVPAD---LSGDE-----WLAALPRALVAGVKTDXKDFQTR---
osai|NP_001047178|NGAAIYTKENLLSNLITAIAPAD---LNRDE-----WLAALPRALVAGVKTDXKDFQTR---
osai|NP_001050291|SAAAIYTKENLLSNVLAAMPD---LNRDE-----WLAALPRALVAGVKTDXKDFQTR---
rcou|30170.m013899|SAAAIYTKENLLSNVLAAMPD---LNRDE-----WLAALPRALVAGVKTDXKDFQTR---
atha|NP_564504|SAAAIYTKENLLSNVLAAMPD---LNRDE-----WLAALPRALVAGVKTDXKDFQTR---
ppat|e.gwl.29.62.1|SAAAIYTKENLLSNVLAAMPD---LNRDE-----WLAALPRALVAGVKTDXKDFQTR---
atha|NP_001031252|KAAAVYTRNLLNHVLSALPSG---LSRDE-----WLAALPRALVAGVKTDXKDFQTR---
atha|NP_177008|KAAAVYTRNLLNHVLSALPSG---LSRDE-----WLAALPRALVAGVKTDXKDFQTR---
rcou|30066.m000733|NAAAIYTRNLLNHVLSALPSG---LSRDE-----WLAALPRALVAGVKTDXKDFQTR---
                                     ...* . * . . . . .
                                     . . . . .

otau|estExt_fggeneshl_pg.C_Chrr_06|SGSTATVCAVRGRMVTTAAVGDLSATLDLPGPILVRLSVEHRLDSSSEKRIIEAGGE
osai|NP_001052931|---GTTVTFFVIIDGVVTVASVGDSCRVLAE---SYYLHSAHHRFDASKEEVGRVTECGGE
osai|NP_001047178|SCTTTVTFFVIIDGVVTVASVGDSCRVLAE---SYYLHSAHHRFDASKEEVGRVTECGGE
osai|NP_001050291|SCTTTVTFFVIIDGVVTVASVGDSCRVLAE---SYYLHSAHHRFDASKEEVGRVTECGGE
rcou|30170.m013899|SCTTTVTFFVIIDGVVTVASVGDSCRVLAE---SYYLHSAHHRFDASKEEVGRVTECGGE
atha|NP_564504|SCTTTVTFFVIIDGVVTVASVGDSCRVLAE---SYYLHSAHHRFDASKEEVGRVTECGGE
ppat|e.gwl.29.62.1|SCTTTVTFFVIIDGVVTVASVGDSCRVLAE---SYYLHSAHHRFDASKEEVGRVTECGGE

```

5. **Exploring a specific OrthoMCL group - examining the cluster graph.** (Use <http://orthomcl.org> for this exercise).

- Visit the orthomcl group OG5\_127676. You can either type the ID in the group quick search option at the top of the page or follow this link: [http://orthomcl.org/group/OG5\\_127676](http://orthomcl.org/group/OG5_127676)

- b. *Examine the “Sequences & Statistics” tab:* Based on the EC description and the product descriptions of the members of this group, what kind of a protein does this group represent? What is the phylogenetic distribution of the members of this group?
- c. *Examine the “PFam Domains (graphic)” tab:* How many PFam domains are represented in this group? What is the most common one? Which one is the least common one?
- d. *Examine the “Cluster Graph” tab:* Modify the E-value cutoff slider. What happens when you increase the E-value? What happens when you decrease the E-value? Can you identify subclusters?

