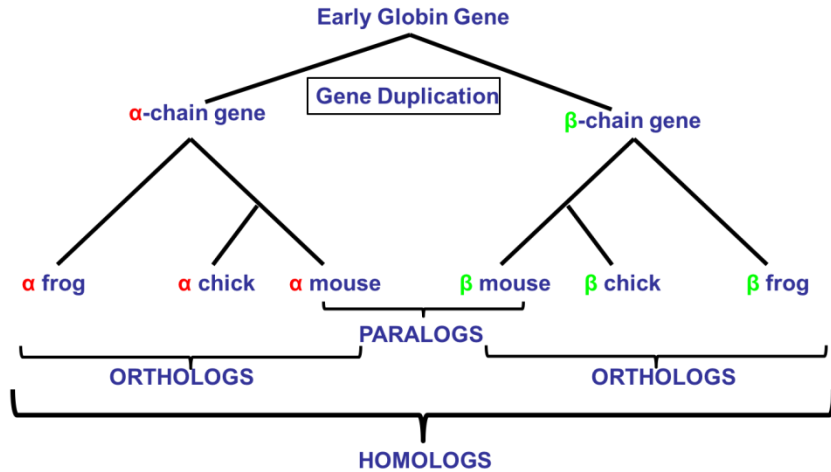


Orthology and Phyletic Patterns

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.
- 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: click on the Ortholog Group link above the table).

7 Orthology and synteny

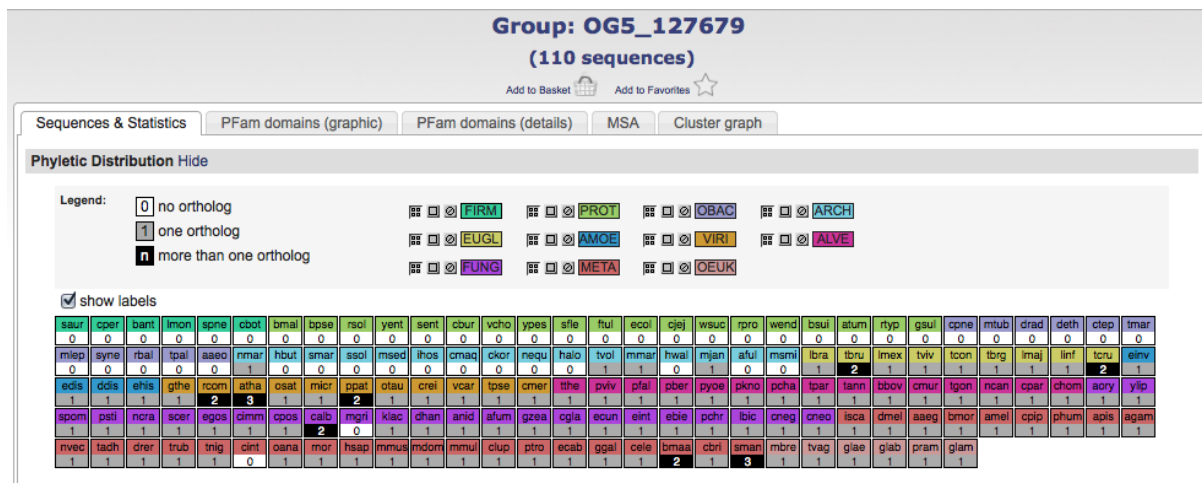
Ortholog Group [OG5_127679](#)

Orthologs and Paralogs within CryptoDB [Data sets](#)

Search this table... Showing 9 rows

Gene	Organism	Product	is syntenic	has comments
CHUDEA7_2290	Cryptosporidium hominis UdeA01	unspecified product	yes	no
CMU_034340	Cryptosporidium muris RN66	hypothetical protein, conserved	yes	no
ChTU502y2012_407g1140	Cryptosporidium hominis isolate TU502_2012	Fcf1	yes	no
Chro.70261	Cryptosporidium hominis TU502	hypothetical protein	yes	no
cand_030400	Cryptosporidium andersoni isolate 30847	hypothetical protein	yes	no
cubi_02904	Cryptosporidium ubiquitum isolate 39726	hypothetical protein	yes	no
Cvel_467	Chromera velia CCMP2878	rRNA-processing protein FCF1 homolog, putative	no	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

- d. 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:

OrthoMCL DB Version 5 10 May 13 A EuPathDB Project

Groups Quick Search: synth Sequences Quick Search: synth

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

Similarity/Pattern

Identify Protein Sequences

Text, IDs

Sequence ID(s)

Group ID(s)

Text Terms

Function

PFam ID or Keyword

Enzyme Commission Assignment

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: ISACT=OT AND ARCH=OT

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

☒ Root (ALL):

☒ Bacteria (BACT):

☒ Archaea (ARCH):

☒ Eukaryota (EUKA):

Get Answer

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

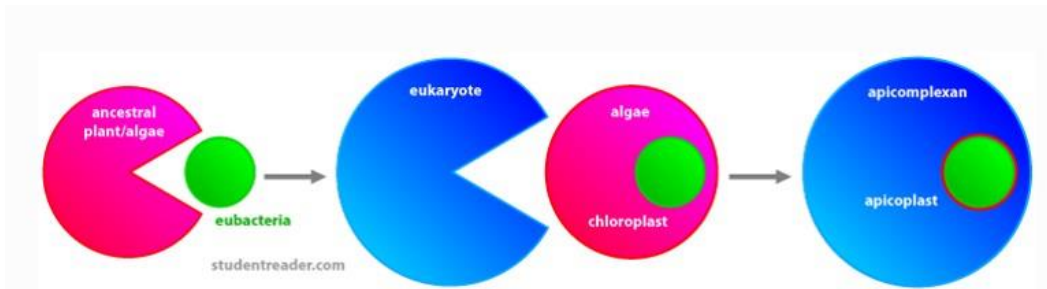
go to the “Phyletic Pattern” search in the Evolution section of the “Identify Ortholog Groups” category). To specify a phyletic pattern click on the icon next to the taxonomic group or species to include or exclude it.

- 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.

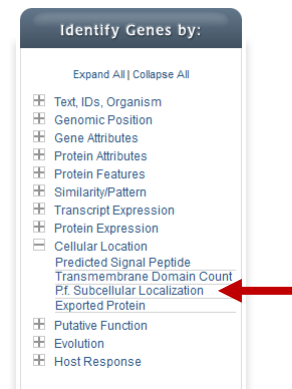
3. Using the orthology transform tool to identify apicomplast targeted genes in *Toxoplasma* and *Neospora*.

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

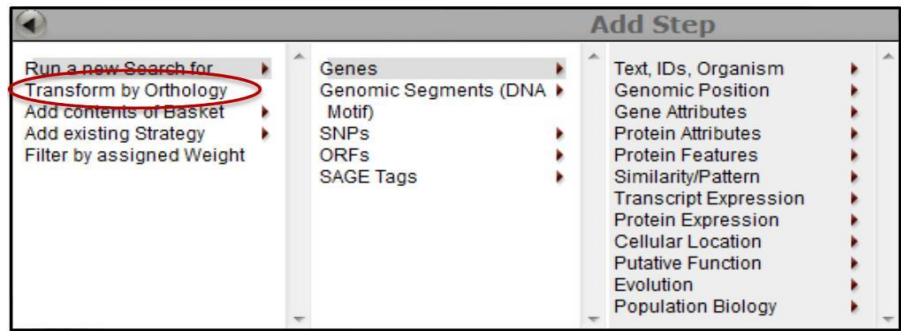


The apicomplast likely became encased in four membranes via a double endosymbiotic event. The chloroplast arose by engulfment of a cyanobacteria by a plant/algae ancestor. An algae was then engulfed by the ancestor of all apicomplexans. Thus an apicomplast organelle arose with four membranes.

- a. Start by finding genes in *Plasmodium* that are predicted to target to the apicomplast. Hint: click on “Cellular Location” then on “P.f. Subcellular Localization”.
- b. Transform the results of the above search to their *Toxoplasma* orthologs.

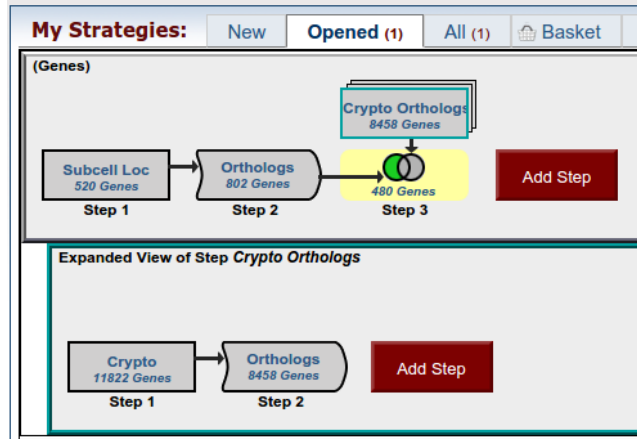


Hint: add a step, then select “Transform by Orthology”. On the search page, select all *Toxoplasma* and *Neospora*.



- 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.



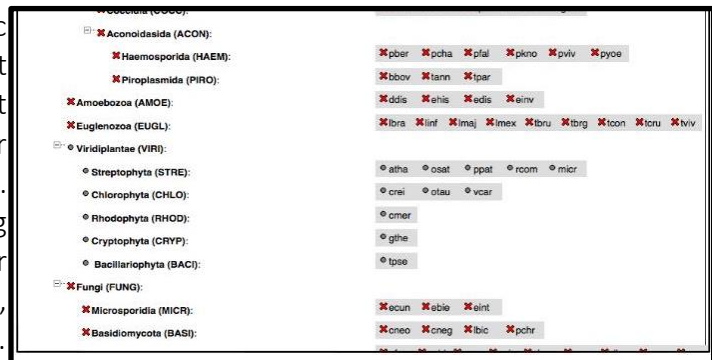
4. Combining searches in OrthoMCL (Use <http://orthomcl.org> for this exercise).

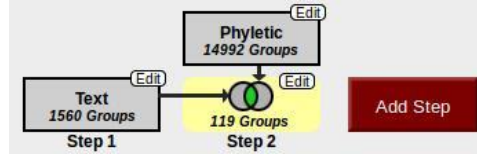
Find all plant proteins that are likely phosphatases that do not have orthologs outside of plants.

- 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).



- 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. 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)

Sequences & Statistics | PFam domains (graphic) | PFam domains (details) | **MSA** | Cluster graph

Phyletic Distribution Hide

MUSCLE (3.7) multiple sequence alignment

```

otau|estExt_fgernes1_pg_C_Ch1_06|-----MSRF-----HSDGFLA----RTVDDDDLD-----
osat|NP_001052931|MAAAAAATVEAVGVAGRRR-----SGSVLGDLLRRESASASASAGGRERE
osat|NP_001047178|-----MSTRSKVPVPGGGAATVPLAVLRREVVSERKTAE-----
osat|NP_001050291|-----MVGQVEV-----GQVPLAVLLKRELCKQVE-----
rcm|30170.m013899|-----MNAARE-----HQTVFLSVLLKRESANERIE-----
atha|NP_564504|-----MSTKGE-----HHTVFLSVLLKRESANEKID-----
ppat|e_gw1.29.62.1|-----MVPLASLAVELKNEIVE-----
atha|NP_001031252|-----MASREGKRRNHHDDEKLVLALISRETKAKME-----
atha|NP_177008|-----MASREGKRRNHHDDEKLVLALISRETKAKME-----
rcm|30066.m000733|-----MASGEGRCRR-----DNLVFLALISREMKMEKME
                          *
    ---ASIVARSRKVQGEDFDLSVPALTWTTPRATGEDAPAV-----MFLYLDFDGHGG
osat|NP_001052931|RRPSVAAGQACRAKGEDFALLKPCACERLPAGGA-----PFSAFALFDGHNG
osat|NP_001047178|---RPELVGLFSQAKGEDYTLFKPCERLPGVPS-----SFSAPGLFDGHNG
osat|NP_001050291|---RPMFLGEASQSKGEDFTLLPKCSRRPGQAGDEGAGGDDTI SVFAFDGHNG
rcm|30170.m013899|---KPEILHGQASQSKGEDFTLLKTECQRVLDGVS-----TYSVFLFDGHNG
atha|NP_564504|---NPELHGQHQSKKGEDFTLVKTECQRVMGDGVT-----TFSVFLFDGHNG
ppat|e_gw1.29.62.1|---NPLRLGLALQPRGEDFALVKTDCRIPDGSS-----TFAVGIFDGHNG
atha|NP_001031252|---KPIVRFQQAQSKKEDYVLKTDLSRVPSNST-----AFSVFAFDGHNG
atha|NP_177008|---KPIVRFQQAQSKKEDYVLKTDLSRVPSNST-----AFSVFAFDGHNG
rcm|30066.m000733|---KPIVRFQQAQSKKEDYVLKTDLSRVPSNST-----AFSVFAFDGHNG
                          :.*****:
    KACASHC AETTFAGEVTRGLDGDGALREDEDAGDAFERRIPEALRRAFVVDFTVAMDVH
osat|NP_001052931|SGAAVYAKENILSNMCCVPAD--LSGDE-----WLAALPRALVAGFKTDKDFQTR--
osat|NP_001047178|NGAAIYTKENLLSNILTAIPAD--LNKDE-----WLAALPRALVAGFKTDKDFQTKARS
osat|NP_001050291|SAAAIYTKENLLSNVLAIPFN--LTSQE-----WTALPRALVAGFKTDKDFQTKAAR
rcm|30170.m013899|SAAAIYTKENLLSNVLAAMPFD--LNKDE-----WVAALPRALVAGFKTDKDFQMRAGT
atha|NP_564504|SAAAIYTKENLLSNVLAIPSD--LNKDE-----WVAALPRALVAGFKTDKDFQERART
ppat|e_gw1.29.62.1|SAAAIYTKENLLSNVMSALSPG--LKRDD-----WLAALPRALVAGFKTDKDFQAKGRT
atha|NP_001031252|KAAAVYTKENLLSNVLSALPSG--LSRDE-----WHLALPRALVAGFKTDKDFQSRGET
atha|NP_177008|KAAAVYTKENLLSNVLSALPSG--LSRDE-----WHLALPRALVAGFKTDKDFQSRGET
rcm|30066.m000733|NAAAIYTKENLLSNVLSALPRG--LGRDE-----WQALPRALVAGFKTDKDFQSRGET
      *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *
    SGSTATVCVRGMVTTAAVGDLSLATLGLGPIVLRLSVREHLLDSSERRRIEAGGE
osat|NP_001052931|---GTVTFVIIDGVVTVASVGDSCRVLQ-AEG-TIYHLSADHRFDASEEVRVTECGGE
osat|NP_001047178|SGTTFVTVIIDGLITVASVGDSCRVLQ-AEG-SITHLSADHRFDASKEEVRVTECGGD
osat|NP_001050291|SGTTFVTVIIDGWVTVASVGDSCRILESAEG--SVYLSADHRLECNVEERERTASGGE
rcm|30170.m013899|SGTTFVTVIEGWVTVASVGDSCRILESAEG--DVYLSADHRLECNVEERERTASGGE
atha|NP_564504|SGTTFVTVIIEGWVTVASVGDSCRILESAEG--GTYLSADHRLEINNEERERTASGGE
ppat|e_gw1.29.62.1|SGTTFVTVIIEGWVTVASVGDSCRILESAEG--VYTDLTVDRHLLDNEERERTASGGE

```

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

