Orthology and Phyletic Patterns
Exercise 9

1. **Getting to OrthoMCL from EuPathDB databases**
   Note: For this exercise use http://cryptodb.org and http://orthomcl.org

   a. Go to the gene page for the *Cryptosporidium parvum* gene with the ID: cgd7_2290.
   b. What does this gene do? It is annotated as a hypothetical protein!
   c. 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).

   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/

   a. How many protein groups in OrthoMCL <u>do not</u> 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 the icon next to the taxonomic group or species to include or exclude it.

   b. How many protein groups <u>do not</u> 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.

2.1. 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.
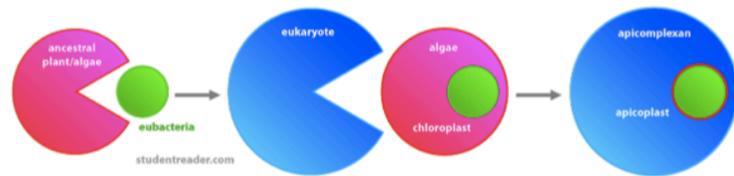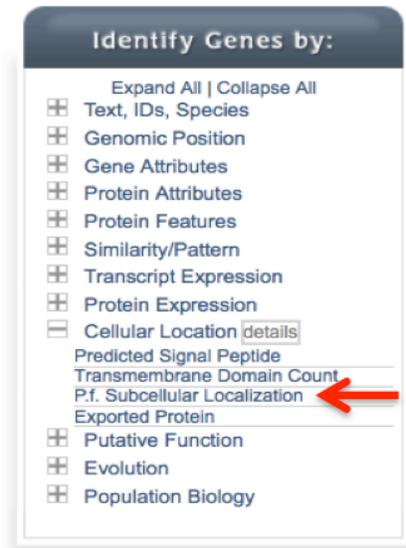
b. Transform the results of the above search to their *Toxoplasma* orthologs. Hint: add a step, then select "Transform by Orthology". On the search page, select all *Toxoplasma* and *Neopspora*.
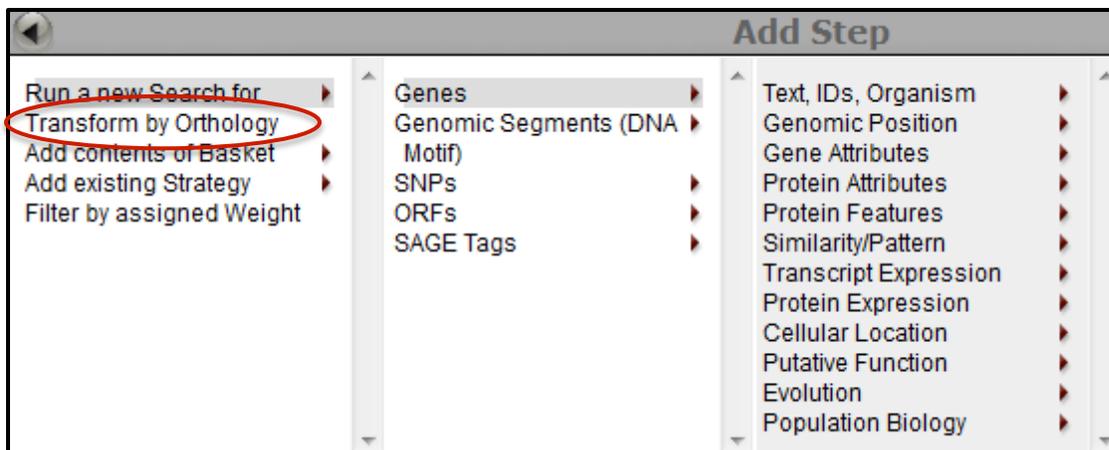
c. Although *Cryptosporidium* is an apicomplexan parasite it has actually lost its apicoplast! Can you use this fact to refine your results from the above search?

Hint: try subtracting out any orthologs present in *Cryptosporidium*. You will need to use a nested strategy.
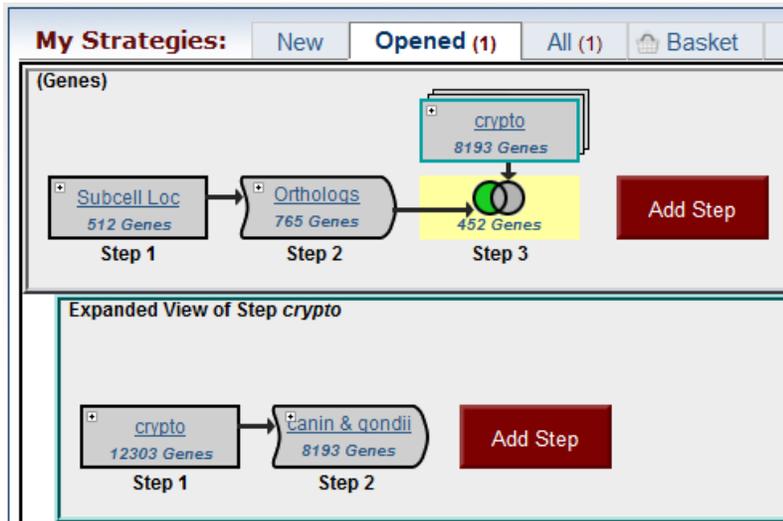
4. **Combining searches in OrthoMCL …. starts on the next page.**

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 groups that contain the word "*phosphatase*".

    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)

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_fgenesh1_pg.C_Chr_06   --------------MSRP-----------HSDGPLA----RTVDDDDLD-------------
osat NP_001052931                  MAAAAAAATVEAVGVAGGRRRR------SGSVALGDLLRREASAERASASASAGAGGRERE
osat NP_001047178                  -------------MSTRSKSVPVPAGGGAATVPLAVLLRREVVSEKTAAE-----------
osat NP_001050291                  -----------MEVQGEVD--------GSGVPLAVLLKRELCNQKVE-------------
rcom 30170.m013899                 -----------MMAARE----------HQTVPLSVLLKRELANERVE-------------
atha NP_564504                     -----------MSTKGE----------HHTVPLSVLLKRESANEKID-------------
ppat e_gw1.29.62.1                 --------------------------MVPLASLLAYELKNEIVE---------------
atha NP_001031252                  ------------MASREGKRRNHNHDDEKLVPLAALISRETKAAKME-------------
atha NP_177008                     ------------MASREGKRRNHNHDDEKLVPLAALISRETKAAKME-------------
rcom 30066.m000733                 ------------MASGEGRCRR-----DNLVPLAALISREMKNEKME-------------
                                                                    .*.

otau estExt_fgenesh1_pg.C_Chr_06   ---ASIVARSRKVQKGEDFDLSVPALTWTPRRATGEDAPAV------MFGLYGLFDGHGG
osat NP_001052931                  RRPSVAAGQACRAKKGEDFALLKPACERLPAGGA-------------PFSAFALFDGHNG
osat NP_001047178                  -RPELQVGLFSQAKKGEDYTFLKPDCERLPGVPSS------------SFSAFGLFDGHNG
osat NP_001050291                  -RPDMLFGEASKSKKGEDFTFLLPKCSRRPGQAQADGEDAGGAGDDDTISVFAIFDGHNG
rcom 30170.m013899                 -KPEILHGQASQSKKGEDFTLLKTECQRVLGDGVS------------TYSVFGLFDGHNG
atha NP_564504                     -NPELIHGQHNQSKKGEDFTLVKTECQRVMGDGVT------------TFSVFGLFDGHNG
ppat e_gw1.29.62.1                 -NPLLRCGLALQPRKGEDFALVKTDCQRIPGDGSS------------TFAVFGIFDGHNG
atha NP_001031252                  -KPIVRFGQAAQSRKGEDYVLIKTDSLRVPSNSST------------AFSVFAVFDGHNG
atha NP_177008                     -KPIVRFGQAAQSRKGEDYVLIKTDSLRVPSNSST------------AFSVFAVFDGHNG
rcom 30066.m000733                 -KPTVRCGSAAQSRKGEDYFLIKTDCQRVPGNSSS------------TFSVFAIFDGHNG
                                    .  . .****: :  .  .           : .:.:****.*

otau estExt_fgenesh1_pg.C_Chr_06   KACASHCAETFAGEVTRGLDGDGALREDEDAGDAFERRIPEALRRAFKVVDETFVAMDVH
osat NP_001052931                  SGAAVYAKENILSNVMCCVPAD--LSGDE-----WLAALPRALVAGFVKTDKDFQTR---
osat NP_001047178                  NGAAIYTKENLLSNILTAIPAD--LNRED-----WLAALPRAMVAAFVKTDKDFQTKARS
osat NP_001050291                  SAAAIYTRENLLNNVLAAIPPN--LTSEE-----WTTALPRALVAGFVKTDKEFQTKAAR
rcom 30170.m013899                 SAAAIYTKENLLNNVIAAMPPD--LNRDE-----WVAALPRALVAGFVKTDKDFQMRAQT
atha NP_564504                     SAAAIYTKENLLNNVLAAIPSD--LNRDE-----WVAALPRALVAGFVKTDKDFQERART
ppat e_gw1.29.62.1                 SAAAIYTKENLLNDVMSALSPG--LKRDD-----WLAALPRALVAGFVKTDKEFQAKEQT
atha NP_001031252                  KAAAVYTRENLLNHVISALPSG--LSRDE-----WLHALPRALVSGFVKTDKEFQSRGET
atha NP_177008                     KAAAVYTRENLLNHVISALPSG--LSRDE-----WLHALPRALVSGFVKTDKEFQSRGET
rcom 30066.m000733                 NAAAIFTRENLLNHILGAIPRG--LGREE-----WLQALPRALVAGFVKTDKEFQSRGET
                                    ...* . *.: .: : . *  ::        : :* *: .* .*: *

otau estExt_fgenesh1_pg.C_Chr_06   SGSTATVCAVRGRMVTTAAVGDSLATLDLGPGIPVLRLSVEHRLDSSESERKRIEEAGGE
osat NP_001052931                  -GTTVTFVIIDGYVVTVASVGDSRCVLE-AEG-TIYHLSADHRFDASEEEVGRVTECGGE
osat NP_001047178                  SGTTVTFVIIDGLFITVASVGDSRCVLE-AEG-SIYHLSADHRFDASKEEVDRVTESGGD
osat NP_001050291                  SGTTVTFVIIDGWVVTVASVGDSRCILESAEG-SVYFLSADHRLDTNEEEVERVTASGGD
rcom 30170.m013899                 SGTTVTFVIIEGWVITVASVGDSRCILESAEG-DVYYLSADHRLECNVEEERRITASGGE
atha NP_564504                     SGTTVTFVIVEGWVVSVASVGDSRCILEPAEG-GVYYLSADHRLEINEEERDRVTASGGE
ppat e_gw1.29.62.1                 SGTTVTFVVIDGWTITVACVGDSRCVLDSQ-G-VVTDLTVDHRLDDNEEERERVRAAGGE
```