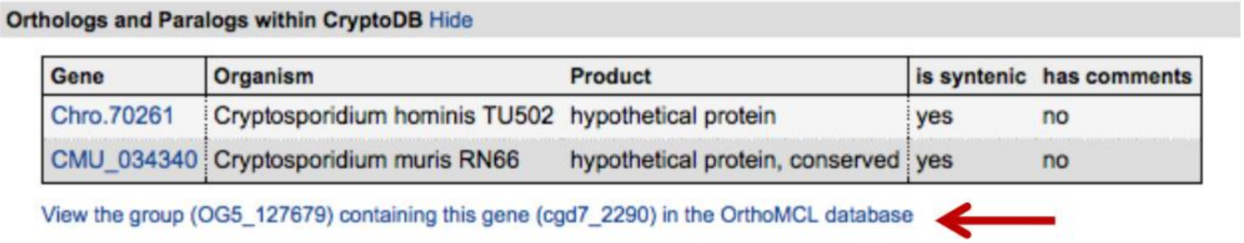
Orthology and Phyletic Patterns

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).

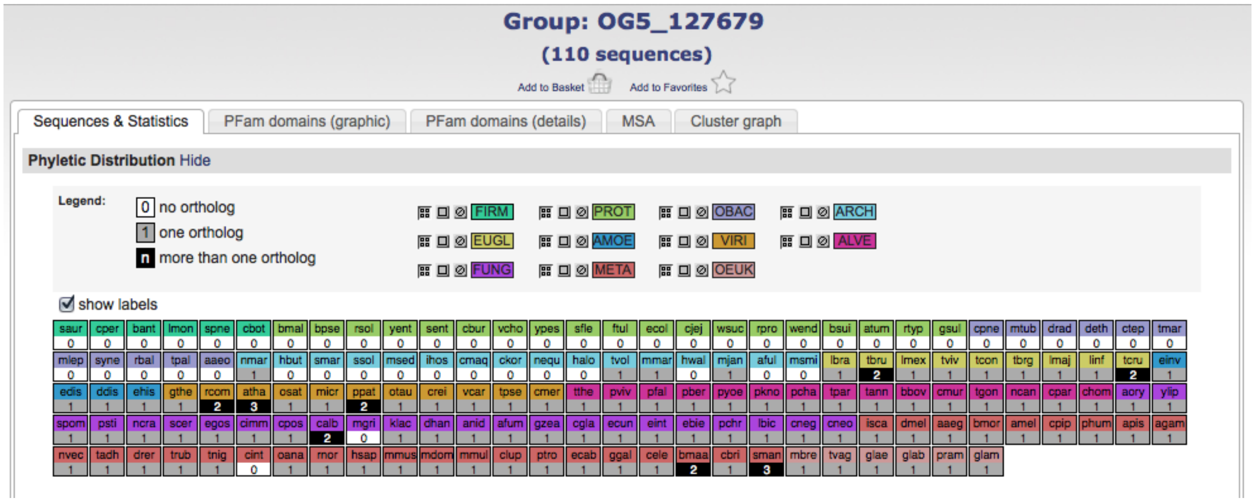


Orthologs and Paralogs within CryptoDB Hide

Gene	Organism	Product	is syntenic	has comments
Chro.70261	Cryptosporidium hominis TU502	hypothetical protein	yes	no
CMU_034340	Cryptosporidium muris RN66	hypothetical protein, conserved	yes	no

View the group (OG5_127679) containing this gene (cgd7_2290) in the OrthoMCL database

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).



Group: OG5_127679

(110 sequences)

Phyletic Distribution Hide

Legend:
0 no ortholog
1 one ortholog
n more than one ortholog

⊞ □ ⊘ FIRM ⊞ □ ⊘ PROT ⊞ □ ⊘ OBAC ⊞ □ ⊘ ARCH
⊞ □ ⊘ EUGL ⊞ □ ⊘ AMOE ⊞ □ ⊘ VIRI ⊞ □ ⊘ ALVE
⊞ □ ⊘ FUNG ⊞ □ ⊘ META ⊞ □ ⊘ OEUK

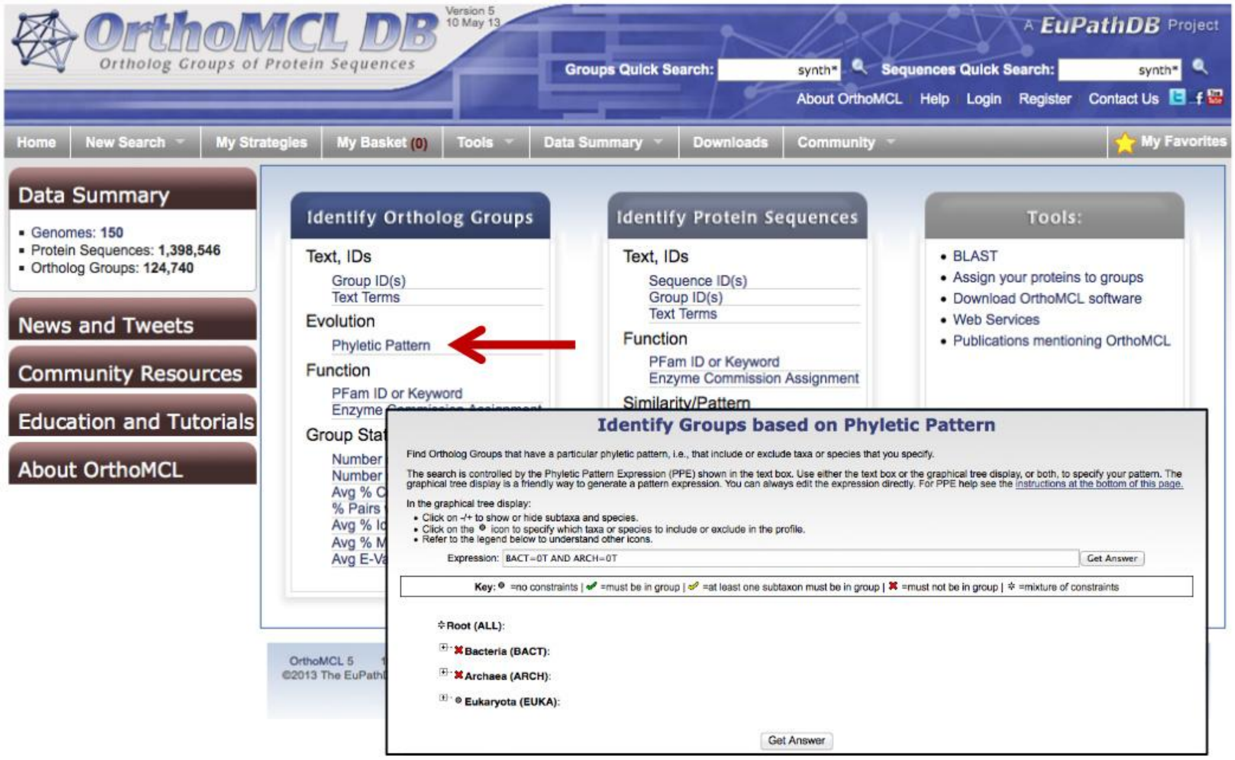
☑ show labels

saur 0	cper 0	bant 0	lmon 0	spne 0	cbot 0	bmal 0	bpse 0	rsol 0	yent 0	sent 0	cbur 0	vcho 0	ypes 0	sfle 0	ftul 0	ecol 0	cjej 0	wsuc 0	rpro 0	wend 0	bsui 0	atum 0	rtyp 0	gsul 0	cpne 0	mtub 0	drad 0	deth 0	ctep 0	tmar 0
mlep 0	syne 0	rbal 0	tpal 0	aaeo 0	nmar 1	hbut 0	smar 0	ssol 0	msed 0	ihos 0	cmaq 0	ckor 0	nequ 0	halo 0	tvol 1	mmar 1	hwal 0	mjan 1	aful 0	msmi 0	lbra 1	tbru 2	lmex 1	tviv 1	tcon 1	tbrg 1	lmaj 1	linf 1	tcru 2	einv 1
edis 1	ddis 1	ehis 1	gthe 1	rcom 2	atha 3	osat 1	micr 1	ppat 2	otau 1	crei 1	vcar 1	tpse 1	cmer 1	tthe 1	pviv 1	pfal 1	pber 1	pyoe 1	pkno 1	pcha 1	tpar 1	tann 1	bbov 1	cmur 1	tgon 1	ncan 1	cpar 1	chom 1	aory 1	ylip 1
spom 1	psti 1	ncra 1	scer 1	egos 1	cimm 1	cpos 1	calb 2	mgri 0	klac 1	dhan 1	anid 1	afum 1	gzea 1	cgla 1	ecun 1	eint 1	ebie 1	pchr 1	lbic 1	cneg 1	cneo 1	isca 1	dmel 1	aaeg 1	bmor 1	amel 1	cpip 1	phum 1	apis 1	agam 1
nvec 1	tadh 1	drer 1	trub 1	tnig 1	cint 0	oana 1	rnor 1	hsap 1	mmus 1	mdom 1	mmul 1	clup 1	ptro 1	ecab 1	ggal 1	cele 1	bmaa 2	cbri 1	sman 3	mbre 1	tvag 1	glae 1	glab 1	pram 1	glam 1					

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



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

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.

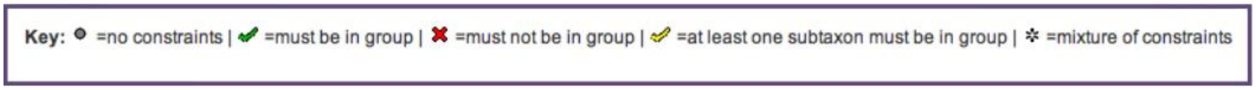
In the graphical tree display:
- Click on -/+ to show or hide subtaxa and species.
- Click on the ● icon to specify which taxa or species to include or exclude in the profile.
- Refer to the legend below to understand other icons.

Expression: BACT=0T AND ARCH=0T [Get Answer]

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):
⊞ ✖ 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

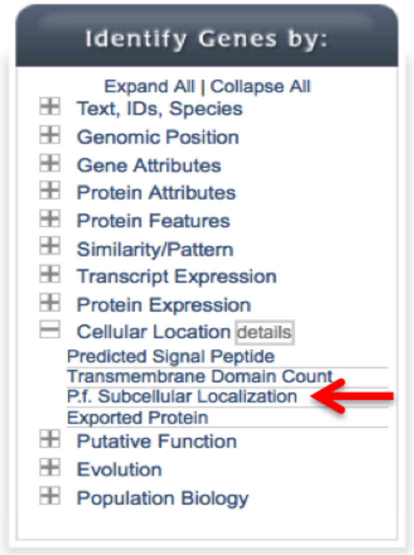
   b. How many protein groups do not contain orthologs from eukaryotes?

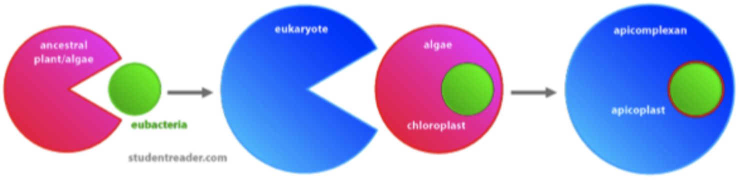
   c. Find all groups that contain orthologs from at least one species of *Cryptosporidium* and *Giardia* but not from bacteria or archaea.

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.

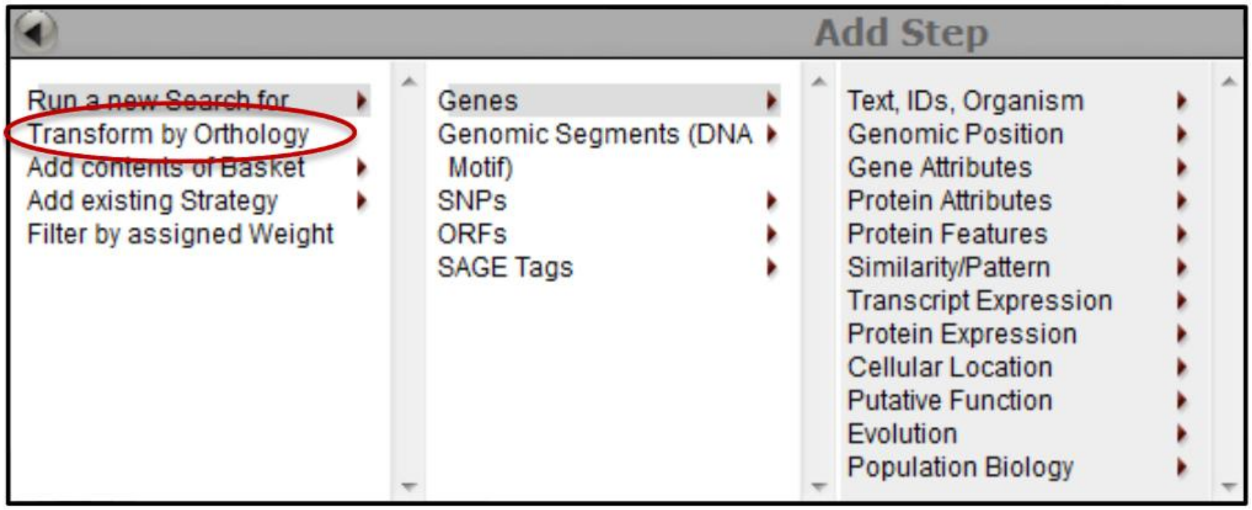
3. **Using the orthology transform tool to identify apicoplast targeted genes in *Toxoplasma* and *Neospora*.**
   Note: For this exercise use http://eupathdb.org

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.



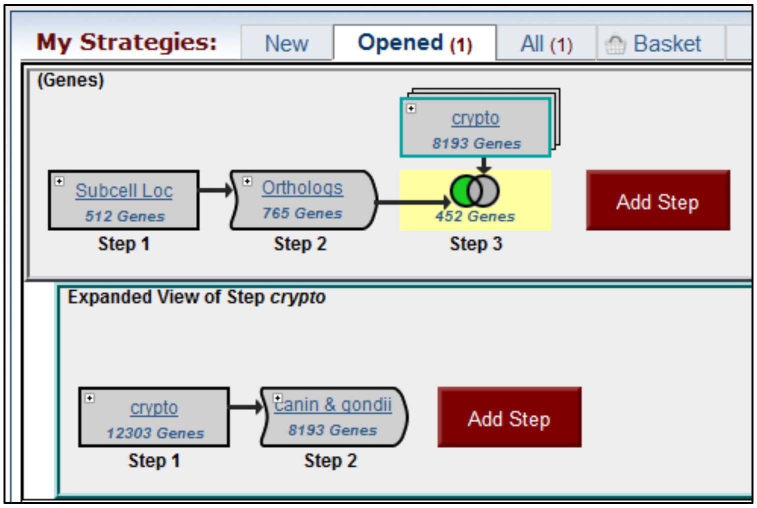


b. Transform the results of the above search to their *Toxoplasma* orthologs.
Hint: add a step, then select "Transform by Orthology". On the search page, select all *Toxoplasma* and *Neospora.*



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?

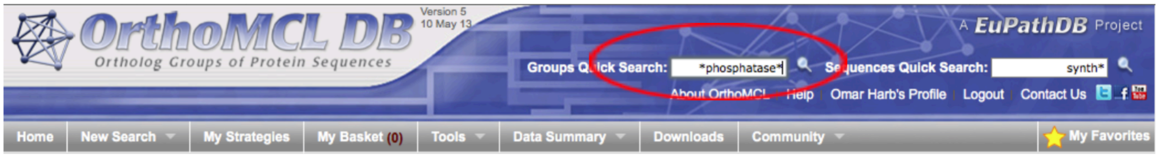
Hint: try subtracting out any orthologs present in *Cryptosporidium*. You will need to use a nested strategy.



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.

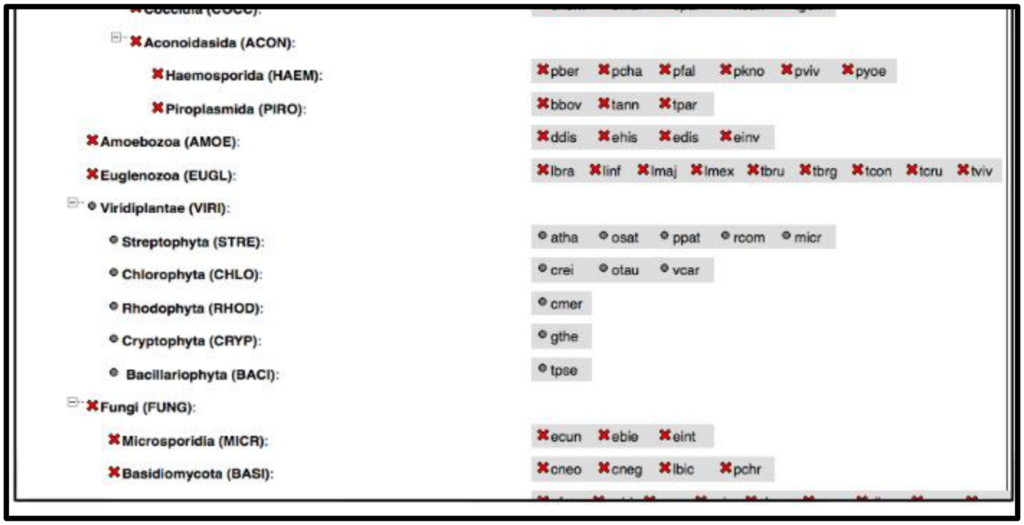
a. Use the text search to find OrthoMCL groups that contain the word "*phosphatase*" (note that the search should be run without the quotation marks but with the asterisks).



Groups Quick Search: *phosphatase*

Sequences Quick Search: synth*

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).



✖ Aconoidasida (ACON):

✖ Haemosporida (HAEM): ✖pber ✖pcha ✖pfal ✖pkno ✖pviv ✖pyoe

✖ Piroplasmida (PIRO): ✖bbov ✖tann ✖tpar

✖ Amoebozoa (AMOE): ✖ddis ✖ehis ✖edis ✖einv

✖ Euglenozoa (EUGL): ✖lbra ✖linf ✖lmaj ✖lmex ✖tbru ✖tbrg ✖tcon ✖tcru ✖tviv

○ Viridiplantae (VIRI):

○ Streptophyta (STRE): ○atha ○osat ○ppat ○rcom ○micr

○ Chlorophyta (CHLO): ○crei ○otau ○vcar

○ Rhodophyta (RHOD): ○cmer

○ Cryptophyta (CRYP): ○gthe

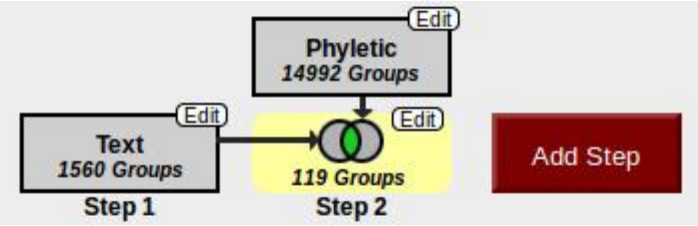
○ Bacillariophyta (BACI): ○tpse

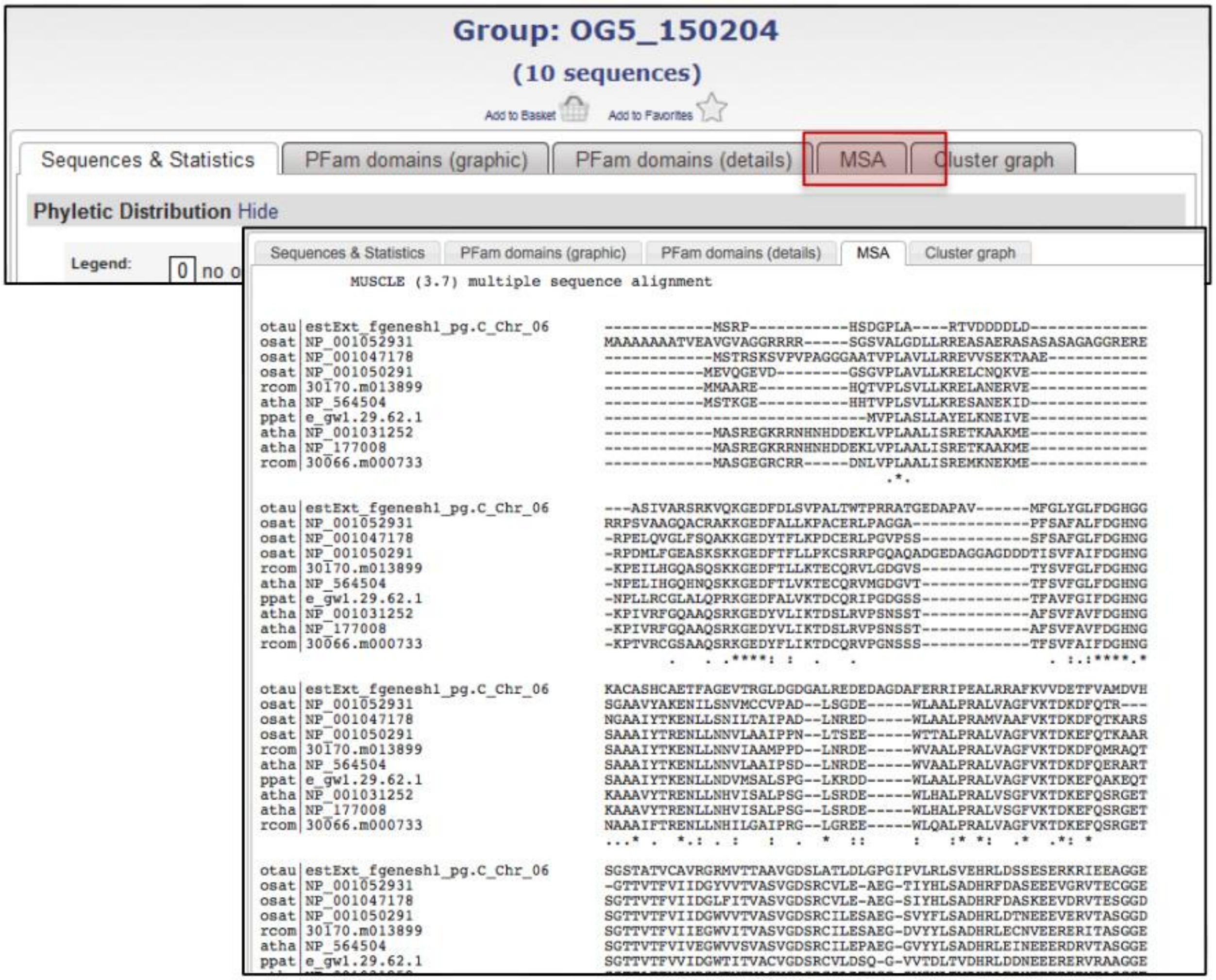
✖ Fungi (FUNG):

✖ Microsporidia (MICR): ✖ecun ✖ebie ✖eint

✖ Basidiomycota (BASI): ✖cneo ✖cneg ✖lbic ✖pchr

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)

```
        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-DVYYLSADHRLECNVEERERITASGGE
atha|NP_564504                        SGTTVTFVIVEGWVVSVASVGDSRCILEPAEG-GVYYLSADHRLEINEEERDRVTASGGE
ppat|e_gw1.29.62.1                    SGTTVTFVVIDGWTITVACVGDSRCVLDSQ-G-VVTDLTVDHRLDDNEEERERVRAAGGE
```

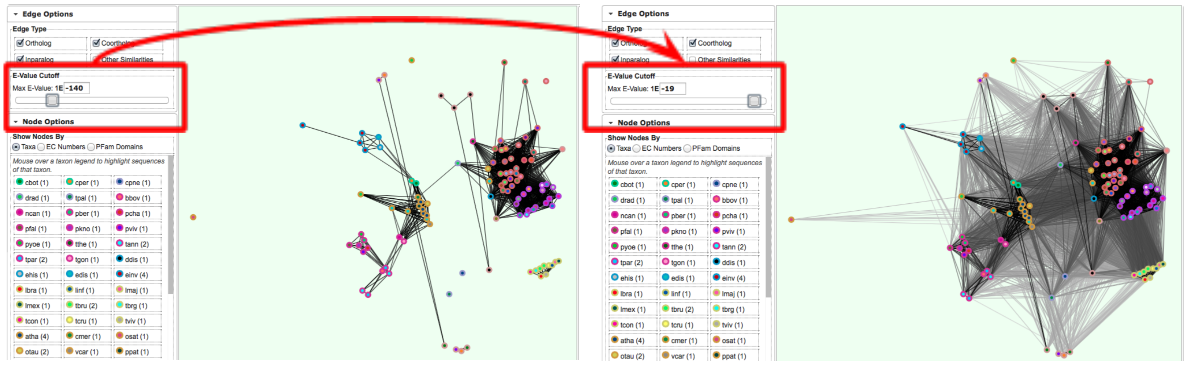
5. **Exploring a specific OrthoMCL group - examining the cluster graph.** (Use http://orthomcl.org for this exercise).

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 of 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 on? 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?



Edge Options

Edge Type

☑ Ortholog ☑ Coortholog

☑ Inparalog ☐ Other Similarities

E-Value Cutoff

Max E-Value: 1E -140

Node Options

Show Nodes By

● Taxa ○ EC Numbers ○ PFam Domains

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

● cbot (1)	● cper (1)	● cpne (1)
● drad (1)	● tpal (1)	● bbov (1)
● ncan (1)	● pber (1)	● pcha (1)
● pfal (1)	● pkno (1)	● pviv (1)
● pyoe (1)	● tthe (1)	● tann (2)
● tpar (2)	● tgon (1)	● ddis (1)
● ehis (1)	● edis (1)	● einv (4)
● lbra (1)	● linf (1)	● lmaj (1)
● lmex (1)	● tbru (2)	● tbrg (1)
● tcon (1)	● tcru (1)	● tviv (1)
● atha (4)	● cmer (1)	● osat (1)
● otau (2)	● vcar (1)	● ppat (1)

Edge Options

Edge Type

☑ Ortholog ☑ Coortholog

☑ Inparalog ☐ Other Similarities

E-Value Cutoff

Max E-Value: 1E -19

Node Options

Show Nodes By

● Taxa ○ EC Numbers ○ PFam Domains

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

● cbot (1)	● cper (1)	● cpne (1)
● drad (1)	● tpal (1)	● bbov (1)
● ncan (1)	● pber (1)	● pcha (1)
● pfal (1)	● pkno (1)	● pviv (1)
● pyoe (1)	● tthe (1)	● tann (2)
● tpar (2)	● tgon (1)	● ddis (1)
● ehis (1)	● edis (1)	● einv (4)
● lbra (1)	● linf (1)	● lmaj (1)
● lmex (1)	● tbru (2)	● tbrg (1)
● tcon (1)	● tcru (1)	● tviv (1)
● atha (4)	● cmer (1)	● osat (1)
● otau (2)	● vcar (1)	● ppat (1)