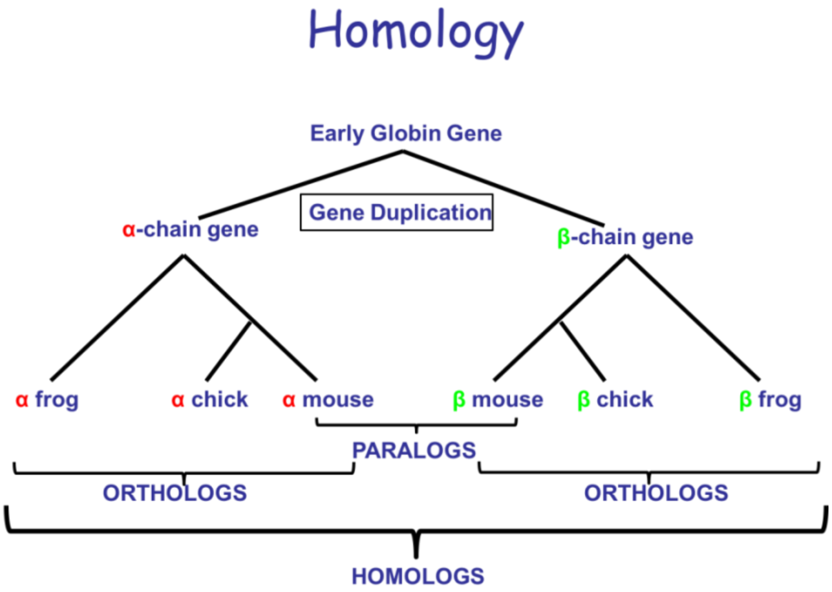
Orthology and Phyletic Patterns

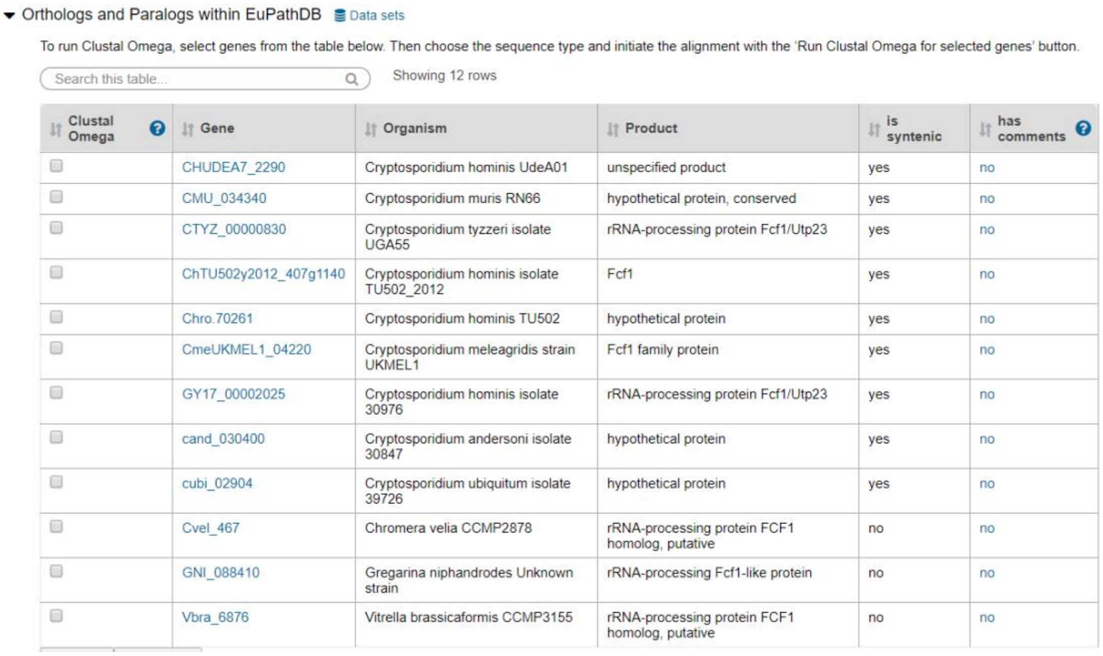


1. **Getting to OrthoMCL from EuPathDB databases**
   Note: For this exercise use http://cryptodb.org and http://orthomcl.org/

a. Go to the gene page for the *Cryptosporidium muris* gene with the ID: CMU_034340

b. 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 or take a look at InterPro domains.

c. 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).



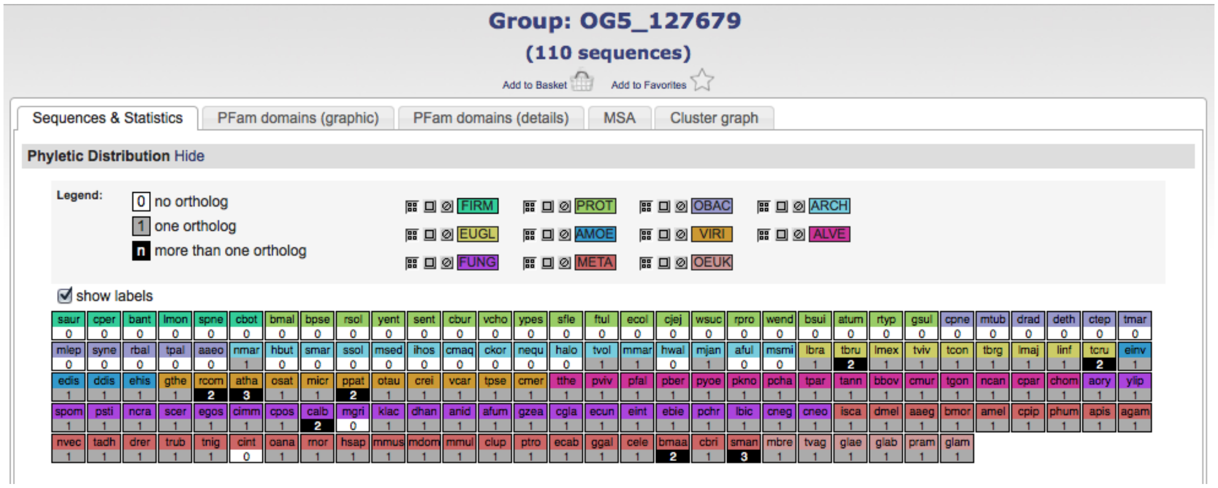
▾ Orthologs and Paralogs within EuPathDB Data sets

To run Clustal Omega, select genes from the table below. Then choose the sequence type and initiate the alignment with the 'Run Clustal Omega for selected genes' button.

Search this table... Showing 12 rows

Clustal Omega	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
□	CTYZ_00000830	Cryptosporidium tyzzeri isolate UGA55	rRNA-processing protein Fcf1/Utp23	yes	no
□	ChTU502y2012_407g1140	Cryptosporidium hominis isolate TU502_2012	Fcf1	yes	no
□	Chro.70261	Cryptosporidium hominis TU502	hypothetical protein	yes	no
□	CmeUKMEL1_04220	Cryptosporidium meleagridis strain UKMEL1	Fcf1 family protein	yes	no
□	GY17_00002025	Cryptosporidium hominis isolate 30976	rRNA-processing protein Fcf1/Utp23	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. What about orthologs in organisms not in EuPathDB? (hint: click on the Ortholog Group link above the table). 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).



Group: OG5_127679

(110 sequences)

Add to Basket Add to Favorites ☆

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

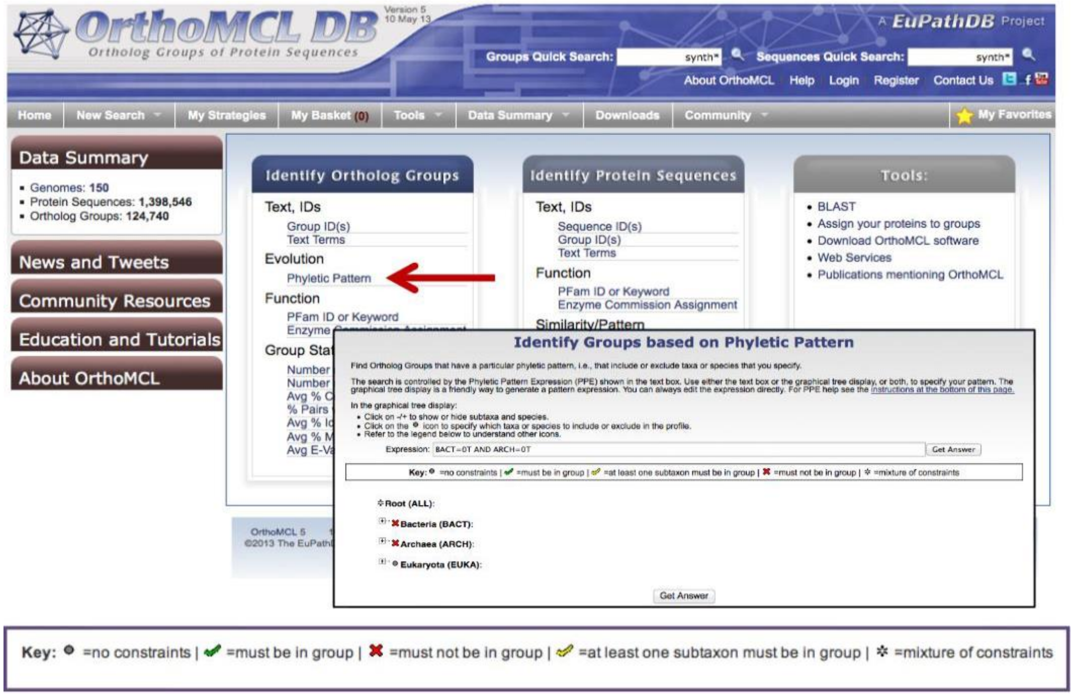
saur	cper	bant	lmon	spne	cbot	bmal	bpse	rsol	yent	sent	cbur	vcho	ypes	sfle	ftul	ecol	cjej	wsuc	rpro	wend	bsui	atum	rtyp	gsul	cpne	mtub	drad	deth	ctep	tmar
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
mlep	syne	rbal	tpal	aaeo	nmar	hbut	smar	ssol	msed	ihos	cmaq	ckor	nequ	halo	tvol	mmar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

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.

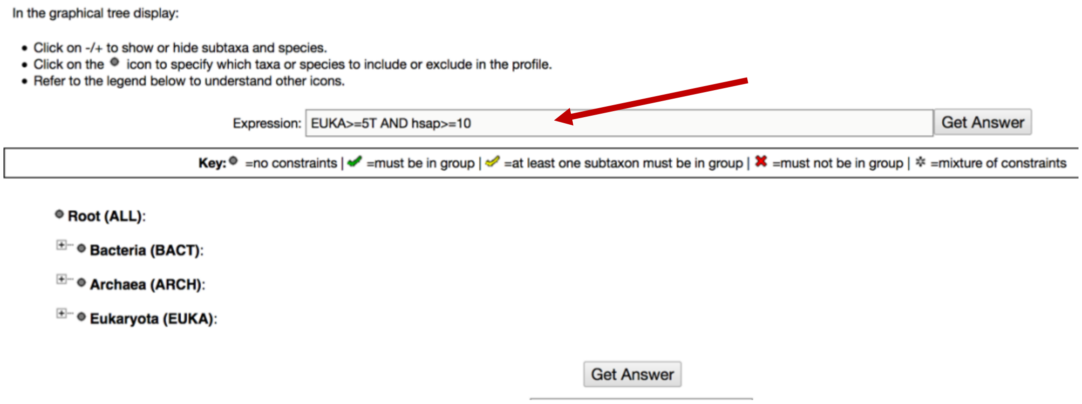
Expression: BACT=0T AND ARCH=0T

Key: ● =no constraints | ✓ =must be in group | ✓ =at least one subtaxon must be in group | ✖ =must not be in group | ✲ =mixture of constraints

Key: ● =no constraints | ✓ =must be in group | ✖ =must not be in group | ✓ =at least one subtaxon must be in group | ✲ =mixture of constraints

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. If you are getting frustrated trying to figure this one out you have a right to be! You cannot answer this question by using the check boxes (we will discuss why). However, OrthoMCL has an added feature that allows you to enter an expression to define the phyletic pattern. This option provides additional flexibility. Can

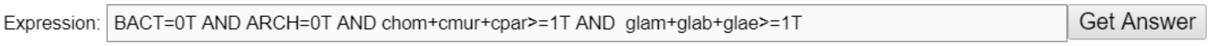


In the graphical tree display:
- Click on -/+ to show or hide subtaxa and species.
- Click on the ● icon to specify which taxa or species to include or exclude in the profile.
- Refer to the legend below to understand other icons.

Expression: EUKA>=5T AND hsap>=10

Key: ● =no constraints | ✓ =must be in group | ✓ =at least one subtaxon must be in group | ✖ =must not be in group | ✲ =mixture of constraints you figure out what expression to use to answer this question? (hint: scroll down to the bottom of the page to find additional information about expression parameters.

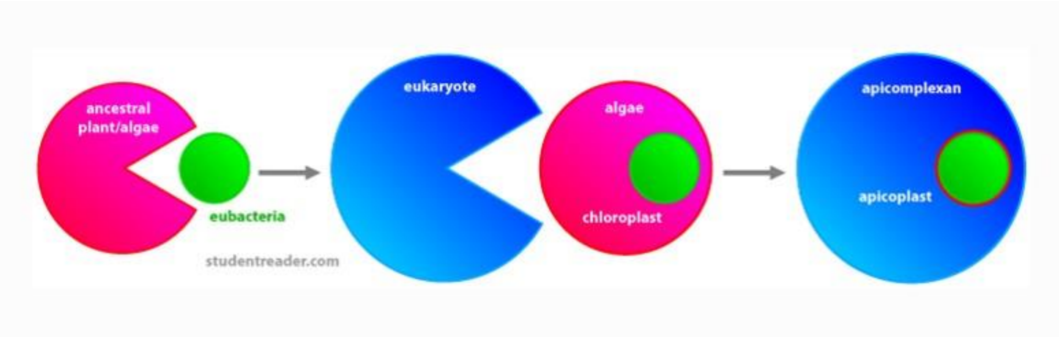
Before looking at the answer below, try this on your own or with the people sitting next to you.



Expression: BACT=0T AND ARCH=0T AND chom+cmur+cpar>=1T AND glam+glab+glae>=1T

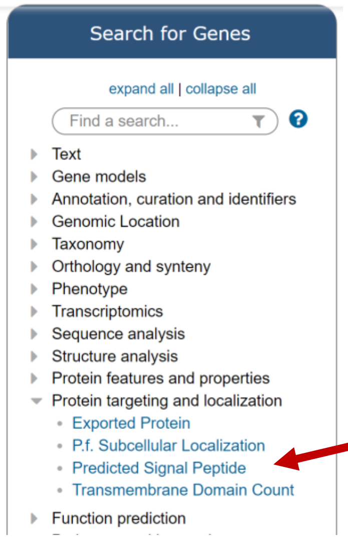
All EuPathDB sites also have a phyletic pattern search that uses OrthoMCL data under Genes -> Evolution -> Orthology Phylogenetic Profile. This search is very useful to identify genes in your organism of interest that are restricted in their profile. For example, you frequently want to identify genes 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.

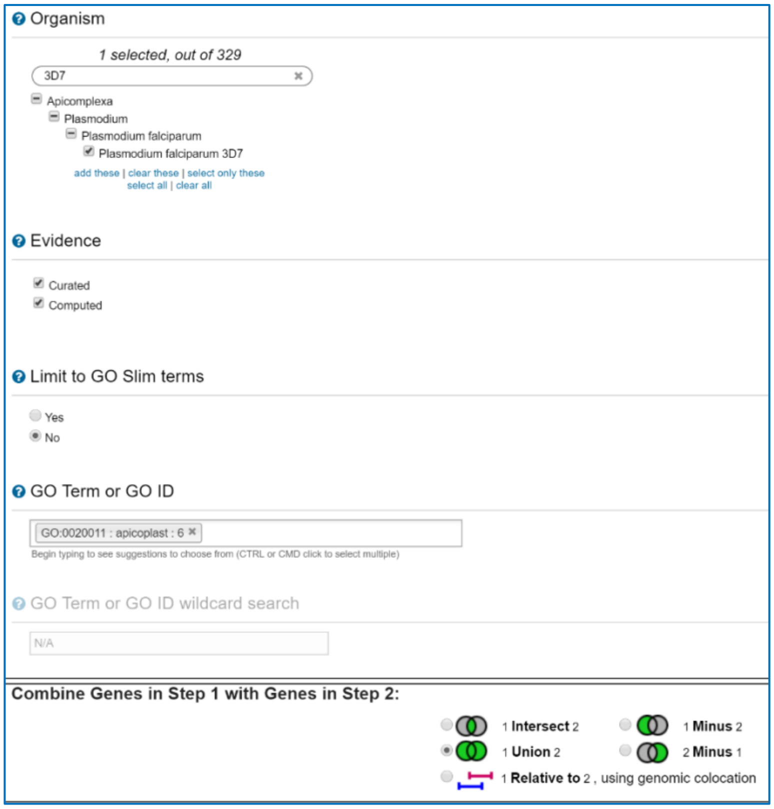
3. **Using the orthology transform tool to identify apicoplast targeted genes in *Toxoplasma* and *Neospora*.**
   Note: For this exercise use http://eupathdb.org



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 "Protein targeting and localization" then on "P.f. Subcellular Localization". You can stop with this list of apicoplast genes or you can union these results with a GO term search for GO:0020011, apicoplast : 6 in *P falciparum* 3D7





Organism

1 selected, out of 329

3D7 ×

⊟ Apicomplexa
⊟ Plasmodium
⊟ Plasmodium falciparum
☑ Plasmodium falciparum 3D7

add these | clear these | select only these
select all | clear all

Evidence

☑ Curated
☑ Computed

Limit to GO Slim terms

○ Yes
● No

GO Term or GO ID

GO:0020011 : apicoplast : 6 ×

Begin typing to see suggestions to choose from (CTRL or CMD click to select multiple)

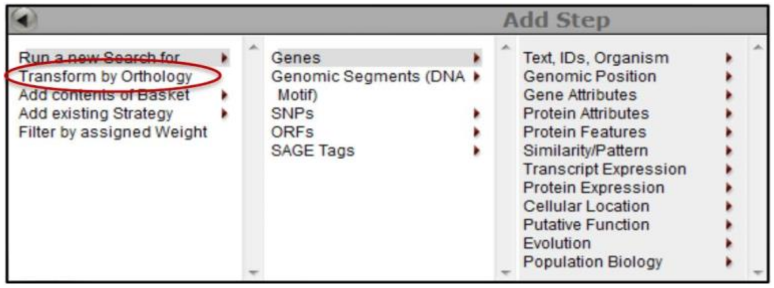
GO Term or GO ID wildcard search

N/A

Combine Genes in Step 1 with Genes in Step 2:

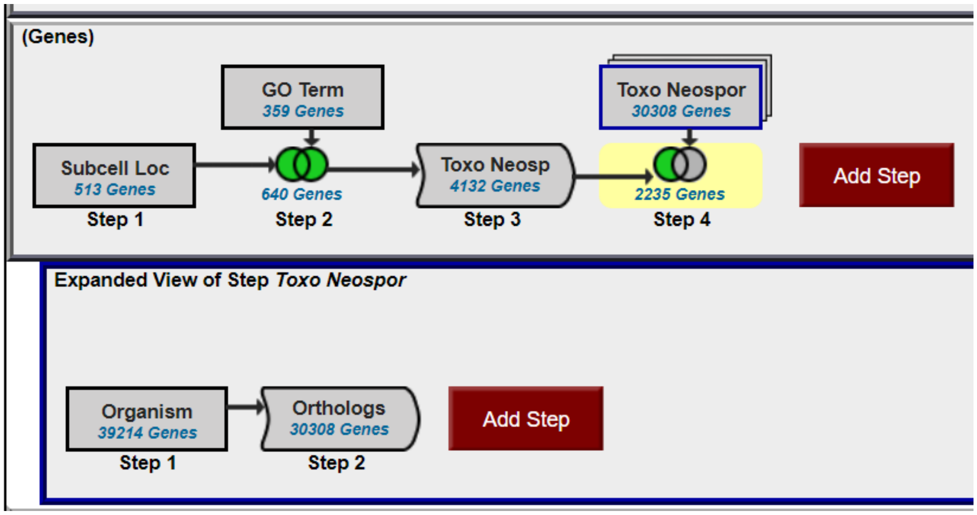
○ 1 **Intersect** 2 ○ 1 **Minus** 2
● 1 **Union** 2 ○ 2 **Minus** 1
○ 1 **Relative to** 2 , using genomic colocation

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



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

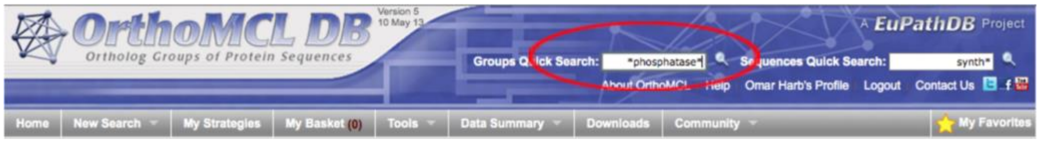
c. Although *Cryptosporidium* is an apicomplexan parasite it has actually lost its apicoplast! Can you use this fact to refine your results from the above search? Hint: try subtracting out any orthologs present in *Cryptosporidium*. You will need to use a nested strategy and use the ortholog transform back to Toxoplasma and Neospora genes for the subtraction to complete.



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.

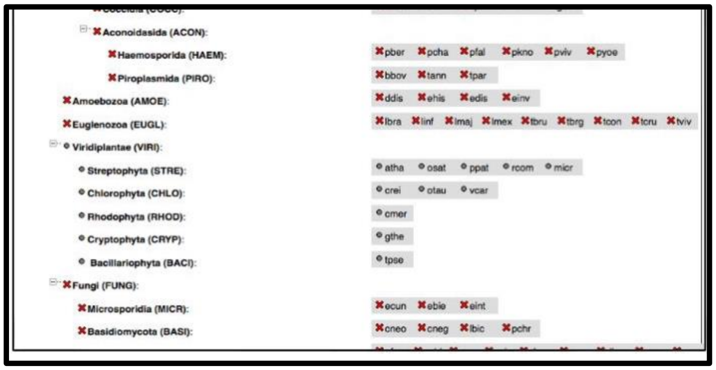
a. Use the text search **to find OrthoMCL groups** that contain the word “*phosphatase*” (note that the search should be run without the quotation marks but with the asterisks).



Groups Quick Search: *phosphatase*

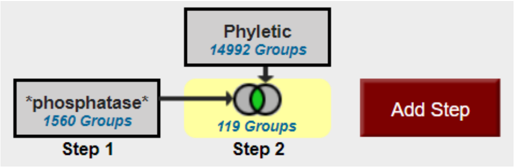
Sequences Quick Search: synth*

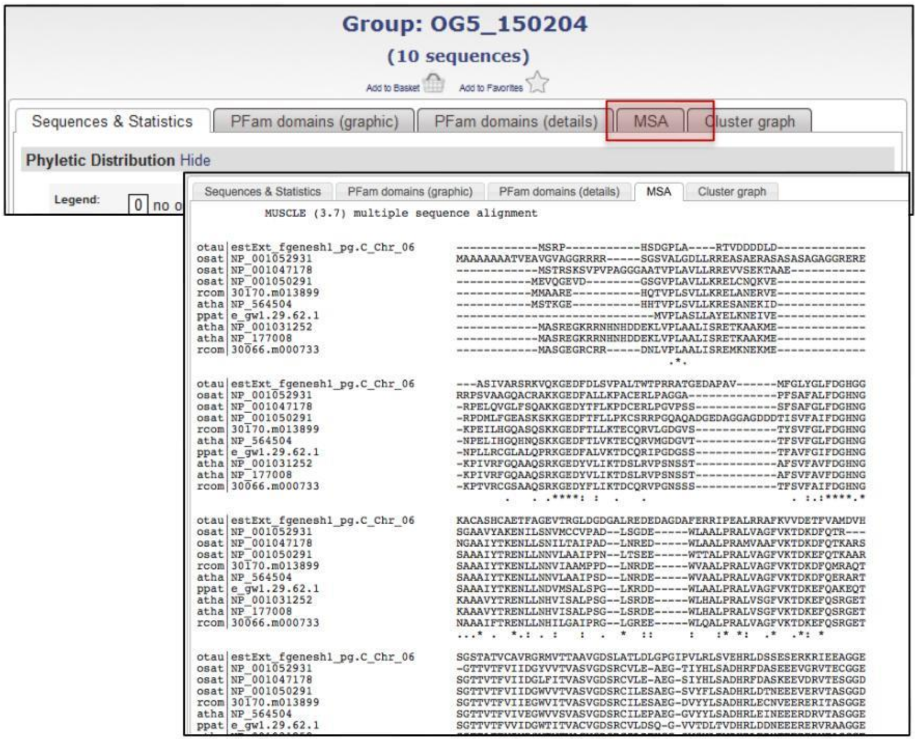
b. Add a step and run a phyletic pattern search for groups that contain any plant protein but do not contain any other organism outside plants. (hint: make sure everything has a red x on it except for plants (Viridiplantae (VIRI)), which should be a grey circle).



✖ 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).



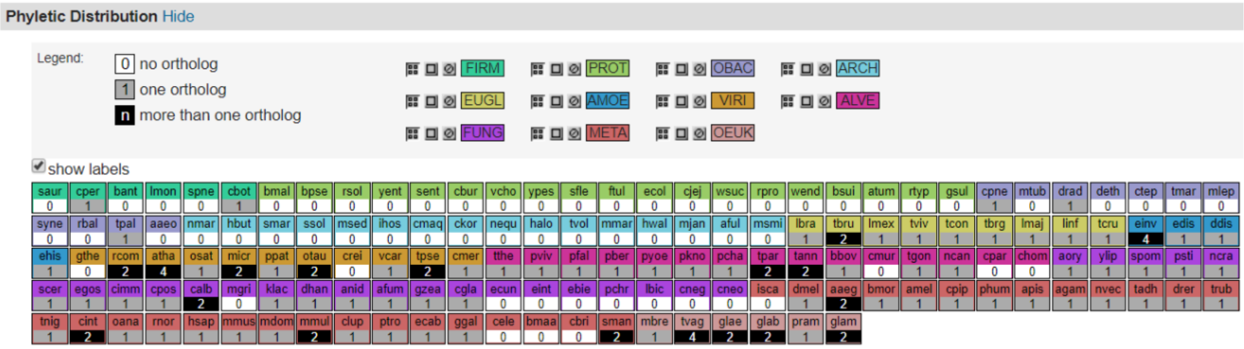


```
          MUSCLE (3.7) multiple sequence alignment

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 proteins are in this group? What is the phylogenetic distribution of the members of this group?



Phyletic Distribution Hide

Legend:

0 no ortholog

1 one ortholog

n more than one ortholog

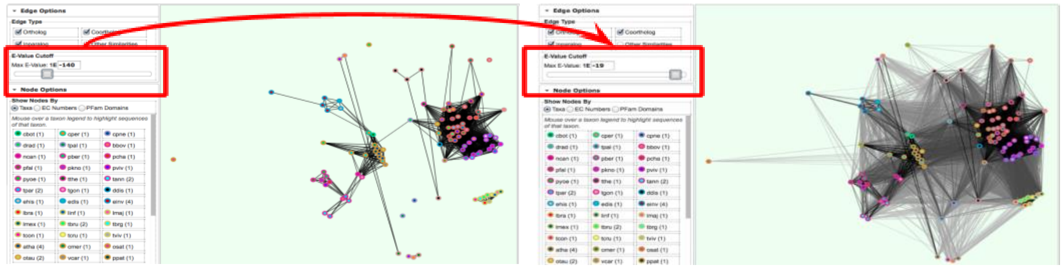
▦ □ ⊘ FIRM	▦ □ ⊘ PROT	▦ □ ⊘ OBAC	▦ □ ⊘ ARCH
▦ □ ⊘ EUGL	▦ □ ⊘ AMOE	▦ □ ⊘ VIRI	▦ □ ⊘ ALVE
▦ □ ⊘ FUNG	▦ □ ⊘ META	▦ □ ⊘ OEUK	

☑ show labels

saur	cper	bant	lmon	spne	cbot	bmal	bpse	rsol	yent	sent	cbur	vcho	ypes	sfle	ftul	ecol	cjej	wsuc	rpro	wend	bsui	atum	rtyp	gsul	cpne	mtub	drad	deth	ctep	tmar	mlep
0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0
syne	rbal	tpal	aaeo	nmar	hbut	smar	ssol	msed	ihos	cmaq	ckor	nequ	halo	tvol	mmar	hwal	mjan	aful	msmi	lbra	tbru	lmex	tviv	tcon	tbrg	lmaj	linf	tcru	einv	edis	ddis
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	1	1	1	1	1	1	1	4	1	1
ehis	gthe	rcom	atha	osat	micr	ppat	otau	crei	vcar	tpse	cmer	tthe	pviv	pfal	pber	pyoe	pkno	pcha	tpar	tann	bbov	cmur	tgon	ncan	cpar	chom	aory	ylip	spom	psti	ncra
1	0	2	4	1	2	1	2	0	1	2	1	1	1	1	1	1	1	1	2	2	1	0	1	1	0	0	1	1	1	1	1
scer	egos	cimm	cpos	calb	mgri	klac	dhan	anid	afum	gzea	cgla	ecun	eint	ebie	pchr	lbic	cneg	cneo	isca	dmel	aaeg	bmor	amel	cpip	phum	apis	agam	nvec	tadh	drer	trub
1	1	1	1	2	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1	2	1	1	1	1	1	1	1	1	1	1
tnig	cint	oana	rnor	hsap	mmus	mdom	mmul	clup	ptro	ecab	ggal	cele	bmaa	cbri	sman	mbre	tvag	glae	glab	pram	glam										
1	2	1	1	1	1	1	2	1	1	1	1	0	0	0	2	1	4	2	2	1	2										

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 or decrease the E-value? Can you identify subclusters?



Max E-Value: 1E -140

Max E-Value: 1E -19