

Orthology and Phyletic Patterns Exercise 8

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 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	<i>Cryptosporidium hominis</i> TU502	hypothetical protein	yes	no
CMU_034340	<i>Cryptosporidium muris</i> RN66	hypothetical protein, conserved	yes	no

[View the group \(OG5_127679\) containing this gene \(cgd7_2290\) in the OrthoMCL database](#)



- 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 pylum names - see image below).

Group: OG5_127679
(110 sequences)

[Add to Basket](#) [Add to Favorites](#)

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

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	imon	spne	icbot	bmali	bpose	rsoli	yenti	sent	cbur	vcho	ypes	sfla	tfui	ecol	cje	wsuc	rpro	wend	bsul	atum	rtyp	gsul	cpne	mtub	drad	deth	ctep	tmari	
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	
mlap	syne	rba	tpal	aaed	nmar	hbut	ssol	msed	ihos	cmaq	ckor	nequ	halo	tol	mmar	hwal	mjan	aful	msmi	lbra	ltru	lmex	tliv	lcon	lbrg	lmaj	lmf	ltru	lmv	lmv	
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	
edis	odis	ehis	gthe	room	atha	osat	micr	ppat	otau	crel	vcar	tpse	cmer	the	piv	pai	pber	pyoe	pkno	pcha	tpar	lann	bbov	cmur	lgon	ncan	cpar	chom	aory	ylp	
1	1	1	1	1	1	1	1	1	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	dimm	cpes	calb	mgd	klac	chan	anid	afum	gzea	cglia	ecun	eint	ebie	pchr	lbic	cneg	cned	lsca	dmet	aaeg	bmor	amel	cpip	phum	apis	agam	
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
nvec	tach	drer	trub	tnig	scint	oana	mor	hsap	mmus	mdom	mmul	clup	ptro	ecab	ggal	cele	bmaz	cbri	smari	mbre	tvag	glae	glab	pram	glam						
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

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

The screenshot shows the OrthoMCL DB website interface. The top navigation bar includes links for Home, New Search, My Strategies, My Basket (0), Tools, Data Summary, Downloads, and Community. The main content area is divided into three columns: 'Identify Ortholog Groups', 'Identify Protein Sequences', and 'Tools'. The 'Identify Ortholog Groups' column has a red arrow pointing to the 'Phyletic Pattern' link under the 'Evolution' section. The 'Tools' column lists various services like BLAST, protein assignment, and software download. A modal window titled 'Identify Groups based on Phyletic Pattern' is open, showing instructions on how to use the Phyletic Pattern Expression (PPE) tool. It includes a text box for the expression, a key for symbols (no constraints, must be in group, must not be in group, at least one subtaxon must be in group, mixture of constraints), and a list of taxonomic groups (Bacteria, Archaea, Eukaryota) with checkboxes to include or exclude them. A 'Get Answer' button is at the bottom of the modal.

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

Groups Quick Search: synth*
Sequences Quick Search: synth*

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
Number
Avg % C
% Pairs
Avg % ID
Avg % M
Avg E-Value

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: 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):

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?

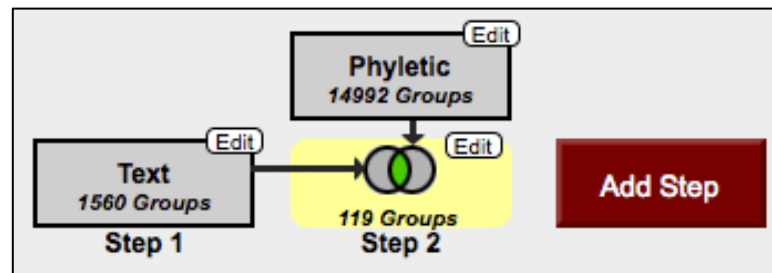
- c. Find all groups that contain orthologs from at least one species of *Cryptosporidium* and *Giardia* but not from bacteria or archaea.

NOTE: All EuPathDB sites also have a phyletic pattern search that uses OrthoMCL data under Genes -> Evolution -> Orthology Phylogenetic Profile.

3. Combining searches in OrthoMCL

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

- Use the text search to find groups that contain the word “*phosphatase*”.
- 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.).)



✧ Root (ALL):	
✧ Bacteria (BACT):	
✧ Archaea (ARCH):	
✧ Eukaryota (EUKA):	
✧ Alveolates (ALVE):	
✧ Ciliates (CILI):	✧ tthe
✧ Apicomplexa (APIC):	
✧ Coccidia (COCC):	✧ chom ✧ cmur ✧ cpar ✧ ncan ✧ lgon
✧ Aconoidasida (ACON):	
✧ Haemosporida (HAEM):	✧ pber ✧ pcha ✧ pfal ✧ pkno ✧ pviv ✧ pyoe
✧ Piroplasmida (PIRO):	✧ bbov ✧ lann ✧ tpar
✧ Amoebozoa (AMOE):	✧ ddis ✧ ehis ✧ edis ✧ einv
✧ Euglenozoa (EUGL):	✧ lbra ✧ llnf ✧ lmaj ✧ lmex ✧ tbru ✧ tbrg ✧ tcon ✧ tcru ✧ tliv
✧ Viridiplantae (VIRI):	
✧ Streptophyta (STRE):	✧ atha ✧ osat ✧ ppat ✧ room ✧ 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

- 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

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

MUSCLE (3.7) multiple sequence alignment

```

oatou estExt_fgenneshl_pg_c_Chrr_06
oatou 001052931
oatou 001047178
oatou 001050291
rcou 10170 0013899
atha 564504
ppat 0 gw1.29.62.1
atha 001031252
atha 177008
rcou 30066.m000733

oatou estExt_fgenneshl_pg_c_Chrr_06
oatou 001052931
oatou 001047178
oatou 001050291
rcou 10170 0013899
atha 564504
ppat 0 gw1.29.62.1
atha 001031252
atha 177008
rcou 30066.m000733

oatou estExt_fgenneshl_pg_c_Chrr_06
oatou 001052931
oatou 001047178
oatou 001050291
rcou 10170 0013899
atha 564504
ppat 0 gw1.29.62.1
atha 001031252
atha 177008
rcou 30066.m000733

oatou estExt_fgenneshl_pg_c_Chrr_06
oatou 001052931
oatou 001047178
oatou 001050291
rcou 10170 0013899
atha 564504
ppat 0 gw1.29.62.1
atha 001031252
atha 177008
rcou 30066.m000733

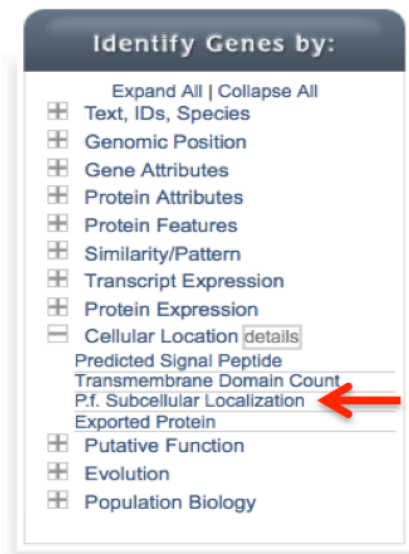
oatou estExt_fgenneshl_pg_c_Chrr_06
oatou 001052931
oatou 001047178
oatou 001050291
rcou 10170 0013899
atha 564504
ppat 0 gw1.29.62.1
atha 001031252
atha 177008
rcou 30066.m000733

```

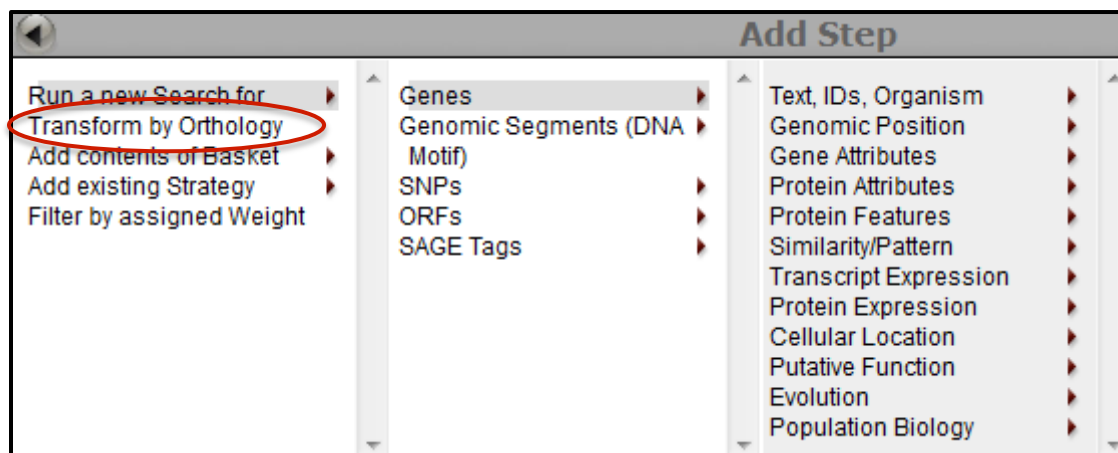
4. (Optional) Using the orthology transform tool to identify apicoplast targeted genes in *Toxoplasma* and *Neospora*.

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.

