## **Orthology and Phyletic Patterns**

## Homology



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

- 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 Seta 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...
 Q
 Showing 12 rows

It Omega	<b>↓† Gene</b>	<b>↓† Organism</b>	<b>‡† Product</b>	lt is syntenic	바as comments ⓒ
	CHUDEA7_2290	Cryptosporidium hominis UdeA01	unspecified product	yes	no
8	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
0	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
0	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).

												G	ou	p: (	OG	5_1	127	67	9											
(110 sequences)																														
Add to Basket Add to Favorites																														
Sequen	ces &	Statis	tics	P	Fam o	doma	ins (g	raphic	c)	PFa	m dor	nains	(deta	ils)	M	SA	Clu	ster g	graph											
hyletic	Distr	ibutic	n Hid	e																										
ingiotic	Disti	Dutie																												
Lege	nd:	<mark>0</mark> r	no orti	holog							Ø FIF	RM		IØP	ROT	88		OBAC			AR	CH								
		1 0	one or	tholo	g					III 🗆	Ø EU	GL		104	MOE		٥D	VIRI												
		n r	nore t	than (	one or	rtholo	g																							
												NG		1 Ø 🛛	IE IA	188	DØ	UEUR												
🗹 s	now la	abels																												
saur 0	cper 0	bant 0	Imon 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	syne	rbal		aaeo		hbut	smar		msed			_	nequ		tvol		hwal	يت	aful	msmi	Ibra	tbru		tviv	tcon	tbrg	Imai	linf	toru	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	pyce	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	inee	1 denal	1	1	1		1	1	1
spom 1	psti 1	ncra 1	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	1	1	isca 1	dmel 1	aaeg	bmor 1	amel 1	1	phum 1	apis 1	agan 1
nvec	tadh	drer	trub	tnig	cint	oana		hsap	mmus	mdom	mmul	clup	ptro	ecab	ggal	cele	bmaa	cbri	sman	mbre	tvag	glae	glab	pram	glam					
	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 <u>http://orthomcl.org/</u>

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.



- 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 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	hom+cmur+cpar>=1T AND_glam+glab+glae>=1T	Get Answer										
ı	In the graphical tree display:												
	<ul> <li>Click on -/+ to show or hide subtaxa and spec</li> <li>Click on the <sup>®</sup> icon to specify which taxa or s</li> <li>Refer to the legend below to understand other</li> </ul>	pecies to include or exclude in the profile.											
	Expression:	EUKA>=5T AND hsap>=10 Ge	t Answer										
	Key: 🔍 =no constraints   🖋 =must be in group   🥓 =at least one subtaxon must be in group   🗮 =must not be in group   🌣 =mixtur												
	Root (ALL):												
	<sup></sup> .● Bacteria (BACT):												
	+  Archaea (ARCH):												
	🖅 🗣 Eukaryota (EUKA):												
		Get Answer											

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 <a href="http://eupathdb.org">http://eupathdb.org</a>

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 further



expand your list of potentially Apicoplast targeted proteins by running a GO terms search for the term "apicoplast" or the GO ID: GO:0020011 in *P falciparum* 3D7 (hint, click on add step the go to the function prediction category and select the GO term search). Which Boolean operation did you use? Union or intersect?

(3D7 *)		
Apicomplexa		
Plasmodium		
<ul> <li>Plasmodium falciparum</li> <li>Plasmodium falciparum 3D7</li> </ul>		
Masmodium talciparum 3D7 add these   clear these   select only these		
select all clear all		
Evidence		
Curated		
Computed		
Limit to GO Slim terms		
Liniit to GO Siini tenns		
Ves		
• No		
- 10		
GO Term or GO ID		
GO Term or GO ID		
GO Term or GO ID		
GO Term or GO ID     [G0:0000011 appropriate: @ W]     Brigh typing to see suggestions to choose from (CTRL or CMD data to select multiple)		
GO Term or GO ID           [G00002011]:epicoplast:@W]           Brigh typing to see suggestions to docer from (CTRL or CMO data to select multiple)           GO Term or GO ID wildcard search		
GO Term or GO ID     [G0:0000011 appropriate: @ W]     Brigh typing to see suggestions to choose from (CTRL or CMD data to select multiple)		
GO Term or GO ID           [G00002011]:epicoplast:@W]           Brigh typing to see suggestions to docer from (CTRL or CMO data to select multiple)           GO Term or GO ID wildcard search		
GO Term or GO ID     [00000011: appropriat: 6 *) Bright typing to see suggestions to choose hore (CTRL or CMD data to select multiple)     GO Term or GO ID wildcard search     [xiA		
GO Term or GO ID     [00000011: appropriat: 6 *) Bright typing to see suggestions to choose hore (CTRL or CMD data to select multiple)     GO Term or GO ID wildcard search     [xiA		
GO Term or GO ID     [000020011: epidooplast: 6 *] Bright typing to see naggestions to choose horn (CTRL or CMO data to select multiple)     GO Term or GO ID wildcard search     [FEA	© 1 Interse	 Minus 2
GO Term or GO ID      [G00020011: epicoplest: @ M] Brigh typing to see neggestions to doose from (CTIR, or CMD dick to select multiple)     GO Term or GO ID wildcard search	I Union:	 Minus

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*.

•	Add Step											
Run a new Search for Transform by Orthology Add coments of Basket Add existing Strategy Filter by assigned Weight	Genes Genomic Segments (DNA Motif) SNPs ORFs SAGE Tags	Text, IDs, Organism Genomic Position Gene Attributes Protein Attributes Protein Features Similarity/Pattern Transcript Expression Protein Expression Cellular Location Putative Function Evolution Population Biology										

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 <u>http://orthomcl.org</u> 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).



E Aconoidasida (ACON):								
Haemosporida (HAEM):	×pber	Xpoha	×pfal	Xpkno	Xpviv	¥ pyce		
#Piroplasmida (PIRO):	×bbov	Xtann	#tpar					
XAmoebozoa (AMOE):	×ddis	Xehis	Xedis	VnieX				
¥ Euglenozoa (EUGL):	Xibra	Xinf 1	Kimaj X	Imex Xtb	ru ¥tbrg	#toon	Mtoru	*
••• • Viridiplantae (VIRI):								
Streptophyta (STRE):	• atha	• osat	• ppat	• room	• micr			
Chlorophyta (CHLO):	• crei	● otau	0 vcar					
Rhodophyta (RHOD):	• omer							
Cryptophyta (CRYP):	• gthe							
Bacillariophyta (BACI):	e tpse	t						
Fungi (FUNG):								
Microsporidia (MICR):	×ecun	Xebie	Xeint					
X Basidiomycota (BASI):	×cneo	Xcneg	Xibic	Xpchr				

- 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 Basiket 🏠 Add to Favorites 🟠	
Sequences & Stati	istics PFam domains (graphic) PFam domains (details) MSA Qiuster grap	h
Phyletic Distributio	on Hide	
	Sequences & Statistics PFam domains (graphic) PFam domains (details) MSA Cluster graph	
Legend: 0 n		
	otau         estExt_fgeneshl_pg.C_Chr_06           osat NP_00155291         MAAAAAATVEANVYAGURARR           osat NP_0015291         MAAAAAATVEANVYAGURARR           osat NP_0015291         MAAAAAATVEANVYAGURARR           osat NP_0015291         MAAAAAATVEANVYAGURARR           osat NP_0015291         MSTRESEVEVPAGGAATVPLANULARELANDVEANDVEANDVEANDVEANDVEANDVEANDVEANDVE	FGLYGLFDGHGG FSAFALFDGHNG FSAFGLFDGHNG ISVFAIFDGHNG ISVFGLFDGHNG
	ppat (e.gwl. 29, 62, 1)         -NPLLRCGLAUOPRKGEDFALVKTDCQRIPEDOSS	FSVFAVFDGHNG FSVFAVFDGHNG
	Ostal         gestber         Cgestbell         Cges	KTDKDFQTR KTDKDFQTKARS KTDKEFQTKAAR KTDKEFQMRAQT KTDKDFQERART KTDKEFQSRGET KTDKEFQSRGET
	otau         estExt_fgeneshl_pg.C_Chr_06         SGSTATVCAVRGRMVTTAAVGDSLATLDLGPGIPVLRLSVEHRLDSSE           osat NP_001632931         -GTTVTVVIIDGUTVTVASVGDSRCVLE-AEC-SITHLSADHRJDAES           osat NP_001647178         SGTTVTVVIIDGUTVTVASVGDSRCVLE-AEC-SITHLSADHRJDAES           osat NP_001630291         SGTTVTVVIIDGUTVTVASVGDSRCLE-AEC-SITHLSADHRJDAES           rcom 30170.m013899         SGTTVTVVIIDGUTVTVASVGDSRCLE-AEC-SITHLSADHRLDTME           ppat e_g41.23.62.1         SGTTVTVVIIGMVVTAAVGDSRCLE-AEC-G-VTLADHRLDDME	EEVGRVTECGGE EEVDRVTESGGD EEVERVTASGGD EERERITASGGE EERDRVTASGGE

- 5. Exploring a specific OrthoMCL group examining the cluster graph. (Use <u>http://orthomcl.org</u> 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																															
Leger	nd:	1	one o	holog rtholo than c	g	rtholo	g		ß		o Fif o Eu o Fu	GL		10F 10A	MOE	88	DØ	OBAC VIRI OEUK	<b> </b>												
show labels																															
saur	cper	bant	Imon	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
syne	rbal	tpal	aaeo	nmar	hbut	smar	ssol	msed		cmaq	ckor	nequ	halo	tvol	mmar	hwal	mjan	aful	msmi	Ibra		Imex	tviv	tcon	tbrg	Imaj	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
enis 1	gtne	rcom 2	atha	osat 1	micr	ppat 1	otau 2	Crei 0	vcar 1	tpse 2	cmer 1	tthe	pviv 1	pfal 1	pber 1	pyoe 1	pkno 1	pcna 1	tpar 2	tann 2	bbov 1	cmur	tgon 1	ncan 1	cpar 0	chom	aory 1	ylip	spom 1	psti 1	ncra 1
scer	egos	cimm		calb	mgri	klac		-	afum	gzea	cgla	ecun	eint	ebie	pchr	lbic	cnea	cneo		dmel	aaeq	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											
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?

