Orthology and Phyletic Patterns Exercise 9

9.1 Getting to OrthoMCL from EuPathDB databases Note: For this exercise use http://www.microsporidiadb.org

- a. Go to the gene page for the *Encephalitozoon cuniculi* gene with the ID: ECU07_0290.
- b. What does this gene do? It is annotated as hypothetical!
- c. Scroll down to the table labeled "Orthologs and Paralogs within MicrosporidiaDB". Does this gene have orthologs in other *Encephalitozoon* species? What about other organisms?

Hint: click on the link below the table that takes you to OrthoMCL.

Inthologs and Paralogs within MicrosporidiaDB Hide							
Gene	Organism	Product	is syntenic	has comments			
ECU10_1730	Encephalitozoon cuniculi GB-M1	hypothetical protein	no	no			
EHEL_070250	Encephalitozoon hellem ATCC 50504	hypothetical protein	yes	no			
EHEL_101770	Encephalitozoon hellem ATCC 50504	hypothetical protein	no	no			
Ein07_0230	Encephalitozoon intestinalis	hypothetical protein	yes	no			
Ein10_1700	Encephalitozoon intestinalis	hypothetical protein	no	no			
EBI_25559	Enterocytozoon bieneusi H348	hypothetical protein	yes	no			
EBI_27325	Enterocytozoon bieneusi H348	hypothetical protein	no	no			
NCER_100684	Nosema ceranae BRL01	hypothetical protein	no	no			
NCER_101866	Nosema ceranae BRL01	hypothetical protein	no	no			
	Find the	group containing ECU0	7_0290 in the	OrthoMCL datab			

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 tables to reveal the full species and pylum names – see image below.

	Group Summary (OG5_127163)																											
			#	Sequ	iences	ences # Taxa # Match Pairs (%)							lue A	Ave % Coverage			Ave	% Ide	ntity				III 🖉 FIRM			II 🖉 PROT		
				13	9	9	6	16552	2 (172	2.5%)	5.46e-08			88.5			32.5						III 🖉 OBAC			📰 🔲 Ø 🗛 RCH		
Links to OG5_127452 0 no ortholog;												Ø AMOE																
	Pfam Domain Architecture 1 one ortholog; I phyletic patterns										Ø ALVE																	
			BIO	Layo	it Ora	apn			n	more	e than	one	orthol	og;		V	patte	ern lab	oels &	coun	its	ľ		o FUN	IG		Ø META	
N	lo I	Nulti	ple S	eq. A	lignn	nent	(too	big)														ľ			JK			
		Ge	t Seo	queno	ces (F	FAST	A)															,						
sa	iur	cper	bant	Imon	spne 0	cbot	bmal	bpse	rsol	yent 0	sent	cbur	vcho	ypes	sfle	ftul	ecol	cjej	wsuc	rpro	wend	bsui	atum	rtyp	gsul	cpne	mtub	
dr	ad	deth	ctep	tmar	mlep	syne	rbal	tpal	aaeo	nmar	hbut	smar	ssol	msed	ihos	cmaq	ckor	nequ	halo	tvol	mmar	hwal	mjan	aful	msmi	Ibra	tbru	
Im) Iex	tviv	tcon	0 tbrg	0 Imaj	linf	0 toru	einv	edis	0 ddis	0 ehis	gthe	rcom	atha	osat	micr	ppat	otau	crei	vcar	0 tpse	cmer	0 tthe	pviv	pfal	1 pber	2 pype	
	1	1	1	1	1	1	2	1	1	1	1	0	1	2	2	1	3	1	1	1	1	1	1	2	2	2	2	
рк	2	2	tpar 1	1	1	2	2 1	2	2	2	aory 1	1 1	spom 1	1	1	1	egos 1	1	1	2	0 mgn	1	1	1	1	gzea 1	1	
ec	un 2	eint 2	ebie 2	pchr 1	lbic 2	cneg	cneo 1	isca 1	dmel	aaeg	bmor 1	amel	cpip 1	phum 1	apis 1	agam 1	nvec 1	tadh 1	drer 1	trub 1	tnig 1	cint 1	oana 1	rnor	hsap 3	mmus 3	mdom 2	
m	nul	clup	ptro	ecab	ggal	cele	bmaa	cbri	sman	mbre	tvag	glae	glab	pram	glam											0	-	
	4	3	3	3	2	1	1	1	1	1	3	1	1	2	1													

- e. Take a look at the PFAM domain architectures. 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?
- 9.2 Using the phyletic pattern tool in OrthoMCL Note: For this exercise use <u>http://www.beta.orthomcl.org/</u>
- a. How many protein groups in OrthoMCL <u>do not</u> have any orthologs in bacteria or archaea?

Hint: go to "Search for Groups by Evolution...Phyletic Pattern".



 b. How many protein groups <u>do not</u> contain orthologs from eukaryotes? Hint: click on the icon to specify which taxa or species to include or exclude in the profile.



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

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

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*.

				Add Step		
Run a new Search for Transform by Orthology Add contents 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	*********	4 >

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.



- 9.4(Optional) Use the <u>orthology transform tool</u> to identify as many *P. falciparum* genes containing signal peptides as possible. Note: For this exercise use <u>http://www.plasmodb.org</u>
- a. How many genes in *P.falciparum* are annotated with signal peptides (use default settings)?
- b. How many *P. vivax* genes are annotated with signal peptides (use default settings)?
- c. How many genes on these two lists are in common? Hint, use the ortholog query to transform between organisms.



- d. How many *P. falciparum* orthologs of vivax genes with signal peptides do not themselves contain signal peptides? Why might this be the case? Look at a couple of these using the synteny viewer to generate some hypotheses.
- e. Using PlasmoDB, generate the most comprehensive list of *P. falciparum* genes that may contain signal peptides. How many did you find?

9.5 (optional) Integrated 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".
- 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.



• 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: 0G5_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 ortholog										