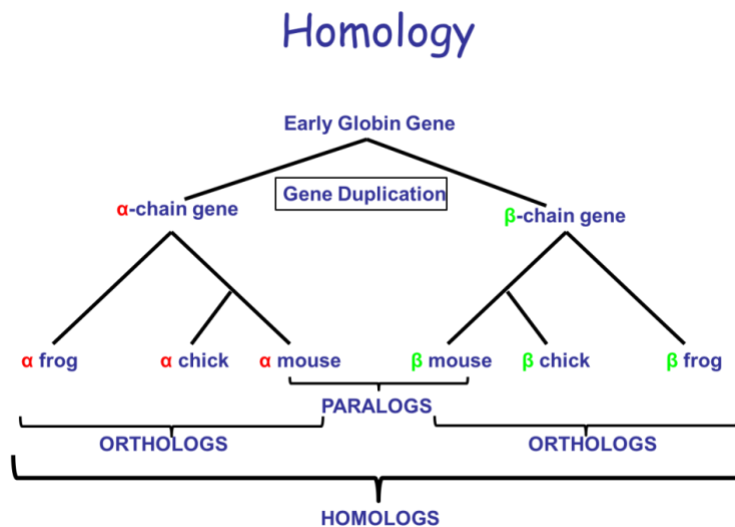


FungiDB & OrthoMCL: Orthology and Phyletic Patterns

Learning objectives:

- Run searches in OrthoMCL
- Run phyletic pattern searches using check boxes or an expression
- Combine searches using the strategy system
- Explore individual ortholog group pages
- Explore the group cluster graphs



1. Examining OrthoMCL output on gene record pages in FungiDB

- Go to the gene record page for the *Cryptococcus gattii* gene CGB_L0350W
- What is the function of this gene? How can you infer its function?
- Click on the “Orthology and Synteny” link on the left. Does this gene have orthologs in other *Cryptococcus* species? What about other organisms outside fungi? (hint: click on the *Ortholog Group OG6_106189*)

Clustal Omega	Gene	Product	Organism	Reference Strain?	Syntenic?	has comments
<input type="checkbox"/>	ALNC14_094800	unspecified product	Albugo laibachii Nc14	yes	no	no
<input type="checkbox"/>	AMAG_06223	cation diffusion facilitator family transporter	Allomyces macrosporus ATCC 38327	yes	no	no

The group page is divided into 5 sections:

1. Phyletic distribution
2. Group summary
3. List of proteins
4. PFam domains
5. Cluster graph

Examine each of the above sections – is it clear what each section contains?

Phyletic distribution: Numbers refer to the number of proteins in that organism or taxonomic group. In order to see organisms and taxonomic groups without proteins in this ortholog group, uncheck 'Hide zero counts.'

Group summary breaks down summary by protein types: A core protein is from one of the 150 core species that were initially used to form 'core' groups. A peripheral protein is from a peripheral species whose entire proteome was mapped into the 'core' groups. Peripheral proteins that do not map into a 'core' group are placed into residuals groups.

- Do all *Cryptococcus* species contain this protein? What is the most common PFAM domain associated with the proteins in this group? How can you look up protein alignments for this group (Hint: run ClustalOmerga tool)

Navigate to OrthoMCL - <http://orthomcl.org/>

OrthoMCL is a genome-scale algorithm for grouping orthologous protein sequences. It provides not only groups shared by two or more species/genomes, but also groups representing species-specific gene expansion families. Thus, it serves as an important utility for automated eukaryotic genome annotation. OrthoMCL starts with reciprocal best hits within each genome as potential in-paralog/recent paralog pairs and reciprocal best hits across any two genomes as potential ortholog pairs. Related proteins are interlinked in a similarity graph. Then, MCL (Markov Clustering algorithm; [Dongen 2000](#); www.micans.org/mcl) is invoked to split mega-clusters. This process is analogous to the manual review in COG construction. MCL clustering is based on weights between each pair of proteins, so to correct for differences in evolutionary distance the weights are normalized before running MCL.

Method for Forming and Expanding Ortholog Groups in OrthoMCL.

Proteins are placed into Ortholog Groups by the following steps:

The OrthoMCL algorithm (see below) is employed on proteins from a set of 150 Core species to form Core ortholog groups. These species were carefully chosen based on proteome quality and widespread placement across the tree of life. Each Core protein is placed by the algorithm into a Core ortholog group consisting of one or more proteins.

Core group names have the format OG6_XXXXXX (e.g., OG6_101327). OG6 refers to OrthoMCL release 6; for each sub-release (e.g., 6.1, 6.2, etc), the Core species and the Core ortholog group names will remain constant.

The proteins from hundreds of additional organisms, termed Peripheral organisms, are mapped into the Core groups. To do this, NCBI BLASTP is used to compare each Peripheral protein to each Core protein in the Core groups. (Note that Peripheral proteins that were previously added to the Core group are NOT used in the BLASTP.) Then, each Peripheral protein is assigned to the Core group containing the Core protein with the best BLAST score, but only if the E-Value is $<1e-5$ and the percent match length is $\geq 50\%$.

All Peripheral proteins that fail to map to a Core group are collected and subjected to independent OrthoMCL analysis, forming Residual groups consisting of one or more proteins. Residual group names have the format OG6r1_XXXXXX (e.g., OG6r1_101327), where OG6 refers to release 6 and r1 refers to sub-release 1.

For each subsequent sub-release (which will occur every ~3 months along with other VEuPathDB sites), proteomes from additional Peripheral organisms will be processed as in steps 2 and 3 above. However, step 3 will differ slightly because the previous set of Residual groups will be disassembled, leaving the previous unmapped Peripheral proteins to be combined with the new unmapped Peripheral proteins. All of these proteins will be used to form new Residual groups (e.g., OG6r2_XXXXXX).

During a sub-release, the proteomes of some species will be updated to the latest version. This can be easily done for a Peripheral species: the old set of proteins are removed from ortholog groups and then the new set is mapped into groups as above. However, this is not possible for Core species because these proteins are used to define Core groups. Thus, the Core species with the older proteome remains on the site but is superficially retired by appending its abbreviation with -old (e.g., aaeg becomes aaeg-old). Then, the latest version of the proteome is mapped in as a peripheral species and obtains the original species abbreviation (e.g., aaeg is a peripheral with a more recent proteome than aaeg-old). These retired species will be eliminated fully when a new set of Core species is defined, as described in the next point.

On occasion, the set of Core species will be re-defined, as more appropriate proteomes become available and/or when a large number of Core species are retired. In this case, new Core groups (e.g., OG7_XXXXXX) and Residual groups (e.g., OG7r1_XXXXXX) will be formed from the latest version of proteomes from a carefully-chosen set of core species.

This design allows for the addition of proteomes at every sub-release (e.g., 6.1, 6.2, etc). Note that Core groups (e.g., OG6_101327) will remain between sub-releases, though these groups will expand as Peripheral proteins are mapped in. In contrast, Residual groups will exist only for that sub-release; thus, Residual groups are useful in allowing the user to find proteins related to their protein(s) of interest, but are not stable groups.

2. Using the Phyletic Pattern search in OrthoMCL

The “Phyletic Pattern” search is an ortholog group search – look under the ortholog groups category and explore the available searches. Can you find the one called “Phyletic Pattern”? There are two ways to specify a phyletic pattern:

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

The screenshot shows the OrthoMCL DB website interface. At the top, there is a search bar and navigation links. The main content area is titled "Overview of Resources and Tools" and features a "Search for..." section on the left with various search options. The central part of the page is a "Phyletic Pattern" search window. It includes a key for constraints, an "Expression" input field containing "BACT-OT AND ARCH-OT", and a tree view of taxonomic groups. The tree view shows a hierarchy of groups with checkboxes next to them, indicating their selection status. The groups listed include Bacteria (BACT), Firmicutes (FIRM), Proteobacteria (PROT), Other Bacteria (OBAC), Archaea (ARCH), Nitrosopumilus maritimus (strain SCM1) (nmar), Euryarchaeota (EURY), Crenarchaeota (CREN), Nanoarchaeota (NANO), Korarchaeota (KORA), Eukaryota (EUKA), Alveolates (ALVE), Amoebozoa (AMOE), Euglenozoa (EUGL), Viridiplantae (VIRI), Fungi (FUNG), Metazoa (META), and Other Eukaryota (OEUK).

1. Using the expression box. Type the expression using hints available at the bottom of the search page.
2. Using the selectable tree menu. Click on the circle next to the taxon you want to include or exclude.
 - a. Using Phyletic pattern search identify how many protein groups do not contain orthologs from bacteria and archaea?
 - b. Find all groups that contain orthologs from at least one species of *Ascomycota fungi* but not from bacteria, archaea or metazoa. Hint: Use the checkboxes to make your selection. Pay attention to the final expression written in the Expression window before clicking the Get Answer button.


In the graphical tree display:

- Click on the icons to show or hide subtaxa and species.
- Click on the icons to specify which taxa or species to include or exclude in the profile.
- Refer to the legend below to understand other icons.

Expression: ASCO=>1T AND META=0T AND ARCH=0T AND BACT=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



Phyletic
110,289 Ortholog Groups
Step 1

+ Add a step

110,289 Ortholog Groups

Revise this search

Ortholog Group Results

1 2 3 ... 5,515 Rows per page: 20

Download Add to Basket Add Columns

Ortholog Group	Total Number Proteins	Keywords	Top PFam Domains	EC Numbers	Archaea	Bacteria	Alveolata
OG6_100719	4502	unknown; source; uniProtKB/TrEMBL;Acc; r1fin; pir protein	PF02009 (4470), PF06024 (3), PF06143 (2)	N/A	0 / 27 (0%)	0 / 47 (0%)	28 / 113 (25%)
OG6_103977	1856	unknown; source; uniProtKB/TrEMBL;Acc	PF13489 (1572), PF13649 (123), PF08241 (20)	2.1.1.163 (7), 2.1.1.- (4), 2.1.1.223 (2), 2.1.1.103 (1), 3.5.1.23	0 / 27 (0%)	0 / 47 (0%)	0 / 113 (0%)

Examine your results and learn how to interpret the graphical representation for each group. For example, if you take a look at the first ortholog group, it occurs in Alveolata (in 23 out of 113 species present in OrthoMCL, which is 25% of the total species currently assigned to Alveolata in OrthoMCL).

- c. Next, revise your search to find groups that do not contain orthologs from Alveolates, Amebozoa, archaea, bacteria and Ascomycetes, but contain at least one ortholog group from *Mucor circinelloides* f. *circinelloides* 1006PhL (mcic) and *Mucor circinelloides* f. *lusitanicus* CBS 277.49 (mcir). Hint: You cannot answer this question by using the check boxes alone.

If you are getting frustrated trying to figure this one out, you have a right to be! 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: start by assigning the “do not contain” parameter (x) using check boxes to Alveolates, Amebozoa, archaea, bacteria and Ascomycetes. Next, use the expression window to add “AND” followed by specific criteria for *Mucor* spp.

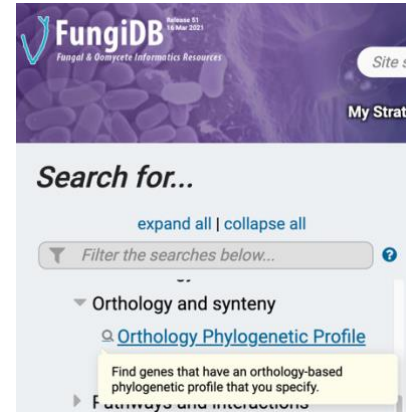


Phyletic
3,259 Ortholog Groups
Step 1

+ Add a step

All VEuPathDB sites also have a phyletic pattern search that uses OrthoMCL data under *Genes -> Orthology and Synteny -> 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 of interest but not present in the host as these genes may make good drug targets or vaccine candidates.



3. Combining searches in OrthoMCL

Find all fungal proteins that are likely phosphatases that do not have orthologs outside of fungi.

- a. Use the site search box at the top of the page in the header **to find OrthoMCL groups** that contain the word “*phosphatase*” (note that the search should be run without the quotation marks but with the asterisks).



- How many proteins sequences did you identify?
- How many ortholog groups did you identify?

Note: that numbers in screen shots will likely be different on the site due to frequent updates.

- b. Click on ortholog groups to filter the results to only show the ortholog groups containing the word phosphatase.

All results matching ***phosphatase*** Export as a Search Strategy to download or mine your results

1 - 20 of 85,180 1 2 3 ... 4,259

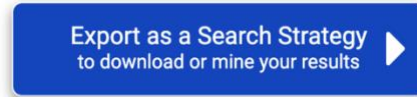
Filter results <input checked="" type="checkbox"/> Hide zero counts	
Genome	
Protein Sequences	80,408
Orthology	
Ortholog Groups	4,772

Filter fields
Select a result filter above

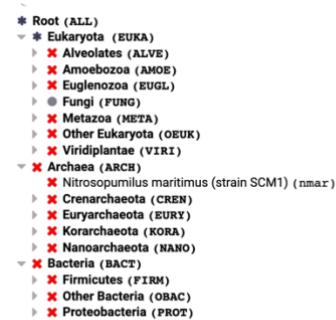
Protein Sequence - aacuASPACDRAFT_10380 Serine/threonine-protein phosphatase [Source:UniProtKB/TrEMBL;Acc:A0A1L9X558] EC Numbers: Protein-serine/threonine phosphatase (3.1.3.16) Ortholog group: OG6_100222 PFam Domains: Calcineurin-like phosphoesterase (PF00149), Serine-threonine protein phosphatase N-terminal domain (PF16891) Taxon Name: Aspergillus aculeatus ATCC 16872 Fields matched: EC Numbers; PFam Domains; Product
Protein Sequence - aacuASPACDRAFT_109194 PPM-type phosphatase domain-containing protein [Source:UniProtKB/TrEMBL;Acc:A0A1L9X7W1] EC Numbers: Protein-serine/threonine phosphatase (3.1.3.16) Ortholog group: OG6_105965 PFam Domains: Protein phosphatase 2C (PF00481) Taxon Name: Aspergillus aculeatus ATCC 16872 Fields matched: EC Numbers; PFam Domains; Product

Notice that you can filter the group results even further by the group fields. Also, notice that the *Export as a Search Strategy* button is now active. This is because all the results are now only one type of record: groups.

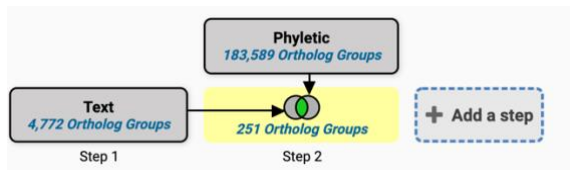
- c. Click on the blue “Export as a Search Strategy” button at the top right of the page to export your results as a strategy.



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



- e. How many groups did you return?



4. **Exploring a specific OrthoMCL group - examining the cluster graph.**

- a. Visit the OrthoMCL record page for the group OG6_115064.

- b. Examine the phyletic distribution tree. What taxa does this group contain?

Phyletic Distribution of Proteins [🔗](#) [Download](#)

Numbers refer to the number of proteins in that organism or taxonomic group.

expand all | collapse all
 Hide zero counts

Type a taxonomic name

- ▼ Eukaryota (EUKA) 150
- ▼ Fungi (FUNG) 150
- Allomyces macrogynus ATCC 38327 (amac) 2
- Catenaria anguillulae PL171 (cang) 1
- Conidiobolus coronatus (strain ATCC 28846 / CBS 209.66 / NRRL 28638) (Delacroixia coronata) (ccor) 1
- Rozella allomycis (strain CSF55) (ral1) 1
- ▶ Ascomycota (ASCO) 103
- ▶ Basidiomycota (BASI) 31
- ▶ Mucoromycota (MUCO) 11

- c. Examine the cluster graph for this group (hint: go to the cluster graph section of the page and then click on the “Click to open the Cluster graph in a new tab”)

You can interact with the cluster graph. For example, move the slide to increase the E-value cutoff stringency (e.g., to a more negative number). Can you identify subclusters? Click on the nodes in the graph – notice how the organism is updated on the right.

The screenshot shows a cluster graph interface with several panels:

- Top Left:** Filter options for Ortholog, Inparalog, Peripheral-Core, Peripheral-Peripheral, and Other Similarities.
- Top Center:** E-Value Cutoff slider set to 1E-132.
- Top Right:** "Show Nodes By" options: Taxa, EC Numbers, PFam Domains, Core/Peripheral.
- Center:** A network graph with nodes of various colors and sizes connected by edges.
- Bottom Left:** "Node Options" panel with "Show Nodes By" set to PFam Domains. A legend shows PF01591 (150) in blue, PF00300 (148) in red, and PF01408 (1) in green.
- Bottom Right:** A table listing sequence information for various proteins, including Source ID, Organism, Description, and BLAST Scores.

Source ID	Organism	Description	Length	E-Value
amacAAAG_10727	amac		425	6PF2K dos
anicAN10824	anic		553	unknown
anigA15g00200	anig		535	6PF2K dos
aninATCC64974_32640	anin		535	6PF2K dos
anovP1740RAFT_145305	anov		535	6PF2K dos
aoryA009070100027	aory		535	unknown
aterATEG_05645	ater		487	6PF2K dos
atheCDV56_105196	athe		411	6PF2K dos
bderBDCC_09157	bder		561	6PF2K dos
blucBion16g02359	bluc		553	unknown
bgliIBDRG_09408	bgli		561	6PF2K dos
cabbC1_02220C_B	cabb		543	Putative p
cabiC1_02220C_A	cabi		543	

On the left of the page in the *Node Options* panel, click on PFam Domains to see which proteins have the various PFam domains.

The screenshot shows a detailed view of a node in the cluster graph. The "Node Options" panel is visible, and a table of BLAST scores is shown on the right.

Node Information: cimcPV07_06804

Sequence Information:

Source ID	Organism	Description	Length
cimcPV07_06804	Cladophialophora immunda strain CBS 83496	6PF2K domain-containing protein [Source: UniProtKB/TrEMBL; Acc: A0A002C7E5]	539

BLAST Scores:

Subject	Type	E-Value
acapiHCAG_06881	Peripheral-Peripheral	1.0E-181
acaqiHCBG_09156	Peripheral-Peripheral	1.0E-181
acarIASPCADRAFT_211354	Peripheral-Peripheral	1.0E-181
aclaIACLA_095440	Peripheral-Peripheral	1.0E-181
afisNFIA_051670	Peripheral-Peripheral	1.0E-181
afsaAFLA_054870	Peripheral-Peripheral	1.0E-181
afubAFUB_093180	Peripheral-Peripheral	1.0E-181
afum-oldAfufg05100	Peripheral-Core	1.0E-181
afumAfufg05100	Peripheral-Peripheral	1.0E-181
agosiQ75CV1	Peripheral-Peripheral	1.0E-103
akawAKAW_10360	Peripheral-Peripheral	1.0E-181
aleniALT_0841	Peripheral-Peripheral	1.0E-181

And in the same *Node Options* panel, click on *Core/Peripheral* to observe which proteins were derived from Core species and which proteins were derived from

Peripheral species. Proteins from Core species were used in the initial OrthoMCL algorithm to form Core ortholog groups. Proteins from Peripheral species were mapped into these Core groups by sequence similarity (determined by BLAST score).

Cluster Graph: OG6_115064 (150 proteins) [?](#)

[Back to Group page](#)

