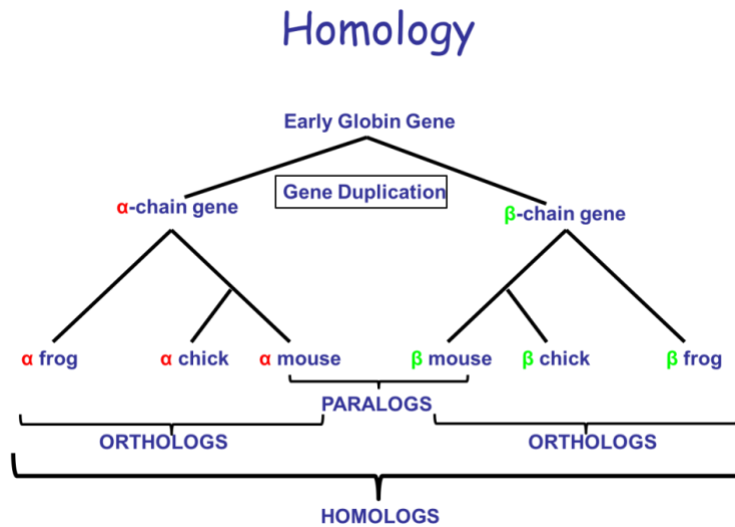


# FungiDB & OrthoMCL: Orthology and Phyletic Patterns

## 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.



## About OrthoMCL.

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](http://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.

## Background on Orthology and Prediction

Orthologs are homologs separated by speciation events. Paralogs are homologs separated by duplication events. Detection of orthologs is becoming much more important with the rapid progress in genome sequencing (Glover et al. 2019).

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](http://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.

OrthoMCL is similar to the INPARANOID algorithm (Remm et al. 2001) but is extended to cluster orthologs from multiple species. OrthoMCL clusters are coherent with groups identified by EGO (Lee et al. 2002), and an analysis using EC number suggests a high degree of reliability (Li et al. 2003).

We evaluated the performance of seven widely-used orthology detection algorithms that use three general prediction strategies: phylogeny-based, evolutionary distance-based and BLAST-based (Chen, et al. 2007). Specifically, we used Latent Class Analysis (LCA), a statistical technique appropriate for testing large data sets when no gold standard is available. Our results show an overall trade-off between sensitivity and specificity among these algorithms, with INPARANOID and OrthoMCL performing best with False Positive (FP) and False Negative (FN) error rates lower than 20%.

### **Method for Forming and Expanding Ortholog Groups in OrthoMCL.**

Proteins are placed into Ortholog Groups by the following steps:

1. 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.
2. 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  $\geq 50\%$ .
3. 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.
4. 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).

5. 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.
6. 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.

## **Examining OrthoMCL output on gene record pages in FungiDB**

- **Go to the gene record page for the CGB\_L0350W, a hypothetical protein CNBL0590.**
  - a. What is the function of this gene? How can you infer its function?
    - i. Click on the “Orthology and Synteny” link in the Contents menu on the left.  
Does this gene have orthologs in other *Cryptococcus* species?

CGB\_L0350W

expand all | collapse all

Search section names...

1 Gene models

2 Annotation, curation and identifiers

3 Link outs

4 Genomic Location

5 Literature

6 Taxonomy

7 Orthology and synteny

8 Phenotype

9 Transcriptomics

10 Sequence analysis

11 Sequences

12 Structure analysis

13 Protein features and properties

14 Function prediction

15 Pathways and interactions

16 Immunology

7 Orthology and synteny

Ortholog Group OG6\_106189

Orthologs and Paralogs within FungiDB 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' button.

Crypto

Clustal Omega	Gene	Product	Organism
<input type="checkbox"/>	D1P53_002977	unspecified product	Cryptococcus cf. gattii MF34
<input type="checkbox"/>	L203_04836	Cation efflux protein [Source:UniProtKB/TrEMBL;Acc:A0A1E3ICE3]	Cryptococcus depauperatus CBS 784
<input type="checkbox"/>	I314_06191	cation antiporter	Cryptococcus gattii CA1873
<input type="checkbox"/>	I306_06271	cation antiporter	Cryptococcus gattii EJB2
<input type="checkbox"/>	I311_05609	cation antiporter	Cryptococcus gattii NT-10

- Examine evidence in the “Function prediction” section.
- What about other organisms outside fungi? (Hint: click on the Ortholog Group OG6\_106189).
- The OrthoMCL group page is divided into 5 sections:
  - Phyletic distribution
  - Group summary
  - List of proteins
  - PFam domains
  - Cluster graph

• **Does this protein have orthologs in Archaea and Bacteria?**

**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.'

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)	254
▶ Alveolates (ALVE)	3
▶ Amoebozoa (AMOE)	1
▶ Euglenozoa (EUGL)	0
▶ Fungi (FUNG)	120
▶ Metazoa (META)	123
▶ Other Eukaryota (OEUK)	3
▶ Viridiplantae (VIRI)	4
▼ Archaea (ARCH)	0
▶ Nitrosopumilus maritimus (strain SCM1) (nmar)	0
▶ Crenarchaeota (CREN)	0
▶ Euryarchaeota (EURY)	0
▶ Korarchaeota (KORA)	0
▶ Nanoarchaeota (NANO)	0
▼ Bacteria (BACT)	0
▶ Firmicutes (FIRM)	0
▶ Other Bacteria (OBAC)	0
▶ Proteobacteria (PROT)	0

**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 currently integrated in FungiDB contain this protein?

☐ Hide zero counts

Cryptod × ?

<b>Eukaryota (EUKA)</b>	<b>254</b>
<b>Fungi (FUNG)</b>	<b>120</b>
<b>Basidiomycota (BASI)</b>	<b>27</b>
Cryptococcus cf. gattii MF34 (ccfg)	1
Cryptococcus depauperatus CBS 7841 (cdep)	1
Cryptococcus gattii CA1873 (cgac)	1
Cryptococcus gattii EJB2 (cgae)	1
Cryptococcus gattii NT-10 (cgan)	1
Cryptococcus gattii VGII R265 (cdeu)	1
Cryptococcus gattii VGIV IND107 (cgai)	1
Cryptococcus gattii WM276 (cgat)	1
Cryptococcus neoformans var. grubii H99 (cneq)	1
Cryptococcus neoformans var. grubii KN99 (cnek)	1
Cryptococcus neoformans var. neoformans B-3501A (cnep)	1
Cryptococcus neoformans var. neoformans JEC21 (cneo)	1
Cryptococcus neoformans var. neoformans JEC21 (old build 2016-06-16) (cneo-old)	1

- What is the most common PFAM domain associated with the proteins in this group?

#### 4 Pfam domains

▼ PFam Legend [Download](#)

Search this table... ?

Accession	Symbol	Description	Count	Legend
PF01545	Cation_efflux	Cation efflux family	251	
PF03645	Tctex-1	Tctex-1 family	2	
PF03102	NeuB	NeuB family	1	
PF01423	LSM	LSM domain	1	

- How can you look up protein alignments for *Cryptococcus*?  
(Hint: run ClustalOmerga tool and use the “Search this table” filter to limit the alignment to “Cryptococcus”).

## 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.

- Find the “Phyletic Pattern” search.

The screenshot displays the OrthoMCL DB website interface. The top navigation bar includes the site logo, release information (Release 8.15, 9 Mar 2023), a search bar, and links to My Strategies, Searches, Tools, My Workspace, Data, About, Help, and Contact Us. A sidebar on the left titled "Search for..." contains a filter dropdown and a list of search categories. The "Ortholog Groups" category is expanded, showing options like % Pairs w/ Similarity, All Groups, Avg % Homology, Avg % Identity, Avg % Match Length, Avg E-Value, EC Number, Group ID(s), Group or Sequence ID, Number of Sequences, Number of Taxa, PFam ID or Keyword, and Phyletic Pattern. The "Phyletic Pattern" option is highlighted with an orange arrow. The main content area, titled "Overview of Resources and Tools", features a row of icons for various tools. Below this, the "Configure Search" section includes a "Reset values to default" button and instructions on how to use the Phyletic Pattern Expression (PPE) text box and graphical tree display. An example expression "EUKA==ST AND heap==10" is shown. A key explains the symbols used in the expression: @ for no constraints, green checkmark for must be in group, yellow star for at least one subtaxon must be in group, red X for must not be in group, and a dot for mixture of constraints. Below the key is a search bar for taxonomic names and a tree view showing the hierarchy of taxonomic groups, including Eukaryota (EUKA), Archaea (ARCS), and Bacteria (BACT).

OrthoMCL DB  
Release 8.15  
9 Mar 2023

Site search, e.g. OG6\_106861 or PF3D7\_1133\* or "binding protein"

My Strategies Searches Tools My Workspace Data About Help Contact Us

Search for...

expand all | collapse all

Filter the searches below...

Ortholog Groups

- % Pairs w/ Similarity
- All Groups
- Avg % Homology
- Avg % Identity
- Avg % Match Length
- Avg E-Value
- EC Number
- Group ID(s)
- Group or Sequence ID
- Number of Sequences
- Number of Taxa
- PFam ID or Keyword
- Phyletic Pattern
- Text Terms

Proteins

Overview of Resources and Tools

OrthoMCL FAQ About OrthoMCL Types of Searches in OrthoMCL Understanding Group Search Results Search Strategies Phyletic Pattern Search Transforming Results Assign Proteins to Groups Downloads

Configure Search Learn More

Reset values to default

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 above under "Learn More".

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: EUKA==ST AND heap==10

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

expand all | collapse all

Type a taxonomic name

Root (ALL)

- Eukaryota (EUKA)
  - Alveolates (ALVE)
  - Amoebozoa (AMOE)
  - Euglenozoa (EUGL)
  - Fungi (FUNG)
  - Metazoa (META)
  - Other Eukaryota (OEUK)
  - Viridiplantae (VZRI)
- Archaea (ARCS)
  - Nitrosopumilus maritimus (strain SCM1) (nmar)
  - Crenarchaeota (CREN)
  - Euryarchaeota (EURY)
  - Korarchaeota (KORA)
  - Nanoarchaeota (NANO)
- Bacteria (BACT)
  - Firmicutes (FIRM)
  - Other Bacteria (OBAC)
  - Proteobacteria (PROT)

There are two ways to specify a phyletic pattern:

1. Using the expression box.

- **Run the default search for `EUKA>=5T AND hsap>=10`.**

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:

[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

- Use the “Learn More” tab to decipher the expression used above.

[Configure Search](#) [Learn More](#)

### Description

Find Ortholog Groups by phyletic pattern.

Phyletic Pattern Expression is a flexible and powerful way to identify ortholog groups with a certain conservation of proteins from specific taxa are present or absent. Also, the pattern finds groups with a certain copy number (

### Examples

These expressions find ortholog groups in which...

<code>hsap&gt;=5</code>	there are five or more human sequences
<code>hsap+ecol=2T</code>	both human and E. coli are present.
<code>hsap+ecol=1T</code>	only one species of human or E. coli is present.

2. Using the selectable tree menu.

You can click on the circle next to the taxon you want to include or exclude it from the search.

[expand all](#) | [collapse all](#)

- \* **Root (ALL)**
  - \* **Eukaryota (EUKA)**
    - Alveolates (ALVE)**
    - Amoebozoa (AMOE)**
    - Euglenozoa (EUGL)**
    - Fungi (FUNG)**
    - Metazoa (META)**
    - Other Eukaryota (OEUK)**
    - Viridiplantae (VIRI)**
  - **Archaea (ARCH)**
    - Nitrosopumilus maritimus (strain SCM1) (nmar)
    - Crenarchaeota (CREN)**
    - Euryarchaeota (EURY)**
    - Korarchaeota (KORA)**
    - Nanoarchaeota (NANO)**
  - **Bacteria (BACT)**
    - Firmicutes (FIRM)**
    - Other Bacteria (OBAC)**
    - Proteobacteria (PROT)**

- Using the “Phyletic pattern” search, identify how many eukaryotic protein groups **do not** contain orthologs from bacteria and archaea.

Hint: leave EUKA class with no constraints.

**Phyletic**  
 882,708 Ortholog Groups

+ Add a step

Step 1

Strategy URL: <https://orthomcl.org/orthomcl/app/workspace/strategies/import/eebc49abcf1d99f>

- Find all groups that contain orthologs from at least one species of *Ascomycota fungi* (1T) but not from bacteria, archaea or metazoan (0T).

**Phyletic**  
 120,871 Ortholog Groups

+ Add a step

Step 1

- Examine your results and learn how to interpret the graphical representation for each group.  
 Scroll to the right of the results table examine graphical representation of the results. You can hover over each graph to learn more about phyletic distribution for each class.

[Download](#)
[Add to Basket](#)
[Add Columns](#)

Archaea	Bacteria	Alveolata	Amoeba	Euglenozoa	Fungi	Metazoa	Viridiplantae
0 / 27 (0%)	0 / 47 (0%)	0 / 137 (0%)	0 / 14 (0%)	0 / 73 (0%)	7 / 309 (2%)	0 / 124 (0%)	0 / 14 (0%)
0 / 27 (0%)	0 / 47 (0%)	0 / 137 (0%)	0 / 14 (0%)	0 / 73 (0%)	1 / 309 (0%)	0 / 124 (0%)	0 / 14 (0%)
0 / 27 (0%)	0 / 47 (0%)	109 / 137 (80%)	4 / 14 (29%)	27 / 73 (37%)	59 / 309 (19%)	0 / 124 (0%)	1 / 14 (7%)
0 / 27 (0%)	0 / 47 (0%)	132 / 137 (96%)	14 / 14 (100%)	72 / 73 (99%)	1 / 309 (0%)	0 / 124 (0%)	1 / 14 (7%)
0 / 27 (0%)	0 / 47 (0%)	1 / 137 (1%)	0 / 14 (0%)	4 / 73 (5%)	1 / 309 (0%)	0 / 124 (0%)	1 / 14 (7%)
0 / 27 (0%)	0 / 47 (0%)	0 / 137 (0%)	0 / 14 (0%)	0 / 73 (0%)	1 / 309 (0%)	0 / 124 (0%)	0 / 14 (0%)

**ALVEOLATA**  
 Ciliates: 0 / 2  
 Apicomplexa  
   Haemosporida: 60 / 60  
   Coccidia: 48 / 51  
   Piroplasmida: 17 / 17  
   Other apicomplexa: 4 / 4  
   Other alveolata: 3 / 3

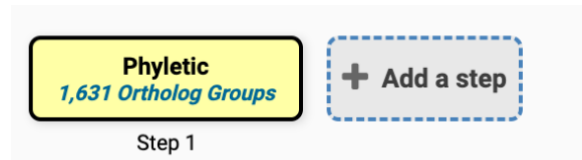
Strategy URL: <https://orthomcl.org/orthomcl/app/workspace/strategies/import/555afd9c529d4927>



- **Revise your search to find groups that:**

- do not contain orthologs from Alveolates, Amoebozoa, archaea, bacteria and Ascomycetes.
- 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. For *Mucor*, use the expression field to finish the parameter set up manually.



If you are getting frustrated trying to figure this one out, you have a right to be! If your results look different, hover over the search step and click to revise the parameter search. The cool thing about OrthoMCL is that 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, Amoebozoa, archaea, bacteria and Ascomycetes. Next, use the expression window to add “AND” followed by specific criteria for *Mucor* spp. Use the learn more tab for more information.

Strategy URL: <https://orthomcl.org/orthomcl/app/workspace/strategies/import/88e60b823cb2c959>

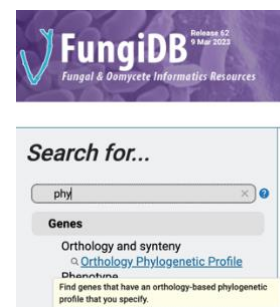
If you ran a search using just check boxes, the search will be configured to look for groups that:

- do not contain orthologs from Alveolates, Amoebozoa, archaea, bacteria and Ascomycetes.
- contain ortholog groups from both *Mucor circinelloides* f. *circinelloides* 1006PhL (mcic) **AND** *Mucor circinelloides* f. *lusitanicus* CBS 277.49 must be present

Strategy URL: <https://orthomcl.org/orthomcl/app/workspace/strategies/326574153/430551723>

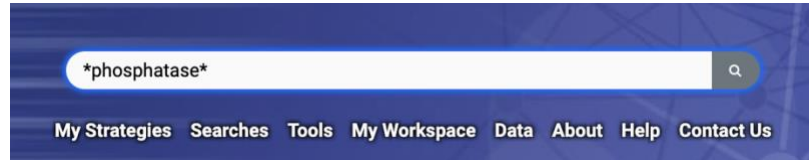
## Useful information:

All VEuPathDB genomics sites (e.g., FungiDB) have an integrated phyletic pattern search that uses OrthoMCL to return lists of genes. For example, you use the “Orthology Phylogenetic Profile” search 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.



## Combining searches in OrthoMCL

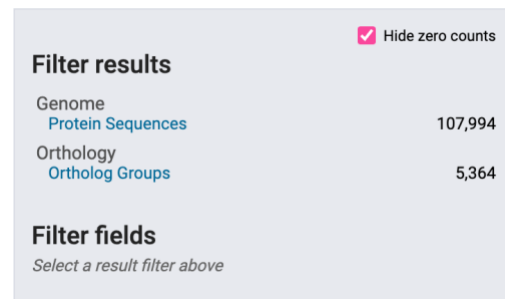
- Find all fungal proteins that are likely to be phosphatases and that do not have orthologs outside of fungal kingdom.
  - Use the site search to look for \*phosphatase\* (use asterisks to find any combination of the word “phosphatase”).



\*phosphatase\*

My Strategies Searches Tools My Workspace Data About Help Contact Us

How many protein sequences were identified? How many ortholog groups did you identify?

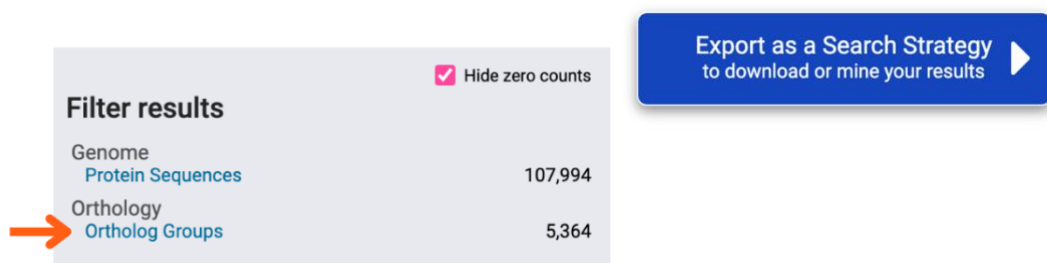


☒ Hide zero counts

<b>Filter results</b>	
Genome	
Protein Sequences	107,994
Orthology	
Ortholog Groups	5,364

**Filter fields**  
Select a result filter above

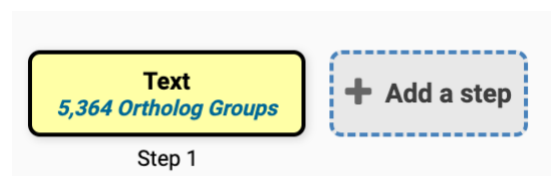
- Display the ortholog groups containing the word phosphatase and export the results as a search strategy.



☒ Hide zero counts

<b>Filter results</b>	
Genome	
Protein Sequences	107,994
Orthology	
Ortholog Groups	5,364

Export as a Search Strategy  
to download or mine your results

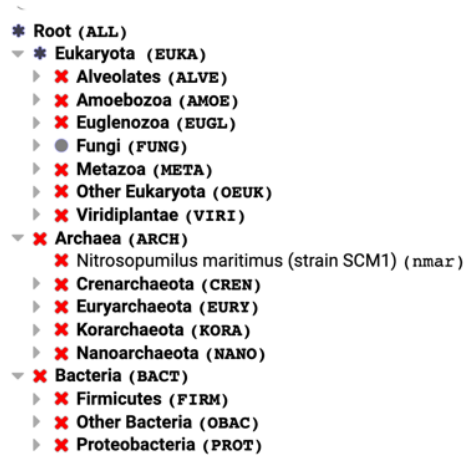


**Text**  
5,364 Ortholog Groups

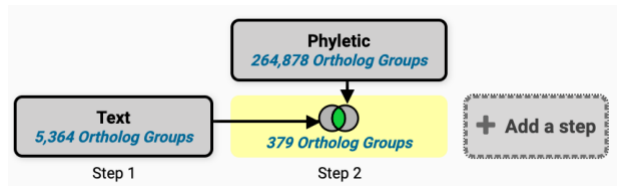
Step 1

+ Add a step

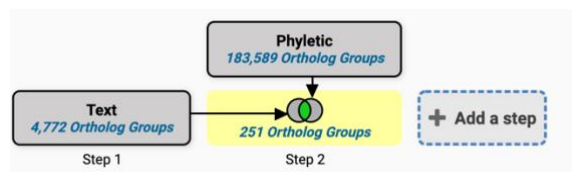
- c. Add a step and run a phyletic pattern search for groups that contain any fungi proteins but do not contain any other organism outside fungi. (hint: make sure everything has a red x on it except for fungi, which should be a grey circle (no constrains)).



How many groups did the search return?

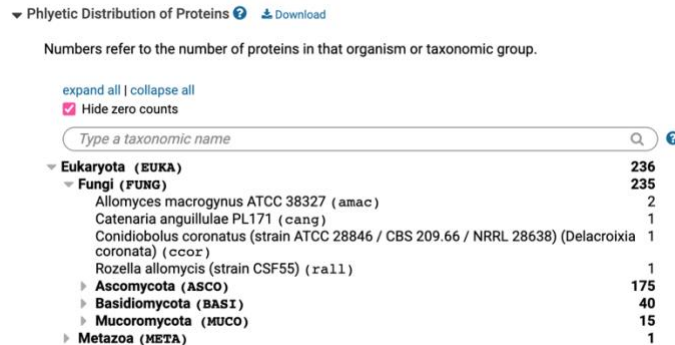


Strategy URL: <https://orthomcl.org/orthomcl/app/workspace/strategies/326574223/430551843>



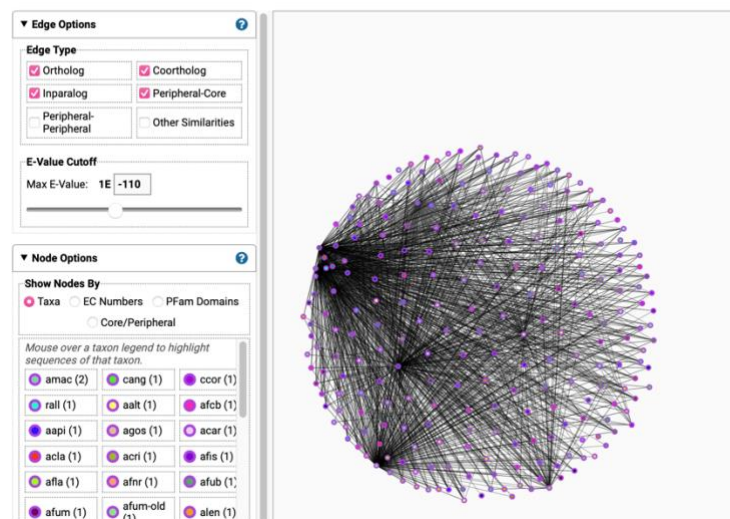
## Exploring a specific OrthoMCL group - examining the cluster graph.

- Visit the OrthoMCL record page for the group OG6\_115064.
- Examine the phyletic distribution tree. What taxa does this group contain?



- Examine the cluster graph for this group (it can be accessed at the bottom of the page)

### Cluster Graph: OG6\_115064 (245 proteins) ⓘ



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.

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

In the *Node Options* panel, you can 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).