
OrthoMCL 7

Introduction & Basic Functionality¹

Understanding OrthoMCL

Terminology	2
What is OrthoMCL 7?	3
Important changes and new features in OrthoMCL-7	4
The OrthoMCL Algorithm	5

Using OrthoMCL

Layout of the Home Page of OrthoMCL.org	6
Tools available on OrthoMCL	7
Ortholog Group Searches	9
Protein Searches	12
Orthogroup Pages	14
Phyletic distribution	14
Group Summary	15
Summary of Pfam domains	17
List of proteins (table + phylogenetic tree)	17

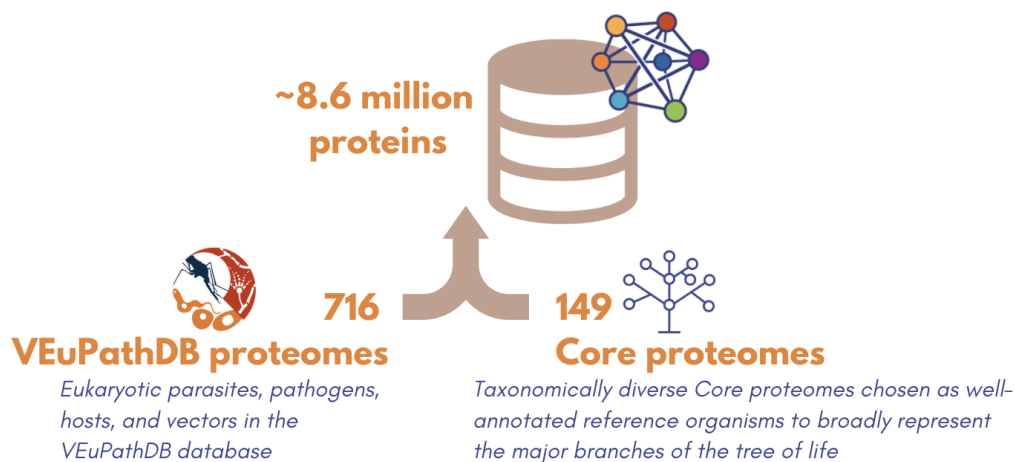
¹ Updated on March 27, 2025

Terminology

- **Orthogroup:** A group of orthologous protein sequences, that is, a set of genes from multiple species that are all descended from a single gene in the last common ancestor (Emms and Kelly, 2019).
 - Orthologs and paralogs constitute two major types of **homologs**.
 - **Orthologs** evolved from a common ancestor by speciation, and **paralogs** are related by gene duplication events (Fitch 1970, 2000).
 - OrthoMCL orthogroups contain both orthologs across multiple species and recently duplicated paralogs, sometimes called in-paralogs.
- **Proteome:** Complete set of proteins from an organism
- **E-value:** The number of random alignments with a score equal to or greater than the observed alignment that one would expect to find by chance in a database of a given size.
 - A small e-value indicates that the quality of a given alignment between 2 sequences (proteins or DNA) is unlikely to occur by chance and is likely an evolutionary relationship
 - A common threshold for judging homology, especially when using BLAST, is an E-value of $1e-5$ or lower ($E \leq 1e-5$).

What is OrthoMCL 7?

OrthoMCL 7 is the latest version of OrthoMCL, a **genome-scale database and website** (orthomcl.org) that uses protein sequence similarity and phylogenetic relationships among proteins to create groups of orthologous protein sequences called orthogroups. Proteins in OrthoMCL orthogroups have been shown to display a high degree of functional conservation.



OrthoMCL is integrated with the VEuPathDB collection of websites, which receive approximately 40,000 unique visitors and provide over 500GB of data downloads per month. The OrthoMCL dataset (as of March 2025) contains 8.6 million proteins, which represent 716 complete proteomes from eukaryotic parasites, pathogens, hosts, and vectors in the VEuPathDB database ([Alvarez-Jarreta, et al. 2024](#)), plus a set of 149 taxonomically diverse Core proteomes that have been chosen as well-annotated reference organisms that broadly represent the major branches of the tree of life including 62 phyla spanning plants, animals, fungi, protists, bacteria, and archaea. For some species, VEuPathDB and OrthoMCL contain proteomes from multiple strains and/or genome assemblies. All OrthoMCL proteins are clustered into 789,914 orthogroups (290,748 orthogroups when orphan groups with only a single protein are excluded).

OrthoMCL provides protein orthology links between any of the organisms in the database. **A protein or a set of proteins in one organism can be transformed into the set of equivalent proteins in another organism.** Given the complex history of gene duplication and speciation in protein families, this relationship may involve a number of proteins rather than just a one-to-one relationship. See the *Shared Ortholog* protein searches for more details.

OrthoMCL can also provide annotation for proteins of unknown function such as predicted proteins in newly sequenced genomes, transcriptome assemblies, metagenome and metatranscriptome data sets by mapping to OrthoMCL groups. Existing proteins in public databases that lack functional annotation (e.g. “hypothetical protein”, “unspecified product”, etc.) may also benefit from assignment to OrthoMCL orthogroups. See the **Map proteins to OrthoMCL** function in the **Tools** menu.

Important changes and new features in OrthoMCL-7

OrthoMCL-7 is a major update to OrthoMCL that improves the accuracy of the clustering algorithm, updates all genomes and protein annotations, removes redundant data, and uses improved computational methods to address the challenge of scale for clustering millions of proteins. OrthoMCL-7 adds new features including approximate maximum-likelihood phylogenetic trees for each orthogroup, and it identifies sets of similar orthogroups which may represent protein superfamilies with related functions. The phylogenetic tree graphics on the orthomcl.org website are interactive, allowing the user to quickly limit the visualization to just chosen species and/or functional domains, so that orthology relationships and the evolutionary history of gene duplications can be more easily studied. To summarize-

- All proteomes have been refreshed with current (March, 2025) UniProt sequences and annotation for each species
- Duplicated "old" genomes from OrthoMCL-6 have been removed
- Orthogroups are now clustered with OrthoFinder, which uses DIAMOND for much faster BLAST analysis, improves normalization of BLAST scores, and includes phylogenetic information to improve clustering.
- Orthogroup web pages now include a phylogenetic tree of all proteins, which can be customized to include only specific organisms, Pfam groups, or words in protein descriptions
- Orthogroup web pages now include Similar Groups, which can identify large super-families of related proteins
- **Map Proteins to OrthoMCL** service is now part of the orthomcl.org website and BLAST results will be saved in a private user workspace.
- A list (or all proteins) from one organism can be transformed into orthologs in another organism.

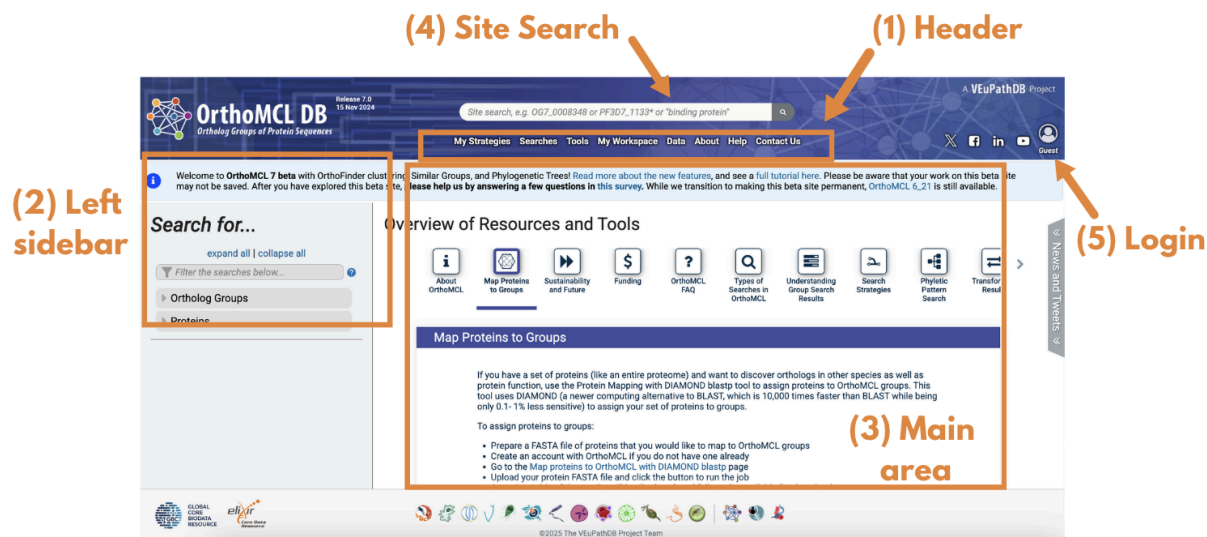
The OrthoMCL Algorithm

OrthoMCL uses OrthoFinder to create clusters of similar proteins.

- Computing the **Core clusters**
 - To avoid an exponential increase in compute time as more genomes are added to the database, 149 taxonomically diverse Core proteomes have been chosen to create a base set of protein clusters.
 - For the Core, all proteins are compared against each other with DIAMOND blastp, a drop-in replacement for NCBI BLAST that is 1000 times faster with minimal loss of sensitivity.
 - BLAST e-values are normalized for protein length and evolutionary distance, then the all-vs-all matrix of similarity values is used to create a graph and the MCL algorithm finds optimal clusters of tightly linked nodes within the graph.
 - Clusters are optimized using phylogenetic trees to create hierarchical ortholog groups.
- Computing the **Peripheral clusters**
 - After the Core clusters are computed, proteomes from additional Peripheral organisms (from VEuPathDB) are added one proteome at a time, by mapping proteins to the most similar cluster.
 - Proteins that do not have a BLAST match with e-value better than 1e-5 become a set of Residuals.
- Computing the **Residual proteomes**
 - After all Peripheral proteomes are processed, the set of Residuals are clustered among themselves with another cycle of all-vs-all BLAST and OrthoFinder to create Residual orthogroups.

Approximate maximum likelihood phylogenetic trees are calculated for the protein sequences in each orthogroup. The computation uses MAFFT for multiple alignment and FastTree to build the trees. Trees are saved as text files in Newick format and the tree graphic is drawn on demand using the tidytree R library (<https://github.com/YuLab-SMU/tidytree>).

Layout of the Home Page of OrthoMCL.org



(1) The **Header** contains menus for

- **Searches** for ortholog groups and proteins, described later in this document.
- **Tools:** These include the **Map proteins to OrthoMCL** tool and a **BLAST** tool, and are described in the next section.
- **Data:** Links to analysis methods and data file downloads.
- **Help:** Links to frequently asked questions (**FAQ**) and **Learn how to use VEuPathDB**, which includes workshops, webinars, tutorials, etc. and user documentation.
- **Contact Us:** Link to a form to message the help desk.
- **My Strategies** and **My Workspace** provide links to a user's previous activity on the site when logged in.

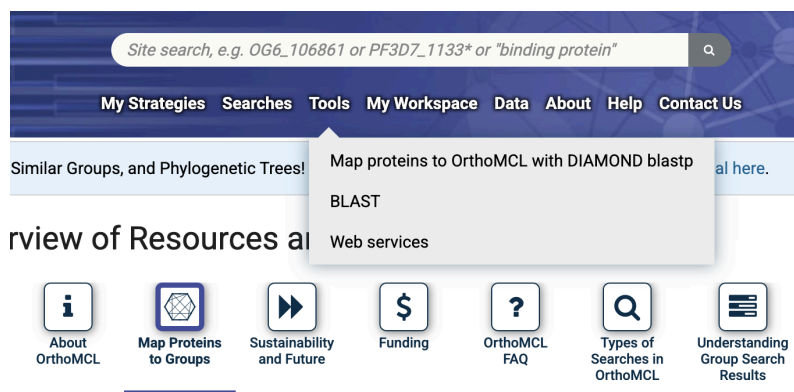
(2) The **Left sidebar** contains links to the same searches as the **Searches** menu in the header.

(3) The **Main area** contains a variety of short help cards and links to longer tutorials.

(4) **Site Search:** Above the menus is a text box for the Site Search which allows users to search for any text in protein descriptions, protein and group IDs, Pfam and EC Number IDs and descriptions.

(5) **Login:** Hovering on the person icon opens access to the free Registration/Login form at the far right of the header. When logged in, users can access their previous work by clicking **My Strategies** and **My Workspace**.

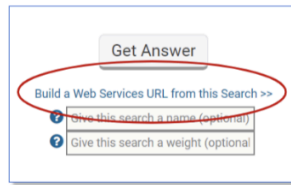
Tools available on OrthoMCL



Hovering on the **Tools** menu in the header allows access to the following tools:

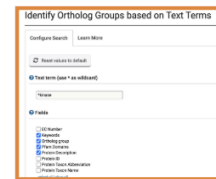
- The **Map proteins to OrthoMCL** tool allows for a bulk analysis of a large set of unknown proteins, such as from automated protein prediction on a newly sequenced and assembled genome or metagenome.
 - A FASTA formatted file of protein sequences can be uploaded, then, with a single button click (there are no user customizable parameters) the input proteins are compared to all OrthoMCL proteins with DIAMOND blastp.
 - The results are provided as a tab delimited text file that includes the following columns:
 - input protein
 - closest matching OrthoMCL protein
 - the protein description
 - the e-value of the match
 - the corresponding OrthoMCL orthogroup
 - The orthogroups provide a broader set of annotations that might not be available for just the closest matching protein, as well as the Similar Groups feature that might provide additional functional insight from related orthogroups.
- **BLAST** search takes a single sequence as input in plain text format, and uses NCBI BLAST for similarity search against all OrthoMCL proteins. Both protein (blastp) and DNA sequence (blastx) input are allowed, with the DNA sequence being automatically translated into amino acids in all 6 reading frames before searching the protein database. The e-value (sensitivity), low complexity filter, and number of matching sequences can be adjusted by the user. BLAST provides a very high sensitivity sequence similarity search, so users should be cautious in the interpretation of low significance matches.
- **Web services** query allows command line scripted REST access to all OrthoMCL searches. The result of a web service request is a list of records (genes, compounds, etc.) in one of various formats (json, csv, etc.). REST services can be executed in a browser by typing a specific URL.

To create a web services URL, go to the desired Search page and fill in the required parameters. Then instead of clicking "Get Answer", click the blue text "Build a Web Services URL for this search>>". Copy the URL to any browser or use it in a command line script.



For example, the URL gives the same result as using the web page for a text term search for "*kinase" in the Ortholog group keywords field.

[https://feature.orthomcl.org/orthomcl.feature/service/record-types/group/searches/GroupsByText/reports/standard?text_expression=*kinase&document_type=group&text_fields=%5B%22keywords%22%2C%22primary_key%22%2C%22PFams%22%2C%22ProteinDescription%22%5D&reportConfig={\"attributes\":\[\"primary_key\",\"number_of_members\",\"keywords\",\"descriptions\",\"ec_numbers\"\],\"tables\":{},\"attributeFormat\":\"text\"}](https://feature.orthomcl.org/orthomcl.feature/service/record-types/group/searches/GroupsByText/reports/standard?text_expression=*kinase&document_type=group&text_fields=%5B%22keywords%22%2C%22primary_key%22%2C%22PFams%22%2C%22ProteinDescription%22%5D&reportConfig={\)



Ortholog Group Searches

The **Ortholog Groups** search can be accessed from the header (Searches menu) or from the left sidebar.

There are **nine different options for searching for ortholog groups**. Search results are made available as a list of ortholog groups in tabular format. Searches can be linked together by VEuPathDB **Search Strategies** ([see this tutorial for an example](#)) to create complex queries that span many types of information.

Search for...

expand all | collapse all

Filter the searches below...

▼ Ortholog Groups

- q All Groups
- q EC Number
- q Group ID(s)
- q Number of Sequences
- q Number of Taxa
- q PFam ID or Keyword
- q Percent Identity
- q Phyletic Pattern
- q Text Terms

- **All Groups:** Clicking on the “All Groups” search simply returns all available ortholog groups in a search strategy as shown below, without any filtering.

Unnamed Search Strategy *

All Groups 789,914 Ortholog Groups

Step 1

789,914 Ortholog Groups

Ortholog Group Results

Ortholog Group	Total Number Proteins	Keywords
OG7_0000011	26208	source; containing protein; uniProtKB/TrEMBL; Acc; domain containing protein; rrm; rrm domain c
OG7_0000397	22264	cytochrome; P450; cytochrome P450; unknown; source

- **EC Number:** This search finds ortholog groups with EC number(s) of interest. An EC number, or Enzyme Commission number, is a numerical classification system for enzymes based on the chemical reactions they catalyze. The search text box recognizes both numerical EC number IDs and text terms found in the descriptions of EC enzymes.

There are a few options for this search.

- Type in the EC number, e.g., 2.7.1.1
- Start typing in the enzyme name, e.g., “kinase” and choose from the drop-down list, 2.7.1.1 Hexokinase
- Type in the enzyme name with wildcard characters (*), e.g., “*kinase*”

Identify Ortholog Groups based on EC Number

Configure Search Learn More

Reset values to default

EC Number or Name

kinase{


- 2.7.1.1 (hexokinase)
- 2.7.1.100 (5-methyl-5-thioribose kinase)
- 2.7.1.101 (tagatase kinase)
- 2.7.1.105 (5-phosphofructo-2-kinase)
- 2.7.1.107 (diacylglycerol kinase (ATP))
- 2.7.1.108 (dialcohol kinase)
- 2.7.1.11 (5-phosphofructokinase)
- 2.7.1.113 (deoxyguanosine kinase)
- 2.7.1.12 (glucokinase)
- 2.7.1.127 (inositol triphosphate 3-kinase)

Read the contents of the “Learn More” tabs in the searches for more information and helpful tips.

- **Group ID(s):** Find Ortholog Groups by ID(s) assigned in the current or previous releases of OrthoMCL. Options include
 - Entering a list of IDs
 - Uploading a text file containing a list of IDs
 - Uploading a text file from a URL
- **Number of Sequences:** Configure the search to find Ortholog Groups that contain a specified number of proteins, or use the Advanced search parameters to specify the number of Core proteins and Peripheral proteins.
 - a. If the number of Core proteins is set to 0, then the search will find only Residual groups.
 - b. If at least 1 Core protein is required, then the search will find only Core groups.
 - c. If groups are set to a minimum of 2 proteins, then singleton groups will be removed from the search results.

Identify Ortholog Groups based on Number of Sequences

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

 Reset values to default

Number of All Proteins

to

Advanced Parameters

OrthoMCL contains two sets of genomes. A Core set of 150 genomes have been chosen as well annotated reference organisms that broadly represent the major branches of the tree of life. All of the additional non-core VEuPathDB organisms (pathogens, hosts, and vectors) are called Peripheral organisms, in some cases including multiple strains and genome assemblies for the same species. The search ranges below can be used to filter the number of proteins (taxa) in each orthogroup by the number of Core and Peripheral (minimum and maximum) in each group. The number of Core and Peripheral will sum to the number of all proteins (taxa) in the group.

Number of Core Proteins

to

Number of Peripheral Proteins

to

[Get Answer](#)

- **Number of Taxa:** Configure the search to find Ortholog Groups that contain a specified number of all taxa, or use Advanced parameters to specify the number of Core and Peripheral taxa.
- **Pfam ID or Keyword:** Pfam is a database of protein functional domains. Search with Pfam IDs and description terms (keywords) to find ortholog groups that contain proteins with these domains.
- **Percent Identity:** Find Ortholog Groups by the median percent identity between all protein sequences within the group. This is a measure of group cohesiveness. High median percent identity indicates a tighter group with few outliers, while lower median identity indicates a group that is more dispersed.

- **Phyletic Pattern:** The Phyletic pattern is the taxonomic distribution of the proteins in an orthogroup. The Phyletic pattern search specifies particular taxa using a selectable tree menu. Click on the grey circles to include or exclude individual organisms or entire clades. Multiple clicks change the type of selection for that term. A click on the grey circle cycles through the options: must be in group (green check ✓), at least one subtaxon must be in group (yellow check ✓), must not be in group (red ✗), and no constraints (grey circle ○). A grey asterisk (*) is shown when multiple constraints are selected for sub-groups.

The phyletic search can be controlled more precisely (both for the number of taxa in a clade and the number of proteins in those taxa) by typing a set of search terms in the expression box. The phyletic expression syntax is shown at the bottom of the search page and explained in more detail in the **Learn More** tab of the search.

Expression:

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

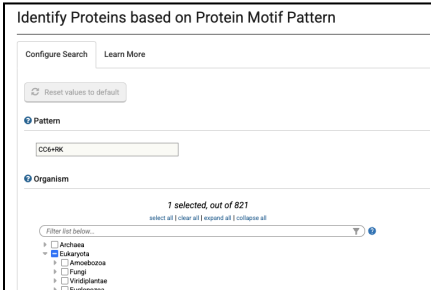
- * Root (ALL)
 - * Eukaryota (EUKA)
 - * Alveolates (ALVE)
 - Chromera velia CCMP2878 (cvel)
 - Vitrella brassicaformis CCMP3155 (vbra)
 - * Apicomplexa (APIC)
 - Gregarina niphandrodes Unknown strain (gnip)
 - Porospora cf. gigantea A (pcga)
 - Porospora cf. gigantea B (pcgb)
 - * Aconoidasida (ACON)
 - ▶ ✓ Haemosporida (HAEM)
 - ▶ ● Piroplasmida (PIRO)
 - ▶ ● Coccidia (COCC)
 - ▶ ● Ciliates (CILI)
 - ▶ ● Amoebozoa (AMOE)
 - ▶ ● Euglenozoa (EUGL)
 - ▶ ● Fungi (FUNG)
 - * Metazoa (META)
 - ▶ ● Arthropoda (ARTH)
 - * Chordata (CHOR)
 - Branchiostoma floridae (bflo)
 - Xenopus tropicalis (xtro)
 - ▶ ● Actinopterygii (ACTI)
 - ▶ ● Aves (AVES)
 - ▶ ✗ Mammalia (MAMM)
 - ▶ ● Tunicates (TUNI)
 - ▶ ● Nematodes (NEMA)
 - ▶ ● Other Metazoa (OMET)
 - ▶ ● Other Eukaryota (OEUK)
 - ▶ ✗ Viridiplantae (VIRI)
 - ▶ ● Archaea (ARCH)
 - ▶ ● Bacteria (BACT)

- **Text Terms:** This general search allows text terms search among a choice of any or all of 8 fields, including keywords, protein description, etc. Wildcard characters can be used to find compound words, so that *kinase will find both hexokinase and fructokinase.

Protein Searches


The **Proteins** searches can be accessed from the header (**Searches** menu) or from the “**Search for...**” left sidebar.


There are **11 different options for searching for proteins of interest**. In each case, results for the search are made available as a list of proteins in tabular format. Searches can be linked together by VEuPathDB **Search Strategies** to create complex queries that span many types of information.


1. **All Proteins:** This search simply returns all available proteins without filtering.
 2. **BLAST:** This search can be used to find sequences that have BLAST similarity to your input query sequence (a protein in plain text). This search uses NCBI-BLAST to determine sequence similarity. This is the same as the BLAST search in the Tools menu.
 3. **EC Number:** This search finds proteins with EC number(s) of interest. An EC number, or Enzyme Commission number, is a numerical classification system for enzymes based on the chemical reactions they catalyze. The search text box recognizes both numerical EC number IDs and text terms found in the descriptions of EC enzymes. There are a few options for this search.
 - o Type in the EC number, e.g., 2.7.1.1
 - o Start typing in the enzyme name, e.g., "kinase" and choose from the drop-down list, 2.7.1.1 Hexokinase
 - o Type in the enzyme name with wildcard characters (*), e.g., "*kinase*"
 4. **Group or Protein Name:** Search by protein name or description term (* wildcards are allowed).
 5. **Pfam ID or Keyword:** Pfam is a database of protein functional domains. Search with Pfam IDs and description terms (keywords) to find proteins with these domains.
 6. **Protein ID(s):** Find proteins by ID(s) assigned in the current or previous releases of OrthoMCL. Options include
 - o Entering a list of IDs
 - o Uploading a text file containing a list of IDs
 - o Uploading from a URL
 7. **Protein Motif Pattern:** Find Proteins that contain specific amino acids, either as a simple string of single letter amino acid abbreviations or with a motif pattern in a regular expression format, such as CC6+RK, which means "two cysteines followed by one or more hydrophobic amino acids, followed by arginine, then lysine". The specified protein motif can be searched
- [Protein Motif Pattern](#)
[Shared Orthologs By Organism](#)
[Shared Orthologs From List](#)
[Taxonomy](#)
[Text Terms](#)
- 


Search for...


expand all | collapse all


 Filter the searches below...





 Ortholog Groups


 Proteins


 All Proteins


 BLAST


 EC Number


 Group or Protein Name


 PFam ID or Keyword


 Protein ID(s)

 Protein Motif Pattern

 Shared Orthologs By Organisms

 Shared Orthologs From List

 Taxonomy

 Text Terms

Identify Proteins based on Protein Motif Pattern

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

Reset values to default

Pattern

Organism

1 selected, out of 621

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

Filter list below...

- ☒ Archaea
- ☒ Bacteria
 - ☐ Actinobacteria
 - ☐ Annelidobacteria
 - ☐ Fungi
 - ☐ Verrucosporidia
 - ☐ Euglenozoa
 - ☐ Metazoa
 - ☒ Alveolates
 - ☐ Chromera vella CCMP2378
 - ☐ Ciliates
 - ☒ Ascomycetes
 - ☐ Coccioidia
 - ☐ Porospora of gigantes B
 - ☒ Basidiomycetes
 - ☒ Hemionsporidia
 - ☐ Phragmotridia
 - ☐ Porospora of gigantes A
 - ☐ Gregarina raphanobrodes Unknown strain
 - ☐ Vanilla brissaciformis CCMP1155
 - ☐ Other Eukaryota
 - ☐ Biocena

[Get Answer](#)

within particular species by picking from the organism tree.

8. **Shared Orthologs by Organism:** This new search identifies all orthologous proteins between two OrthoMCL organisms and shows the orthology relationships.

The screenshot shows a web interface titled "Identify Proteins based on Shared Orthologs By Organisms". It features two tabs: "Configure Search" and "Learn More". Below the tabs is a "Reset values to default" button. The main section is divided into two parts: "Query Organism" and "Target Organism". The "Query Organism" field contains the text "Plasmodium falciparum 3D7". The "Target Organism" field contains the text "Trypanosoma vivax Y486". At the bottom right, there is a "Get Answer" button.

9. **Shared Orthologs from List:** This search takes a list of proteins (from one or more OrthoMCL organisms) as input and finds orthologs for each query in a target organism. This allows users to transform an interesting list of genes found in one organism into the equivalent genes in another organism, showing the orthology relationships.

The screenshot shows a web interface titled "Identify Proteins based on Shared Orthologs From List". It features two tabs: "Configure Search" and "Learn More". Below the tabs is a "Reset values to default" button. The main section is divided into two parts: "Query Sequence List" and "Target Organism". The "Query Sequence List" field contains the text "pfalPF3D7_1474700". The "Target Organism" field contains the text "Trypanosoma vivax Y486". At the bottom right, there is a "Get Answer" button.

10. **Taxonomy:** This search allows you to choose particular taxa and return a list of OrthoMCL proteins belonging to those taxa.
11. **Text Terms:** This general search allows you to use text terms and search among your choice of any or all of 8 fields, including keywords, protein description, etc.

Orthogroup Pages

Each orthogroup has its own web page at **OrthoMCL.org**. For example, the orthogroup OG7_0001789 has its page [here](#).

The screenshot shows the OrthoMCL.org interface for orthogroup OG7_0001789. It is divided into three main sections:

- (1) Summary information:** Located at the top, it provides key statistics: Group Type (Core), Total Number of Proteins (647), Number of Core Proteins (124), Number of Peripheral Proteins (523), Keywords (transporter, zinc; unknown; zinc transporter; source), EC Numbers (1.3.1.74(3)), Top Pfam Domains (PF01545 (621), PF03645 (2), PF03102 (1), PF07993 (1)), and Previous Groups (OG3_10134, OG4_10082, OG5_127157, OG6_100508, OG6_105399, OG6_106189...).
- (2) Table of contents:** A sidebar on the left with a search bar and expand/collapse controls. It lists sections: Phyletic distribution, Group summary, Summary of Pfam domains, and List of proteins.
- (3) Main page:** The central area showing the 'Phyletic distribution' section, which includes a 'Download' button, a note about protein counts, and a taxonomic tree with counts for various groups.

1. Summary information:

The top of the page provides some summary information about the group including **Group Type**, whether the group contains proteins from Core organisms (or just Residual proteins from Peripheral organisms), the **Total Number of Proteins** (647 in this example), **Number of Core Proteins**, **Number of Peripheral Proteins**, **Keywords** from the protein descriptions (zinc transporter in this example), **EC Numbers** and **Top Pfam Domains** in the proteins (if any).

- 2. Table of contents:** Below the summary is a table of contents for the page. This section can be collapsed by clicking the << icon to provide more horizontal space to view the other features of the group page.
- 3. Main page:** The various sections are described below.

Phyletic distribution

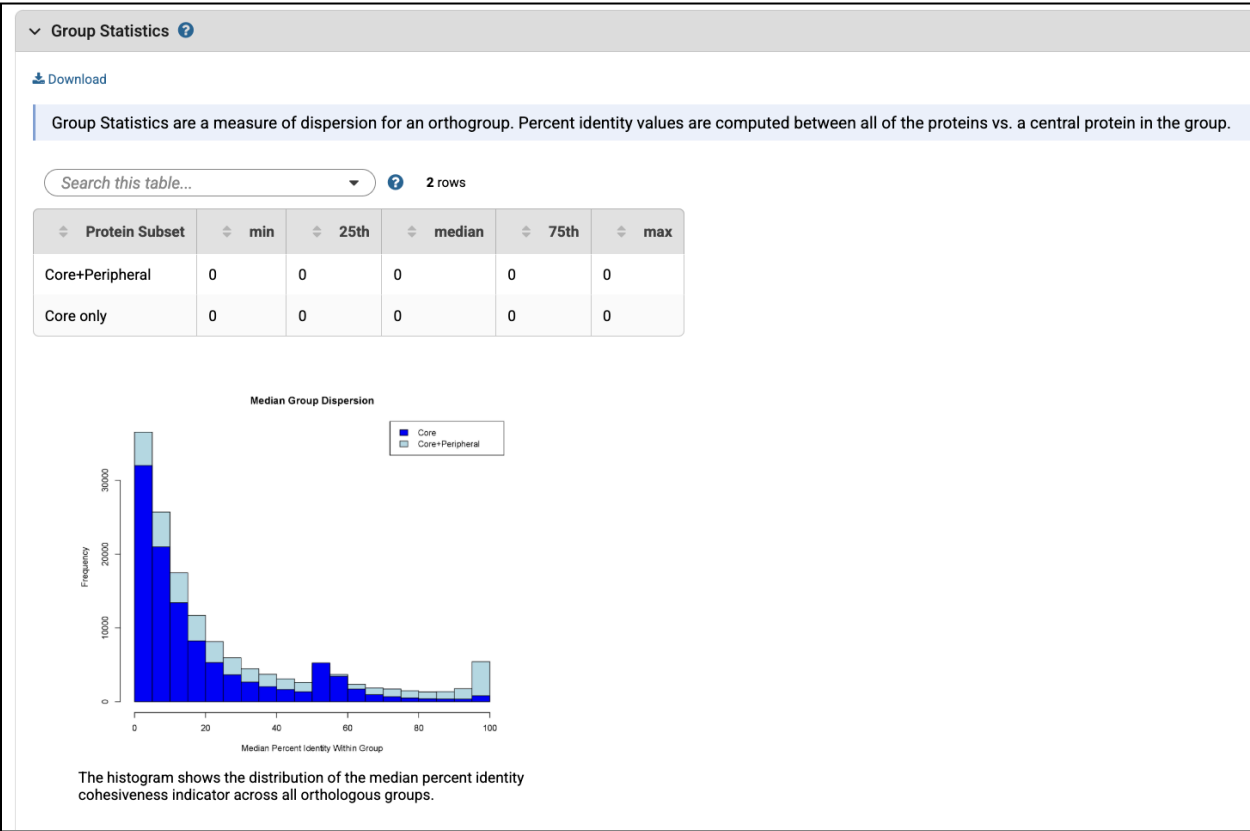
The Phyletic distribution of the proteins- referring to the evolutionary history and the pattern of occurrence of the proteins across different lineages or taxonomic groups- in the orthogroup is shown across the tree of life with **counts** (right side of page) for each clade and individual taxon. By default, empty branches are hidden (with the **Hide zero counts** button), but the full tree contains all VEuPathDB and OrthoMCL Core organisms.

Phyletic distribution	
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)	641
Alveolates (ALVE)	3
Amoebozoa (AMOE)	17
Euglenozoa (EUGL)	66
Fungi (FUNG)	282
Metazoa (META)	211
Other Eukaryota (OEUK)	48
Viridiplantae (VIRI)	14
Bacteria (BACT)	6
Other Bacteria (OBAC)	2
Proteobacteria (PROT)	4

Group Summary

The **Group Summary** contains four sub-sections-

- **All previous groups:** This section provides a reference list of orthogroup numbers for this particular orthogroup from all previous iterations of OrthoMCL.
- **Group Statistics** summarize the within-group dispersion of the proteins. The intra-group e-values of group members is shown as a 5-value table (median, maximum, minimum, 25th and 75th percentile). The median scores can be used as a search to find more tightly cohesive groups vs. more dispersed groups that might contain sub-groups or outliers.



- **Similar Groups** are listed which share significant BLAST similarity between the central proteins in the two groups. Groups with pairwise BLAST scores (**E-value**) better than 1e-5 are considered similar.

Similar Groups

Download

Similar groups are computed by a Diamond BLASTX comparison of the Representative Sequences (centroids) of all groups with each other. Groups with pairwise BLAST scores better than 1e-5 are considered similar.

Search this table... 10 rows

Similar Group ID	# Proteins	Keyword	Pfam Domain	E-value
OG7_0060299	3	uncharacterized protein	PF01545	1.27E-58
OG7_0001787	800	zinc; transporter; unknown; source; zinc transporter	PF00076, PF01545, PF13418, PF13793, PF13854, PF14572	1.07E-37
OG7_0184122	2	proton-coupled zinc antiporter slc3	PF01545	4.38E-33
OG7_0292593	3	efflux	PF01545	2.68E-30
OG7_0001788	36	cation; efflux; cation efflux; transporter	PF01545, PF16916	1.2E-28
OG7_0010944	117	member 6; solute carrier; 30 member 6; transporter	PF01545	1.54E-17
OG7_0004656	4	cation; cation efflux; transporter	PF01545, PF16916	1.26E-10
OG7_0004658	19	cation; efflux; cation efflux; efflux system; cation efflux system; efflux system protein; transporter; cation efflux system protein	PF01545, PF02579, PF16916	7.33E-7
OG7_0004657	7	efflux; cation efflux; efflux system protein	PF01545, PF16916	0.0000897
OG7_0103976	2	uncharacterized protein	PF01545	0.000229

- **Summary of EC numbers** is available wherever EC numbers are available, i.e., for enzymes.

Summary of EC Numbers

Download

EC Number	EC Description	# Proteins
1.3.1.74	2-alkenal reductase [NAD(P)(+)]	3

Summary of Pfam domains

This is a table that shows a list of Pfam domains (**Accession**) for all proteins in the group, a **Description** of each domain, the number of proteins that contain each domain (**Count**), and the cartoon that is used as a label for each domain in the Protein Table below (**Legend**).

Summary of Pfam domains

Summary of Pfam domains

Download

Search this table... 4 rows

Accession	Description	Count	Legend
PF01545	Cation efflux family	621	
PF03645	Tctex-1 family	2	
PF07993	Male sterility protein	1	
PF03102	NeuB family	1	

List of proteins (table + phylogenetic tree)

The **List of All Proteins** section opens up to show a table of all proteins in the orthogroup.

List of proteins

List of All Proteins

Download

This section features a Clustal Omega alignment tool (max 1000 sequences) and a tree visualization of proteins in this group. The proteins can be filtered in a number of ways.

- Filter via text search on any/specific columns in the table.

Read more

Search this table... 647 rows

Filters: Proteins Pfam domains Core/Peripheral Organism

	Domain architecture	Accession	Description	Organism	Clade	Core/Peripheral	Length	EC Numbers
		gtheIL1JV76	Uncharacterized protein	Guillardia theta	Viridiplantae	Core	327	N/A
		mbalMBAL_002101	unknown	Mastigamoeba balamuthi ATCC 309	Amoebozoa	Peripheral	774	N/A
		ngruNAEGRDRAFT_60718	unknown	Naegleria gruberi strain NEG-M	Other Eukaryota	Peripheral	312	N/A
		nlovIC9374_005754	unknown	Naegleria lovaniensis strain ATCC 30	Other Eukaryota	Peripheral	460	N/A
		nftyINFTy_066570	solute carrier family 30 (zinc transp	Naegleria fowleri strain Ty	Other Eukaryota	Core	335	N/A
		nfoaIFDP41_005479	unknown	Naegleria fowleri strain ATCC 30894	Other Eukaryota	Peripheral	496	N/A
		cbraIA0A388KN58	Uncharacterized protein	Chara braunii	Viridiplantae	Core	754	N/A
		ppatIA0A2K1KFV0	Uncharacterized protein	Physcomitrium patens	Viridiplantae	Core	422	N/A

- The **table** contains a cartoon of Pfam domains in each protein in the **Domain architecture** column (Pfam IDs are shown if the pointer is hovered over the cartoon), the protein ID (**Accession**), protein **Description**, **Organism**, **Clade**, protein **Length**, and **EC numbers** (if any).
- A **phylogenetic tree** is shown at the left of the table. (*The tree is not shown if more than 1000 proteins are listed in the table- to see the tree, please filter as shown below.*) The tree is calculated directly from the proteins in the orthogroup by maximum likelihood analysis, so it does not represent a reference taxonomy. Branch points in the tree represent speciation and gene duplication events. If all of the sub-branches contain genes in a single species (or a single clade), then this branch point represents a recent duplication event and the genes can be considered in-paralogs.

- A number of **Filters** are available above the table that can be used to reduce the number of proteins in the tree.
 - The **Pfam domains** filter can limit the tree (and table) to only show proteins that contain selected Pfam domains.
 - The **Core/Peripheral** filter can show proteins from only Core (or only Peripheral) species.
 - The **Organism** filter can limit the tree to only show proteins from selected species. The organism filter can be very useful to study the pattern of orthology between a small number of species.
 - Individual proteins can be selected in the table by clicking on the checkbox, then the tree can be filtered to just the selected proteins with the **Proteins** filter.
 - The **text** filter (**Search this table...**) on the left can be used to limit the table to proteins that contain a specific word in the description.
 - All filters can be removed by clicking the **Reset** arrow ↶ to the right of the **Organism** filter.

Search this table... 10 rows (filtered from a total of 1,170)

Filters: Proteins* Pfam domains Core Organism ↶

* You have checked 6 proteins in the table.

Filter to keep only these proteins

	Domain architecture	Accession	Description	Organism	Clade	Core/Perip
<input type="checkbox"/>		rirr GLOIN_2v1434745	Non-specific serine/threonine pro	Rhizophagus irregularis DAOM 11	Fungi	Core
<input type="checkbox"/>		rirr GLOIN_2v1426031	Pkinase_fungal domain-containin	Rhizophagus irregularis DAOM 11	Fungi	Core
<input checked="" type="checkbox"/>		sscl sscle_16g110010	unknown	Sclerotinia sclerotiorum 1980 UF	Fungi	Core
<input type="checkbox"/>		sscl sscle_14g101660	unknown	Sclerotinia sclerotiorum 1980 UF	Fungi	Core
<input checked="" type="checkbox"/>		sscl sscle_06g054860	Atypical serine/threonine protein	Sclerotinia sclerotiorum 1980 UF	Fungi	Core
<input type="checkbox"/>		sscl sscle_15g105640	Atypical serine/threonine protein	Sclerotinia sclerotiorum 1980 UF	Fungi	Core
<input checked="" type="checkbox"/>		sscl sscle_16g110000	Protein kinase domain-containin	Sclerotinia sclerotiorum 1980 UF	Fungi	Core
<input type="checkbox"/>		sscl sscle_04g040280	Atypical serine/threonine protein	Sclerotinia sclerotiorum 1980 UF	Fungi	Core
<input type="checkbox"/>		sscl sscle_11g085040	Protein kinase domain-containin	Sclerotinia sclerotiorum 1980 UF	Fungi	Core

If proteins are selected in the table by clicking the checkbox located between the tree and the Pfam cartoon, then the **Run Clustal Omega for selected proteins** button below the table becomes active.

By clicking the button, the selected proteins are aligned with Clustal Omega.

- The Neighbor Joining tree created for this alignment by Clustal can be sent to ITOL for visualization.
- The entire tree for the orthogroup can be downloaded as a newick file for visualization with any phylogenetic tool.

Questions? Comments?

Contact us- help@veupathdb.org
