Basic Functionality of OrthoMCL 7

Contents

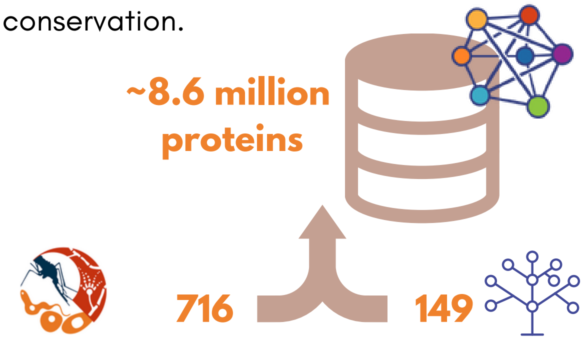
Terminology

- **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

- **Proteome**: Complete set of proteins from an organism

A. Introduction

OrthoMCL (OMCL) is 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 (**orthogroups**). Proteins in OMCL orthogroups have been shown to display a high degree of functional conservation.



~8.6 million proteins

716 VEuPathDB proteomes

Eukaryotic parasites, pathogens, hosts, and vectors in the VEuPathDB database

149 Core proteomes

Taxonomically diverse Core proteomes chosen as well-annotated reference organisms to broadly represent the major branches of the tree of life

OMCL 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.

OMCL 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 OMCL groups. Existing proteins in public databases that lack functional annotation (e.g. "hypothetical protein", "unspecified product", etc.) may also benefit from assignment to OMCL orthogroups. See the *Map proteins to OrthoMCL* function in the Tools menu.

B. Background on orthology

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)
- An orthogroup 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).

OMCL orthogroups contain both orthologs across multiple species and recently duplicated paralogs, sometimes called in-paralogs.

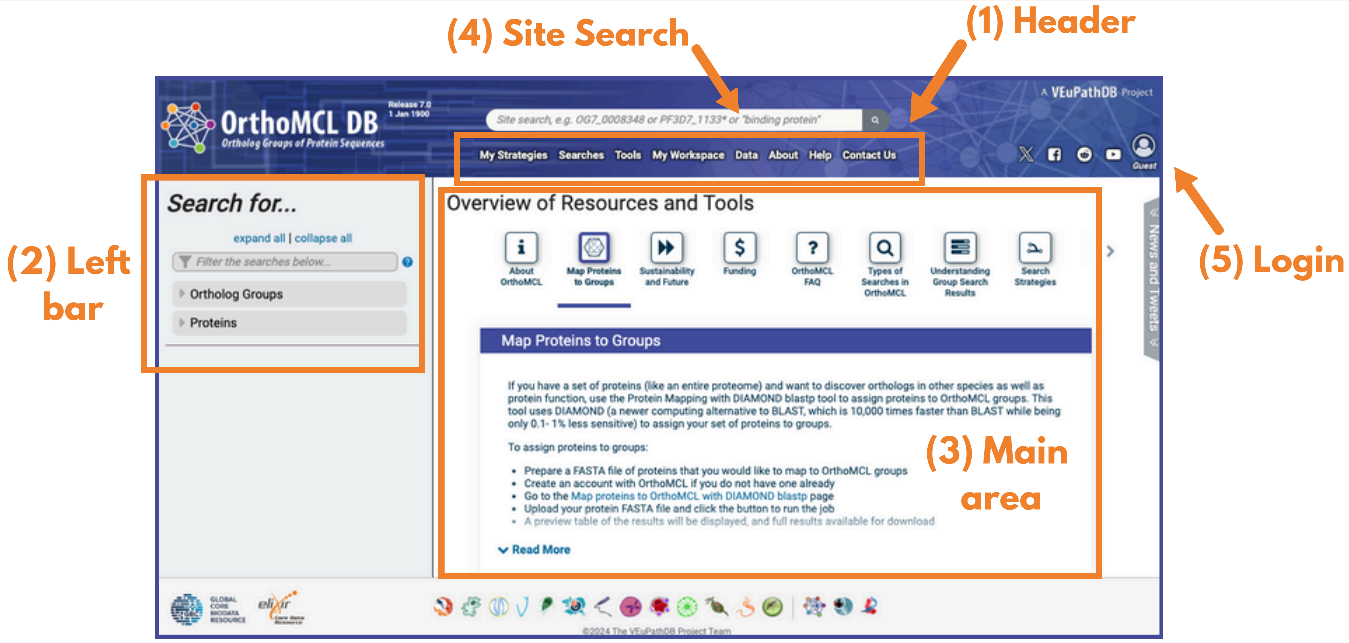
C. The OrthoMCL algorithm

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).

D. Explore the OrthoMCL Home page



If you have a set of proteins (like an entire proteome) and want to discover orthologs in other species as well as protein function, use the Protein Mapping with DIAMOND blastp tool to assign proteins to OrthoMCL groups. This tool uses DIAMOND (a newer computing alternative to BLAST, which is 10,000 times faster than BLAST while being only 0.1- 1% less sensitive) to assign your set of proteins to groups.

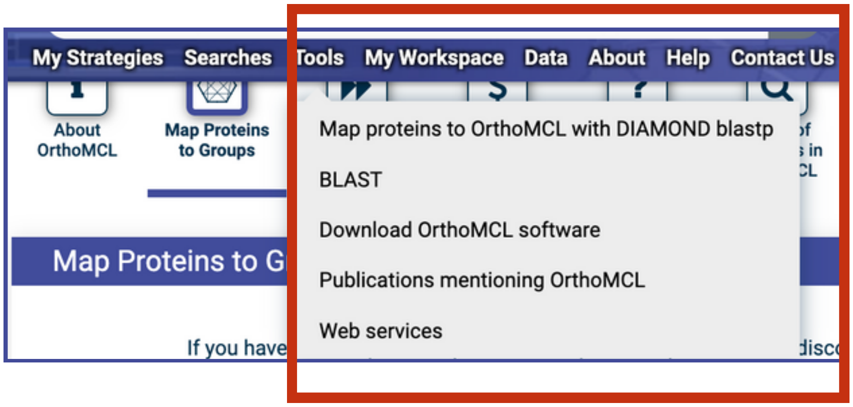
To assign proteins to groups:

- Prepare a FASTA file of proteins that you would like to map to OrthoMCL groups
- Create an account with OrthoMCL if you do not have one already
- Go to the Map proteins to OrthoMCL with DIAMOND blastp page
- Upload your protein FASTA file and click the button to run the job
- A preview table of the results will be displayed, and full results available for download

1. The **Header** contains menus for
   a. **Searches** for ortholog groups and proteins, described later in this document.
   b. **Tools**: These include the *Map proteins to OrthoMCL Tool* and a BLAST tool, described on the next page.
   c. **Data**: Links to analysis methods and data file downloads are provided.
   d. **Help**: Links to frequently asked questions (FAQs), workshops, webinars, tutorials, etc. and user documentation.
   e. **Contact Us**: Link to a form to message the help desk.
   f. **My Strategies** and **My Workspace** provide links to a user's previous activity on the site.
2. The **Left Bar** 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 finds any text in protein descriptions, protein and group IDs, Pfam and EC Number IDs and descriptions.
5. **Login**: To make use of *My Strategies* and *My Workspace*, use the person icon to access the free Registration/Login form at the far right of the header.

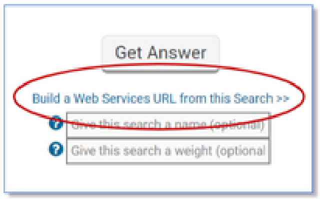
E. Tools

The **Tools** drop-down menu can be accessed from the header.



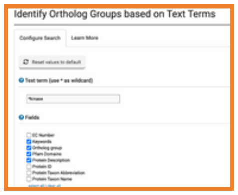
The following tools are available:

1. 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. 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.
    b. The results are provided as a tab delimited text file with the input protein, closest matching OrthoMCL protein, the protein description, the e-value of the match, and the corresponding OrthoMCL orthogroup.
    c. 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.
2. **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.
3. **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 fille 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.


Get Answer
Build a Web Services URL from this Search >>
Give this search a name (optional)
Give this search a weight (optional

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"}


Identify Ortholog Groups based on Text Terms

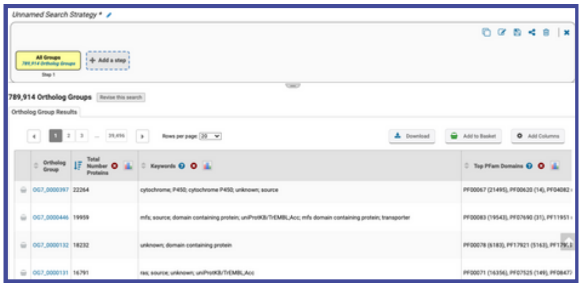
Text term (use * as wildcard): *kinase

Fields:
- ☐ EC Number
- ☑ Keywords
- ☑ Ortholog group
- ☑ Pfam Domains
- ☑ Protein Description
- ☐ Protein ID
- ☐ Protein Taxon Abbreviation
- ☐ Protein Taxon Name

F. Ortholog Group search

The ortholog group search can be accessed from the header (Searches menu) or from the left bar. There are **nine different options** for searching for ortholog groups of interest. In each case, results for the search are made available in a search strategy with the result presented as a list of ortholog groups in tabular format.

All groups: This search simply returns all available ortholog groups without filtering in a search strategy as shown below



*Unnamed Search Strategy **

All Groups
789,914 Ortholog Groups
Step 1

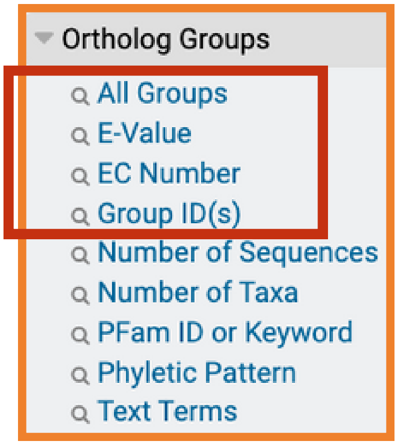
+ Add a step

789,914 Ortholog Groups [Revise this search]

Ortholog Group Results

‹ 1 2 3 … 39,496 › Rows per page 20 | Download | Add to Basket | Add Columns

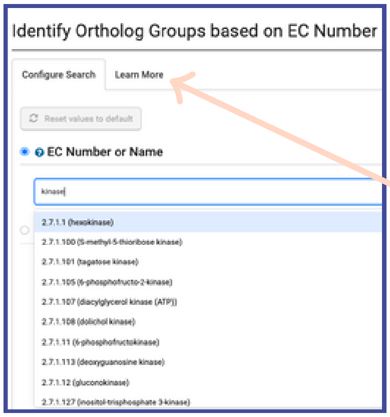
	Ortholog Group	Total Number Proteins	Keywords	Top PFam Domains
	OG7_0000397	22264	cytochrome; P450; cytochrome P450; unknown; source	PF00067 (21495), PF00620 (14), PF04082
	OG7_0000446	19959	mfs; source; domain containing protein; uniProtKB/TrEMBL;Acc; mfs domain containing protein; transporter	PF00083 (19543), PF07690 (31), PF11951
	OG7_0000132	18232	unknown; domain containing protein	PF00078 (6183), PF17921 (5163), PF179
	OG7_0000131	16791	ras; source; unknown; uniProtKB/TrEMBL;Acc	PF00071 (16356), PF07525 (149), PF0847



E-value: This search finds ortholog groups that have internal differences (dispersion) between proteins that fall within a range of BLAST e-values. E-values are negative exponents that run from -4 to -200 with the smaller value being more significant, indicating higher group cohesiveness, i.e., a tighter group. A group with more dispersion (less significant e-value) might contain two distinct sub-groups, or might contain outliers.

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

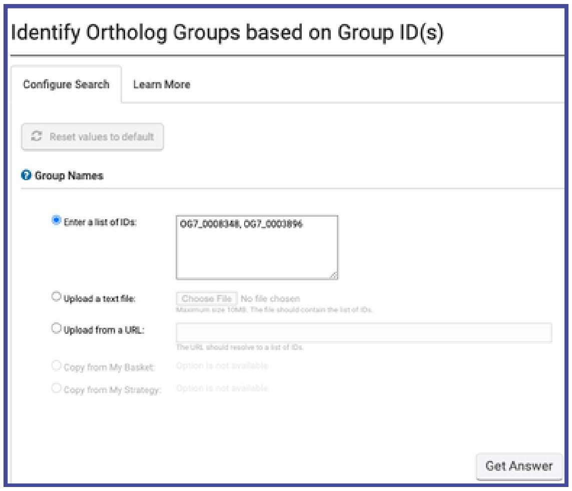
kinase

2.7.1.1 (hexokinase)
2.7.1.100 (S-methyl-5-thioribose kinase)
2.7.1.101 (tagatose kinase)
2.7.1.105 (6-phosphofructo-2-kinase)
2.7.1.107 (diacylglycerol kinase (ATP))
2.7.1.108 (dolichol kinase)
2.7.1.11 (6-phosphofructokinase)
2.7.1.113 (deoxyguanosine kinase)
2.7.1.12 (gluconokinase)
2.7.1.127 (inositol-trisphosphate 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 from a URL



Identify Ortholog Groups based on Group ID(s)

Configure Search | Learn More

Reset values to default

Group Names

● Enter a list of IDs: OG7_0008348, OG7_0003896

○ Upload a text file: Choose File No file chosen
Maximum size 10MB. The file should contain the list of IDs.

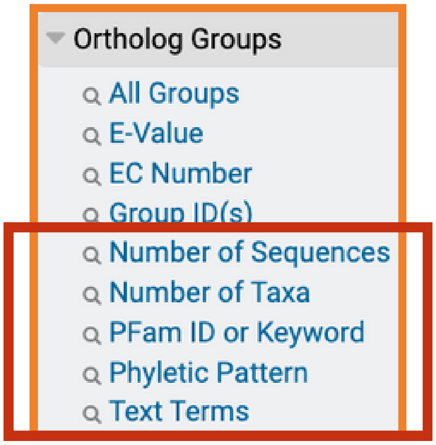
○ Upload from a URL:
The URL should resolve to a list of IDs.

○ Copy from My Basket: Option is not available

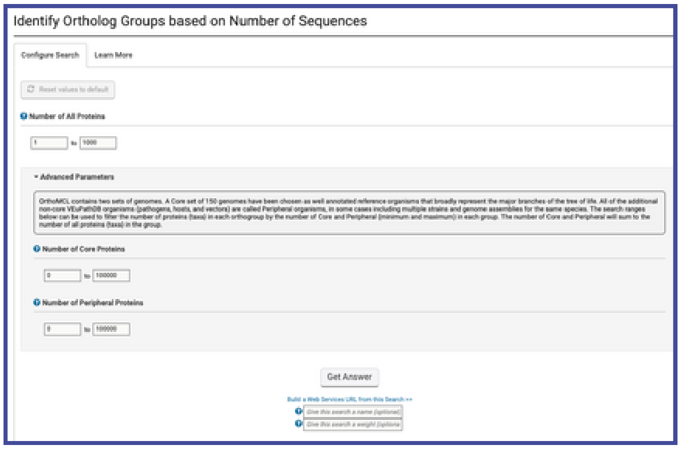
○ Copy from My Strategy: Option is not available

Get Answer

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. If the number of Core proteins is set to 0, then the search will find only Residual groups. If at least 1 Core protein is required, then the search will find only Core groups. If groups are set to a minimum of 2 proteins, then singleton groups will be removed from the search results.



Number of taxa: Configure the search to find Ortholog Groups that contain a specified number of all taxa, or used Advanced parameters to specify the number of Core and Peripheral taxa.


Identify Ortholog Groups based on Number of Sequences

Configure Search | Learn More

Reset values to default

Number of All Proteins

1 to 1000

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

0 to 100000

Number of Peripheral Proteins

0 to 100000

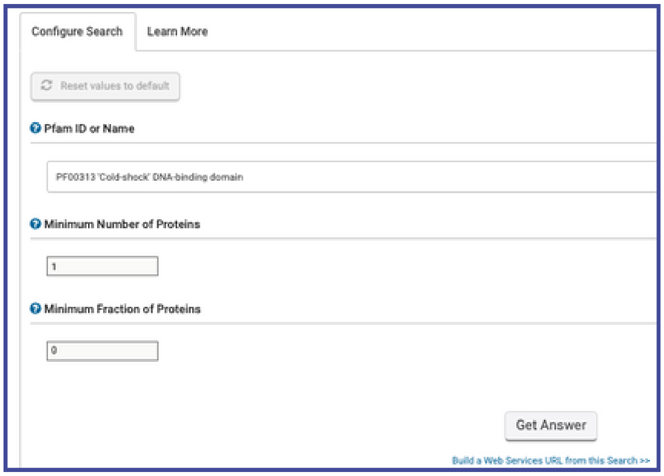
Get Answer

Build a Web Services URL from this Search >>

Give this search a name (optional)

Give this search a weight (optional)

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.


Configure Search | Learn More

Reset values to default

Pfam ID or Name

PF00313 'Cold-shock' DNA-binding domain

Minimum Number of Proteins

1

Minimum Fraction of Proteins

0

Get Answer

Build a Web Services URL from this Search >>

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. 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 section.

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.

Identify Ortholog Groups based on Phyletic Pattern

Configure Search | Learn More

Get Answer

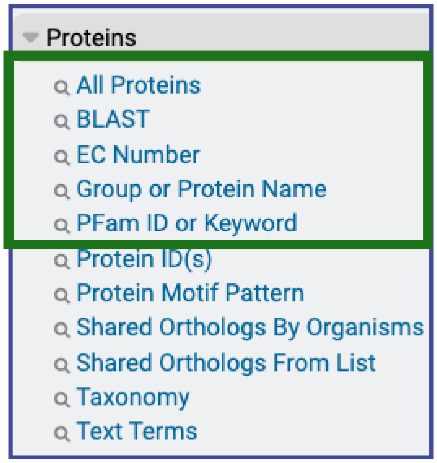
G. Protein search

The protein search can be accessed from the header (Searches menu) or from the left bar. There are **11 different options** for searching for proteins of interest. In each case, results for the search are made available in a search strategy with the result presented as a list of proteins in tabular format.

All proteins: This search simply returns all available proteins without filtering.

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.

- 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*"



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.

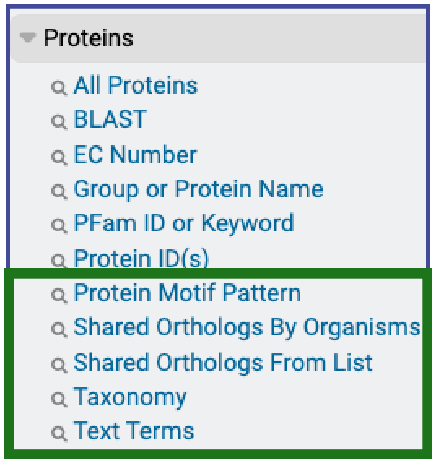
Group or Protein name: Search by protein name or description terms.

Protein ID(s): Find proteins 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 from a URL

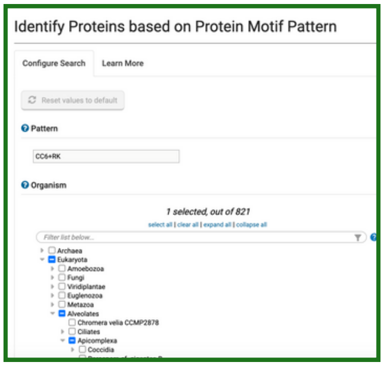
Pfam ID or Keyword: Search with Pfam IDs and description terms (keywords) to find proteins with these domains.

Protein motif pattern: Find Proteins that contain an amino acid either as a simple string of single letter amino acid abbreviations or with 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 ve searched within particular species by picking from the organism tree.



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

CC6+RK

Organism

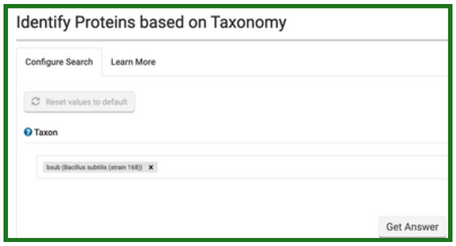
1 selected, out of 821

select all | clear all | expand all | collapse all

Filter list below...

- □ Archaea
- ☑ Eukaryota
  - □ Amoebozoa
  - □ Fungi
  - □ Viridiplantae
  - □ Euglenozoa
  - □ Metazoa
  - ☑ Alveolates
    - □ Chromera velia CCMP2878
    - □ Ciliates
    - ☑ Apicomplexa
      - □ Coccidia

Taxonomy: This search allows you to choose particular taxa and return a list of proteins belonging to those taxa.



Identify Proteins based on Taxonomy

Configure Search | Learn More

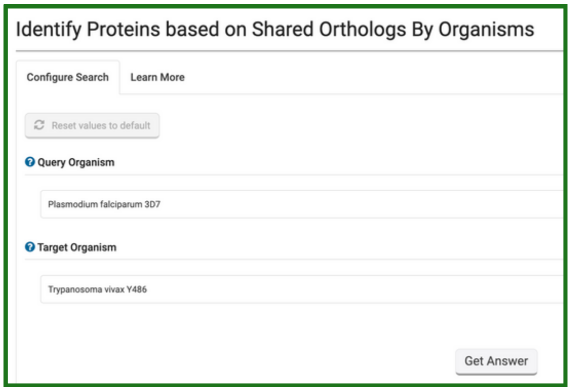
Reset values to default

Taxon

bsub (Bacillus subtilis (strain 168)) ×

Get Answer

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



Identify Proteins based on Shared Orthologs By Organisms

Configure Search | Learn More

Reset values to default

Query Organism

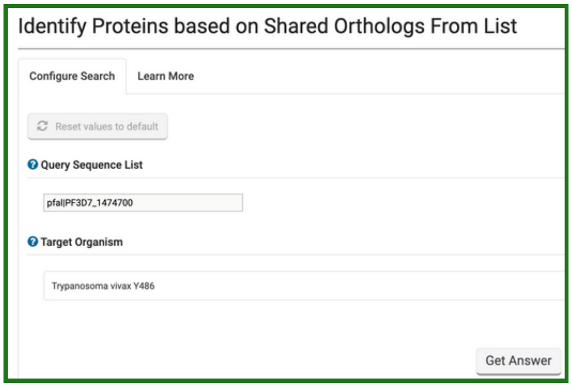
Plasmodium falciparum 3D7

Target Organism

Trypanosoma vivax Y486

Get Answer

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.



Identify Proteins based on Shared Orthologs From List

Configure Search | Learn More

Reset values to default

Query Sequence List

pfal|PF3D7_1474700

Target Organism

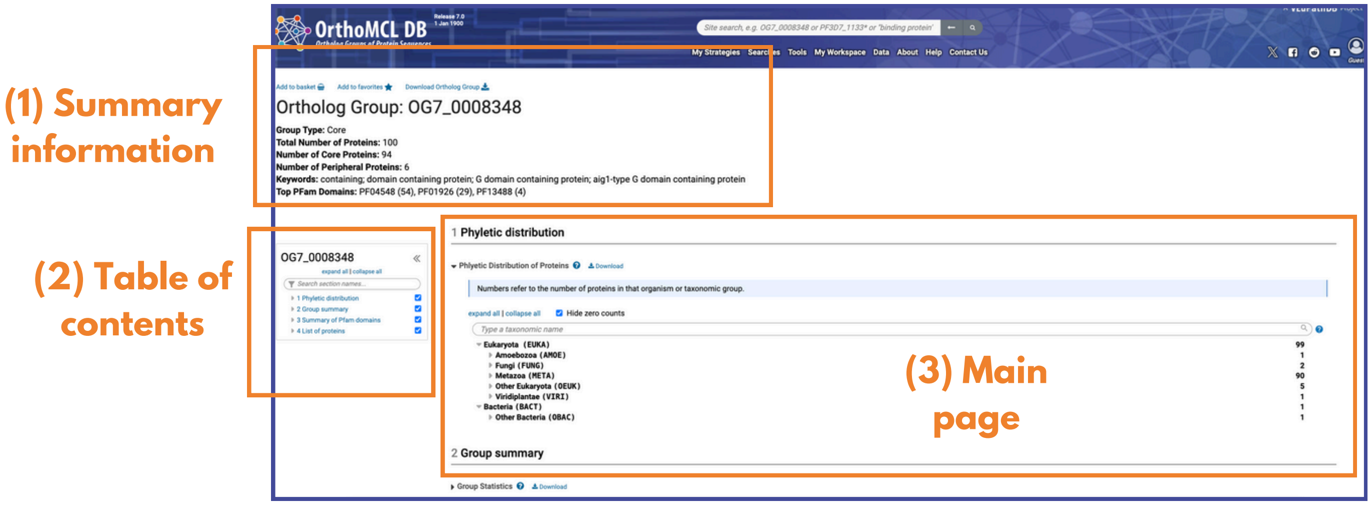
Trypanosoma vivax Y486

Get Answer

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.

H. Orthogroup Page

Each orthogroup has its own web page in at OrthoMCL.org.



Ortholog Group: OG7_0008348

Group Type: Core
Total Number of Proteins: 100
Number of Core Proteins: 94
Number of Peripheral Proteins: 6
Keywords: containing; domain containing protein; G domain containing protein; aig1-type G domain containing protein
Top PFam Domains: PF04548 (54), PF01926 (29), PF13488 (4)

OG7_0008348 «

- 1 Phyletic distribution ☑
- 2 Group summary ☑
- 3 Summary of Pfam domains ☑
- 4 List of proteins ☑

1 Phyletic distribution

▾ Phlyetic Distribution of Proteins

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

expand all | collapse all ☑ Hide zero counts

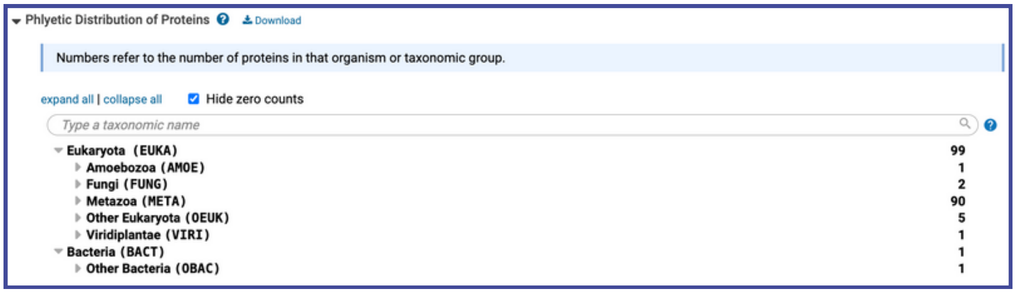
Eukaryota (EUKA)	99
Amoebozoa (AMOE)	1
Fungi (FUNG)	2
Metazoa (META)	90
Other Eukaryota (OEUK)	5
Viridiplantae (VIRI)	1
Bacteria (BACT)	1
Other Bacteria (OBAC)	1

2 Group summary

▸ Group Statistics

1. **Summary information**: The top of the page provides some summary information about the group including if the group contains proteins from Core organisms (or just Residual proteins from Peripheral organisms), the total number of proteins, number of Core proteins, number of Peripheral proteins, keywords from the protein descriptions, EC numbers and 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.

    1. The **Phyletic distribution of the proteins** in the orthogroup is shown across the tree of life with counts 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 OMCL Core organisms.



▾ Phlyetic Distribution of Proteins

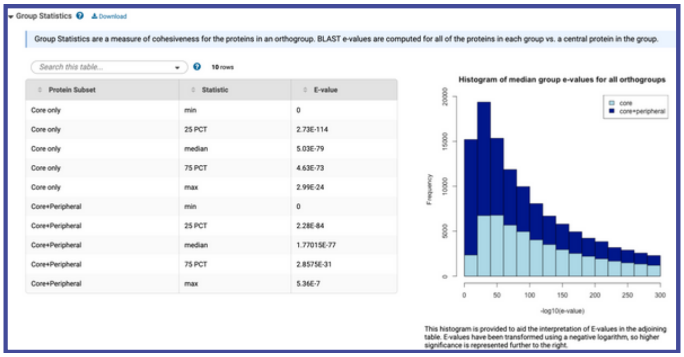
Numbers refer to the number of proteins in that organism or taxonomic group.

expand all | collapse all ☑ Hide zero counts

Eukaryota (EUKA)	99
Amoebozoa (AMOE)	1
Fungi (FUNG)	2
Metazoa (META)	90
Other Eukaryota (OEUK)	5
Viridiplantae (VIRI)	1
Bacteria (BACT)	1
Other Bacteria (OBAC)	1

2. The **Group Summary** contains three sub-sections

- **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.



Group Statistics ❓ Download

Group Statistics are a measure of cohesiveness for the proteins in an orthogroup. BLAST e-values are computed for all of the proteins in each group vs. a central protein in the group.

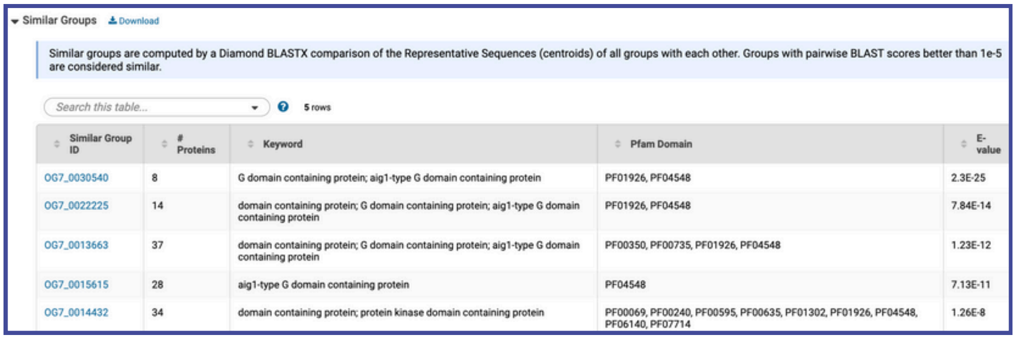
Search this table... ❓ 10 rows

Protein Subset	Statistic	E-value
Core only	min	0
Core only	25 PCT	2.73E-114
Core only	median	5.03E-79
Core only	75 PCT	4.63E-73
Core only	max	2.99E-24
Core+Peripheral	min	0
Core+Peripheral	25 PCT	2.28E-84
Core+Peripheral	median	1.77015E-77
Core+Peripheral	75 PCT	2.8575E-31
Core+Peripheral	max	5.36E-7

Histogram of median group e-values for all orthogroups

This histogram is provided to aid the interpretation of E-values in the adjoining table. E-values have been transformed using a negative logarithm, so higher significance is represented further to the right.

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



Similar Groups Download

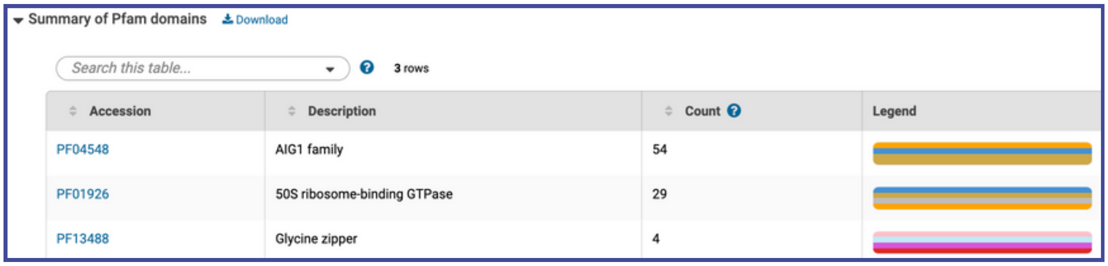
Similar groups are computed by a Diamond BLASTX comparison of the Representative Sequences (centroids) of all groups with each other. Groups with pairwise BLAST scores better than 1e-5 are considered similar.

Search this table... ❓ 5 rows

Similar Group ID	# Proteins	Keyword	Pfam Domain	E-value
OG7_0030540	8	G domain containing protein; aig1-type G domain containing protein	PF01926, PF04548	2.3E-25
OG7_0022225	14	domain containing protein; G domain containing protein; aig1-type G domain containing protein	PF01926, PF04548	7.84E-14
OG7_0013663	37	domain containing protein; G domain containing protein; aig1-type G domain containing protein	PF00350, PF00735, PF01926, PF04548	1.23E-12
OG7_0015615	28	aig1-type G domain containing protein	PF04548	7.13E-11
OG7_0014432	34	domain containing protein; protein kinase domain containing protein	PF00069, PF00240, PF00595, PF00635, PF01302, PF01926, PF04548, PF06140, PF07714	1.26E-8

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

3. **Summary of Pfam domains** is a table that shows the Pfam domains for all proteins in the group, a description of each, the number of proteins that contain each domain, and the cartoon that is used as a label for each domain in the Protein Table below.



Summary of Pfam domains Download

Search this table... ❓ 3 rows

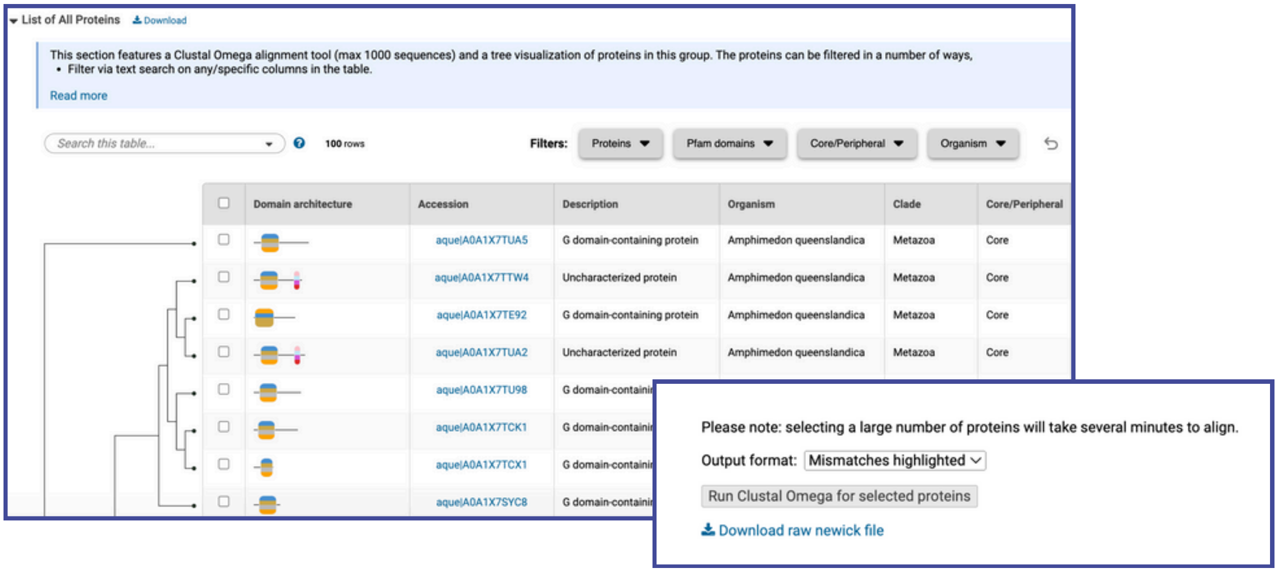
Accession	Description	Count ❓	Legend
PF04548	AIG1 family	54	▬
PF01926	50S ribosome-binding GTPase	29	▬
PF13488	Glycine zipper	4	▬

I. Orthogroup protein table and tree

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

- The table contains a cartoon of Pfam domains in each protein (Pfam IDs are shown if the mouse pointer is moved over the cartoon), the protein ID, protein description, organism, clade, protein length, and EC number (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- 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 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 Core/Peripheral filter can show proteins from only Core (or only Peripheral) species.
    - The Pfam filter can limit the tree (and table) to only show proteins that contain selected Pfam domains.
    - 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.
    - The text filter can be used to limit the table to proteins that contain a specific word in the description.

If proteins are selected in the table by clicking the checkbox located between the tree and the Pfam cartoon, then the Clustal Omega 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.



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.

□	Domain architecture	Accession	Description	Organism	Clade	Core/Peripheral
□		aque|A0A1X7TUA5	G domain-containing protein	Amphimedon queenslandica	Metazoa	Core
□		aque|A0A1X7TTW4	Uncharacterized protein	Amphimedon queenslandica	Metazoa	Core
□		aque|A0A1X7TE92	G domain-containing protein	Amphimedon queenslandica	Metazoa	Core
□		aque|A0A1X7TUA2	Uncharacterized protein	Amphimedon queenslandica	Metazoa	Core
□		aque|A0A1X7TU98	G domain-containin			
□		aque|A0A1X7TCK1	G domain-containin			
□		aque|A0A1X7TCX1	G domain-containin			
□		aque|A0A1X7SYC8	G domain-containin			

Please note: selecting a large number of proteins will take several minutes to align.

Output format: Mismatches highlighted