OrthoMCL DB
Ortholog Groups of Protein Sequences

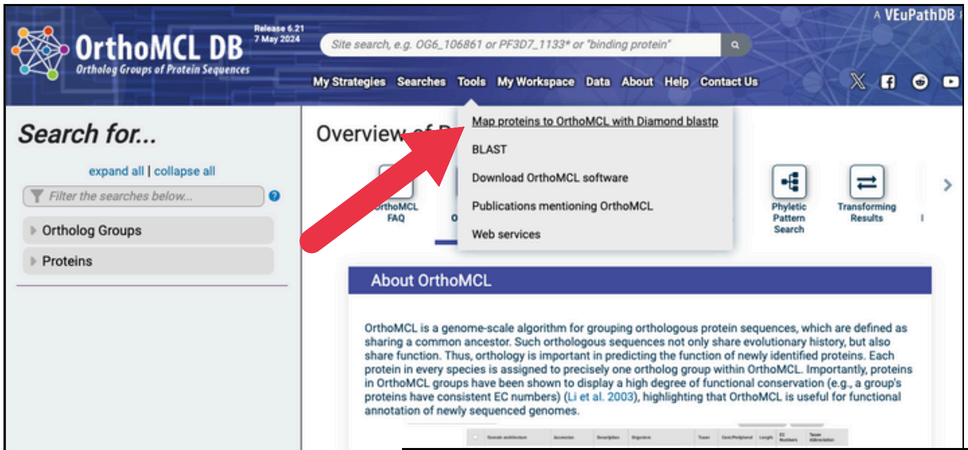
VEuPathDB
Eukaryotic Pathogen, Vector & Host Informatics Resources

Map Proteins to OrthoMCL with Diamond BLAST- A Tutorial

Learning Objectives

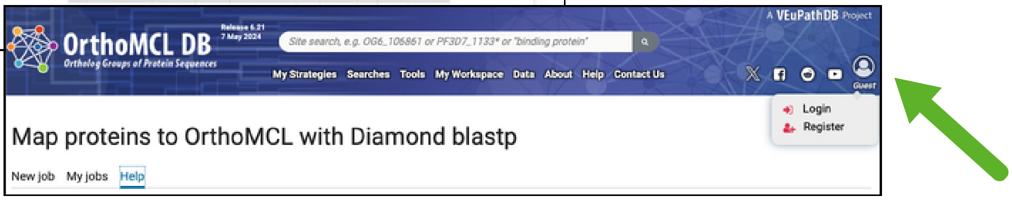
- Understand the purpose of the protein mapping tool in OrthoMCL
- Learn how to prepare and upload data
- Explore the output and understand the Diamond job result page

1. Introduction
2. OrthoMCL is a genome-scale algorithm for grouping orthologous protein sequences. The OrthoMCL algorithm is employed on proteins from a set of 150 Core species to form Core and Residual ortholog groups.
3. Purpose of the protein mapping tool: The purpose of this tool is to allow users to map a set of proteins of interest, usually a complete proteome from an organism, to existing 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.
4. Access the tool from the **Tools menu** in the header > Map proteins to OrthoMCL with Diamond BLASTP (red arrow below)


OrthoMCL DB
Ortholog Groups of Protein Sequences
Release 6.21
7 May 2024
Site search, e.g. OG6_106861 or PF3D7_1133* or "binding protein"
VEuPathDB
My Strategies
Searches
Tools
My Workspace
Data
About
Help
Contact Us
Search for...
expand all | collapse all
Filter the searches below...
Ortholog Groups
Proteins
Overview
Map proteins to OrthoMCL with Diamond blastp
BLAST
Download OrthoMCL software
Publications mentioning OrthoMCL
Web services
OrthoMCL FAQ
Phyletic Pattern Search
Transforming Results
About OrthoMCL

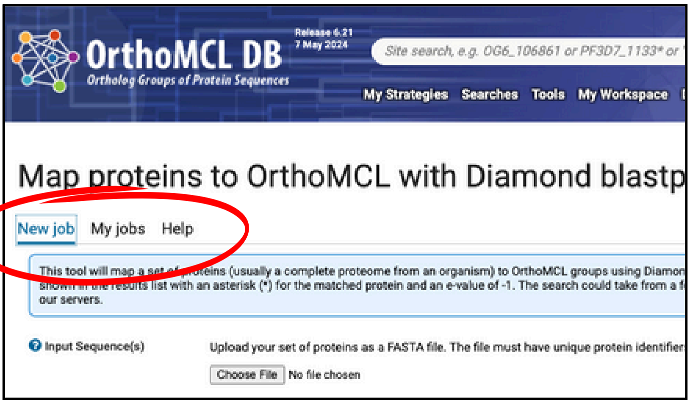
OrthoMCL is a genome-scale algorithm for grouping orthologous protein sequences, which are defined as sharing a common ancestor. Such orthologous sequences not only share evolutionary history, but also share function. Thus, orthology is important in predicting the function of newly identified proteins. Each protein in every species is assigned to precisely one ortholog group within OrthoMCL. Importantly, proteins in OrthoMCL groups have been shown to display a high degree of functional conservation (e.g., a group's proteins have consistent EC numbers) (Li et al. 2003), highlighting that OrthoMCL is useful for functional annotation of newly sequenced genomes.

Note that you must be logged into your free OrthoMCL account to use this tool. Click on the person icon at top right (green arrow below) to log in or register.



Layout of the Diamond BLASTP protein mapping tool: There are three tabs (circled in red on right)

1. **New job**: Upload data here
2. **My jobs**: Table of all your previous jobs; these are saved in your account and persist between sessions
3. **Help**: Tips for using the tool


OrthoMCL DB
Ortholog Groups of Protein Sequences
Release 6.21
7 May 2024
Site search, e.g. OG6_106861 or PF3D7_1133* or
My Strategies
Searches
Tools
My Workspace

Map proteins to OrthoMCL with Diamond blastp

New job My jobs Help

This tool will map a set of proteins (usually a complete proteome from an organism) to OrthoMCL groups using Diamon
shown in the results list with an asterisk (*) for the matched protein and an e-value of -1. The search could take from a f
our servers.

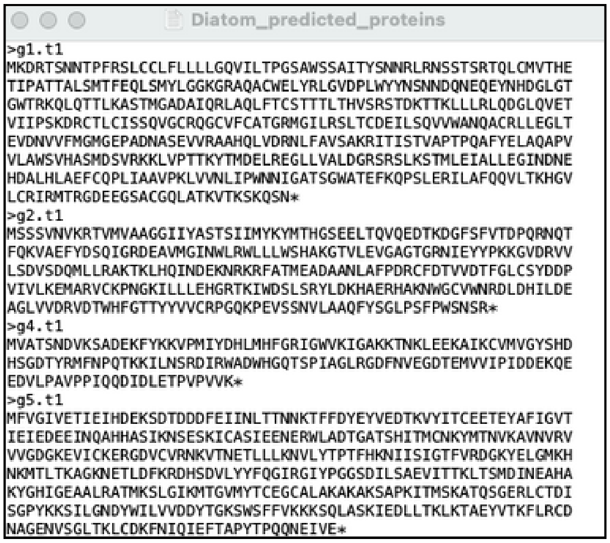
Input Sequence(s) Upload your set of proteins as a FASTA file. The file must have unique protein identifier

Choose File No file chosen

Preparing your data: Your set of proteins must be formatted as a plain text FASTA file.

- Each protein in the FASTA file must have unique protein identifiers as the first word in its definition/header line.
- Header line must start with a greater than (>) symbol and end with a carriage return.

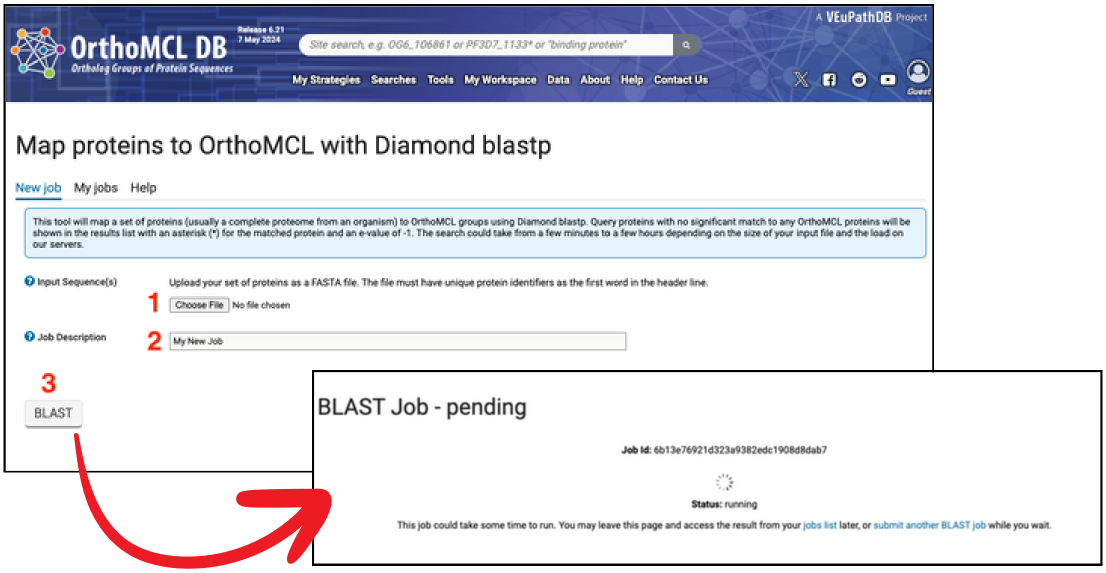
For example, the FASTA file on the right shows a set of predicted proteins from a newly sequenced Diatom species.



```
>g1.t1
MKDRTSNNTPFRSLCCLFLLLLGQVILTPGSAWSSAITYSNNRLRNSSTSRTQLCMVTHE
TIPATTALSMTFEQLSMYLGGKGRAQACWELYRLGVDPLWYYNSNNDQNEQEYNHDGLGT
GWTRKQLQTTLKASTMGADAIQRLAQLFTCSTTTLTHVSRSTDKTTKLLLRLQDGLQVET
VIIPSKDRCTLCISSQVGCRQGCVFCATGRMGILRSLTCDEILSQVVWANQACRLLEGLT
EVDNVVFMGMGEPADNASEVVRAAHQLVDRNLFAVSAKRITISTVAPTPQAFYELAQAPV
VLAWSVHASMDSVRKKLVPTTKYTMDELREGLLVALDGRSRSLKSTMLEIALLEGINDNE
HDALHLAEFCQPLIAAVPKLVVNLIPWNNIGATSGWATEFKQPSLERILAFQQVLTKHGV
LCRIRMTRGDEEGSACGQLATKVTKSKQSN*
>g2.t1
MSSSVNVKRTVMVAAGGIIYASTSIIMYKYMTHGSEELTQVQEDTKDGFSFVTDPQRNQT
FQKVAEFYDSQIGRDEAVMGINWLRWLLLWSHAKGTVLEVGAGTGRNIEYYPKKGVDRVV
LSDVSDQMLLRAKTKLHQINDEKNRKRFATMEADAANLAFPDRCFDTVVDTFGLCSYDDP
VIVLKEMARVCKPNGKILLLEHGRTKIWDSLSRYLDKHAERHAKNWGCVWNRDLDHILDE
AGLVVDRVDTWHFGTTYYVVCRPGQKPEVSSNVLAAQFYSGLPSFPWSNSR*
>g4.t1
MVATSNDVKSADEKFYKKVPMIYDHLMHFGRIGWVKIGAKKTNKLEEKAIKCVMVGYSHD
HSGDTYRMFNPQTKKILNSRDIRWADWHGQTSPIAGLRGDFNVEGDTEMVVIPIDDEKQE
EDVLPAVPPIQQDIDLETPVPVVK*
>g5.t1
MFVGIVETIEIHDEKSDTDDDFEIINLTTNNKTFFDYEYVEDTKVYITCEETEYAFIGVT
IEIEDEEINQAHHASIKNSESKICASIEENERWLADTGATSHITMCNKYMTNVKAVNVRV
VVGDGKEVICKERGDVCVRNKVTNETLLLKNVLYTPTFHKNIISIGTFVRDGKYELGMKH
NKMTLTKAGKNETLDFKRDHSDVLYYFQGIRGIYPGGSDILSAEVITTKLTSMDINEAHA
KYGHIGEAALRATMKSLGIKMTGVMYTCEGCALAKAKAKSAPKITMSKATQSGERLCTDI
SGPYKKSILGNDYWILVVDDYTGKSWSFFVKKKSQLASKIEDLLTKLKTAEYVTKFLRCD
NAGENVSGLTKLCDKFNIQIEFTAPYTPQQNEIVE*
```

Uploading data: Do the following steps in the "New job" tab (refer to figure below)

1. **Input sequence**: Choose a FASTA-formatted data file with protein sequences from your computer
2. **Job description:** Add brief text describing your set of proteins
3. **BLAST**: Click on the button to start the job. You will see a message with a job ID assignment.



Map proteins to OrthoMCL with Diamond blastp

New job My jobs Help

This tool will map a set of proteins (usually a complete proteome from an organism) to OrthoMCL groups using Diamond blastp. Query proteins with no significant match to any OrthoMCL proteins will be shown in the results list with an asterisk (*) for the matched protein and an e-value of -1. The search could take from a few minutes to a few hours depending on the size of your input file and the load on our servers.

Input Sequence(s) Upload your set of proteins as a FASTA file. The file must have unique protein identifiers as the first word in the header line.

Choose File No file chosen

Job Description: My New Job

BLAST

BLAST Job - pending

Job Id: 6b13e76921d323a9382edc1908d8dab7

Status: running

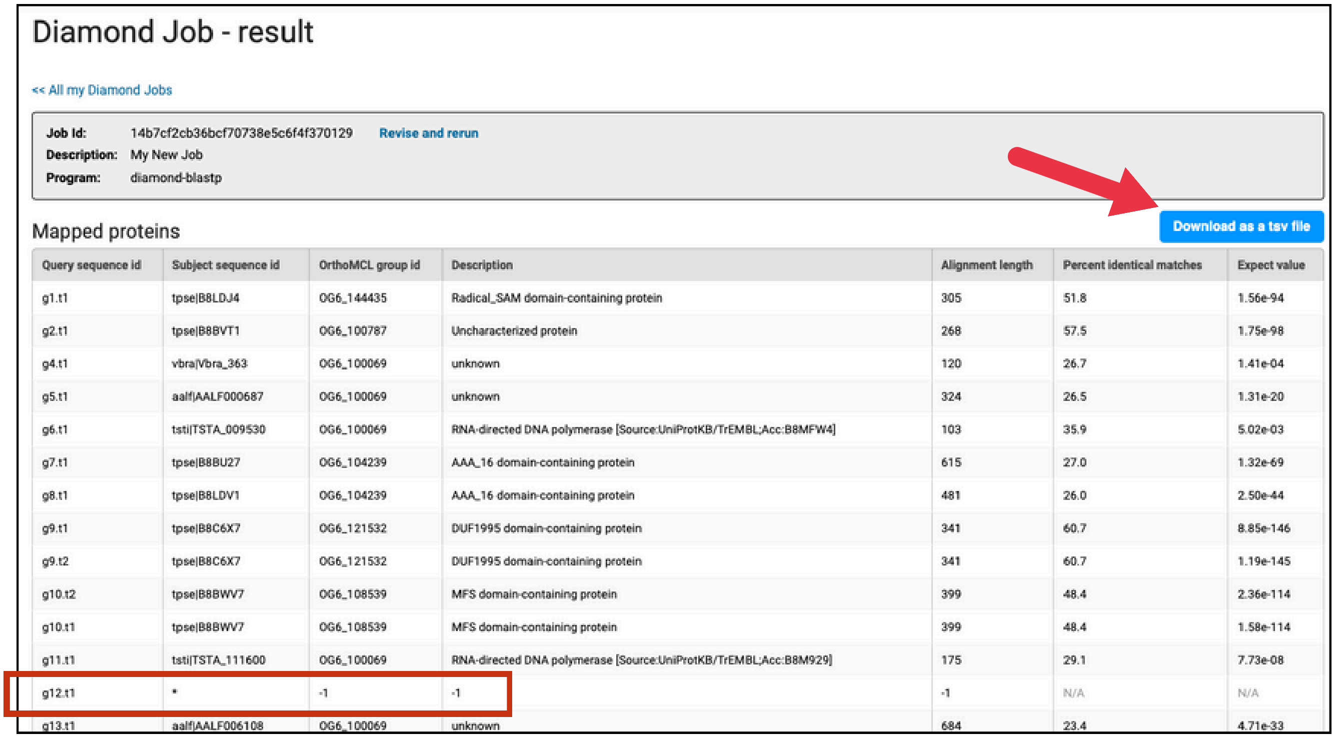
This job could take some time to run. You may leave this page and access the result from your jobs list later, or submit another BLAST job while you wait.

Understanding the output:

The output page has two components

1. **The results table** (see below). This is a preview of the matching results for the first 100 sequences in your query file.
2. **A blue download button** at the top right (see red arrow below). The complete result can be downloaded as a tab delimited file with one best match for each query protein with the following columns:
    - Query_ID
    - Subject_ID
    - Orthogroup
    - Subject_description
    - Alignment_length
    - Percent_identity
    - e-value

Note: Unmatched query proteins are included in the results file without an OrthoMCL protein or group listed. For example, see red rectangle below.



Diamond Job - result

<< All my Diamond Jobs

Job Id: 14b7cf2cb36bcf70738e5c6f4f370129 Revise and rerun
Description: My New Job
Program: diamond-blastp

Mapped proteins

Query sequence id	Subject sequence id	OrthoMCL group id	Description	Alignment length	Percent identical matches	Expect value
g1.t1	tpse|B8LDJ4	OG6_144435	Radical_SAM domain-containing protein	305	51.8	1.56e-94
g2.t1	tpse|B8BVT1	OG6_100787	Uncharacterized protein	268	57.5	1.75e-98
g4.t1	vbra|Vbra_363	OG6_100069	unknown	120	26.7	1.41e-04
g5.t1	aalf|AALF000687	OG6_100069	unknown	324	26.5	1.31e-20
g6.t1	tsti|TSTA_009530	OG6_100069	RNA-directed DNA polymerase [Source:UniProtKB/TrEMBL;Acc:B8MFW4]	103	35.9	5.02e-03
g7.t1	tpse|B8BU27	OG6_104239	AAA_16 domain-containing protein	615	27.0	1.32e-69
g8.t1	tpse|B8LDV1	OG6_104239	AAA_16 domain-containing protein	481	26.0	2.50e-44
g9.t1	tpse|B8C6X7	OG6_121532	DUF1995 domain-containing protein	341	60.7	8.85e-146
g9.t2	tpse|B8C6X7	OG6_121532	DUF1995 domain-containing protein	341	60.7	1.19e-145
g10.t2	tpse|B8BWV7	OG6_108539	MFS domain-containing protein	399	48.4	2.36e-114
g10.t1	tpse|B8BWV7	OG6_108539	MFS domain-containing protein	399	48.4	1.58e-114
g11.t1	tsti|TSTA_111600	OG6_100069	RNA-directed DNA polymerase [Source:UniProtKB/TrEMBL;Acc:B8M929]	175	29.1	7.73e-08
g12.t1	*	-1	-1	-1	N/A	N/A
g13.t1	aalf|AALF006108	OG6_100069	unknown	684	23.4	4.71e-33