

---

## OrthoMCL 7

# Map Proteins to OrthoMCL with Diamond blastp: A Tutorial<sup>1</sup>

---

### Learning Objectives

- Understand the purpose of the OrthoMCL protein mapping tool
- Learn how to prepare and upload sets of proteins for mapping
- Explore the output and understand the DIAMOND job result page

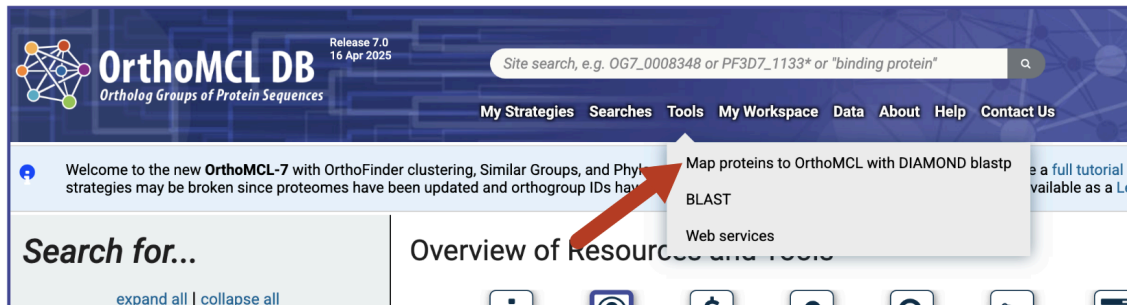
### Introduction

- [OrthoMCL](#) is a genome-scale algorithm that uses protein sequence similarity and phylogenetic relationships to create groups of orthologous protein sequences both within and across species. OrthoMCL includes all [VEuPathDB](#) species plus additional Core species that broadly represent the diversity across the Tree of Life.
- **Purpose**
  - The protein mapping tool allows users to **map a set of proteins of interest**, usually a complete proteome from an organism, to existing OrthoMCL groups.
  - This tool can also be used to **annotate a set of translated proteins** from a transcriptome or metagenome.

---

<sup>1</sup> Updated on April 18, 2025

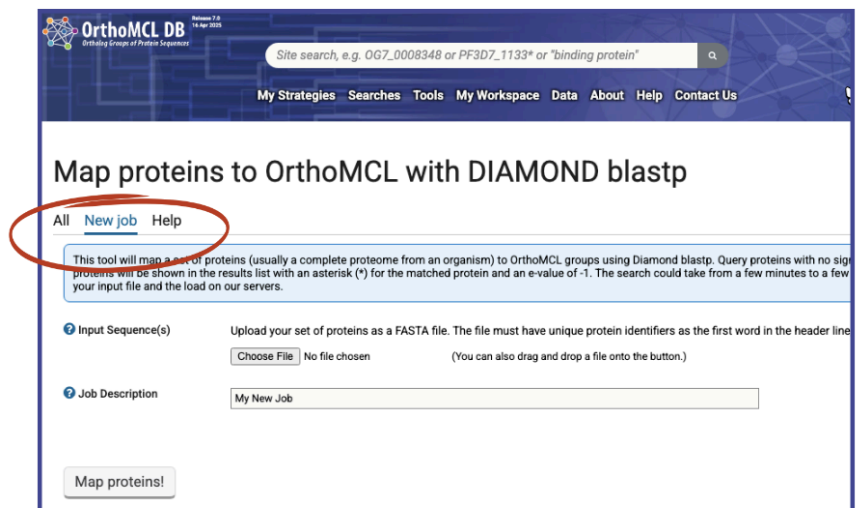
- This tool uses **DIAMOND blastp**, an alternative to NCBI BLAST which is ~1,000 times faster while being only 0.1- 1% less sensitive.
- Access the tool from the **Tools** menu in the header > **Map proteins to OrthoMCL with DIAMOND blastp** (red arrow below)



## Layout of the DIAMOND blastp protein mapping page

There are **three tabs** (circled in red)

- **All:** A list of all of your previous jobs if you were logged in. These are saved in your account and persist across visits to the website.
- **New job:** Interface that allows you to upload a FASTA formatted file of protein sequences
- **Help:** Tips for using the tool



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

- The single-line description/header for each protein sequence in the FASTA file must begin with a unique protein identifier.
- Header line must start with a greater than (>) symbol and end with a carriage return.

```

Diatom_predicted_proteins -- Edited

>g1.t1
MKDRTSNNTPFRLCCLFLLLLGQVILTPGSAWSSAITYSNNRLNRSSTSRTQLCMVTHE
TIPATTALSMTFEQLSMYLGKGRAQACWELRYLGDPLWYNNNDQNEQYVNDHGLGT
GWTRKQLQTTLKASTMGADAIQRLAQLFTCSSTTLTHVSRSDTKTKLLRLQDGLQVET
VIIPSKDRCTLCISSQVGCRCQGVCFATGRMGILRSLTCDIELSQVVMANQACRLLEGLT
EVDNVVFMGMPADNASEVVRAAHQLVDRNLFVSAKRITISTVAPTQAFYELAQAPV
VLAWSVHASMDSVRKKLVPPTTKYTMDELREGLLVALDGRSRLKSTMLEIALLEGINDNE
HDALHLAEFCQPLIAAVPKLVNLIWNNIGATSGWATEFKQPSLERILAFQVLTKHGV
LCIRIMTRGDEEGSACGLATKVTQSKQSN*
>g2.t1
MSSSVNVKRTVMVAAGGIIYASTSIIMYKMTGHEELTQVQEDTKDGF5FVTDPOPNQT
FKQVAFYDSQIGRDEAVMGINWLRWLLWSHAKGTLEVAGAGTGRNIEYYPKGVDRVV
LSDVSDQMLLRKTKLHQINDEKNRKFATMEADANLAFDRCDFTVVDTFGLCSYDDP
VTVLKEMARVCKPNKILLLEHGRTKIWDLSRLYDKHAERHAKNMGCVMNRDLHILDE
AGLVVDRVDTWHFGTTYVVCRCRPGQKPEVSNVLAQFYSGPLSPFWNSNR*
>g4.t1
MVATSDNVKSADEKFYKVPIMYDHLMHFGRIGWVKIGAKKTNKLEKAIAKCVMGYSHD
HSGDTRYMFNPOTKKILNSRDIRWADWHGQTSPIAGLRGDFNVGDETEMVVIIDDEKQE
EDVLPAPVPIQQIDLETVPVVK*
>g5.t1
MFVGVIVETIEIHDEKSDTDDDFEINLTNNKTFDFYEVVEDTKVYITCEETEYAFIGVT
IEIEDEEINQAHASIKNSESKEICASIEENERWLADTGATSHITMCNKYMTNVKAVNRV
VVGDKVEICKERGDCVCRNKVTNETLLKNVLYPTTFHKNIIISIGTFVRDQKYLEGMKH
NMKTLTKAGKNETLDFKRDHSDVLYYFQGIIRGIYPGGSIDLSAEVITTKLTSMDINEAHA
KYGHIGEALRATMKSGLIKMTGVMYTCGALAKAKAKSAPKATMSKATQSGERLCTDI
SGPYKKSILGNDYWLVDVDTGKSWFFVKKSQLASKIEDLLTKLTAEVYTKFLRCD
NAGENVSLTKLCLKFNIQIEFTAPYTPQQNEIVE*
>g6.t1
ATKADIQAALVASITSSDNKKELSMSPFRNGDPTKFNEFWMTLTRLSTPVRWHILAS
PDDPSYESLSQALYLVLIHLDQAADHWSRNDILHNGIGLLAEVTKYRVSHSYSLIN
LLTAWDNLFQDKETPLQLSFRTRRELVSKAADAGQLFTEPFICYRFLALGPAFHAFVQP
YLLHQDKITMTLVLTASAKTYTDTPGFDSKMAHTGHRSAAPSSSDSSSPSTAWIGS
TSFNTSQARAMQFKCPIHRCNHDLAACSLVTSRFSIVPKPLSSSGGGPSTAPRPTTS

```

The figure on the right shows a properly formatted FASTA file.

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

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

The screenshot shows the OrthoMCL DB web interface. The main heading is "Map proteins to OrthoMCL with DIAMOND blastp". Below this, there is a "New job" tab selected. The form has two main sections: "Input Sequence(s)" and "Job Description". The "Input Sequence(s)" section has a "Choose File" button and a note: "Upload your set of proteins as a FASTA file. The file must have unique protein identifiers as the first word in the header line, and be no larger than 30MB. (You can also drag and drop a file onto the button.)". The "Job Description" section has a text input field with "My New Job" entered. A red arrow points from the "Map proteins!" button to a confirmation box on the right. The confirmation box is titled "DIAMOND Mapping Job - pending..." and displays the "Job id: ca0b3bc15bc76c8b779cab30c37f2ddd". It also shows a "Status: running" with a loading spinner and a note: "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 job while you wait."

**Understanding the output:** The output page has two components

- **The results table** (see below). This is a preview of the matching results for the first 100 sequences in your query file.
- **A blue download button** at the top right (see red arrow below). The complete result can be downloaded as a tab delimited file (tsv) with one best match for each query protein with the following columns:
  - **Query sequence id:** the identifier for the sequence in your input file
  - **Subject sequence id:** the identifier for the best matching OrthoMCL sequence
  - **OrthoMCL group id:** the orthogroup containing the best matching OrthoMCL sequence
  - **Description:** description of the best matching OrthoMCL sequence
  - **Alignment length:** length of the aligned region between Query and Subject sequences
  - **Percent identical matches:** percent identity between Query and Subject sequences
  - **Expect value:** BLAST significance score for the alignment between Query and Subject sequences. The expect value (E-value) cutoff is 0.05, allowing you to filter the output file more stringently if required.

**Note:** Unmatched query proteins (no significant match) are included in the results file without an OrthoMCL protein or group listed. For example, see the red rectangle below.

DIAMOND Mapping Job - result

<< All my DIAMOND Mapping Jobs

Job Id: ca0b3bc15bc75c8b779cab30c37f2ddd  
 Description: My New Job  
 Program: diamond-blastp

Showing all 75 sequences in your query file. [Download as a tsv file](#)

Query sequence id	Subject sequence id	OrthoMCL group id	Description	Alignment length	Percent identical matches	Expect value
g1.t1	tpse B8LDJ4	OG7_0006976	gene=B8LDJ4 product=Radical SAM core domain-containing protein (Fragment)	305	51.8	1.65e-94
g2.t1	tpse B8BV11	OG7_0001781	gene=B8BV11 product=Methyltransferase type 11 domain-containing protein	268	57.5	1.85e-98
g4.t1	vbra Vbra_363	OG7_0001280	gene=Vbra_363 product=unknown	120	26.7	1.49e-04
g5.t1	aalf AALF000687	OG7_0001280	gene=AALF000687 product=unknown	324	26.5	1.39e-20
g6.t1	tsti TSTA_009530	OG7_0001280	gene=TSTA_009530 product=RNA-directed DNA polymerase [Source:UniProtKB/TrEMBL;Acc:B6MFV4]	103	35.9	5.32e-03
g7.t1	tpse B8BU27	OG7_0009784	gene=B8BU27 product=Orc1-like AAA ATPase domain-containing protein	615	27.0	1.40e-69
g8.t1	tpse B8LDV1	OG7_0009784	gene=B8LDV1 product=Orc1-like AAA ATPase domain-containing protein (Fragment)	481	26.0	2.65e-44
g9.t1	tpse B8C6X7	OG7_0020726	gene=B8C6X7 product=DUF1995 domain-containing protein	341	60.7	9.37e-146
g9.t2	tpse B8C6X7	OG7_0020726	gene=B8C6X7 product=DUF1995 domain-containing protein	341	60.7	1.26e-145
g10.t2	tpse B8BWV7	OG7_0003575	gene=B8BWV7 product=Major facilitator superfamily (MFS) profile domain-containing protein (Fragment)	399	48.4	2.50e-114
g10.t1	tpse B8BWV7	OG7_0003575	gene=B8BWV7 product=Major facilitator superfamily (MFS) profile domain-containing protein (Fragment)	399	48.4	1.68e-114
g11.t1	tsti TSTA_111600	OG7_0001280	gene=TSTA_111600 product=RNA-directed DNA polymerase [Source:UniProtKB/TrEMBL;Acc:B6M929]	175	29.1	8.19e-08
g12.t1	*	-1	-1	-1	N/A	N/A
g13.t1	aalf AALF006108	OG7_0001280	gene=AALF006108 product=unknown	664	23.4	4.99e-33

Questions? Comments?

Contact us- [help@veupathdb.org](mailto:help@veupathdb.org)