Map Proteins to OrthoMCL with DIAMOND blastp
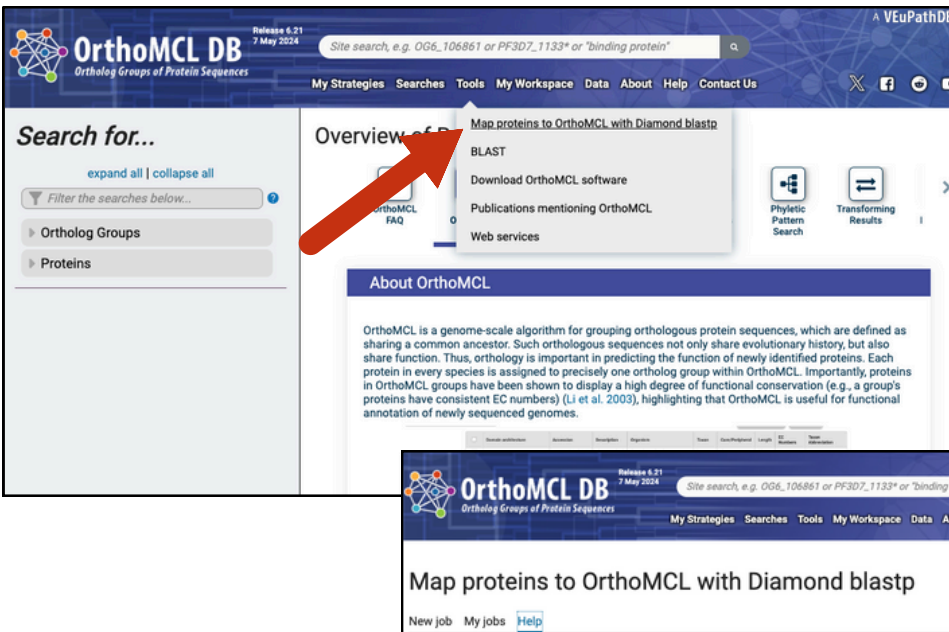## *A Tutorial*

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

1. 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.

2. **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.

3. This tool uses DIAMOND blastp, an alternative to NCBI BLAST which is 10,000 times faster 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)


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 page:

There are three tabs (circled in red on right)

1. **New job**: Upload a FASTA formatted file of protein sequences
2. **My jobs**: A list of all of your previous jobs. These are saved in your account and persist across visits to the website
3. **Help**: Tips for using the tool
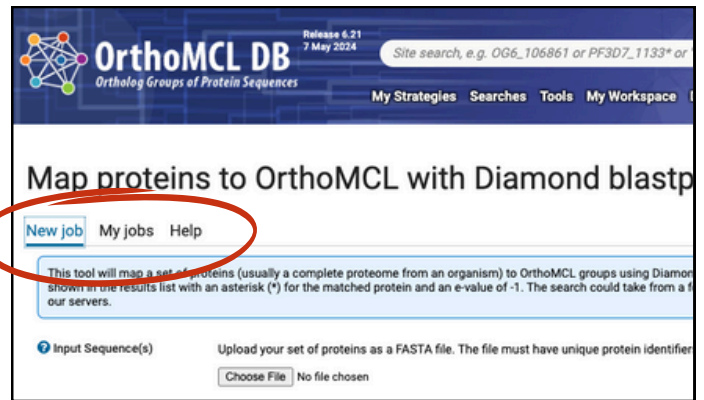
Preparing your data: Your set of proteins must be formatted as a plain text FASTA file.

- 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.

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)

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 BLAST button to start the job. You will see a message with a job ID assignment.

2
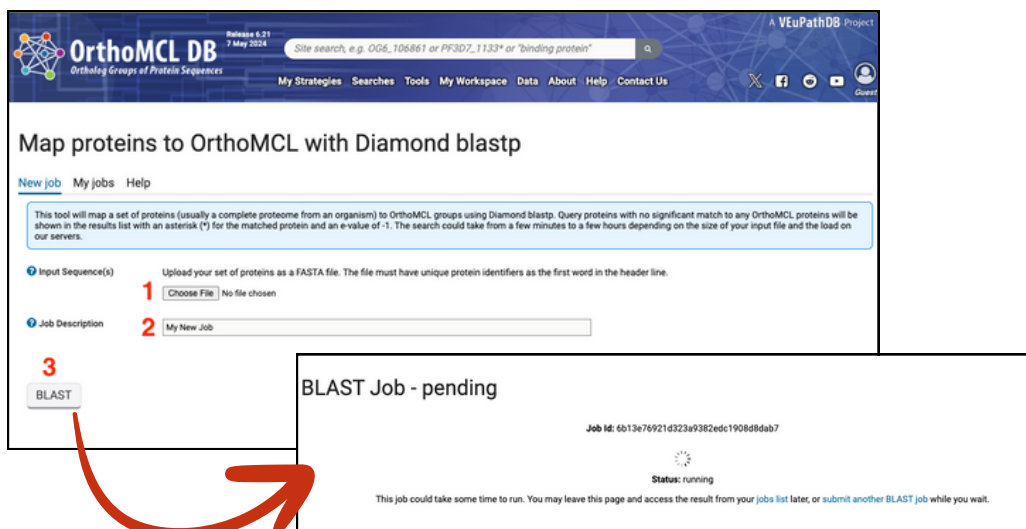
Understanding the output:

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 (tsv) with one best match for each query protein with the following columns:

   - **Query_ID**: the identifier for the sequence in your input file
   - **Subject_ID**: the identifier for the best matching OrthoMCL sequence
   - **Orthogroup**: the orthogroup containing the best matching OrthoMCL sequence
   - **Subject_description**: description of the best matching OrthoMCL sequence
   - **Alignment_length**: length of the aligned region between Query and Subject sequences
   - **Percent_identity**: percent identity between Query and Subject sequences
   - **e-value**: BLAST significance score for the alignment between Query and Subject sequences. The expectation 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 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

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	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

Questions? Comments? Write to

help@veupathdb.org