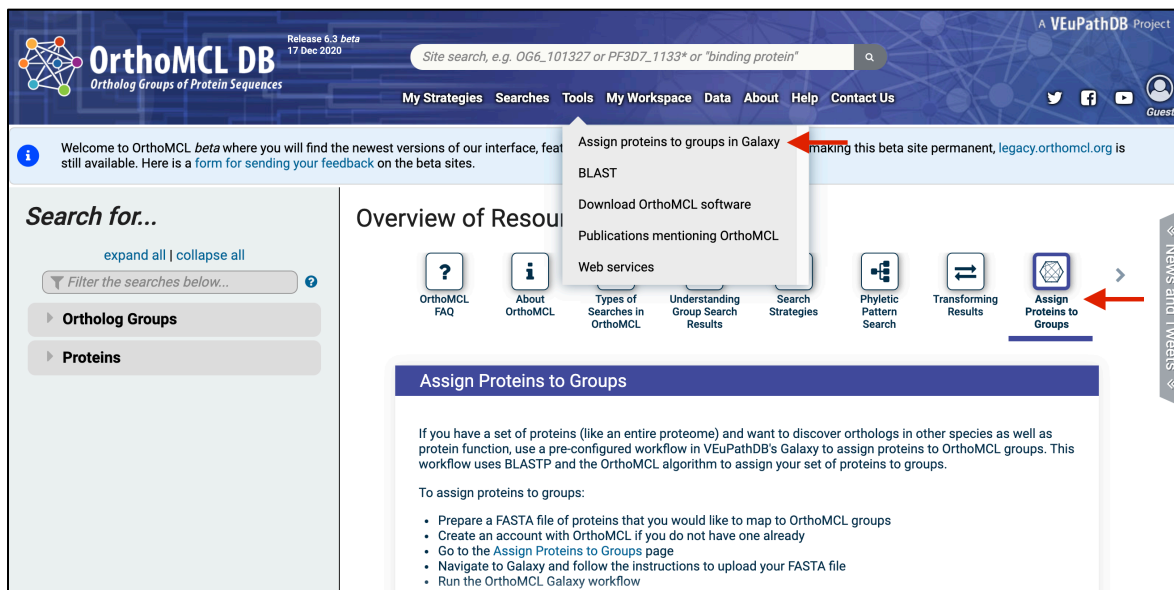


Map your proteins to OrthoMCL groups

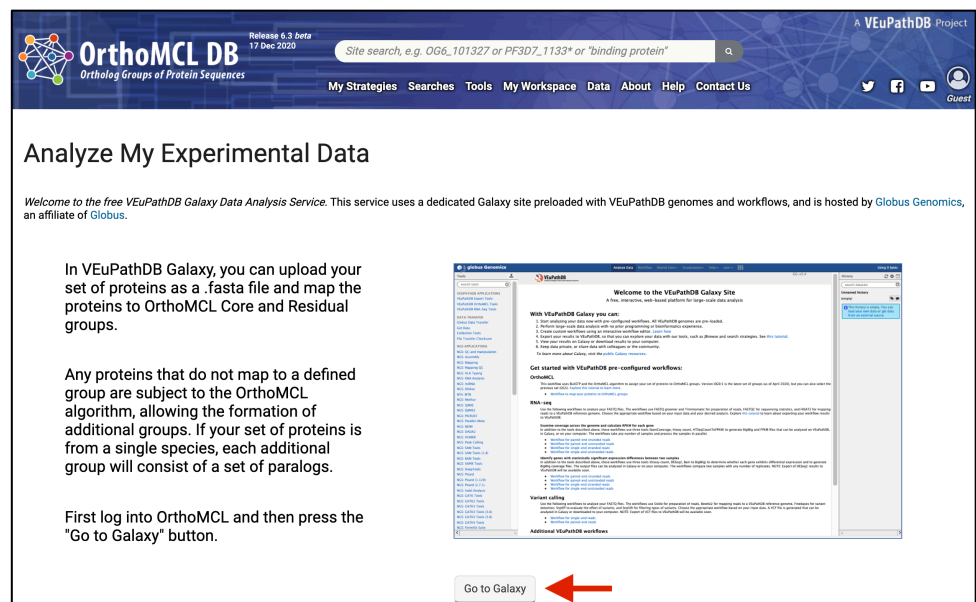
This tool allows you to map proteins from a FASTA file to OrthoMCL groups. The tool has been implemented as a workflow in the VEuPathDB Galaxy workspace. To use this tool, follow these steps:

1. There are two ways to access this feature: (a) Click on the “Tools” item in the grey menu and then select “Assign your proteins to groups in Galaxy”. (b) On the home page, click on the “Assign Proteins to Groups” entry in the Overview of Resources and Tools section. Then, select the link entitled “Assign Proteins to Groups”.



2. The next page provides instruction and a link to the VEuPathDB Galaxy server. Click on the “Go to Galaxy” button.

Note: to use this service, you must create a VEuPathDB account.



- Once in Galaxy, import your FASTA file containing protein sequences by clicking on the “Get Data” option in the left panel and selecting an import option, such as “Upload file from your computer”.

The screenshot shows the VEuPathDB Galaxy Site interface. The left sidebar contains a 'Tools' section with a search bar and a list of categories: 'VEUPATHDB APPLICATIONS', 'DATA TRANSFER', and 'NGS APPLICATIONS'. Under 'DATA TRANSFER', the 'Get Data' option is highlighted with a red arrow. The main content area features a 'Welcome to the VEuPathDB Galaxy Site' message, a description of the platform, and a list of tasks: '1. Start analyzing your data now with pre-configured workflows...', '2. Perform large-scale data analysis with no prior programming or bioinformatics experience.', '3. Use dataset Collections, to process multiple samples in parallel and/or to group paired-end files.', '4. Create custom workflows using an interactive workflow editor.', '5. Export your results to VEuPathDB...', '6. View your results on Galaxy or download results to your computer.', and '7. Keep data private, or share data with colleagues or the community.' The right sidebar shows a 'History' panel with an empty list and a message: 'This history is empty. You can load your own data or get data from an external source'.

- Your imported file will appear as a step in the History panel located in the right-hand side of the page. The color of the step indicates its status: grey indicates *in queue*, yellow *in process*, and green *completed*.
- Once your file has been successfully uploaded into Galaxy, select the “Workflow to map your proteins to OrthoMCL groups” from the middle section of the page.

The screenshot shows the VEuPathDB Galaxy Site interface after a file upload. The left sidebar is the same as in the previous screenshot. The main content area now displays 'Get started with VEuPathDB pre-configured workflows:' followed by 'OrthoMCL'. Below this, it states: 'This workflow uses BLASTP and the OrthoMCL algorithm to assign your set of proteins to OrthoMCL groups. Version OG6r1 is the latest set of groups (as of April 2020), but you can also select the previous set (OG5). Explore this OrthoMCL workflow tutorial to learn more.' A red arrow points to the link 'Workflow to map your proteins to OrthoMCL groups'. The right sidebar shows the 'History' panel with a list of datasets. The first dataset, '1: MyProteins.fasta', is highlighted in green, indicating it is completed. A red arrow points to this dataset.

6. The pre-written workflow will be imported into your workspace, where you can process your protein file. On the workflow page, there are three options that you need to set: (a) Proteome FASTA file, (b) Protein BLAST database, and (c) OrthoMCL Groups file. For (b) and (c), it is recommended to choose the most recent version, which is 6r1 or later; version 5 is an older version that is kept for those that would like to compare their current analysis with their old analyses.

Workflow: imported: Map your proteins to OrthoMCL groups (v.2) Run workflow

History Options

Send results to a new history

OrthoMCL Clean FASTA file Clean a FASTA file for use in OrthoMCL. The input is a pre-existing proteome FASTA file. The output is a proteome FASTA file that conforms to OrthoMCL requirements (Galaxy Version 1.0.0)

Proteome FASTA file:

1: MyProteins.fasta

A Proteome FASTA file. Must be a set of proteins from exactly one proteome. The file must contain only protein sequences, where each sequence starts with a definition line. The first field in the definition line must be a unique identifier.

NCBI BLAST+ blastp Search protein database with protein query sequence(s) (Galaxy Version 0.3.0)

Protein query sequence(s)

Output dataset 'output' from step 1 (-query)

Subject database/sequences

Locally installed BLAST database

Protein BLAST database

OrthoMCL 6r1 proteins blastDB

OrthoMCL 5 proteins blastDB

OrthoMCL Map Proteome to Groups Use BLAST results to map a proteome to OrthoMCL groups. (Galaxy Version 1.0.0)

Self-self BLAST result (OrthoMCL format):

Output dataset 'output' from step 6

OrthoMCL BLAST result (OrthoMCL format):

Output dataset 'output' from step 5

OrthoMCL Groups file

OG6r1_Groups

7. After setting up the workflow, press the “Run workflow” button at the top of the page. It may take a few minutes before the workflow starts to run.

globus Genomics Analyze Data Workflow Shared Data Visualization Help User Using 24.8 GB

Tools

search tools

VEUPATHDB APPLICATIONS

VEuPathDB Export Tools

VEuPathDB OrthoMCL Tools

VEuPathDB RNA-Seq Tools

DATA TRANSFER

Globus Data Transfer

Get Data

Upload File from your computer

Get single dataset from BaseSpace

Transfer dataset from BaseSpace by file ID

Workflow: imported: Map your proteins to OrthoMCL groups (v.2) Run workflow

History Options

Send results to a new history

OrthoMCL Clean FASTA file Clean a FASTA file for use in OrthoMCL. The input is a pre-existing proteome FASTA file. The output is a proteome FASTA file that conforms to OrthoMCL requirements (Galaxy Version 1.0.0)

Proteome FASTA file:

1: MyProteins.fasta

A Proteome FASTA file. Must be a set of proteins from exactly one proteome. The file must contain only protein sequences, where each sequence starts with a definition line. The first field in the definition line must be a unique identifier.

History

search datasets

Unnamed history

1 shown

1.04 KB

1: MyProteins.fasta

8. Workflow steps will queue up in the right-hand panel. The entire workflow may take up to 24 hours or more to run, depending on the number of proteins in your FASTA file.

The screenshot shows the globus Genomics interface. The top navigation bar includes 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The 'Workflow' tab is active. On the left, the 'Tools' panel lists various applications under 'VEUPATHDB APPLICATIONS' and 'NGS APPLICATIONS'. The main area displays a green success message: 'Successfully invoked workflow imported: Map your proteins to OrthoMCL groups (v.2). You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.' On the right, the 'History' panel shows a list of workflow steps, with the first step '1: MyProteins.fasta' highlighted in green.

Tools

search tools

VEUPATHDB APPLICATIONS

- VEuPathDB Export Tools
- VEuPathDB OrthoMCL Tools
- VEuPathDB RNA-Seq Tools

DATA TRANSFER

- Get Data
 - Upload File from your computer
 - Get single dataset from BaseSpace
 - Transfer dataset from BaseSpace by file ID
 - Get Sample set Data from BaseSpace
 - Transfer data from BaseSpace for a sample set
- Collection Tools
- File Transfer Checksum

NGS APPLICATIONS

- NGS: QC and manipulation
- NGS: Assembly
- NGS: Mapping
- NGS: Mapping QC
- NGS: HLA Typing
- NGS: RNA Analysis
- NGS: miRNA
- NGS: DNase
- NYS: MTB

History

search datasets

Unnamed history

9 shown, 1 hidden

1.04 KB

- 10: MCS on data 9
- 9: paralogPairs.txt
- 8: proteinsMappedToGroups.txt
- 7: reformat blastp MyProteins.fasta.clean vs __cn__sq__protein BLAST database from data 2__sq__cn__
- 6: reformat blastp MyProteins.fasta.clean vs __cn__sq__OrthoMCL_6r1_proteins_blast_db__sq__cn__
- 5: blastp MyProteins.fasta.clean vs 'protein BLAST database from data 2'
- 4: blastp MyProteins.fasta.clean vs 'OrthoMCL_6r1_proteins_blast_db'
- 3: protein BLAST database from data 2
- 1: MyProteins.fasta

9. Once the workflow has completed, there are two important tab-delimited output files that can be downloaded to your computer: (a) proteinsMappedToGroups.txt and (b) MCS.
10. File (a) contains your proteins that mapped to OrthoMCL groups. The columns are: your_protein_id, orthomcl_group_id, most_similar_orthomcl_protein_id, evalule_mantissa, evalule_exponent, percent_identity, percent_match.
11. Those proteins that do not map to OrthoMCL group are processed by the OrthoMCL algorithm to determine whether there is any pairwise similarity which would indicate paralogous groups. Each row of file (b) contains the proteins that make up a paralogous group.