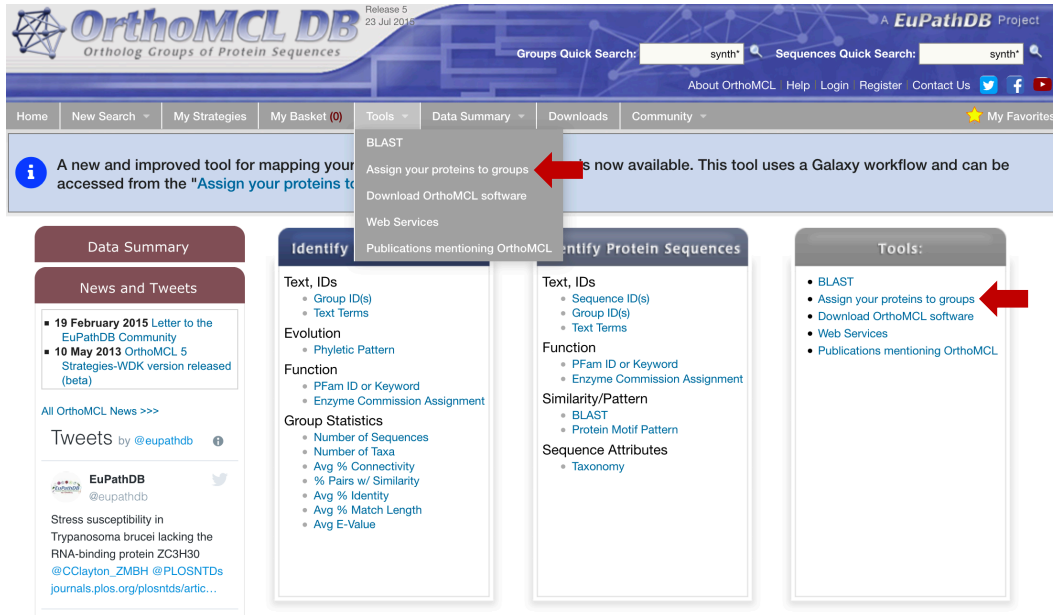


## Map your proteins to OrthoMCL groups

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

1. Click on the “Tools” item in the grey menu then select “Assign your proteins to groups” or click on the “Assign your proteins to groups” link from the right-hand side of the main OrthoMCL page.



2. The next page provides some instruction and a link to the EuPathDB Galaxy server. Click on the EuPathDB Galaxy Server link to access this service.  
**Note:** to use this service you will have to create a EuPathDB account.



## Map your proteins to OrthoMCL groups

If you have a .fasta file with a set of proteins, you can map the proteins to OrthoMCL Groups.

If your .fasta file contains the proteins from a single proteome, you can additionally find paralog groups for proteins that do not map to OrthoMCL Groups.

First please log in as a EuPathDB user.

Once logged in, go to the **EuPathDB Galaxy server**:

1. Use the **Get Data** tool (in the left panel at the top) to upload your .fasta file
2. On the Galaxy home page, click **Map a proteome to OrthoMCL groups**, under "Get started with pre-configured workflows:"
3. Then provide your .fasta file, and run the workflow!

**Note:** The Galaxy workflow run may take 24 hours or more to complete, depending on the size of the job queue

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

The screenshot shows the EuPathDB Galaxy Site interface. The left sidebar contains a 'Tools' section with a 'Get Data' button highlighted by a red box. Below 'Get Data' are four options: 'Get Data via Globus High speed file upload', 'Get Data via Globus from the EBI server using your unique file identifier', 'Upload File from your computer', and 'Send Data via Globus Transfers data via Globus'. The main content area displays a welcome message and a list of pre-configured workflows. The right sidebar shows a 'History' panel with a search bar and a message indicating that the history is empty.

- Your imported file will appear as a step in the right hand history panel. The color of the step indicates its status. The step is grey when it is in queue, yellow when in process and green when completed.
- Once your protein FASTA file has been successfully uploaded into Galaxy, select the “Map your proteins to OrthoMCL groups” from the middle section.



The screenshot shows the EuPathDB Galaxy Site interface after a file has been uploaded. The left sidebar now shows a 'Get Data' button. The main content area displays the 'Map your proteins to OrthoMCL groups' workflow, which is highlighted by a red box. The right sidebar shows a 'History' panel with a search bar and a list of steps. The first step, 'AssignToOrthoMCL', is shown with a green status icon, indicating it is completed. Below it, a step labeled '1: MyProteins' is also shown with a green status icon.

6. This will import the workflow into your workspace and allow you to run it on your protein file. In most cases you will not need to change any of the default parameters. Simply click on the “Run workflow” button at the bottom of the middle section. Note that there is a few second lag between clicking on the button and the workflow starting to run – please be patient.

### Running workflow “imported: Map your proteins to OrthoMCL groups”

[Expand All](#)[Collapse](#)

Assign your set of proteins to OrthoMCL groups. This workflow uses BLASTP and the OrthoMCL algorithm to (1) map proteins to OrthoMCL groups using BLAST similarity and (2) form paralog groups from proteins with no significant similarity to any OrthoMCL proteins. The workflow produces a file with the mapping from protein ID to group ID, along with similarity metrics. It also produces a file of paralog groups. The latter file is only valid if your input proteins all belong to a single proteome. This workflow might take 24 hours or more to run, depending on the size of the job queue.

<b>Step 1: OrthoMCL Clean FASTA file</b> (version 1.0.0) 1 <b>Proteome FASTA file:</b> 1: MyProteins  <b>Maximum allowed number of input sequences</b> 100000  <b>Action:</b> Hide output 'output'.
<b>Step 2: NCBI BLAST+ makeblastdb</b> (version 0.3.0) 7
<b>Step 3: NCBI BLAST+ blastp</b> (version 0.3.0) 3
<b>Step 4: NCBI BLAST+ blastp</b> (version 0.3.0) 2
<b>Step 5: OrthoMCL Reformat Blast</b> (version 1.0.0) 5
<b>Step 6: OrthoMCL Reformat Blast</b> (version 1.0.0) 4
<b>Step 7: OrthoMCL Map Proteome to Groups</b> (version 1.0.0) 6
<b>Step 8: MCL</b> (version 14.137) 7

☐ Send results to a new history[Run workflow](#)

7. 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 size.

The screenshot shows the globus Genomics interface. On the left is a sidebar with various tools categorized under 'Tools', 'Get Data', and 'NGS APPLICATIONS'. The main area displays a green notification box stating: 'Successfully ran workflow "Imported: Map your proteins to OrthoMCL groups". The following datasets have been added to the queue: 2: MyProteins.clean, 3: protein.BLAST database from data 2, 4: blastp MyProteins.clean vs 'orthomcl\_v5\_proteins\_blast\_db', 5: blastp MyProteins.clean vs 'protein.BLAST database from data 2', 6: reformat blastp MyProteins.clean vs 'cn\_sq\_orthomcl\_v5\_proteins\_blast\_db\_sq\_cn\_', 7: reformat blastp MyProteins.clean vs 'cn\_sq\_protein.BLAST database from data 2\_sq\_cn\_', 9: paralogPairs.txt, 8: proteinsMappedToGroups.txt, 10: MCS on data 9'. On the right, the 'History' panel shows a list of workflow steps, with '2: MyProteins.clean' highlighted in yellow.

8. Once the workflow has completed, you can work with the output files. Click on the name of the step in your history in the right-hand panel to expand it. This provides additional options including the option to download file.

This screenshot shows the 'History' panel with the step '10: MCS on data 9' expanded. It displays details for the 'proteinsMappedToGroups.txt' file, including its size (7,002 lines) and format (tabular). Below this, there is a section for scanning files, listing paths like '/mnt/galaxyIndices2/genomes/miscellaneous/OrthoMCL\_Map\_Proteome\_to\_Groups' and '/scratch/galaxy/files/051/dataset\_51041.dat'. At the bottom, a 'Download' button is visible, along with a list of protein IDs and their corresponding gene names, such as 'mRNA1\_NF00000110-p1 OG5\_129304 ddisl dhk' and 'mRNA1\_NF0000020-p1 OG5\_197203 creil156'.