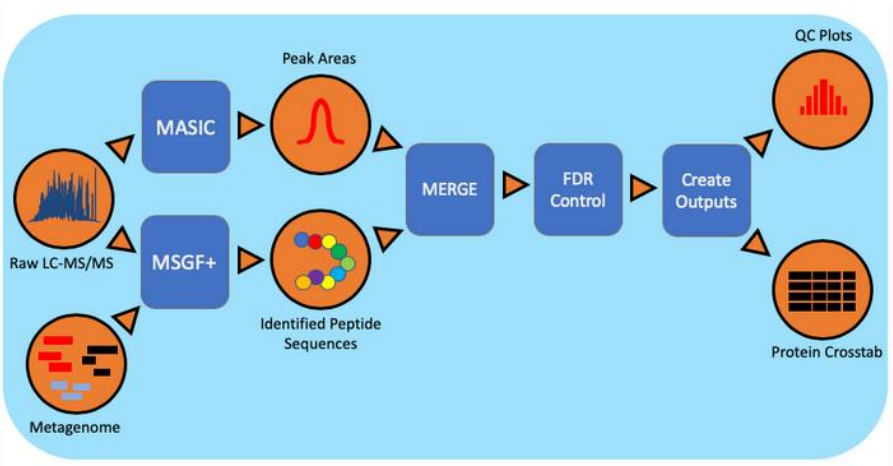
Metaproteomics Workflow (v1.0.0)



Overview

The metaproteomics workflow is an end-to-end data processing and analyzing pipeline for studying proteomes i.e studying protein identification and characterization using MS/MS data.

Running the Workflow

Currently, this workflow can be run in NMDC EDGE or from the command line (CLI instructions and requirements are found here).

Tutorial videos on how to run each workflow in NMDC EDGE are found here.

Input

Metaproteomics requires xxx.

- **Acceptable file formats:** .raw, xxx

Details

The metaproteomics workflow/pipeline is an end-to-end data processing workflow for protein identification and characterization using MS/MS data. Briefly, mass spectrometry instrument generated data files(.RAW) are converted to mzML, an open data format, using MSConvert. Peptide identification is achieved using MSGF+ and the associated metagenomic information in the FASTA (protein sequences) file format. Intensity information for identified species is extracted using MASIC and combined with protein information.

Software Versions

- MSGFPlus (v20190628
- Mzid-To-Tsv-Converter (v1.3.3)
- PeptideHitResultsProcessor (v1.5.7130
- pwiz-bin-windows (x86_64-vc141-release-3_0)20149_b73158966
- MASIC (v3.0.7235
- Sqlite-netFx-full-source (1.0.111.0)
- Conda (3-clause BSD)

Output

The table below lists the primary output files. The main outputs are xxx.

Primary Output Files	Description

Running the Metaproteomics Workflow in NMDC EDGE

Select a workflow

1. From the Metaproteomics category in the left menu bar, select 'Run a Single Workflow'.
2. Enter a ***unique*** project name with no spaces (underscores are fine).
3. A description is optional, but helpful.
4. Select 'Metaproteomics' from the dropdown menu under Workflow.

Input

The metaproteome workflow requires xxx. **Acceptable file formats:** xxx.

5. Click the button to the right of the blank for Input Raw File. A box called 'Select a File' will open to allow the user to find the desired file from a previously run assembly project, the public data folder, or a file uploaded by the user.
6. Click the button to the right of the blank for Input Fasta File. A box called 'Select a File' will open to allow the user to find the read mapping file from a previously run assembly project, the public data folder, or a file uploaded by the user.
7. Click the button to the right of the blank for Input GFF File. A box called 'Select a File' will open to allow the user to find the desired file(s) from a previously run annotation project, the public data folder, or a file uploaded by the user.
8. Select 'True' or 'False' to specify whether your mass spec file comes from a ThermoFisher instrument.
9. Select your desired QValue Threshold for analyzing peptides of interest using the sliding bar.
10. Input the name of your study from its sequencing project. If none, put in any name.
11. Click 'Submit' when ready to run the workflow.

Output

The General section of the output shows which workflow and which tools were run and the run time information.

The Metaproteome Result section includes xxx.

The Browser/Download Output section provides all output files available to download. Xxx.