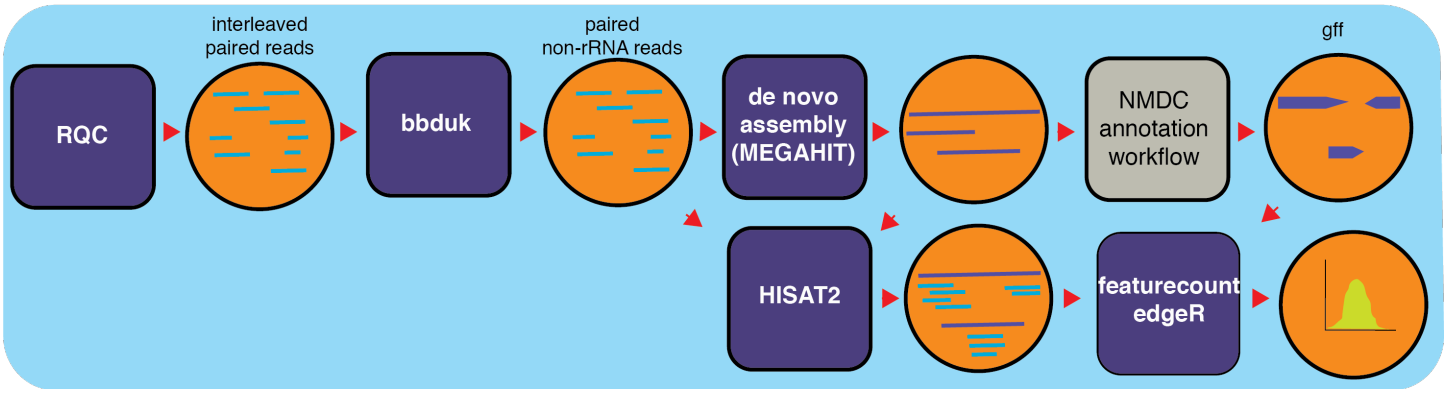
Metatranscriptomics Workflow (v0.0.2)



Overview

The metatranscriptome (metaT) workflow takes in raw metatranscriptome data, filters the data for quality, removes rRNA reads, then assembles and annotates the transcripts. The data is mapped back to the genomic features in the transcripts and RPKMs ((Reads Per Kilobase of transcript per Million mapped reads) are calculated for each feature in the functional annotation file.

Running the Workflow

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

Input

Metatranscriptomics requires paired-end Illumina data as an interleaved file or as separate pairs of FASTQ files.

- **Acceptable file formats:** .fastq, .fq, .fastq.gz, .fq.gz

Details

MetaT is a workflow designed to analyze metatranscriptomes, and this workflow builds upon other NMDC workflows for processing input sequencing data. The metatranscriptomics workflow takes in raw RNA sequencing data and quality filters the reads using the ReadsQC workflow. Then the MetaT workflow filters out ribosomal RNA reads (using the SILVA rRNA database) and separates interleaved files into separate pairs of files using bbduk (BBTools). After the filtering steps, the reads are assembled into transcripts using MEGAHIT and transcripts are annotated using the Metagenome Annotation NMDC Workflow which produces GFF functional annotation files. Features are counted with Subread's featureCounts which assigns mapped reads to genomic features and generates RPKMs for each feature in a GFF file for sense and antisense reads.

Software Versions

- BBTools v38.44
- hisat2 v2.1
- Python v3.7.6.
- featureCounts v2.0.1
- R v3.6.0
- edgeR v3.28.1
- pandas v1.0.5
- gffutils v0.10.1

Output

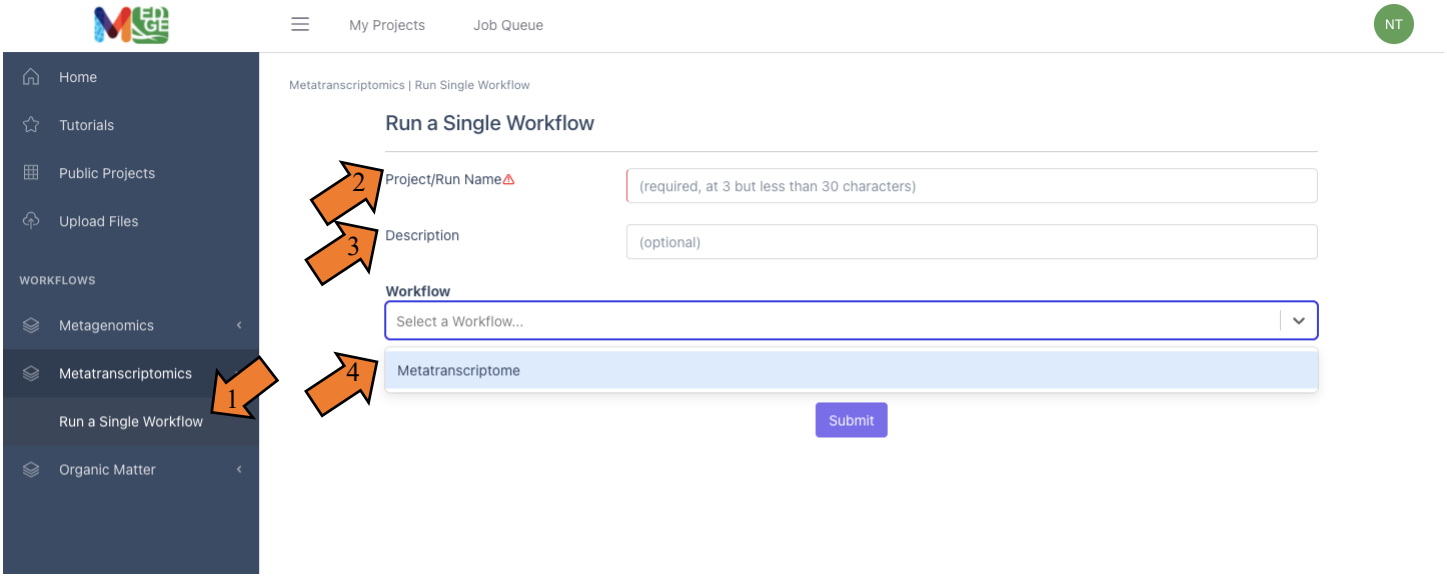
The table below lists the primary output files. The main outputs are the assembled transcripts and annotated features file. Several annotation files are also available to download.

Primary Output Files	Description
INPUT_NAME.contigs.fa	Assembled transcripts
rpkm_sorted_features.tsv	Feature table sorted by RPKM

Running the Metatranscriptomics Workflow in NMDC EDGE

Select a workflow

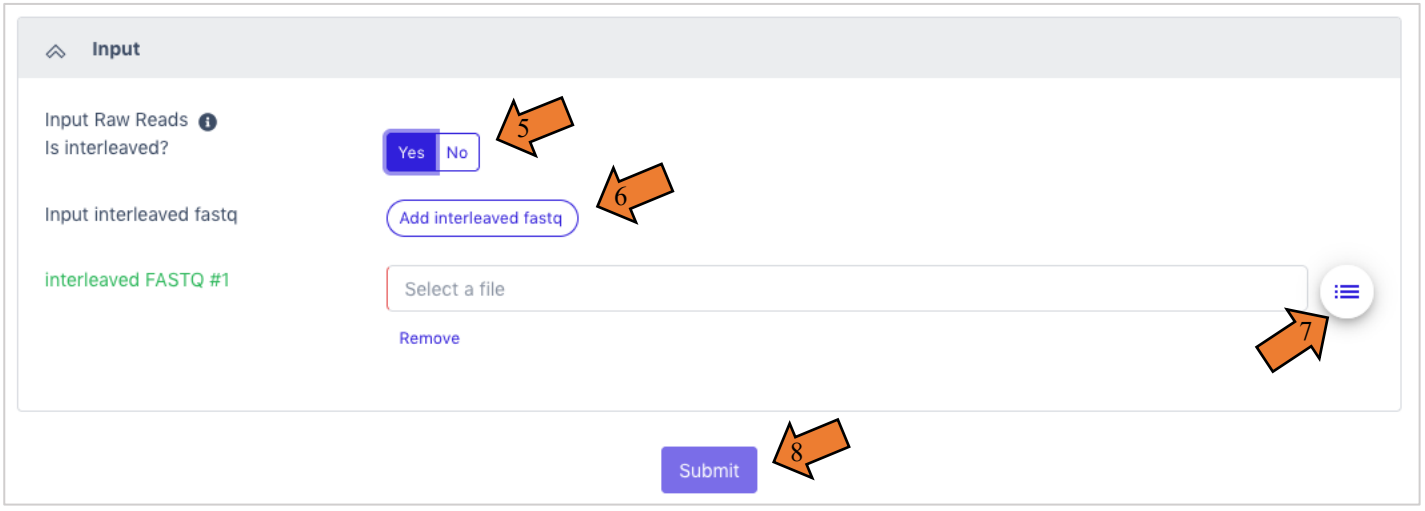
1. From the Metatranscriptomics category in the left menu bar, select 'Run a Single Workflow'.
2. Enter a ***unique*** project name with no spaces (underscores are fine).
3. A description is optional, but helpful.
4. Select 'Metatranscriptome' from the dropdown menu under Workflow.



Input

The metatranscriptome workflow requires paired-end Illumina data in FASTQ format as the input; the file can be interleaved and can be compressed. **Acceptable file formats:** .fastq, .fq, .fastq.gz, .fq.gz

5. The default setting is for the raw data to be in an interleaved format (paired reads interleaved into one file). If the raw data is paired reads in separate files (forward and reverse), click 'No'.
6. Additional data files (of the same type–interleaved or separate) can be added with the button below.
7. Click the button to the right of the input blank to select the data file for the analysis. (If there are separate files, there will be two input blanks.) A 'Select a File' box will open to allow the user to find the desired file(s) from previously run projects, the public data folder, or files uploaded by the user.
8. Click 'Submit' when ready to run the workflow.



Output

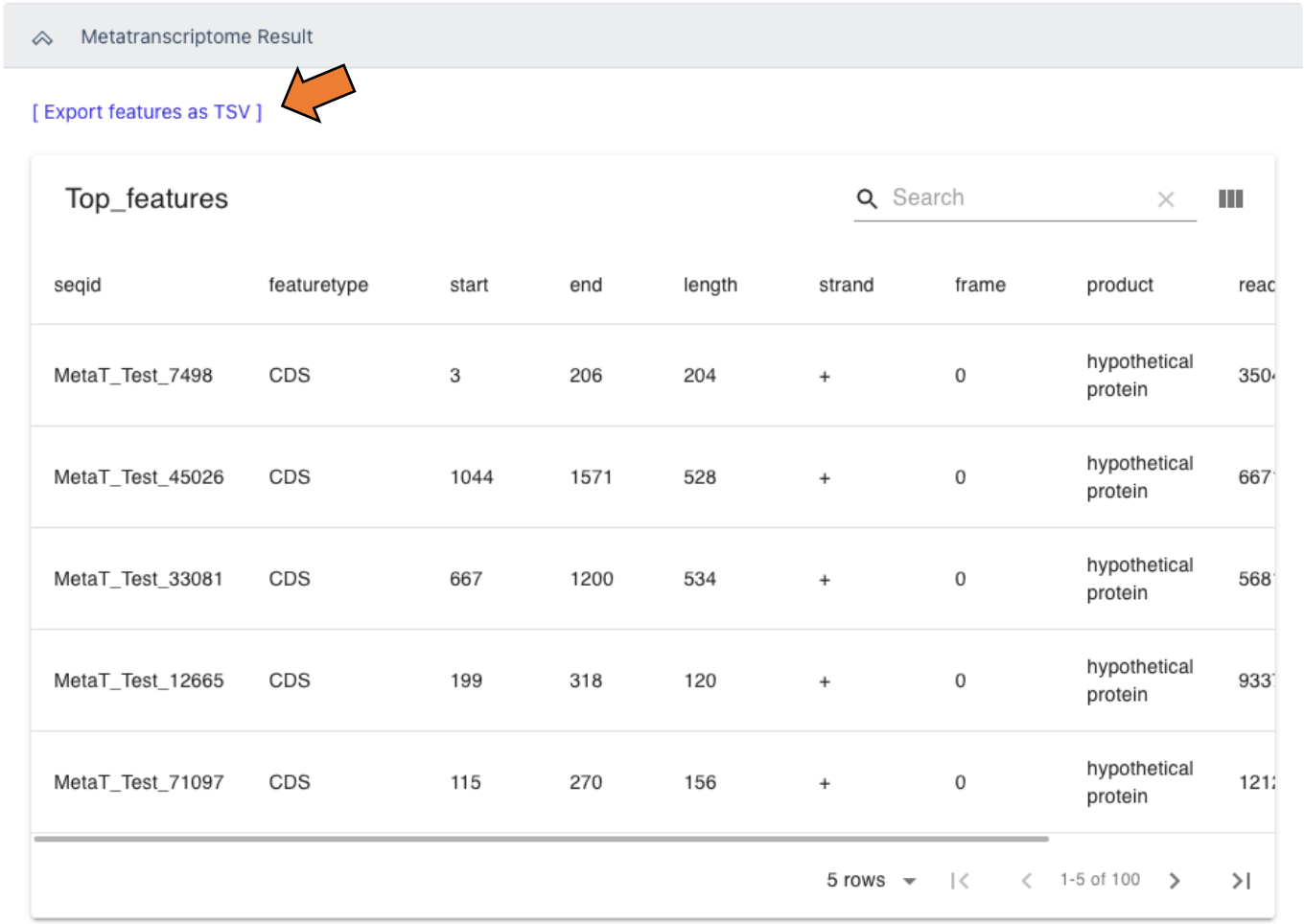
The General section of the output shows which workflow was run, the run time information, and the Project Configuration

General

Workflow	Run	Status	Running Time	Start	End
Metatranscriptome	On	Done	04:02:07	2022-01-14 17:06:24	2022-01-14 21:08:31

"Project Configuration" : {...}

The Metatranscriptome Result section includes a table of the top 100 RPKM results from the overall metatranscriptome data file sorted by RPKM. Selecting the header of each column will sort this data by that column. This section also includes a button to quickly download a tsv file of all detected features in the input dataset for further analysis.



The Browser/Download Output section provides all output files available to download. The output contigs can be found in the assembly folder and the tsv file of all detected features sorted by RPKM is available under the metat_output folder.