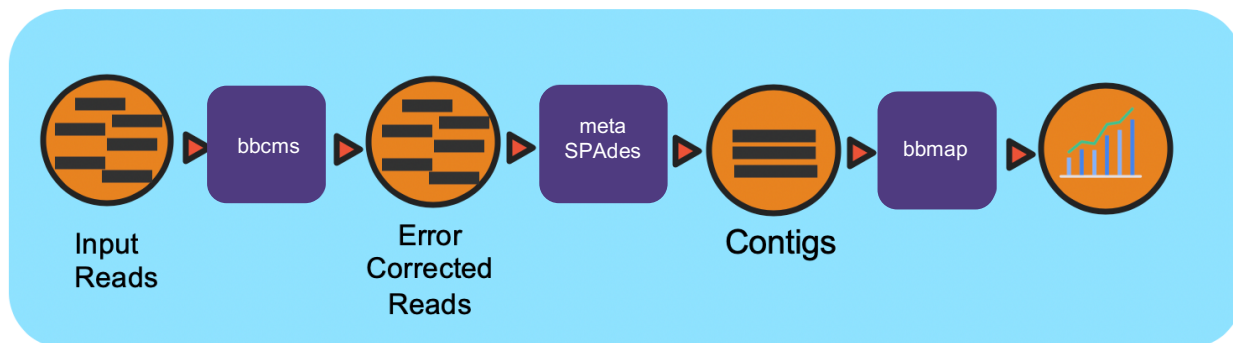


Metagenome Assembly Workflow (v1.0.1)



Overview

This workflow takes in paired-end Illumina data, runs error correction, assembly, and assembly validation.

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

Metagenome Assembly requires paired-end Illumina data as an interleaved file or as separate pairs of FASTQ files. The recommended input is the output from the ReadsQC workflow.

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

Details

This workflow takes in paired-end Illumina reads and performs error correction using bbcms (BBTools). Then the corrected reads are assembled using metaSPAdes. After assembly, the reads are mapped back to the contigs by bbmap (BBTools) for coverage information.

Software Versions and Parameters

- bbcms (BBTools v38.94)
- metaSpades v3.15.0
- bbmap (BBTools v38.94)

Output

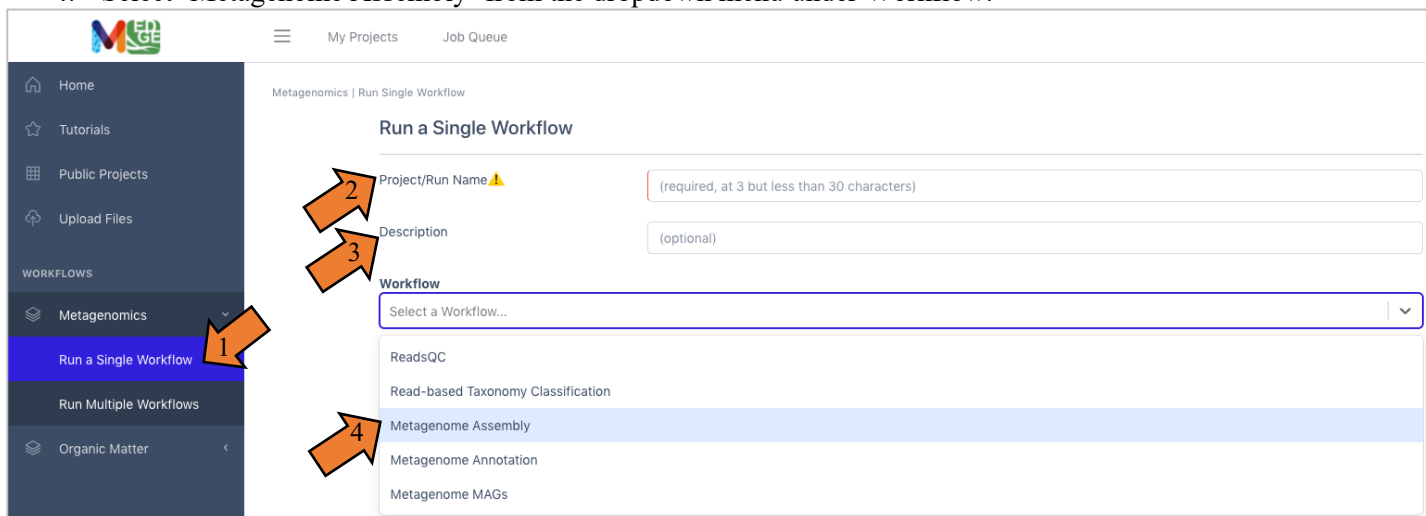
The main output is the assembled contigs file (assembly_contigs.fna).

| Primary Output Files | Description |
|-------------------------|---|
| Assembly Contigs | Final assembly contigs (assembly_contigs.fna) |
| Assembly Scaffolds | Final assembly scaffolds (assembly_scaffolds.fna) |
| Assembly AGP | An AGP format file which describes the assembly |
| Assembly Coverage BAM | Sorted bam file of reads mapping back to the final assembly |
| Assembly Coverage Stats | Assembled contigs coverage information |

Running the Metagenome Assembly Workflow in NMDC EDGE

Select a workflow

1. From the Metagenomics 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 'Metagenome Assembly' from the dropdown menu under Workflow.

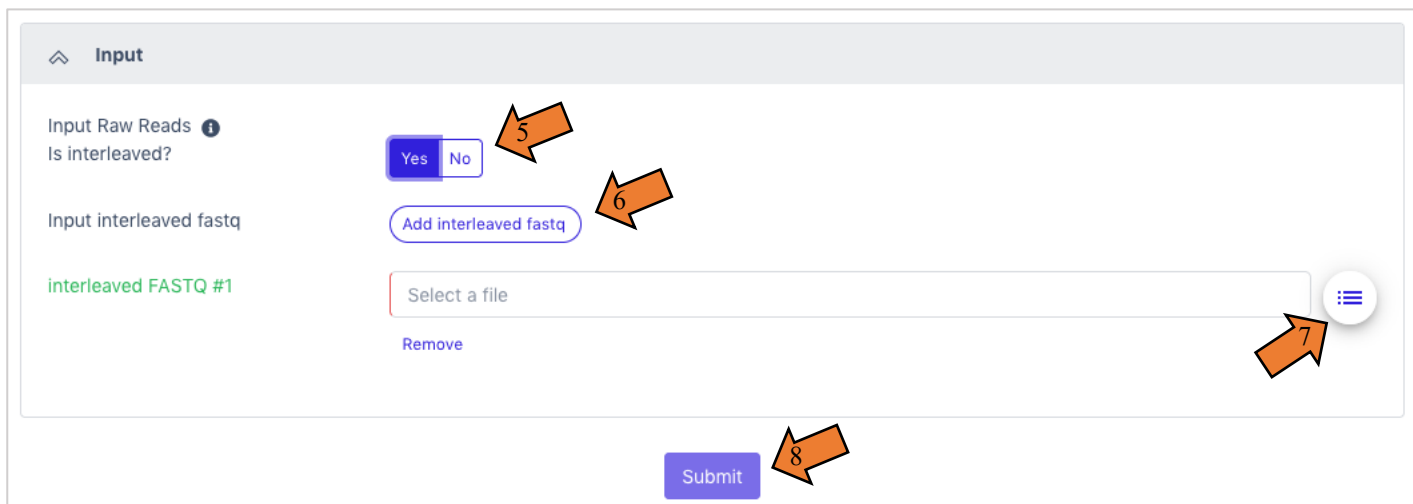


Input

This workflow accepts Illumina data in FASTQ format as the input; the file can be interleaved and can be compressed. (It is highly recommended to input clean data from the ReadsQC workflow.)

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 for data to select the data file for the analysis. (If there are separate files, there will be two input blanks.) A box called 'Select a File' 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. Then click 'Submit'.



Output

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

| General | | | | | |
|-------------------------------------|-----|--------|--------------|---------------------|---------------------|
| Workflow | Run | Status | Running Time | Start | End |
| Metagenome Assembly | On | Done | 02:02:22 | 2021-10-13 22:51:35 | 2021-10-14 00:53:57 |
| "Project Configuration" : { . . . } | | | | | |

The Metagenome Assembly Result section has all of the statistics from the assembly.

| Metagenome Assembly Result | |
|----------------------------|------------------------|
| Name | Status |
| scaffolds | 25,324 |
| contigs | 25,726 |
| scaf_bp | 52,206,897 |
| contig_bp | 52,201,077 |
| gap_pct | 0.011 |
| scaf_N50 | 691 |
| scaf_L50 | 4,103 |
| ctg_N50 | 724 |
| ctg_L50 | 3,971 |
| scaf_N90 | 14,186 |
| scaf_L90 | 726 |
| ctg_N90 | 14,473 |
| ctg_L90 | 716 |
| scaf_logsum | 645,093 |
| scaf_powsum | 120,098 |
| ctg_logsum | 638,015 |
| ctg_powsum | 116,432 |
| asm_score | 33.765 |
| scaf_max | 1,491,105 |
| ctg_max | 859,644 |
| scaf_n_gt50K | 96 |
| scaf_l_gt50K | 20,678,937 |
| scaf_pct_gt50K | 39.61 |
| gc_avg | 0.473 |
| gc_std | 0.062 |
| filename | assembly_scaffolds.fna |

The Browser/Download Output section provides output files available to download. The primary result is the assembly_contigs.fna file which can also be the input for the Metagenome Annotation workflow. The pairedMapped_sorted.bam file along with the assembled contigs file can be the input for the MAGs Generation workflow.

| Browser/Download Outputs | | |
|--------------------------|------------|---------------|
| File | Size | Last Modified |
| MetagenomeAssembly | | |
| assembly.agp | 1.72 MB | 21 days ago |
| assembly_contigs.fna | 51.30 MB | 21 days ago |
| assembly_scaffolds.fna | 51.22 MB | 21 days ago |
| covstats.txt | 1.92 MB | 21 days ago |
| pairedMapped.sam.gz | 2338.54 MB | 21 days ago |
| pairedMapped_sorted.bam | 2130.75 MB | 21 days ago |
| stats.json | 619 B | 21 days ago |