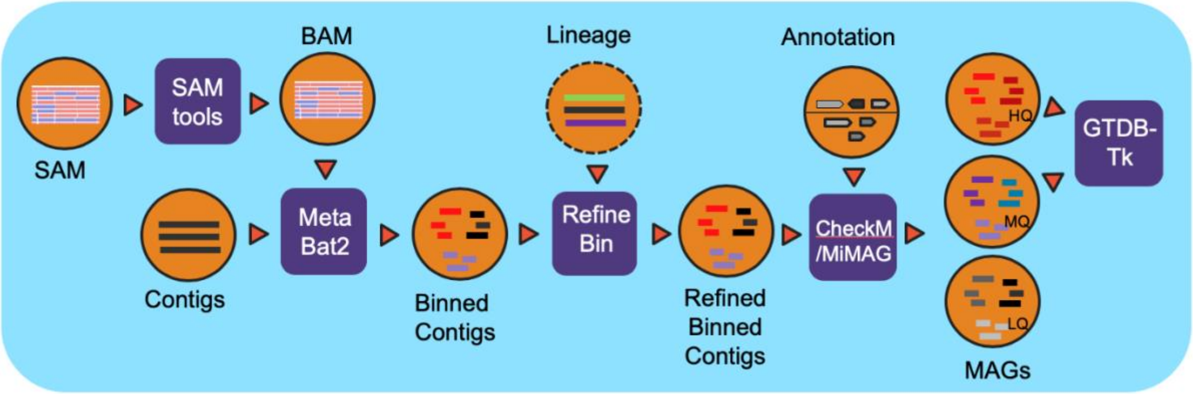
Metagenome Assembled Genomes (MAGs) Workflow (v1.0.2)



Overview

This workflow classifies contigs into bins and the resulting bins are refined using the functional annotation file. The bins are evaluated for completeness and contamination. The quality of the bins is determined, and a lineage is assigned to each bin of high or medium quality.

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

This workflow requires assembled contigs in a FASTA file, the read mapping file from the assembly (SAM or BAM), a functional annotation of the assembly in a GFF file.

- **Acceptable file formats:** assembled contigs (.fasta, .fa, or .fna); read mapping to assembly (.sam.gz or .bam); Functional annotation (.gff)

Details

The workflow is based on IMG metagenome binning pipeline and has been modified specifically for the NMDC project. For all processed metagenomes, it classifies contigs into bins using MetaBat2. Next, the bins are refined using the functional Annotation file (GFF) from the Metagenome Annotation workflow and optional contig lineage information. The completeness of and the contamination present in the bins are evaluated by CheckM and bins are assigned a quality level (High Quality (HQ), Medium Quality (MQ), Low Quality (LQ)) based on MiMAG standards. In the end, GTDB-Tk is used to assign lineage for HQ and MQ bins. The required GTDB-Tk database is incorporated into NMDC EDGE.

Software Versions

- Biopython v1.74
- Sqlite
- Pymysql
- requests
- samtools > v1.9 (License: MIT License)
- Metabat2 v2.15
- CheckM v1.1.2
- GTDB-TK v1.2.0
- FastANI v1.3
- FastTree v2.1.10

Output

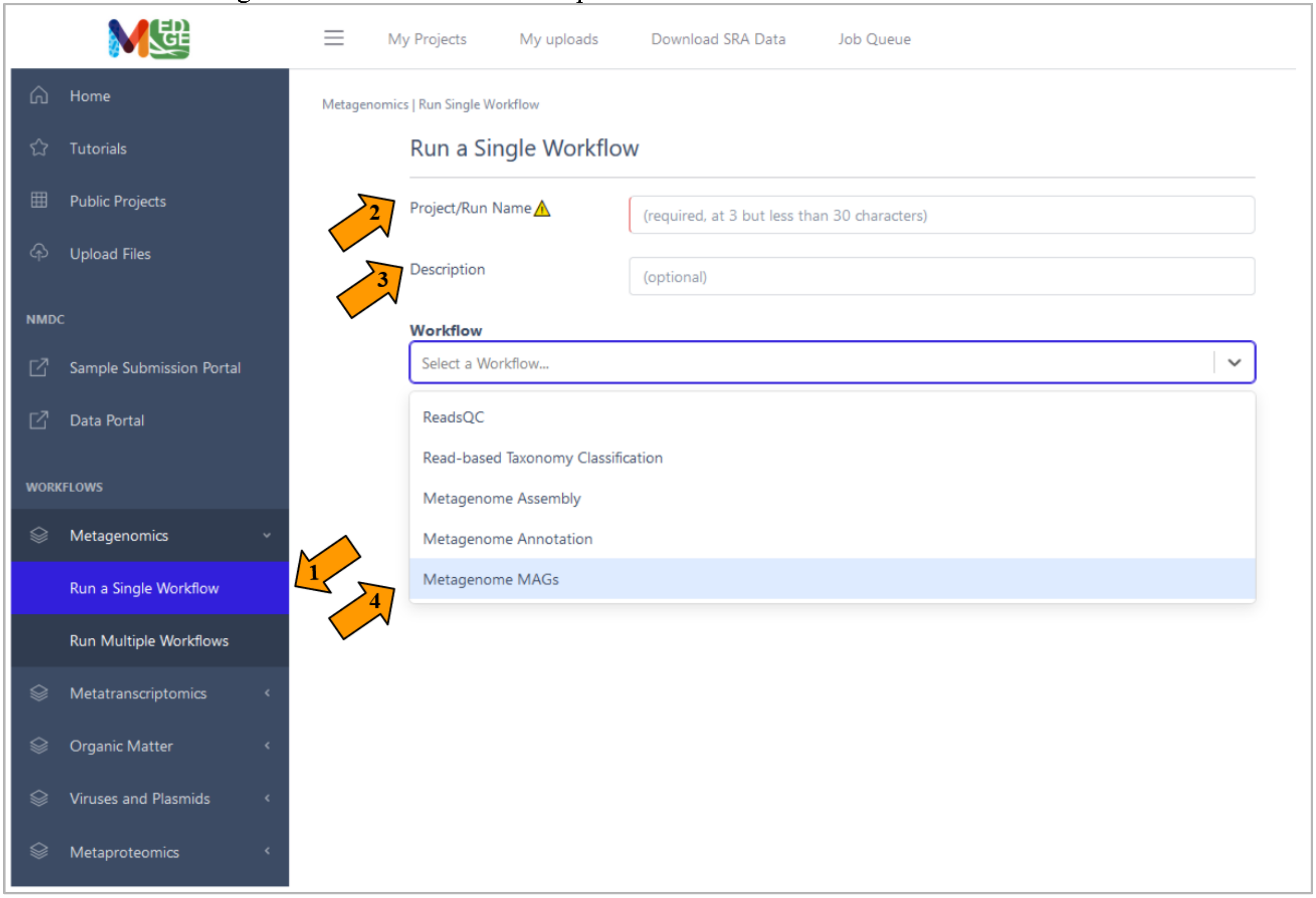
The primary output is the zipped file of high quality (HQ) and medium quality (MQ) bins. Unbinned contigs, and low-quality bins are also available.

Primary Output Files	Description
hqmq-metabat-bins.zip	Bins of contigs rated high or medium quality

Running the Metagenome Assembled Genomes (MAGs) Workflow in NMDC EDGE

Select a workflow

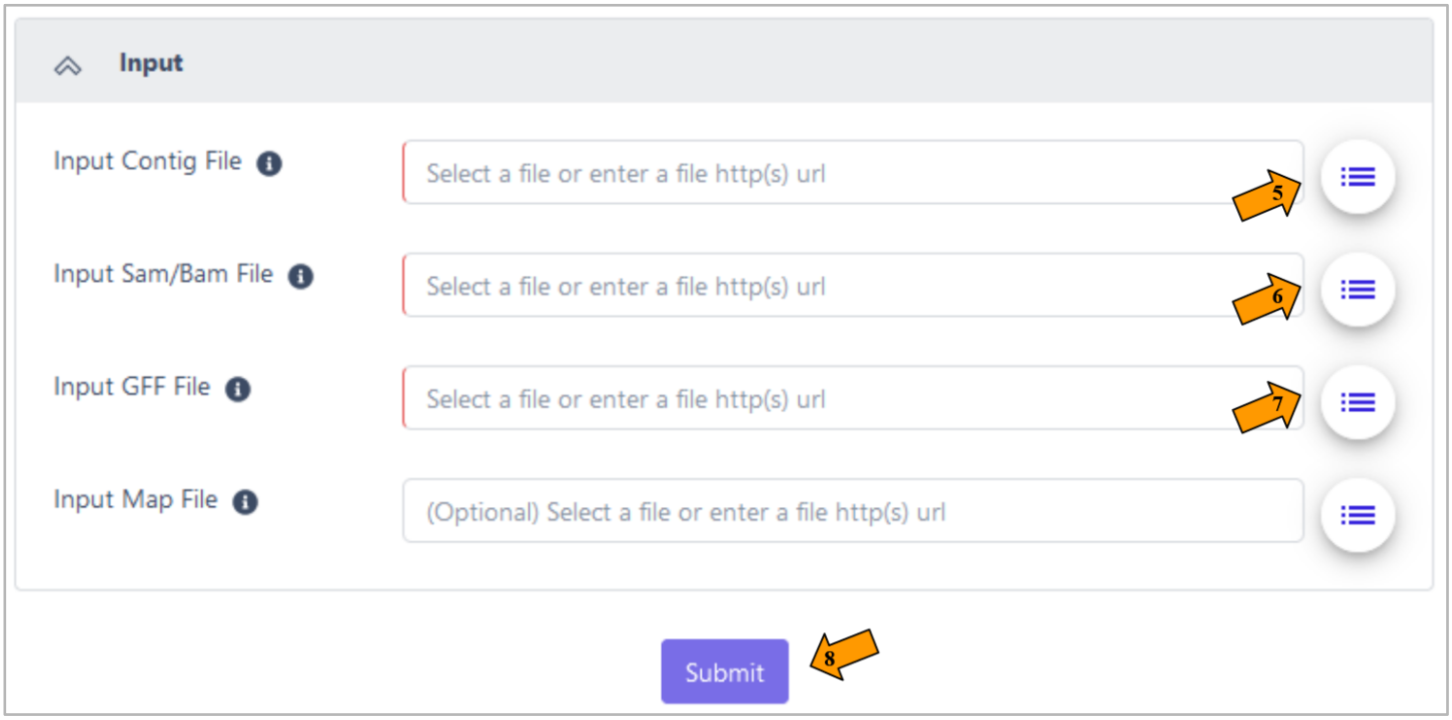
1. From the Metagenomics category in the left menu bar, select ‘Run a Single Workflow’.
2. Enter a ***unique*** project name with no spaces (underscores are fine).
3. A description is optional, but helpful.
4. Select ‘Metagenome MAGs’ from the dropdown menu under Workflow.



Input

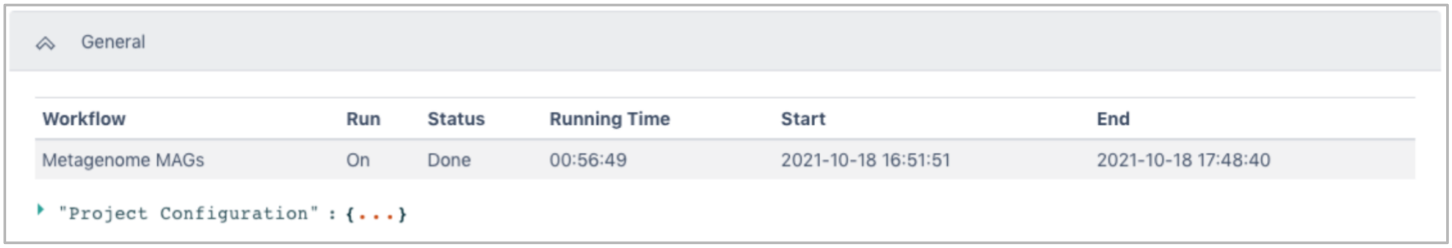
Metagenome MAGs requires assembled contigs, the read mapping file of reads to assembled contigs, and a functional annotation file. The recommended input would be from the NMDC assembly and annotation workflows**. Acceptable file formats:** assembled contigs (.fasta, .fa, or .fna); read mapping to assembly (.sam.gz or .bam); functional annotation (.gff).

5. Click the button to the right of the blank for Input Contig 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 Sam/Bam 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. Then click ‘Submit’.


Input
Input Contig File
Select a file or enter a file http(s) url
5
Input Sam/Bam File
Select a file or enter a file http(s) url
6
Input GFF File
Select a file or enter a file http(s) url
7
Input Map File
(Optional) Select a file or enter a file http(s) url
Submit
8

Output

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

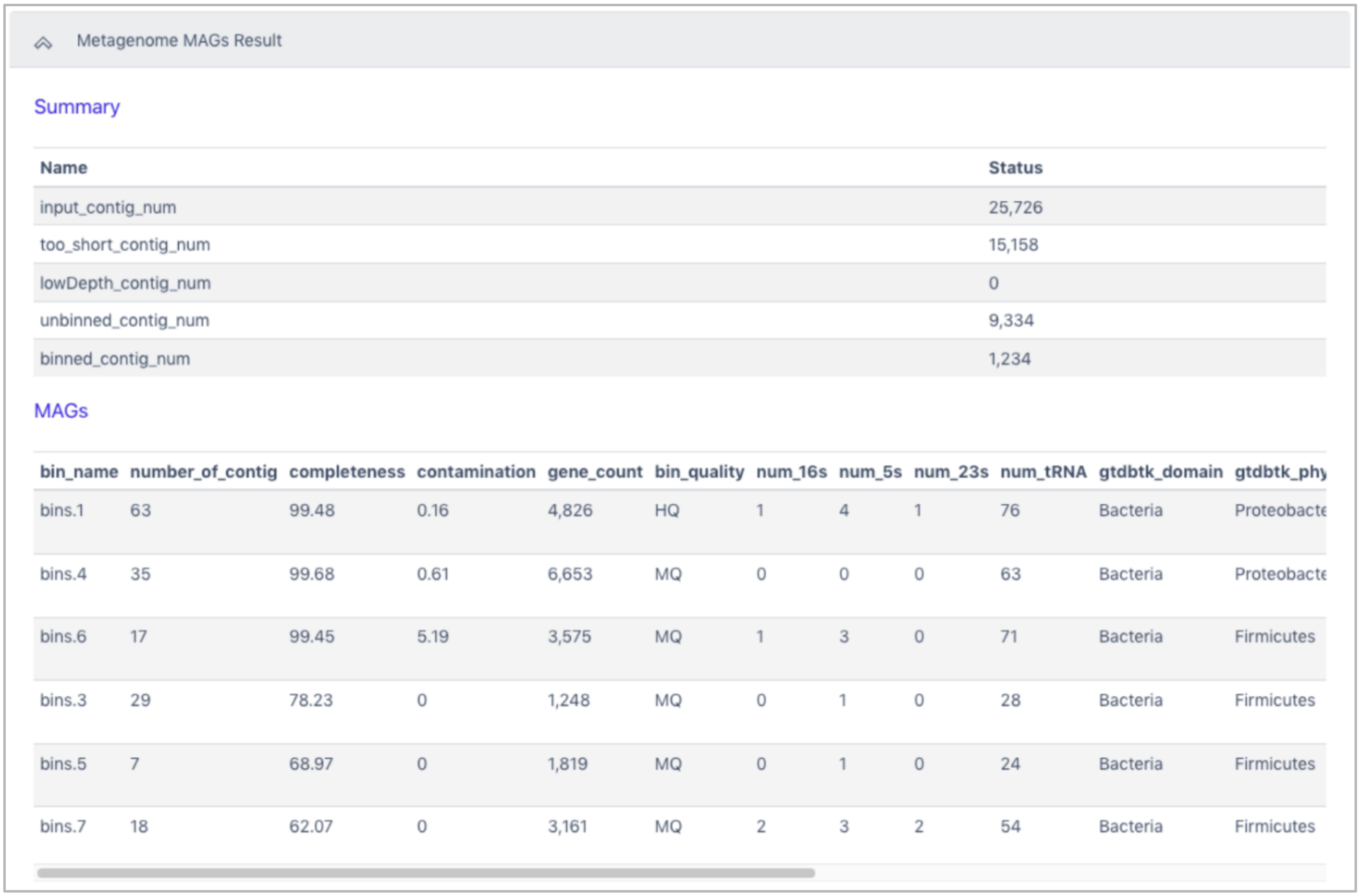


General

Workflow	Run	Status	Running Time	Start	End
Metagenome MAGs	On	Done	00:56:49	2021-10-18 16:51:51	2021-10-18 17:48:40

```
▸ "Project Configuration" : {...}
```

The Metagenome MAGs Result section provides a Summary section with information on binned and unbinned contigs. The MAGs section provides information such as the completeness of the genome, amount of contamination, and number of genes present on all bins determined to be high quality or medium quality.



Metagenome MAGs Result

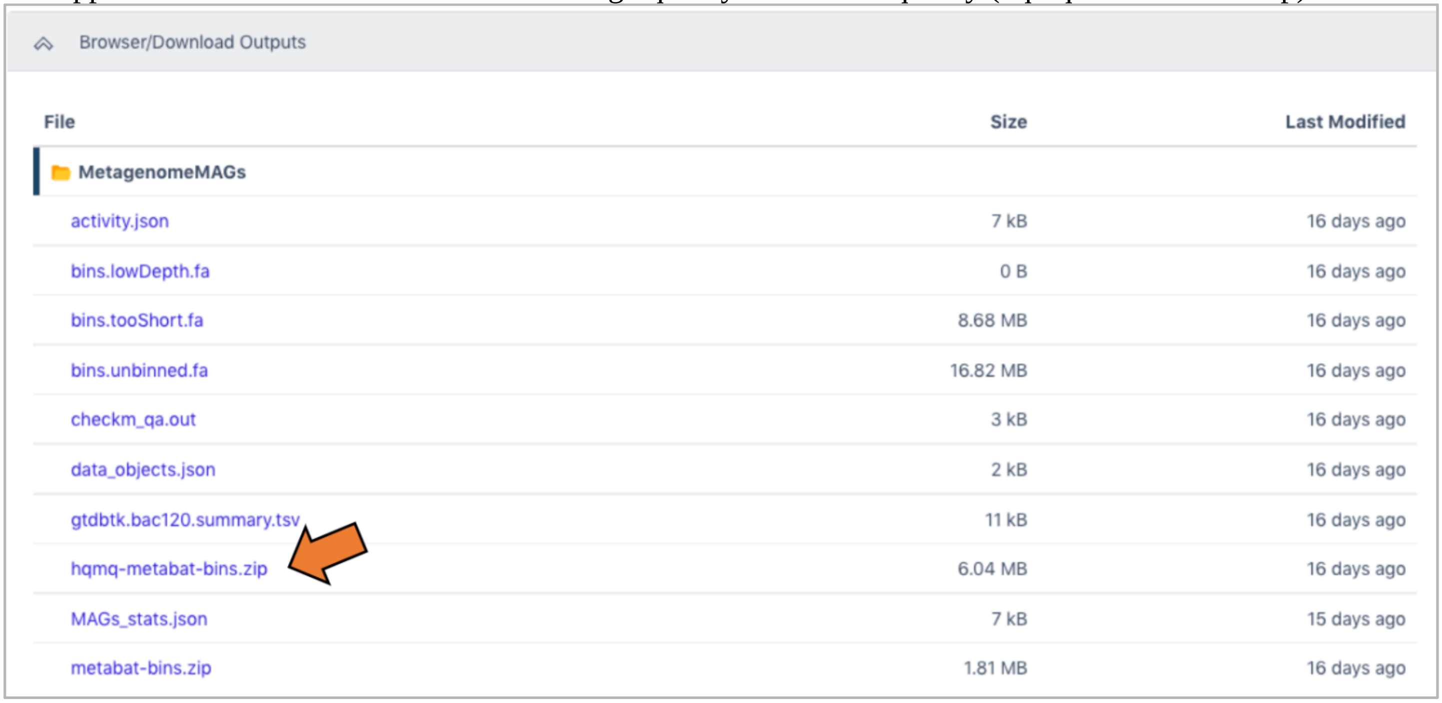
Summary

Name	Status
input_contig_num	25,726
too_short_contig_num	15,158
lowDepth_contig_num	0
unbinned_contig_num	9,334
binned_contig_num	1,234

MAGs

bin_name	number_of_contig	completeness	contamination	gene_count	bin_quality	num_16s	num_5s	num_23s	num_tRNA	gtdbtk_domain	gtdbtk_phy
bins.1	63	99.48	0.16	4,826	HQ	1	4	1	76	Bacteria	Proteobacte
bins.4	35	99.68	0.61	6,653	MQ	0	0	0	63	Bacteria	Proteobacte
bins.6	17	99.45	5.19	3,575	MQ	1	3	0	71	Bacteria	Firmicutes
bins.3	29	78.23	0	1,248	MQ	0	1	0	28	Bacteria	Firmicutes
bins.5	7	68.97	0	1,819	MQ	0	1	0	24	Bacteria	Firmicutes
bins.7	18	62.07	0	3,161	MQ	2	3	2	54	Bacteria	Firmicutes

The Browser/Download Output section provides output files available to download. The primary output file is the zipped file with all bins determined to be high quality or medium quality (hqmq-metabat-bins.zip).



Browser/Download Outputs

File	Size	Last Modified
MetagenomeMAGs		
activity.json	7 kB	16 days ago
bins.lowDepth.fa	0 B	16 days ago
bins.tooShort.fa	8.68 MB	16 days ago
bins.unbinned.fa	16.82 MB	16 days ago
checkm_qa.out	3 kB	16 days ago
data_objects.json	2 kB	16 days ago
gtdbtk.bac120.summary.tsv	11 kB	16 days ago
hqmq-metabat-bins.zip	6.04 MB	16 days ago
MAGs_stats.json	7 kB	15 days ago
metabat-bins.zip	1.81 MB	16 days ago