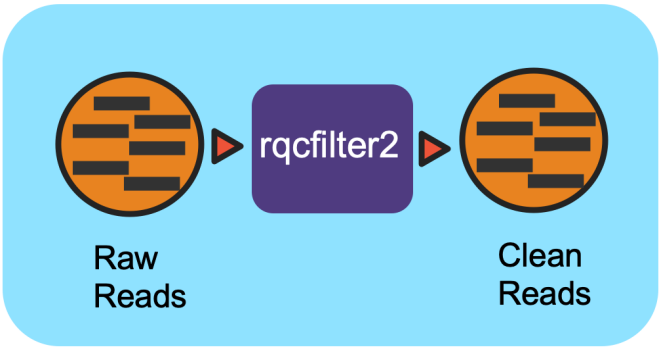
Reads QC Workflow (v1.0.1)



Overview

This workflow performs quality control on raw Illumina reads to trim/filter low quality data and to remove artifacts, linkers, adapters, spike-in reads and reads mapping to several hosts and common microbial contaminants.

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

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

Details

This workflow utilizes the program "rqcfilter2" from BBTools to perform quality control on raw Illumina reads. The workflow performs quality trimming, artifact removal, linker trimming, adapter trimming, and spike-in removal (using BBDuk), and performs human/cat/dog/mouse/microbe removal (using BBMap).

Software Versions

- rqcfilter2 (BBTools v38.94)
- bbduk (BBTools v38.94)
- bbmap (BBTools v38.94)

Output

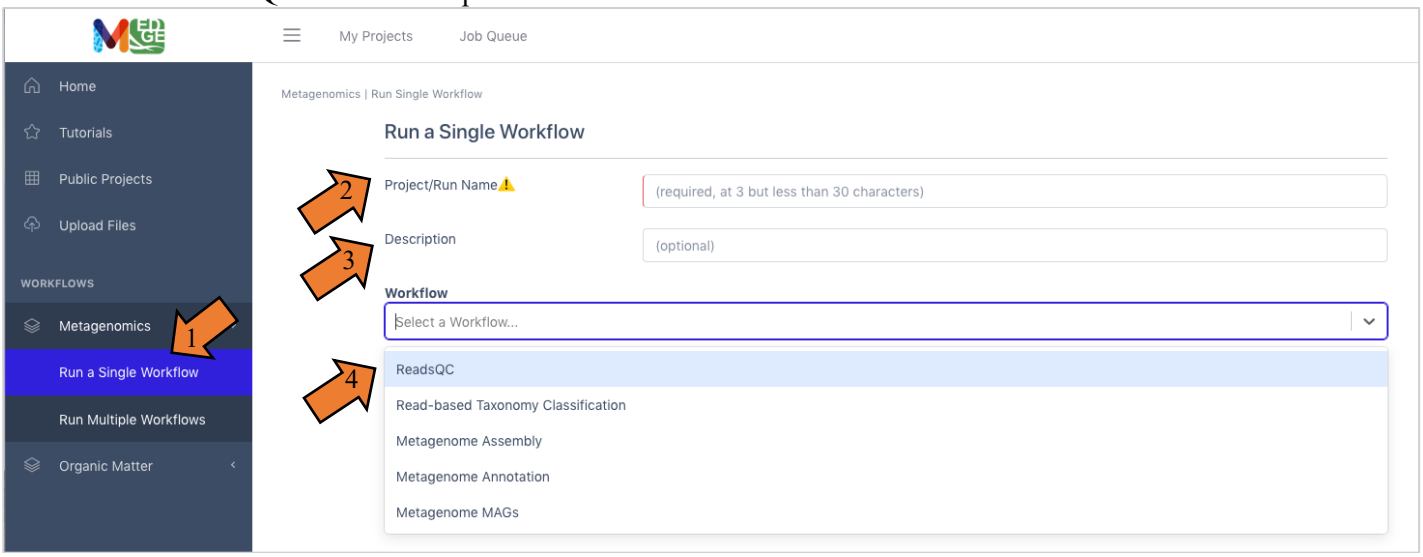
The main output is the cleaned data as an compressed interleaved fastq file (.fq.gz). There are also general statistics and more detailed statistics from the QC workflow in text files (.txt).

Primary Output Files	Description
Filtered Sequencing Reads	Cleaned paired-end data in interleaved format (.fastq.gz)
QC statistics (2 files)	Reads QC summary statistics (.txt)

Running the Reads QC Workflow in NMDC EDGE

Select a workflow

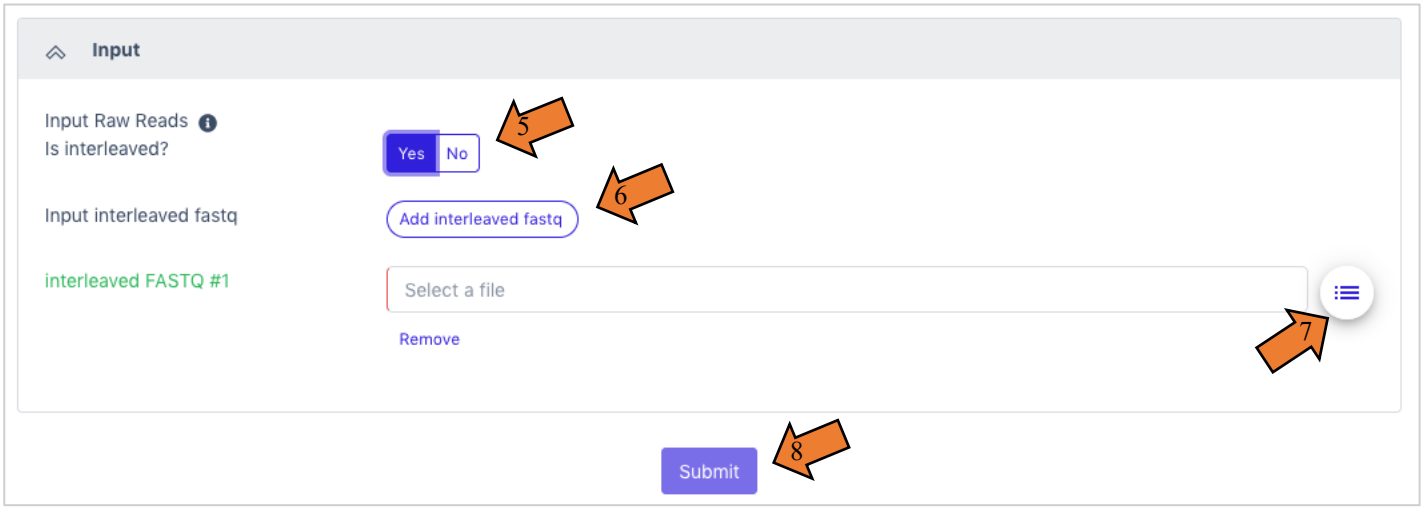
1. From the Metagenomics category in the left menu bar, select ‘Run a Single Workflow’.
2. Enter a ***unique*** project name with no spaces (underscores are fine).
3. A description is optional, but helpful.
4. Select ‘ReadsQC’ from the dropdown menu under Workflow.



Input

ReadsQC 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 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
ReadsQC	On	Done	03:03:03	2021-10-20 20:33:14	2021-10-20 23:36:17

"Project Configuration" : {...}

The ReadsQC Result section shows the data input and provides a variety of metrics including the number of reads and bases before and after trimming and filtering.

ReadsQC Result

Input: SRR7877884-int

Reads	Status
inputReads	44,943,418
inputBases	6,741,512,700
qtrimmedReads	8,583
qtrimmedBases	8,690
qfilteredReads	200,626
qfilteredBases	29,786,796
ktrimmedReads	6,186,690
ktrimmedBases	354,478,706
kfilteredReads	100,360
kfilteredBases	14,684,762
outputReads	33,510,668
outputBases	4,868,925,674
gcPolymerRatio	0.42

The Browser/Download Output section provides output files available to download. The clean data will be in an interleaved .fq.gz file. General QC statistics are in the filterStats.txt file.

Browser/Download Outputs

File	Size	Last Modified
ReadsQC		
SRR7877884-int		
filterStats2.txt	706 B	5 days ago
filterStats.json	337 B	5 days ago
filterStats.txt	287 B	5 days ago
SRR7877884-int.anqdpht.fq.gz	1713.58 MB	5 days ago