

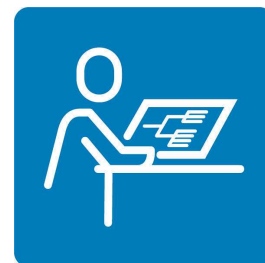
Jan 09, 2026

Version 8

Quality control assessment for microbial genomes: GalaxyTrakr MicroRunQC workflow V.8

DOI

<https://dx.doi.org/10.17504/protocols.io.5jyl8mj16g2w/v8>



Candace Bias¹, Ruth Timme¹, Yesha Shrestha², Tina Pfefer³, Paul Morin⁴, Maria Balkey³, Errol Strain³

¹US Food and Drug Administration; ²Center for Veterinary Medicine, US Food and Drug Administration;

³Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration, College Park, Maryland, USA;

⁴U.S. Food and Drug Administration, Jamaica, New York, USA

GenomeTrakr

Tech. support email: genomeTrakr@fda.hhs.gov



Ruth Timme

US Food and Drug Administration

Create & collaborate more with a free account

Edit and publish protocols, collaborate in communities, share insights through comments, and track progress with run records.

Create free account

OPEN  ACCESS



DOI: <https://dx.doi.org/10.17504/protocols.io.5jyl8mj16g2w/v8>



Protocol Citation: Candace Bias, Ruth Timme, Yesha Shrestha, Tina Pfefer, Paul Morin, Maria Balkey, Errol Strain 2026. Quality control assessment for microbial genomes: GalaxyTrakr MicroRunQC workflow. **protocols.io**
<https://dx.doi.org/10.17504/protocols.io.5jyl8mj16g2w/v8> Version created by **Maria Balkey**

License: This is an open access protocol distributed under the terms of the **Creative Commons Attribution License**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: Working

We use this protocol and it's working

Created: January 09, 2026

Last Modified: January 09, 2026

Protocol Integer ID: 238343

Keywords: WGS, Quality Control, GalaxyTrakr, GenomeTrakr, microbial pathogen surveillance, galaxytrakr microrunqc workflow, galaxytrakr microrunqc workflow purpose, quick access to genometrakr sequence quality threshold, wgs sequence quality for bacterial pathogen, quality control assessment for microbial genome, genometrakr sequence quality threshold, check against genometrakr qc, microrunqc workflow, microrunqc, genometrakr qc, microbial genome, checking wgs sequence quality, quality assessments for raw read, genometrakr, microbial pathogen, galaxytrakr, most microbial pathogen, bacterial pathogen, sequence type for each isolate, sequence type definition file, de novo assembly, available in sequence type definition file, end fastq file, pathogen, galaxytrakr account, raw read, cronobacter threshold, account in galaxytrakr, custom galaxy instance, mlst method, quality control assessment, assembly qc, added enterobacter qc, additional mlst data field, galaxytrakr upgrade, nextseq, entire miseq, miseq

Disclaimer

Please note that this protocol is public domain, which supersedes the CC-BY license default used by protocols.io.



Abstract

PURPOSE: Step-by-step instructions for checking WGS sequence quality for bacterial pathogens. The MicroRunQC workflow, implemented in a custom Galaxy instance, will produce quality assessments for raw reads (Illumina paired-end fastq files) and draft de novo assemblies, along with reporting the sequence type for each isolate. This workflow will work on most microbial pathogens, so we advise laboratories to upload their entire MiSeq/NextSeq run through this workflow.

SCOPE: This protocol covers the following tasks:

1. Quick access to GenomeTrakr sequence quality thresholds by organism
2. Create a GalaxyTrakr account
3. Set up an account in GalaxyTrakr
4. Create a new history/workspace
5. Upload data
6. Execute the MicroRunQC workflow
7. Interpret the results - check against GenomeTrakr QC thresholds

Version updates:

V7: Edits to incorporate GalaxyTrakr upgrades and new interface.

V6: Minor edits, including section reorganization and addition of clarifying notes

V5: New column in the output table to capture additional mlst data fields when available in Sequence Type definition files (not available for all species)

V4: MicroRunQC updated to V1.1 Includes updates to skeza and mlst methods, as well as adjusted assembly QC thresholds for E.coli. Added *Enterobacter* QC thresholds to threshold table.

V3: updated with *Cronobacter* thresholds

Quick Access to QC Benchmarks

- 1 This protocol will walk the user through various aspects of the quality assessment of bacterial genome sequences, from setting up a GalaxyTrakr account to the quality control (QC) benchmarks GenomeTrakr uses for its sequencing efforts. For quick access, GenomeTrakr QC benchmarks are included in the table below.

These are also relevant for NARMS and VetLIRN contributors.

*MicroRunQC users should follow QC threshold guidelines established by their respective surveillance coordinating body(s).

	A	B	C	D	E	F	G	H	I	J
Quality metric		<i>Salmonella</i>	<i>Listeria</i>	<i>E. coli</i>	<i>Shigella</i>	<i>Campylobacter</i>	<i>Vibrio para.</i>	<i>Cronobacter</i>	<i>Enterococcus faecium</i>	<i>Enterococcus faecalis</i>
Average read quality Q score for R1 and R2		>=30	>=30	>=30	>=30	>=30	>=30	>=30	>=30	>=30
Average coverage		>=30X	>=20X	>=40X	>=40X	>=20X	>=40X	>=20X	>=50X	>=40X
De novo assembly: Seq. length (Mbp)		~4.3-5.2	~2.7-3.2	~4.5-5.9	~4.0-5.0	~1.5-1.9	~4.8-5.5	~4-5	~2.5-3.5	~2.5-3.25
De novo assembly: no. contigs		<=300	<=300	<=400	<=550	<=300	<=300	<=500	<=350	<=200

Account set up

- 2 1. Create a GalaxyTrakr account here: <https://account.galaxytrakr.org/Account/Register>



Note

This is a more detailed form than what is available by clicking "Register here" on the genometrakr.org main page. Please use the form linked here when creating a new account.

User Registration Form

Location

California Department of Public Health - Food and Drug Laboratory Branch

[Add New Location](#)

First Name

Enter First Name, Do not use characters: \[] ; : | ' = + - * ? < > @ .

Last Name

Enter First Name, Do not use characters: \[] ; : | ' = + - * ? < > @ .

Email

Email will be used for automated messages to include registration information!

Primary Phone

Please enter number with country code, without dashes, for example +17035456789

If possible please use a mobile number than can accept text messages, only used for support

Title

Requirements

Please annotate intended use of Galaxy and Analysis tools. List specific tools you would like to see deployed in Galaxy.

Register

2.1 Log into your GalaxyTrakr account: <https://galaxytrakr.org>

GalaxyTrakr

Workflow Visualize Shared Data Help Login or Register

Using 0 bytes

Welcome to Galaxy, please log in

Public Name or Email Address

Password

Forgot password? Click here to reset your password.

Login

Don't have an account? Register here.

Welcome to GalaxyTrakr: open-source bioinformatics for public health.

This site is intended for use by GenomeTrakr laboratories and their collaborators to assist in the analysis of genomic data for foodborne pathogens. This instance of Galaxy is hosted in a public environment and no personally identifiable (PII) or commercial confidential information should be uploaded.

--!!--Information and Announcements--!!--

Please let us know if you have any issues with the new version of Galaxy. Thank you.

Access CFSAN SNP Pipeline workflows in the shared workflows screen.

Post in the official Galaxy GenomeTrakr board on the Redmine Site: Click here

Click here to access the Galaxy Trakr User Guide

Click here to access the Galaxy Trakr FAQ Document

Create a new history

3 Create a new history.

We recommend creating a new history for each new MiSeq Run and including the flow-cell ID and date in the history name.

Save your MicroRunQC output here and any other relevant analyses, like serotyping, or AMR detection.

After all the analysis output from this run is saved to your internal data network or computer, older histories should be purged/deleted so as not to occupy the limited storage space in your account. In some cases it may be useful to save, for a limited time, multiple histories or to run analyses concurrently in multiple histories. In these cases you need to pay attention to your % usage bar (shows % used of allocated storage space) in the upper right corner of the GalaxyTrakr page. If you need additional space you can contact galaxytrakrsupport@fda.hhs.gov and request additional storage.

3.1 Click on the + icon in the upper right History panel

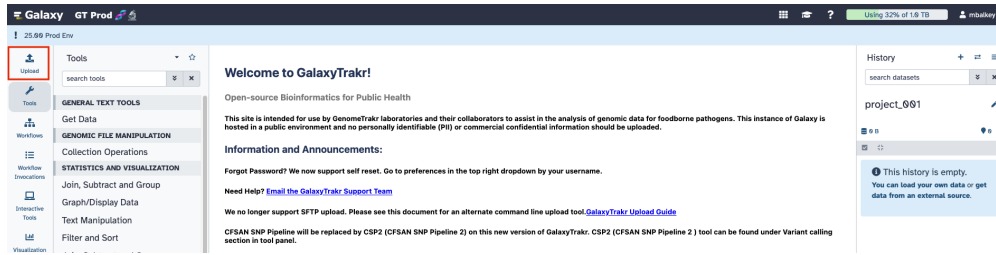


3.2 Name your new History by clicking on the "Unnamed history" text, type in desired name, and hit Enter. We recommend including the run cell ID and the date the run was started.

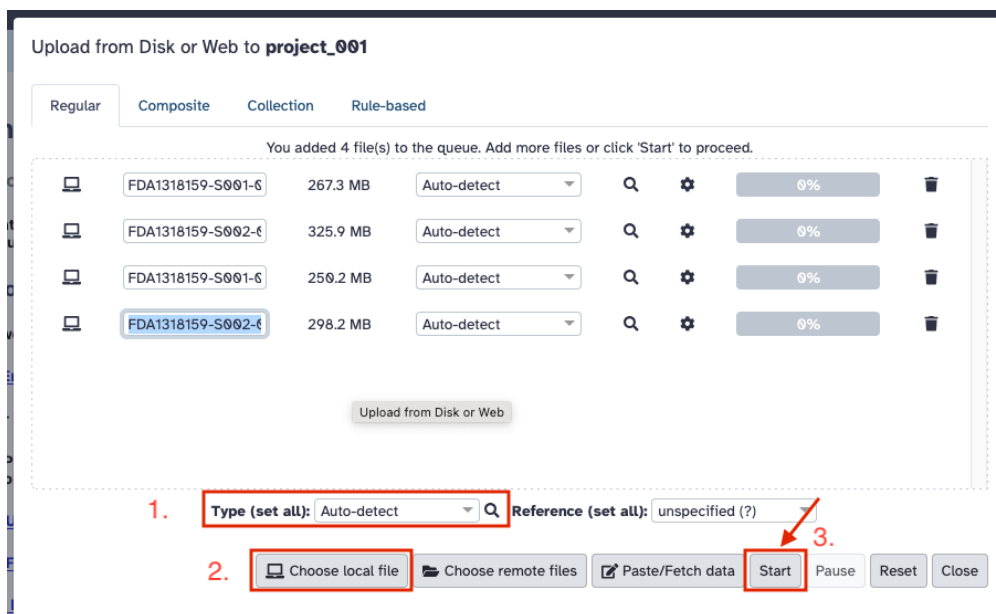


Upload data

- 4 This section will describe the process for uploading raw fastq files into your active History panel. After the files have been uploaded they will stay in your account until they are deleted.
- 4.1 Click on the Upload icon in the GalaxyTrkr menu to start an upload process.



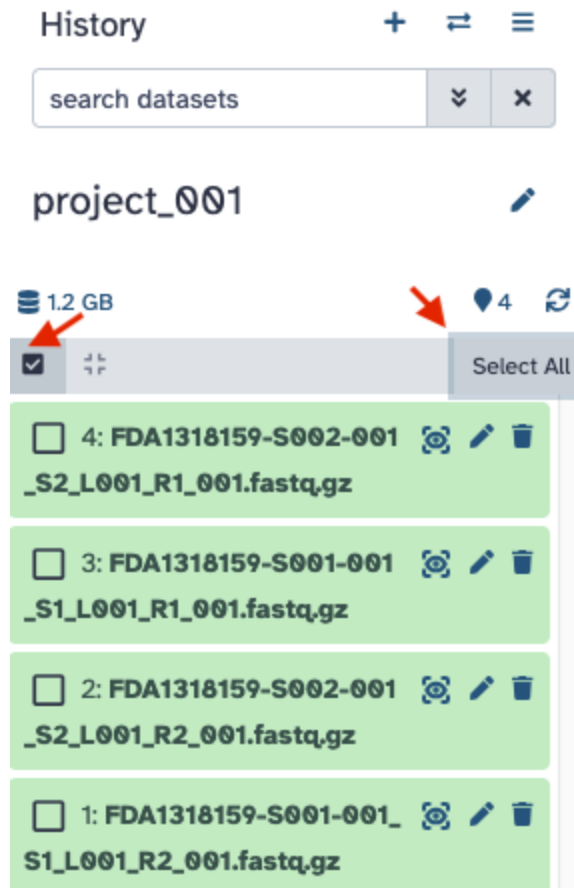
- 4.2 Select "**Type (set all):auto-detect.**" Click "**Choose local files**" button and navigate to the desired fastq files, then click "**Start**" to upload files. These files should be paired (two per sample/isolate).



- 4.3 As the file uploads complete, each row will turn green. Samples in yellow are still in process.

You have just upload a set of forward and reverse reads. For further analysis, these files need to be paired properly so the platform knows which R1 and R2 files go together. GalaxyTrakr does this by creating a List of Dataset Pairs.

Within your newly created History panel, click the check box, then select all the files you just uploaded by clicking "All" or by individually selecting the ones you want to pair.



Screenshot of History panel showing recently uploaded files. Note the way the files are named, using R1 and R2 to identify the paired reads. This will be important in the next step. Some naming conventions can be slightly different.

- 4.4 Click "Select All" and choose "Advance Build List", select "List of Paired Datasets".



1 What are you building? 2 Auto Pairing 3 Builder

Choose a list builder to match your selected data and the kind of list you'd like to build.

Flat List
This options creates a simple flat list of files. If your data isn't nested and does not contain paired datasets, this is the option to choose.

List of Paired Datasets
This options creates a uniform list of paired datasets, use this option if all of your data should be paired off.

Rule Builder
The rule builder allows building arbitrarily nested collections of data with reproducible metadata parsing and other advanced features.

Mixed List of Paired and Unpaired Datasets
For studies that have a mix of paired and unpaired data, this may be the right choice. Existing tools and workflows may have to be updated to allow this modality though.

Nested List of Datasets
A nested list of datasets. This type of collection is a list where each element is in turn a list of datasets. The outer list identifier might group samples together by condition, sample type, etc..

Nested List of Dataset Pairs
A nested list of datasets. This type of collection is a list where each element is in turn a list of pairs of datasets.

Next

History
search datasets
project_001
1.2 GB
All 4 selected
4: FDA131..._S2_L001_R1_...
3: FDA131..._S1_L001_R1_...
2: FDA131..._S2_L001_R2_...
1: FDA131..._S1_L001_R2_...
With 4 selected...
Hide
Delete
Delete (permanently)
Change Database/Build
Change data type
Add tags
Remove tags
Auto Build List
Advanced Build List

4.5 A new window will open to help you pair the fastq files properly. By default _R1 and _R2 are the selected options to pair fastq files. Note how your paired reads are named.

Click on the filter icon if “_R1” has to be replaced by other suffix.

1 What are you building? 2 Auto Pairing 3 Builder

Configure auto-pairing

Filters: _R1 _R2

Auto-matched 2 pair(s) of datasets from target datasets.

Auto-matched Pairs
These pairs of datasets have been auto-matched up using the filters above. They will be included as dataset pairs in the final list without additional action.
FDA1318159-S002-001_S2_L001_001 (4: FDA1318159-S002-001_S2_L001_R1_001.fastq.gz | 2: FDA1318159-S002-001_S2_L001_R2_001.fastq.gz), FDA1318159-S001-001_S1_L001_001 (3: FDA1318159-S001-001_S1_L001_R1_001.fastq.gz | 1: FDA1318159-S001-001_S1_L001_R2_001.fastq.gz)

No Un-matched Datasets
All datasets successfully auto-paired, this is generally a good sign Galaxy was able to correctly pair off all supplied datasets using their names.

Back Next

History
search datasets
project_001
1.2 GB
All 4 selected
4: FDA1318159-S002-001_S2_L001_R1_001.fastq.gz
3: FDA1318159-S001-001_S1_L001_R1_001.fastq.gz
2: FDA1318159-S002-001_S2_L001_R2_001.fastq.gz
1: FDA1318159-S001-001_S1_L001_R2_001.fastq.gz

Click **Next**, double check the name of the datasets. If the names are accurate, name the list of datasets and click **Build**.

What are you building? Auto Pairing Builder

Assemble, label, and sort your list of pairs.

This interface allows you to build a new Galaxy list of pairs. List of pairs are an ordered list of individual datasets paired together in their own paired collection (often forward and reverse reads).

Auto-matched 2 pair(s) of datasets from target datasets. If this isn't correct, [configure auto-pairing](#).

Dataset(s)	List Identifier	Discard	Status
FORWARD 4: FDA1318159-S002-001_S2_L00 REVERSE 2: FDA1318159-S002-001_S2_L00'	FDA1318159-S002-001_S2_L001_001	✗	✓
FORWARD 3: FDA1318159-S001-001_S1_L00 REVERSE 1: FDA1318159-S001-001_S1_L00'	FDA1318159-S001-001_S1_L001_001	✗	✓

Remove file extensions? Hide original elements

Name: project_001

Back Build

History

search datasets

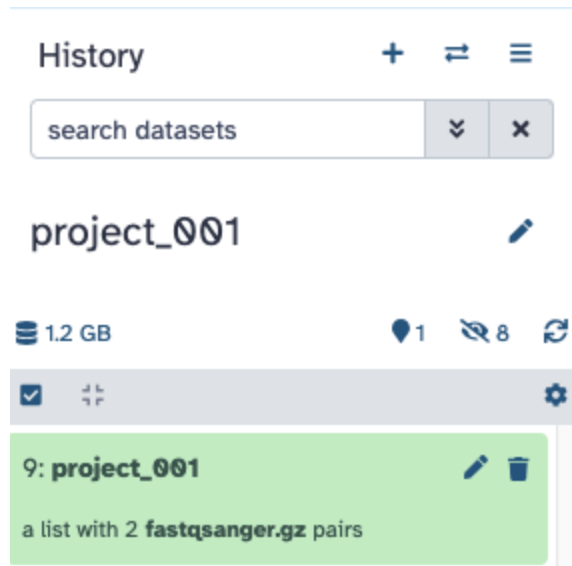
project_001

1.2 GB

All 4 selected

- 4: FDA1318159-S002-001_S2_L001_R1_001.fastq.gz
- 3: FDA1318159-S001-001_S1_L001_R1_001.fastq.gz
- 2: FDA1318159-S002-001_S2_L001_R2_001.fastq.gz
- 1: FDA1318159-S001-001_S1_L001_R2_001.fastq.gz

- 4.6 This paired dataset will now be available for analysis in your history panel. You can run multiple analyses on the same dataset in a history rather than upload the same sequence data to a new history to perform additional analyses. This will help you use your allocated storage space efficiently.



Run the MicroRunQC workflow

- 5 Add the **MicroRunQC** workflow to your own "**Workflows**" panel. You only have to do this step once for each new workflow you need.
- 5.1 Navigate to the "**Workflow**" menu, click "**Public workflows**", and search for "MicroRunQC_v1.2."



The screenshot shows the Galaxy GT Prod interface. The 'Workflows' tab is active, displaying a grid of public workflows. The 'Public workflows' section is highlighted with a red box. The 'Import' button is also highlighted with a red box. The 'Batch MicroRunQC' workflow is visible in the list.

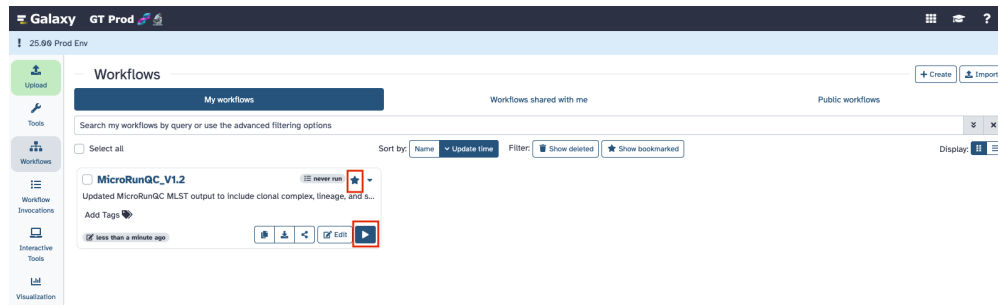
Click "Import" to select MicroRunQC.

The screenshot shows the Galaxy GT Prod interface. The 'Workflows' tab is active, displaying a grid of public workflows. The 'Batch MicroRunQC' workflow is highlighted with a red box. The 'Import' button is also highlighted with a red box. The 'MicroRunQC_v1.2 (prd25)' workflow is visible in the list.

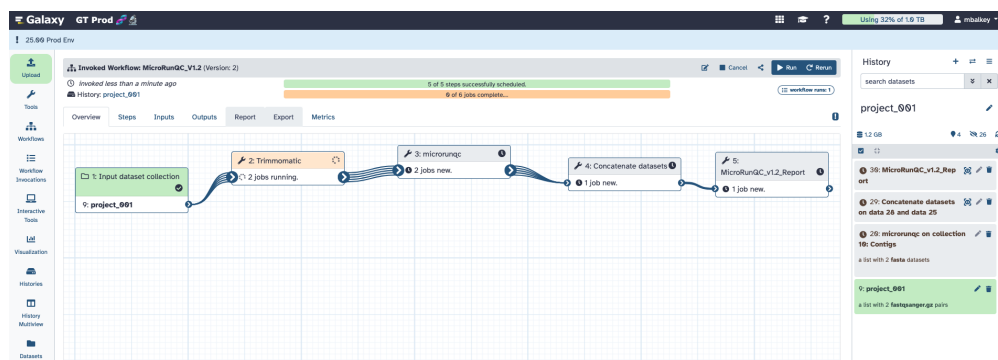
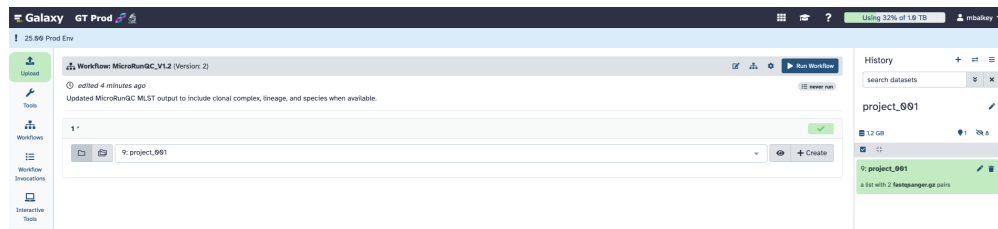
5.2 To see the new imported workflow, click the "Workflow" tab on the main menu, click "My Workflows."

Click the **star** to bookmark the **MicrorunQC_v1.2**

Click the play icon to run the **MicroRunQC_v1.2** and select the dataset collection that was created earlier.



5.3 Click **Run Workflow**. This can take some time depending on the number of samples you are analyzing. If you choose to you can log out of GalaxyTrakr and log back in at a later time to see if the job is completed.



5.4 Upon completion of the pipeline all tiles in the History pane will be green.

In the “**MicroRunQC_v1.2_Report**” tile, click on the “Eye” icon to view the output table in the GalaxyTrakr window.

Interpret the results

6 Download and interpret the results:

6.1 Click MicroRunQC_v1.2_Report and then the floppy disc icon. The tabular file can be opened in a text reader or converted to a format (.txt) that can be opened in Excel.

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9	Column 10	Column 11	Column 12	Column 13	Column 14	Column 15	Column 16	Column 17	Column 18	Column 19
File	Contigs	Length	EstCov	N50	MedianInsert	MeanLength_R1	MeanLength_R2	MeanQ_R1	MeanQ_R2	Schema	ST							
FGA1318109-S982-691_S12_1991_991.fasta	39	3492474	295.1	325318	152	121.8	122.9	36.6	35.6	isteria_2	191	CC+CC181Lineage=II	abc2(7)	bgk(15)	cat(75)	dap(8)	dat(6)	ish(14)
FGA1318109-S981-691_S11_1991_991.fasta	24	3886567	184.1	325318	299	146.4	146.4	37.1	36.4	isteria_2	191	CC+CC181Lineage=II	abc2(7)	bgk(15)	cat(75)	dap(8)	dat(6)	ish(14)

30: MicroRunQC_v1.2_Report
Add Tags
3 lines 20 columns
format **tabular**, database ?
Filtering with c1!='File',
kent 75.00% of 4 valid lines (4 total)
Download

1	2
File	Con'
FDA1318159-S002-001_S2_L001_001.fasta	30
FDA1318159-S001-001_S1_L001_001.fasta	24

20: microrunqc on collection 10:
Contigs
a list with 2 fasta datasets

9: project_001
a list with 2 fastqsanger.gz pairs

6.2 The MicroRunQC output file includes the following columns:

	A	B	C
	Parameter	Input	Description
	Contigs	Assembl y	Number of contigs in the de-novo SKESA assembly. Contigs smaller than 200 base-pairs (bp) are not counted.
	Length	Assembl y	Total length of all contigs > 200bp. This should approximate the size of the genome for the target organism.
	EstCov	Assembl y	Mean coverage for contigs in the SKESA assembly.



A	B	C
N50	Assembly	Sequence length of the shortest contig at 50% of the total genome length
MedianInsert	Read	Distance between forward and reverse reads. Calculated by mapping reads to SKESA assembly using bwa.
MeanLength_R1	Read	Mean length of forward read
MeanLength_R2	Read	Mean length of reverse read
MeanQ_R1	Read	Mean Q-score of forward read
MeanQ_R2	Read	Mean Q-score of reverse read
Scheme	Assembly	PubMLST scheme name (output from mlst application that scans contig files against traditional PubMLST typing schemes).
ST	Assembly	Sequence Type
MLST extra	Assembly	e.g. <i>Listeria</i> clonal complex info
Loci	Assembly	gene (allele number) – for example aroC(118)

MicroRunQC output table headers. This table lists the summary metrics for sequence quality, number of contigs, and estimated genome size, along with other common metrics for reads (Median Insert Size and Mean Length) and assemblies (N50). Additionally, if the Multi-Locus Sequence Type (MLST) for the isolate is available from pubmlst, the workflow also reports Sequence Type (ST) and the associated alleles.

***MLST extra:** Additional data fields reported when available in Sequence Type definition files (not available for all species)

1. clonal_complex – sequences grouped by similarity to central allelic profile (e.g., *Campylobacter* ST-21 complex)
2. CC – clonal_complex – Abbreviation used for organism like *Listeria*, ST profiles are maintained by different groups
3. Lineage – *Listeria monocytogenes* lineage (I,II,III, and IV), *Listeria* species also reported here (e.g. *L. innocua*)
4. species – e.g., *Vibrio alginolyticus*

****This output should be saved either to your LIMS or to a spreadsheet linked to the sequencing run and samples.**

6.3 Example output for 1 *Salmonella* and 5 *Listeria* isolates.

A	B
Srain ID	Lab Confirmation
FDA1216271-C001-001	Listeria mono
FDA817806-S073-001	Listeria mono
FDA746634	Listeria mono
FDA1213377-C001-002	Listeria grayi
FDA933376-S060-005	Listeria innocua
FDA1213835-C001-001	Salmonella

Lab confirmed IDs for 6 isolates

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
	File	Contigs	Length	EstCov	N50	MedianInsert	MeanLength R1	MeanLength R2	MeanQ R1	MeanQ R2	Scheme	ST	MLSTextra							
	FDA1216271-C001-001	16	2911949	36.7	476210	321	148.4	148.4	36.4	34.6	listeria_2	5	CC=CC5;Lineage=I	abcZ(2)	bgIA(1)	cat(11)	dapE(3)	dat(3)	ldh(1)	lhkA(7)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
	FDA817806-S073-001	20	3068354	179.6	525438	329	234.7	235.2	36.7	31.9	listeria ₂	321	CCC=CC3 ²¹ Lin eage=I	abcZ(5)	bgIA(6)	cat(8)	dapE(62)	dat(6)	ldh(7)	lhkA(34)
	FDA746634	30	3052888	41.4	293947	320	148.4	148.4	36.5	36	listeria ₂	-		abcZ(2)	bgIA(1)	cat(11)	dapE(3)	dat(3)	ldh(1)	lhkA(~7)
	FDA1213377-C001-002	20	2672180	155.1	473181	270	147.3	147.3	37.2	36.1	-	-								
	FDA933376-S	9	2881869	213	1498790	303	232.1	232.2	37	36.2	listeria ₂	1489	CCC=CC1489Lin	abcZ(250)	bgIA(21)	cat(83)	dapE(298)	dat(20)	ldh(458)	lhkA(216)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
	060-005												e a g e = L. i n n o c u a							
	F D A 1 2 1 3 8 3 5 - C 0 0 1 - 0 0 1	3 7	4 8 3 2 3 6 5	3 4 . 4	2 9 4 9 3 6	3 5 4	1 4 9	1 4 9	3 6. 6	3 5. 7	s e n t e r i c a - a c h t m a n 2	2 1 4		a r o C (1 4)	d n a N (7 2)	h e m D (2 1)	h i s D (1 2)	p u r E (6)	s u c A (1 9)	t h r A (1 5)

MicroRunQC example report showing mlst ST results for different *Listeria* species.

The mlst *Listeria* database includes multiple species, including *Listeria monocytogenes* and *L. innocua*. When available, the *Listeria* clonal complex (CC) or *L. monocytogenes* lineage is listed alongside the ST.

6.4 For quality control threshold guidelines for the GenomeTrakr surveillance network,

[go to step #1](#) These are also relevant for NARMS and VetLIRN contributors.

*MicroRunQC users should follow QC threshold guidelines established by their respective surveillance coordinating body(s).