

VERSION 6

FEB 26, 2024

OPEN  ACCESS**DOI:**

[dx.doi.org/10.17504/protocols.io.  
5jyl8mj16g2w/v6](https://dx.doi.org/10.17504/protocols.io.5jyl8mj16g2w/v6)

**External link:**

<https://galaxytrakr.org>

**Protocol Citation:** Candace Hope Bias, Ruth Timme, Yesha Shrestha, Tina Lusk Pfefer, Paul Morin, Maria Balkey, Errol Strain 2024. Quality control assessment for microbial genomes: GalaxyTrakr MicroRunQC workflow. **protocols.io** <https://dx.doi.org/10.17504/protocols.io.5jyl8mj16g2w/v6> Version created by [Candace Hope Bias](#)

## Quality control assessment for microbial genomes: GalaxyTrakr MicroRunQC workflow V.6

Candace Hope Bias<sup>1</sup>, Ruth Timme<sup>1</sup>, Yesha Shrestha<sup>2</sup>, Tina Lusk Pfefer<sup>3</sup>,  
Paul Morin<sup>4</sup>, Maria Balkey<sup>3</sup>, Errol Strain<sup>3</sup>

<sup>1</sup>US Food and Drug Administration;

<sup>2</sup>Center for Veterinary Medicine, US Food and Drug Administration;

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

<sup>4</sup>U.S. Food and Drug Administration, Jamaica, New York, USA

GenomeTrakr

Tech. support email: [genomeTrakr@fda.hhs.gov](mailto:genomeTrakr@fda.hhs.gov)



Ruth Timme

US Food and Drug Administration

### DISCLAIMER

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

**MANUSCRIPT CITATION:**

Timme, R. E., W. J. Wolfgang, M. Balkey, S. L. G. Venkata, R. Randolph, M. Allard, and E. Strain. 2020. Optimizing open data to support one health: best practices to ensure interoperability of genomic data from bacterial pathogens. One Health Outlook 2: 20.

**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:** Jan 05, 2024

**Last Modified:** Feb 26, 2024

**PROTOCOL integer ID:** 92989

**Keywords:** WGS, Quality Control, GalaxyTrakr, GenomeTrakr, microbial pathogen surveillance

**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:**

- 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
<i>De novo</i> 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
<i>De novo</i> 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>

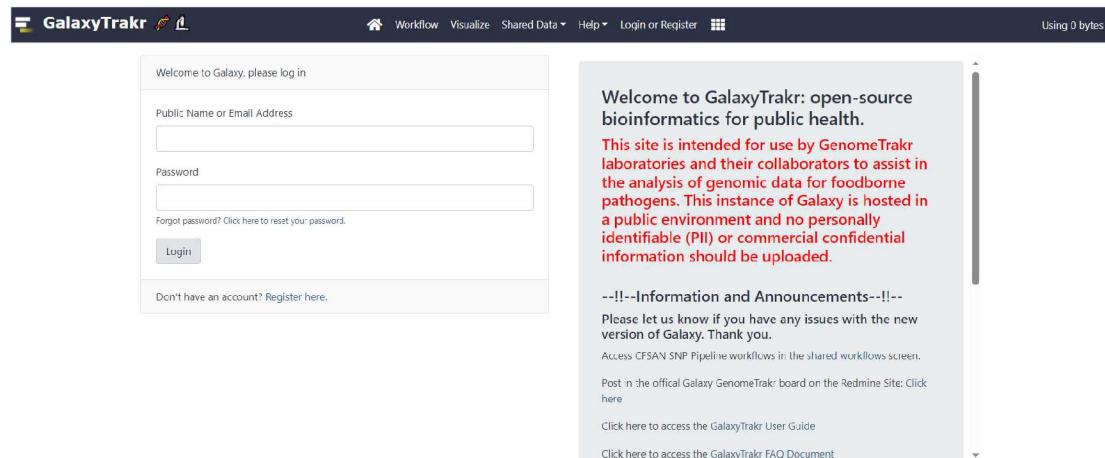
### Note

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

<b>Location</b> <input type="text" value="California Department of Public Health - Food and Drug Laboratory Branch"/> <a href="#">Add New Location</a>	<b>First Name</b> <input type="text" value="Enter First Name, Do not use characters: \[\]; =+-*?&lt;&gt;@."/>
<b>Last Name</b> <input type="text" value="Enter First Name, Do not use characters: \[\]; =+-*?&lt;&gt;@."/>	<b>Email</b> <input type="text"/> <small>Email will be used for automated messages to include registration information!</small>
<b>Primary Phone</b> <input type="text" value="Please enter number with country code, without dashes, for example +17035456789"/> <small>If possible please use a mobile number than can accept text messages, only used for support</small>	<b>Title</b> <input type="text"/>
<b>Requirements</b> <small>Please annotate intended use of Galaxy and Analysis tools. List specific tools you would like to see deployed in Galaxy.</small>	<input type="button" value="Register"/>

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



## Create a new history

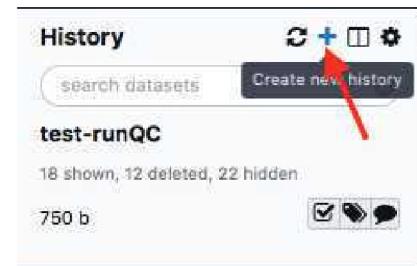
### 3 Create a new history.

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

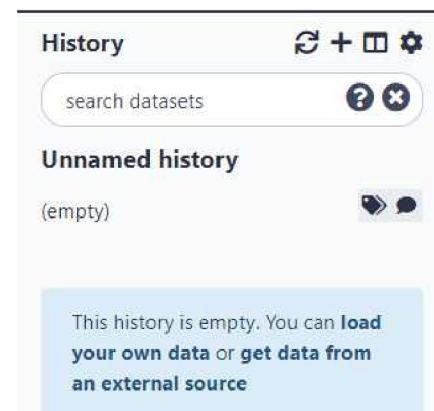
Save your MicroRunQC output here and any other relevant analyses, like serotyping, or AMR detection.

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](mailto: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 Data icon on the top of the left web page 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).

Download from web or upload from disk

Regular    Composite    Collection    Rule-based

You added 4 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status
CFSAN074382_S15_L	151.8 MB	Auto-det...	unspecified (?)	⚙️	0%
CFSAN074382_S15_R	152.5 MB	Auto-det...	unspecified (?)	⚙️	0%
CFSAN074384_S20_L	172.6 MB	Auto-det...	unspecified (?)	⚙️	0%
CFSAN074384_S20_R	181.2 MB	Auto-det...	unspecified (?)	⚙️	0%

1. Type (set all): Auto-detect    2. Choose local file    3. Start

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

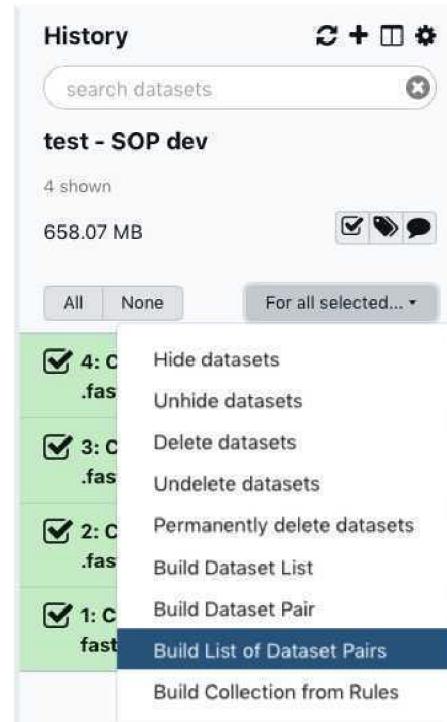
- 4.3** You have just upload a set of forward and reverse reads. For further analysis, these files need to be paired properly so the platform knows which R1 and R2 files go together. GalaxyTrakr does this by creating a **List of Dataset Pairs**.

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.



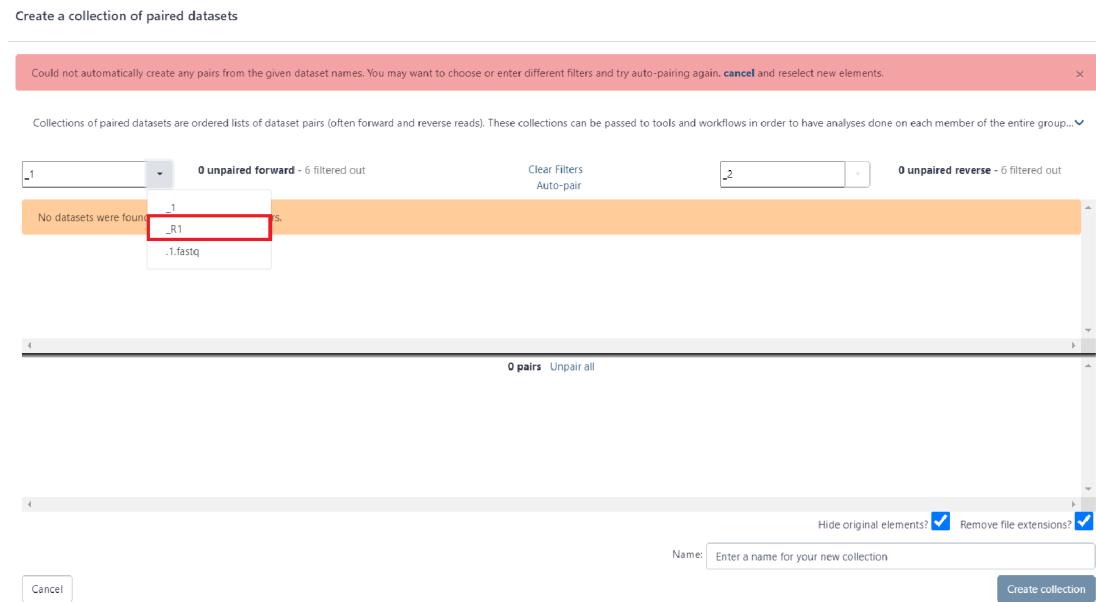
**Screenshot of History panel showing recently uploaded files.** Note the way the files are named, using R1 and R2 to identify the paired reads. This will be important in the next step. Some naming conventions can be slightly different.

#### 4.4 Click "For all selected" and choose "Build List of Dataset Pairs"

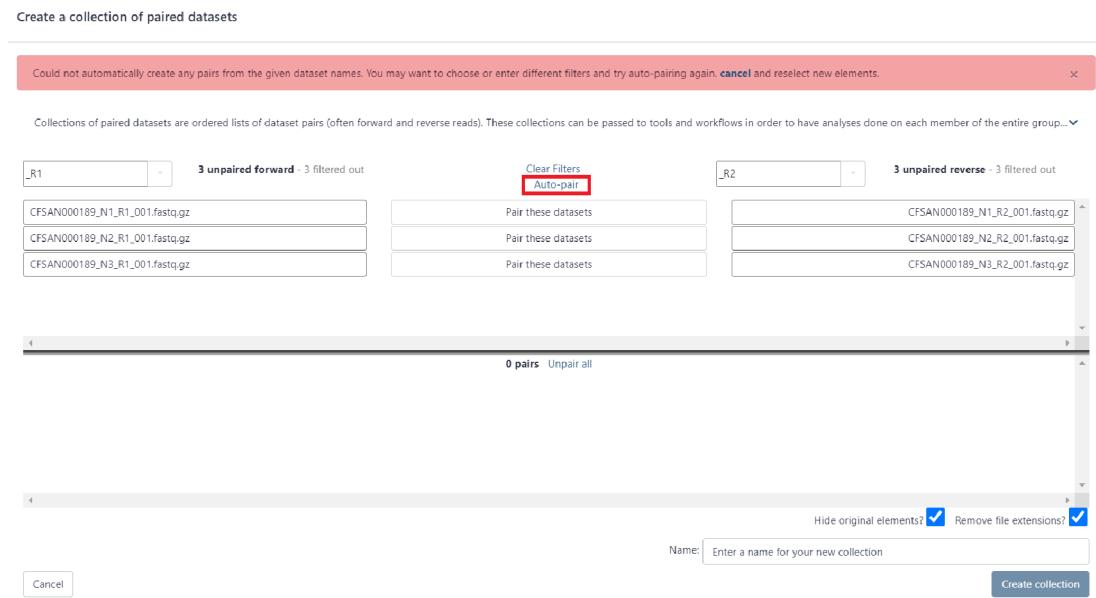


## 4.5 A new window will open to help you pair the fastq files properly. Note how your paired reads are named.

First, click on the drop down arrow and choose “\_R1,” if that is the naming convention your files follow. This automatically populates the corresponding “\_R2” in the next box.



**Click Auto-pair.**



Paired reads will pair in the middle column and turn green.

Unselect "**Hide original elements**" which is the default setting and undesired here.

Create a collection of paired datasets

Could not automatically create any pairs from the given dataset names. You may want to choose or enter different filters and try auto-pairing again. [cancel](#) and reselect new elements. [X](#)

Collections of paired datasets are ordered lists of dataset pairs (often forward and reverse reads). These collections can be passed to tools and workflows in order to have analyses done on each member of the entire group...[▼](#)

\_R1      **0 unpaired forward - 0 filtered out**      [Clear Filters](#) [Auto-pair](#)      \_R2      **0 unpaired reverse - 0 filtered out**

No datasets were found matching the current filters.

3 pairs		Unpair all
CFSAN000189_N1_R1_001.fastq.gz ➔	CFSAN000189_N1_001.fastq	◀ CFSAN000189_N1_R2_001.fastq.gz
CFSAN000189_N2_R1_001.fastq.gz ➔	CFSAN000189_N2_001.fastq	◀ CFSAN000189_N2_R2_001.fastq.gz
CFSAN000189_N3_R1_001.fastq.gz ➔	CFSAN000189_N3_001.fastq	◀ CFSAN000189_N3_R2_001.fastq.gz

Hide original elements?  Remove file extensions? [▼](#)

Name:  Enter a name for your new collection [Create collection](#)

[Cancel](#)

Clear the checkmark by "Hide original elements?" seen here by clicking on it.

Name your dataset: Example, "pairedSet-<FlowCell>-<date>"

Click **Create list**.

- 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.

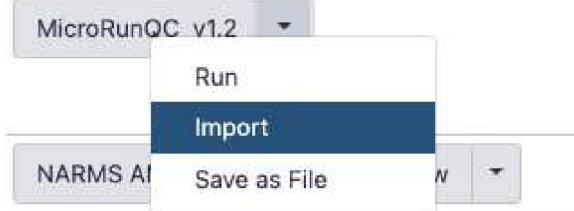
## Run the MicroRunQC workflow

**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 “**Shared Data**” drop down menu, choose “**Workflows**,” and search for “MicroRunQC\_v1.2.”

Name	Annotatio	Owner	Community Rating	Community Tags	Last Updated
MicroRunQC_v1.2		estrain	★★★★★		6 days ago

From the dropdown menu by the MicroRunQC title, select “**Import**.”



## 5.2 To see the new imported workflow, click the “Workflow” tab on the top panel.

Click the box under "Bookmarked" to make it available in the left panel under "Workflows" when MicroRunQC is searched for.

A screenshot of the "Workflow" tab on the top navigation bar. Below the header, there is a search bar labeled "Search Workflows" and two buttons: "+ Create" and "Import". The main area displays a list of workflows. The first workflow in the list is "imported: MicroRunQC\_v1.2", which has a red box drawn around the "Workflow" tab in the header. The list includes columns for Name, Tags, Updated, Sharing, and Bookmarked. The "imported: MicroRunQC\_v1.2" entry has a checked box in the "Bookmarked" column and a play button icon. The other two workflows listed are "imported: MicroRunQC\_v1.1" and "imported: SeqSero2 v1.1.1 collection workflow", both of which have unchecked boxes in the "Bookmarked" column and play button icons.

## 5.3 From the Workflow menu on the left panel, select **MicroRunQC\_v1.2**.

**GalaxyTrakr**

Tools ☆ ⚙

search tools ×

Upload Data

**Metagenomics:Kraken**

**Metagenomics:Mitokmer**

**Metagenomics:Graphlan**

**Metagenomics:Functional Profiling**

**Metagenomics:Assembly**

**Metagenomics:Rpackages**

**Metagenomics:Metaphlan**

**Metagenomics:CPIES**

**test:tools**

**tes**

**blast\_to\_scaffold** Generate DNA scaffold from blastn or tblastx alignment of Contigs

**small\_rna\_maps**

**Clip adapter**

**Normalize By Median** Filter reads using digital normalization via k-mer abundances

**Bowtie2** - map reads against reference genome

**Trim sequences**

**NCBI EFetch** fetch records from NCBI

**Unique** occurrences of each record

**QIIME**

**kraken2**

**QualiMap BamQC** Tool to facilitate the quality control of alignment sequencing data and its derivatives like feature counts.

**NGS:Simulator**

**WORKFLOWS**

All workflows

imported: MicroRunQC\_v1.2

## 5.4 Select paired list dataset you created earlier.

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.

### Note

If you see an orange box at the top of the workflow before running it that states that the tool is out of date or contains errors, please go back to **Step 4.1** of this protocol and import the version of the tool with the most current "Last Updated" date. The version number on the MicroRunQC workflow does not increment when only the MLST database is updated, meaning that though the version number of the workflow may remain the same, re-import may be necessary for the most up-to-date analysis of your genomes.

Some tools in this workflow may have changed since it was last saved or some errors were found. The workflow may still run, but any new options will have default values. Please review the messages below to make a decision about whether the changes will affect your analysis.

Workflow: imported: MicroRunQC\_v1.2 Run Workflow

History Options  
Send results to a new history  No

1: input dataset collection  
45: Nevada\_Clearlabs\_Validation

2: Trimmomatic (Galaxy Version 0.36.4)

3: micorunqc (Galaxy Version 1.0.1)

4: Concatenate multiple datasets (Galaxy Version 0.3)

5: Filter (Galaxy Version 1.1.1)

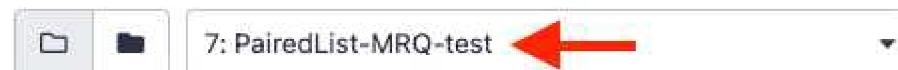
Example of warning box



## Workflow: imported: MicroRunQC\_v1.2

✓ Run Workflow

1



Expand to full workflow form.

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

In the “**Filter on Data ##**” tile, click on the “Eye” icon to view the output table in the GalaxyTrakr window.

Successfully invoked workflow imported: MicroRunQC\_v1.2.

You can check the status of queued jobs and view the resulting data by refreshing the History pane, if this has not already happened automatically.

View Report 1

5 of 5 steps successfully scheduled.  
6 of 6 jobs complete.

Download BioCompute Object

▶ Inputs  
▶ Output Collections  
▶ Steps

History

test - SOP dev

14 shown, 4 deleted, 39 hidden

1.78 GB

71: Filter on data 70

63: microurunqc on collection 5

1: MLST

a list with 2 items

62: microurunqc on collection 5

1: MI ST

## Interpret the results

### 6 Download and interpret the results:

- 6.1 Click **Filter on data ##** 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.

- 6.2 The MicroRunQC output file includes the following columns:

A	B	C
<b>Parameter</b>	<b>Input</b>	<b>Description</b>
<b>Contigs</b>	Assembly	Number of contigs in the de-novo SKESA assembly. Contigs smaller than 200 base-pairs (bp) are not counted.
<b>Length</b>	Assembly	Total length of all contigs > 200bp. This should approximate the size of the genome for the target organism.
<b>EstCov</b>	Assembly	Mean coverage for contigs in the SKESA assembly.
<b>N50</b>	Assembly	Sequence length of the shortest contig at 50% of the total genome length
<b>MedianInsert</b>	Read	Distance between forward and reverse reads. Calculated by mapping reads to SKESA assembly using bwa.
<b>MeanLength_R1</b>	Read	Mean length of forward read
<b>MeanLength_R2</b>	Read	Mean length of reverse read
<b>MeanQ_R1</b>	Read	Mean Q-score of forward read
<b>MeanQ_R2</b>	Read	Mean Q-score of reverse read
<b>Scheme</b>	Assembly	PubMLST scheme name (output from mlst application that scans contig files against traditional PubMLST typing schemes).
<b>ST</b>	Assembly	Sequence Type
MLST extra	Assembly	e.g. Listeria clonal complex info

A	B	C
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
<b>Strain ID</b>	<b>Lab Confirmation</b>
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	Media	Mean Length R1	Mean Length R2	Mean Q1	Mean Q2	Scheme	ST	MLST extra							

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
FDA 121 627 1-C00 1-001	16	29 11 94 9	3 6. 7	47 62 10	32 1	14 8.4	14 8.4	.36 .4	.34 .6	list eria _2	5	CC=C C5,Lin eage= I	ab cZ (2)	bg IA( 1)	cat (1 1)	da pE (3)	d at (3 )	Id h( 1)	lhk A(7 )	
FDA 817 806-S07 3-001	20	30 68 35 4	1 7 9. 6	52 54 38	32 9	23 4.7	23 5.2	.36 .7	.31 .9	list eria _2	3 2 1	CC=C C321, Linea ge=II	ab cZ (5)	bg IA( 6)	cat (8)	da pE (6 2)	d at (6 )	Id h( 7)	lhk A(3 4)	
FDA 746 634	30	30 52 88 8	4 1. 4	29 39 47	32 0	14 8.4	14 8.4	.36 .5	.36	list eria _2	-		ab cZ (2)	bg IA( 1)	cat (1 1)	da pE (3)	d at (3 )	Id h( 1)	lhk A( ~7)	
FDA 121 337 7-C00 1-002	20	26 72 18 0	1 5 5. 1	47 31 81	27 0	14 7.3	14 7.3	.37 .2	.36 .1	-	-									
FDA 933 376-S06 0-005	9	28 81 86 9	2 1 3	14 98 79 0	30 3	23 2.1	23 2.2	37	.36 .2	list eria _2	1 4 8 9	CC=C C148 9,Line age=L : innoc ua	ab cZ (2 50 )	bg IA( 21)	cat (8 3)	da pE (2 98 )	d at (2 0)	Id h( 45 8)	lhk A(2 16)	
FDA 121 383 5-C00 1-001	37	48 32 36 5	3 4. 4	29 49 36	35 4	14 9	14 9	.36 .6	.35 .7	sen teri ca_ ach tm an_ 2	2 1 4		ar oC (1 4)	dn a N( 72 )	he m D( 21)	hi sD (1 2)	p ur E( 6)	su cA (1 9)	thr 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).

