

VERSION 5

JUN 14, 2023

OPEN  ACCESS**DOI:**dx.doi.org/10.17504/protocols.io.5jyl8mj16g2w/v5**External link:**<https://galaxytracr.org>

Protocol Citation: Ruth Timme, Yesha Shrestha, Tina.Pfefer, Paul Morin, Maria Balkey, Errol Strain 2023. Quality control assessment for microbial genomes: GalaxyTracr MicroRunQC workflow. **protocols.io** <https://dx.doi.org/10.17504/protocols.io.5jyl8mj16g2w/v5> Version created by Ruth Timme

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.

Quality control assessment for microbial genomes: GalaxyTracr MicroRunQC workflow V.5

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

¹US Food and Drug Administration ;

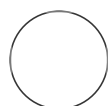
²Center for Veterinary Medicine, US Food and Drug Administration ;

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

GenomeTracr

Tech. support email: genomeTracr@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.

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: Jun 13, 2023

Last Modified: Jun 14, 2023

PROTOCOL integer ID:
83341

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. set up an account in GalaxyTrakr
2. Create a new history/workspace
3. Upload data
4. Execute the MicroRunQC workflow
5. Interpret the results - check against GenomeTrakr QC thresholds

Version updates:

V3: updated with *Cronobacter* thresholds

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.

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)

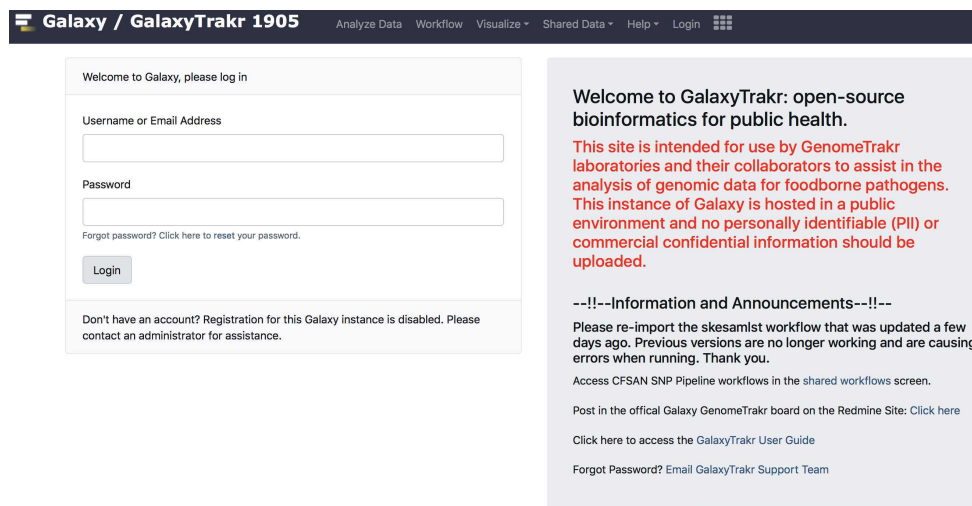
Account set up

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

User Registration Form

Location	<input type="text" value="California Department of Public Health - Food and Drug Laboratory Branch"/> <small>Add New Location</small>
First Name	<input "="" +-?*<>@."="" type="text" value="Enter First Name, Do not use characters: ^[]{}:;'"/>
Last Name	<input "="" +-?*<>@."="" type="text" value="Enter First Name, Do not use characters: ^[]{}:;'"/>
Email	<input type="text"/> <small>Email will be used for automated messages to include registration information!</small>
Primary Phone	<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>
Title	<input type="text"/>
Requirements	<input type="text" value="Please annotate intended use of Galaxy and Analysis tools. List specific tools you would like to see deployed in Galaxy."/> <small>Register</small>

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



Create a new history

2 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 history's 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.

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



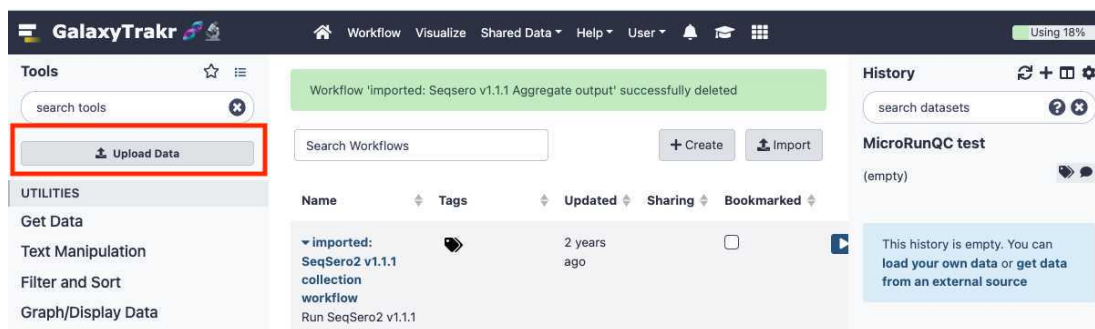
- 2.2 Name your new History by clicking on the “Unnamed history”, type in desired name and hit enter. We recommend including the run cell ID and the date the run was started.



Upload data

- 3 **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.

- 3.1 Click on the Upload Data icon on the top of the left web page to start an upload process.



3.2 Select "Type (set all):auto-detect." Choose local file 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_L	152.5 MB	Auto-det...	unspecified (?)		0%
CFSAN074384_S20_L	172.6 MB	Auto-det...	unspecified (?)		0%
CFSAN074384_S20_L	181.2 MB	Auto-det...	unspecified (?)		0%

1. Type (set all): Auto-detect Genome (set all): unspecified (?)

2. Choose local file Choose FTP file Paste/Fetch data Pause Reset Start Close

3. Start

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

3.3 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 with each sample/isolate. 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.

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



3.5 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 downarrow and choose “_R1”. This automatically chooses “_R2” in the next box.

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.

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. ...

_1 0 unpaired forward - 2 filtered out Clear Filters Auto-pair _2 0 unpaired reverse - 2 filtered out

No datasets were found for _1. .1.fastq

0 pairs Unpair all

Hide original elements? ☒ Remove file extensions? ☒

Name: Enter a name for your new collection

Cancel Create collection

Click **Auto-pair**.

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.

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 1 unpaired forward - 1 filtered out Clear Filters Auto-pair _R2 1 unpaired reverse - 1 filtered out

FDA1142650-C002-001_S3_L001_R1_001.fastq.gz Pair these datasets FDA1142650-C002-001_S3_L001_R2_001.fastq.gz

0 pairs Unpair all

Hide original elements? ☒ Remove file extensions? ☒

Name: Enter a name for your new collection

Cancel Create collection

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

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. ✕

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. ... ▼

0 unpaired forward - 0 filtered out Clear Filters 0 unpaired reverse - 0 filtered out
Auto-pair

No datasets were found matching the current filters.

1 pairs Unpair all

FDA1142650-C002-001_S3_L001_R1_001.fastq.gz → FDA1142650-C002-001_S3_L001_001.fastq ← FDA1142650-C002-001_S3_L001_R2_001.fastq.gz ✕

"Hide original elements?" defaults to being clicked. Unclick it. → Hide original elements? ☒ Remove file extensions? ☒

Name:

Cancel Create collection

Unclick **"Hide original elements"**.

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

Click **Create list**.

Create a collection of paired datasets

2 pairs created: all datasets have been successfully paired

0 unpaired forward - (0 filtered out)

Choose filters Clear filters

0 unpaired reverse - (0 filtered out)

Filter this list

Filter this list

2 paired Unpair all

CFSAN074382_S15_L001_R1_001.fastq.gz	→	CFSAN074382_S15_L001_R_001.fastq	←	CFSAN074382_S15_L001_R2_001.fastq.g	🔍
CFSAN074384_S20_L001_R1_001.fastq.gz	→	CFSAN074384_S20_L001_R_001.fastq	←	CFSAN074384_S20_L001_R2_001.fastq.g	🔍

Remove file extensions from pair names? ☒ Hide original elements? ☐

Name:

Cancel

Create list

3.7 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.

You can re-name this PairedList by clicking on the name.

protocols.io |
<https://dx.doi.org/10.17504/protocols.io.5jyl8mj16g2w/v5>

9



Run the MicroRunQC workflow

- 4 **Add the MicroRunQC workflow to your own "workflows" panel.** You only have to do this step once for each new workflow you need.
 - 4.1 Navigate to the **"Shared Data"** drop down menu, choose **workflows** and locate "MicroRunQC_v1.2"

Published Workflows

search name, annotation, owner

Advanced Search

Workflow Visualize Shared Data Help User

Data Libraries
Histories
Workflows
Visualizations
Access published resources

Name	Ann	Pages	Owner	Community Rating	Community Tags	Last Updated
RISHL-NARMS Unknown or Mixed Run AMR Workflow v4.1	RISHL working version of NARMS mixed-run v4.1		ssierrapatev	★★★★★	narms amr speciation rishl	18 hours ago
SeqSero2 v1.2.1 workflow tabular and row outputs	Run SeqSero2 v1.2.1 on a list-of-pairs collection and it will produce tabular and row output files		cstrittmatter	★★★★★	salmonella seqsero serotyping	May 30, 2023
Metagenomics_Taxonomy_Metaphlan_species_table	Species level taxonomical analysis of multiple metagenomics samples.		jgangiredia	★★★★★	metagenomics fastqc species krona	May 25, 2023
NARMS Unknown or Mixed Run AMR Workflow v4.1	Not bua specific. For mixed MiSeq runs or unknown isolates. Updated to MicroRunQC v1.2		gmartin	★★★★★	narms amr speciation	May 25, 2023
NARMS E. coli AMR Workflow V4.1	E. coli AMR, speciation, and QC. Updated to MicroRunQC v1.2		gmartin	★★★★★	narms amr e.coli speciation	May 25, 2023
NARMS Salmonella AMR Workflow v4.1	Salmonella AMR, serotyping, and QC. Updated to MicroRunQC v1.2.		gmartin	★★★★★	salmonella seqsero narms amr	May 25, 2023
NARMS Campylobacter AMR Workflow v4.1	Campy AMR, speciation, and QC. Updated to MicroRunQC v1.2.		gmartin	★★★★★	narms campylobacter amr speciation	May 25, 2023
MicroRunQC_v1.2	Updated MicroRunQC MLST output to include clonal complex, lineage, and species when available.		estrain	★★★★★		May 24, 2023

From Dropdown, select **"Import"**

MicroRunQC v1.2

Run
Import
Save as File

NARMS A

4.2 To see the new imported workflow, click "Workflow" tab on the top panel.


Click "Bookmarked" box to make it available in the left panel under "Workflows"

Workflow Visualize Shared Data Help User

Search Workflows + Create Import

Name	Tags	Updated	Sharing	Bookmarked
Imported: MicroRunQC_v1.2 Updated MicroRunQC MLST output to include clonal complex, lineage, and species when available.		a few seconds ago		<input checked="" type="checkbox"/>
▼ Imported: MicroRunQC_v1.1		3 months ago		<input type="checkbox"/>
▼ Imported: SeqSero2 v1.1.1 collection workflow Run SeqSero2 v1.1.1 on a list-of-pairs collection with a tabular output		3 years ago		<input type="checkbox"/>

4.3 From the Workflow menu on the left panel, select **MicroRunQC_v1.2**



Tools
☆
☰

✕

📁 Upload Data

Metagenomics:Kraken
Metagenomics:Mitokmer
Metagenomics:Graphlan
Metagenomics:Functional Profiling
Metagenomics:Assembly
Metagenomics:Rpackages
Metagenomics:Metaphlan
Metagenomics:CPIPES
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 to facilitate the quality control of alignment sequencing data and its derivatives like feature counts.
NGS:Simulator

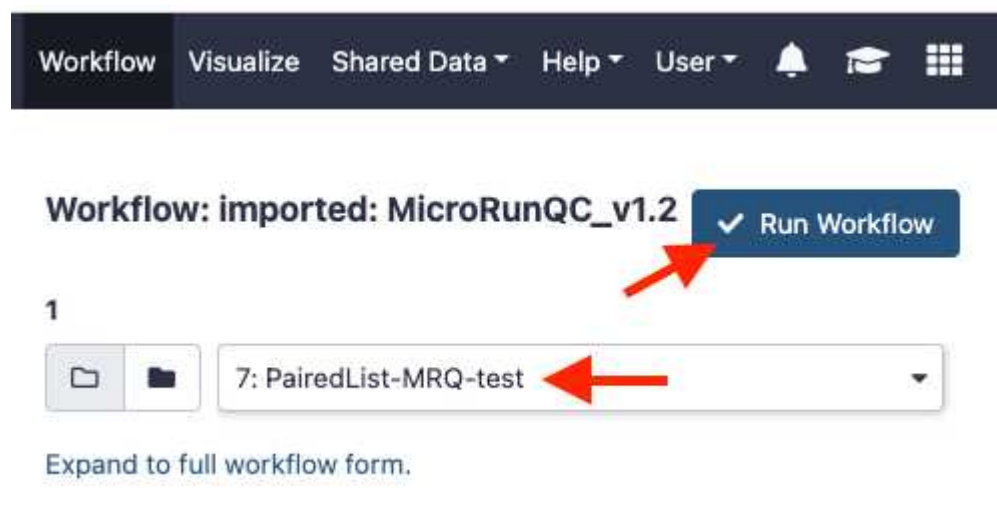
WORKFLOWS

All workflows

imported: MicroRunQC_v1.2

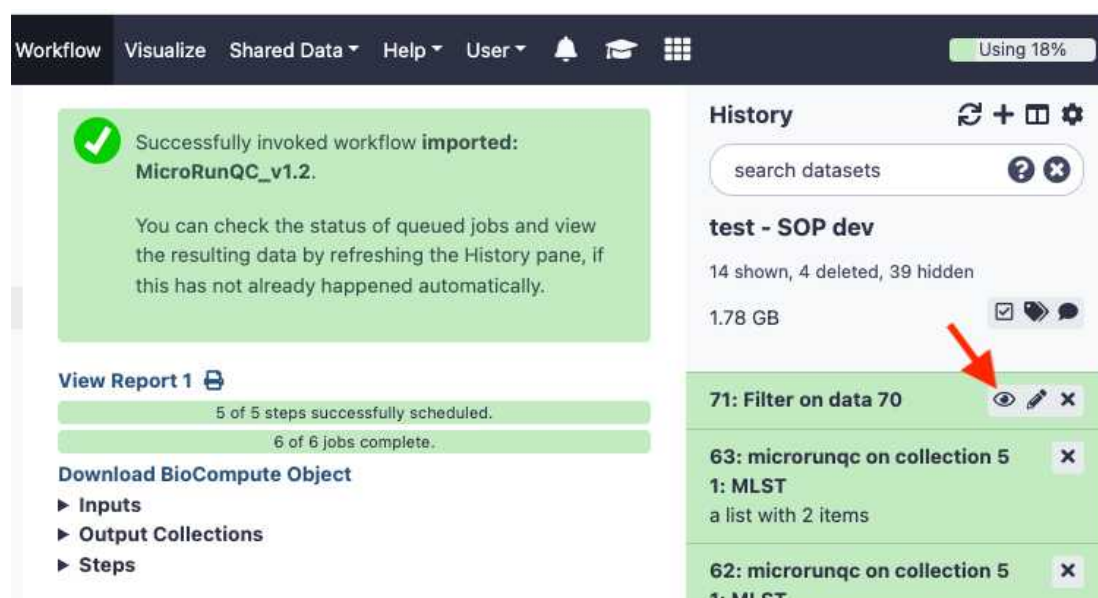
4.4 Select paired list dataset you created earlier.

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.



4.5 Upon completion of the pipeline all tiles in the history bar will be green.

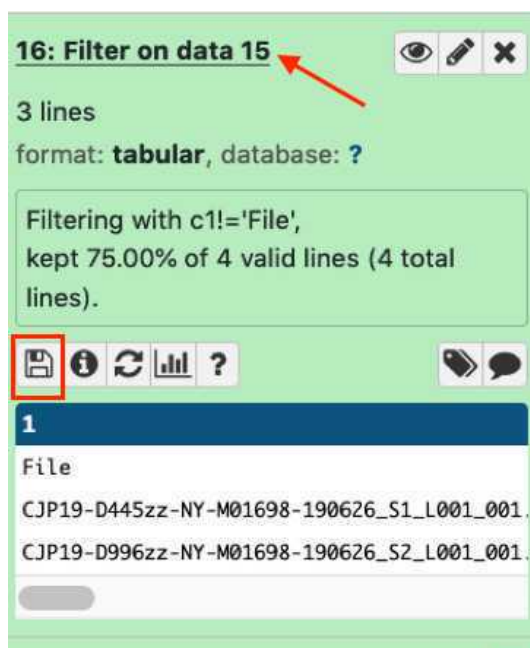
In the “**Filter on Data ##**”, click on the “Eye” icon to view the output table in the GalaxyTrakr window.



Interpret the results

5 Download and interpret the results:

- 5.1 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.



- 5.2 The MicroRunQC output file includes the following columns:

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

A	B	C
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. Listeria 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.

5.3 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

A	B
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	MLST extra							
FDA1216271-C001-001	16	291194	36.7	476210	321	148.4	148.4	36.4	34.6	listeria_2	5	CC=C5,Lineage=I	abcZ(2)	bgIA(1)	cat(11)	daPE(3)	dat(3)	ldh(1)	lhkA(7)
FDA817806-S073-001	20	3068354	179.6	525438	329	234.7	235.2	36.7	31.9	listeria_2	321	CC=C321,Lineage=II	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-S060-005	9	2881869	213	1498790	303	232.1	232.2	37	36.2	listeria_2	1489	CC=C1489,Lineage=Linnocua	abcZ(250)	bgIA(21)	cat(83)	daPE(298)	dat(20)	ldh(458)	lhkA(216)
FDA1213835-C001-001	37	4832365	34.4	294936	354	149	149	36.6	35.7	senterica_achtm_2	214		aroc(14)	dnAN(72)	hemD(21)	hisD(12)	purE(6)	sucA(19)	thrA(15)

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.mono lineage is listed alongside the ST.

5.4 Quality control threshold guidelines for the GenomeTrakr surveillance network. 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 par.</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