

Mar 16, 2023

Version 4

Quality control assessment for microbial genomes: GalaxyTrakr MicroRunQC workflow V.4

 In 1 collection

DOI

dx.doi.org/10.17504/protocols.io.5jyl8mj16g2w/v4

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

Vet LIRN



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/v4>

External link: <https://galaxytrakr.org>



Protocol Citation: Ruth Timme, Yesha Shrestha, Tina Pfefer, Paul Morin, Maria Balkey, Errol Strain 2023. Quality control assessment for microbial genomes: GalaxyTrakr MicroRunQC workflow. **protocols.io**
<https://dx.doi.org/10.17504/protocols.io.5jyl8mj16g2w/v4> 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.

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: March 14, 2023

Last Modified: March 16, 2023

Protocol Integer ID: 78694

Keywords: WGS, Quality Control, GalaxyTrakr, GenomeTrakr, microbial pathogen surveillance, galaxytrakr microrunqc workflow purpose, wgs sequence quality for bacterial pathogen, quality control assessment for microbial genome, microrunqc workflow, microrunqc, microbial genome, quality assessments for raw read, checking wgs sequence quality, bacterial pathogen, galaxytrakr, most microbial pathogen, de novo assembly, fastq file, end fastq file, custom galaxy instance, raw read, sequence type for each isolate, cronobacter, account in galaxytrakr, added enterobacter qc, assembly qc, miseq, entire miseq, quality control assessment

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

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.

Troubleshooting



Account set up

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

User Registration Form

The registration form includes the following fields and options:

- Location:** A dropdown menu showing "California Department of Public Health - Food and Drug Laboratory Branch" with a link to "Add New Location".
- First Name:** A text input field with a placeholder "Enter First Name, Do not use characters: ^[]{}:|'\"=+~*?<>@,.".
- Last Name:** A text input field with the same placeholder as the first name.
- Email:** A text input field with a note: "Email will be used for automated messages to include registration information!".
- Primary Phone:** A text input field with a placeholder "Please enter number with country code, without dashes, for example +17035456789" and a note: "If possible please use a mobile number than can accept text messages, only used for support".
- Title:** A text input field.
- Requirements:** A large text area with a placeholder "Please annotate intended use of Galaxy and Analysis tools. List specific tools you would like to see deployed in Galaxy."
- Register:** A button at the bottom of the form.

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

The page shows the login interface for GalaxyTrakr 1905. The top navigation bar includes links for "Analyze Data", "Workflow", "Visualize", "Shared Data", "Help", and "Login".

Left Panel (Login Form):

- Header: "Welcome to Galaxy, please log in"
- Field: "Username or Email Address" with an input box.
- Field: "Password" with an input box.
- Link: "Forgot password? Click here to reset your password."
- Button: "Login"
- Footer: "Don't have an account? Registration for this Galaxy instance is disabled. Please contact an administrator for assistance."

Right Panel (Welcome Message):

- Header: "Welcome to GalaxyTrakr: open-source bioinformatics for public health."
- Text: "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."
- Section: "--!!--Information and Announcements--!!--"
- Text: "Please re-import the skesamlst workflow that was updated a few days ago. Previous versions are no longer working and are causing errors when running. Thank you."
- Text: "Access CFSAN SNP Pipeline workflows in the shared workflows screen."
- Text: "Post in the official Galaxy GenomeTrakr board on the Redmine Site: Click here"
- Text: "Click here to access the GalaxyTrakr User Guide"
- Text: "Forgot Password? Email GalaxyTrakr Support Team"

Create a new history

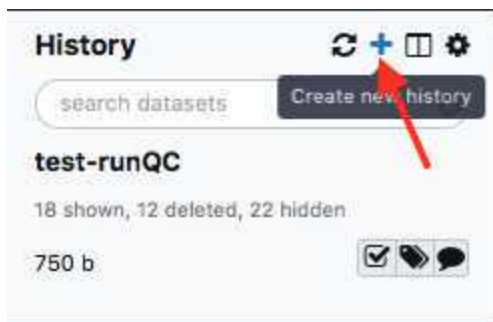
2 Create a new history.

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



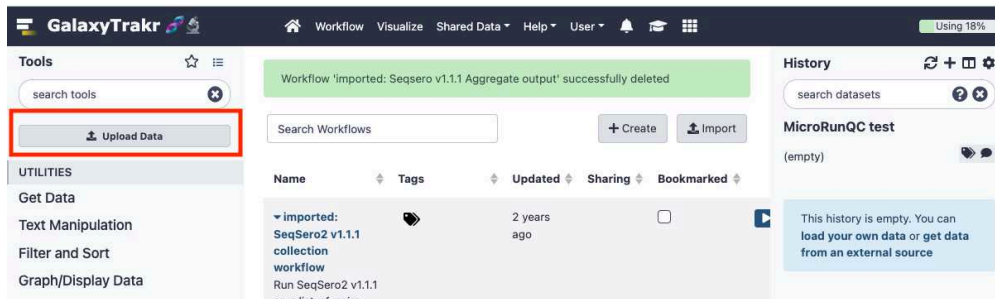
2.2 Name your new History by clicking on the "Unnamed history", type in desired name and hit enter. We recommend including the run cell ID and the date the run was started.



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.

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.

Select **Clear filters**, then click **Auto-pair**.

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. Close this message using the X on the right to view more help.

0 unpaired forward - (4 filtered out) Choose filters **Clear filters** 1 0 unpaired reverse - (4 filtered out)

_1 Auto-pair 2 _2

(no datasets were found matching the current filters)

If auto-pairing does not work, you can click "choose filters" and select the appropriate filter for the pairing:

e.g. choose "_R1 "and "_R2 "

0 unpaired forward - (4 filtered out) **Choose filters** Clear filters 0 unpaired reverse - (4 filtered out)

_1

Choose from the following filters to change which unpaired reads are shown in the display:

Forward: _1, Reverse: _2

Forward: _R1, Reverse: _R2

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

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

Filter this list

0 unpaired reverse - (0 filtered out)

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.fast	←	CFSAN074384_S20_L001_R2_001.fastq.g	⌕

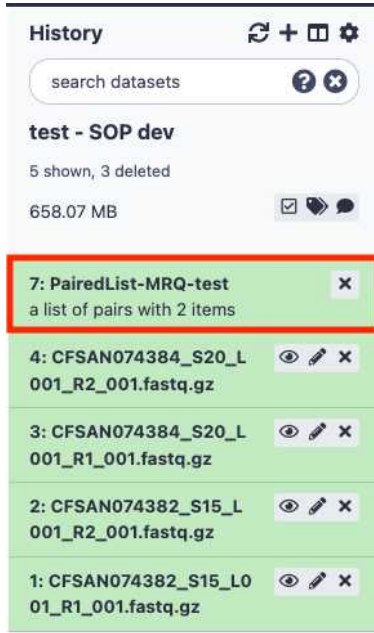
Remove file extensions from pair names? ☒
Hide original elements? ☐

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.

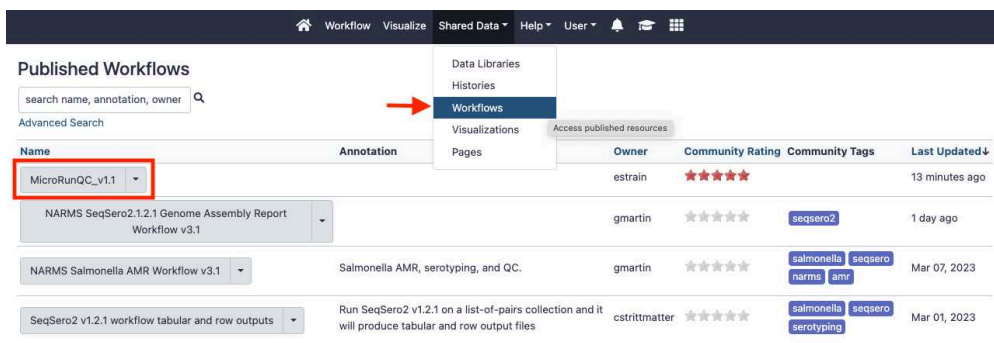
You can re-name this PairedList by clicking on the name.



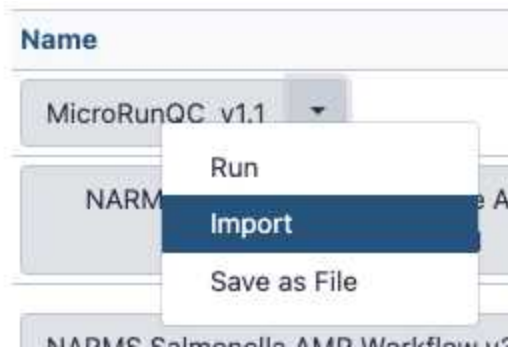
Run the MicroRunQC workflow

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_v1.1

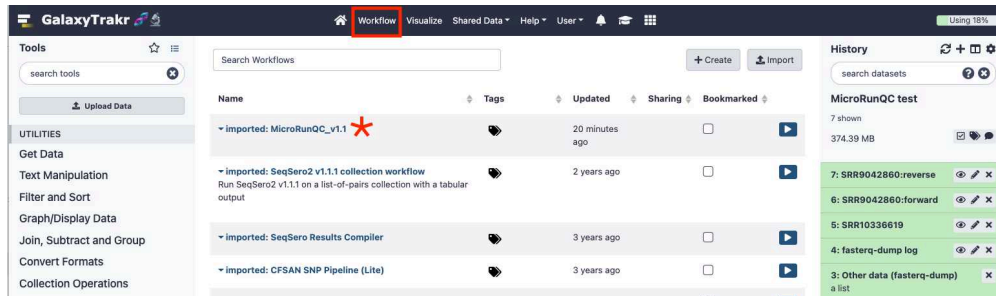


From Dropdown, select **"Import"**





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"





4.3 From the Workflow menu on the left panel, select **MicroRunQC_v1.1**



 **GalaxyTrakr** 

Tools  



 **Upload Data**

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_v1.1 

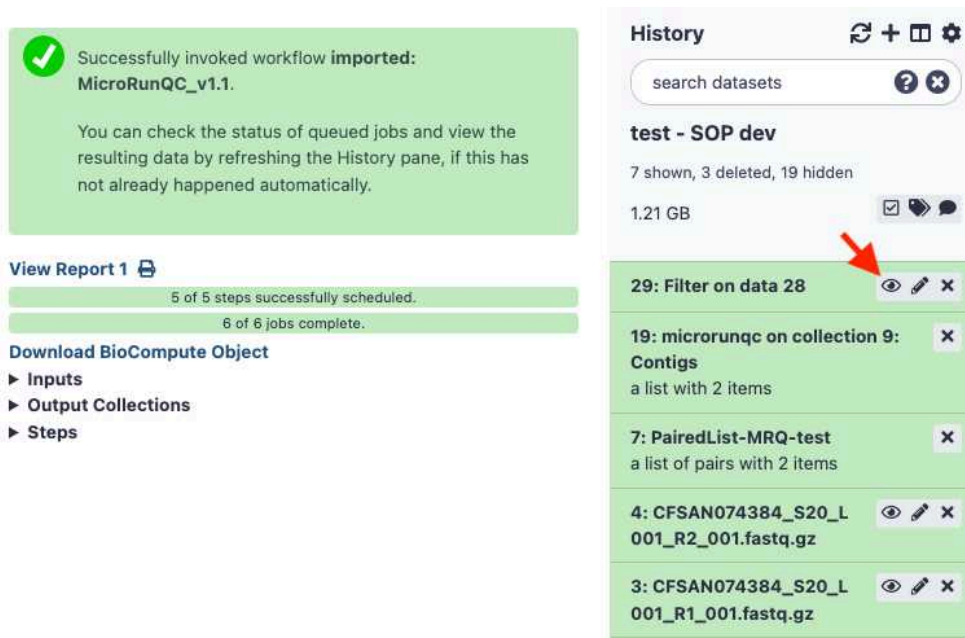
4.4 Select paired list dataset you created earlier.

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.

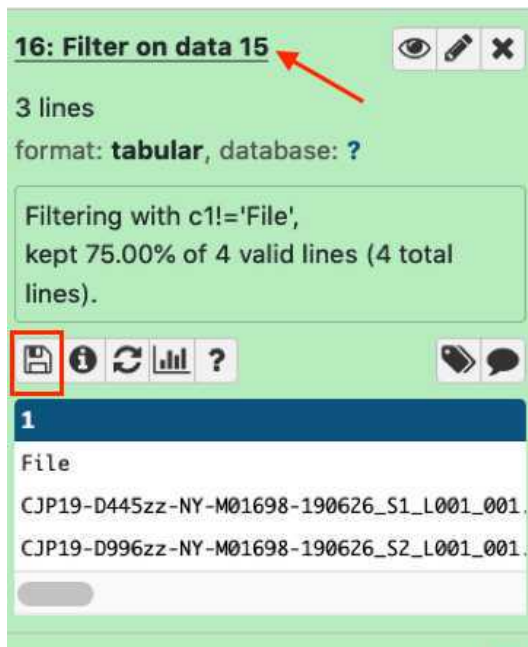
In the "**Filter on Data ##**", click on the "Eye" icon to view the output table in the GalaxyTrakr window.



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
Parameter	Input	Description
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.



A	B	C
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
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
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.

**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

A	B
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
	File	Contigs	Length	Est Cov	N50	Median Insert	Mean Length \bar{R}_1	Mean Length \bar{R}_2	Mean Q \bar{R}_1	Mean Q \bar{R}_2	Scheme	ST							
	FDA1216271-C001-001	16	2911949	36.7	476210	321	148.4	148.4	36.4	34.6	listeria_2	5	abcZ(2)	bgIA(1)	cat(11)	dape(3)	datt(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	abcZ(5)	bgIA(6)	cat(8)	dape(62)	datt(6)	ldh(7)	lhkA(34)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
	F D A 7 4 6 6 3 4	3 0	3 0 5 2 8 8 8	4 1. 4	2 9 3 9 4 7	3 2 0	14 8. 4	14 8. 4	3 6. 5	3 6	lis te ri a_ 2	-	a b c Z(2)	b gl A (1)	c at (1 1)	d a p E (3)	d a t (3)	l d h (1)	l h k A (~ 7)
	F D A 12 13 3 7 7- C 0 01 - 0 0 2	2 0	2 6 7 2 1 8 0	1 5 5 .1	4 7 3 1 8 1	2 7 0	14 7. 3	14 7. 3	3 7. 2	3 6. 1	-	-							
	F D A 9 3 3 3 7 6 - S 0 6 0 - 0 0 5	9	2 8 8 1 8 6 9	2 1 3	1 4 9 8 7 9 0	3 0 3	2 3 2. 1	2 3 2. 2	3 7	3 6. 2	lis te ri a_ 2	1 4 8 9	a b c Z(2 5 0)	b gl A (2 1)	c at (8 3)	d a p E (2 9 8)	d a t (2 0)	l d h (4 5 8)	l h k A (2 1 6)
	F D A 12 13 8 3 5- C 0 01 - 0 01	3 7	4 8 3 2 3 6 5	3 4 .4	2 9 4 9 3 6	3 5 4	14 9	14 9	3 6. 6	3 5. 7	se nt er ic a_ a c ht m a n_ 2	2 1 4	ar o C (1 4)	d n a N (7 2)	h e m D (2 1)	hi 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 listeria database includes multiple species, including *Listeria monocytogenes* and *L. innocua*. If users want to investigate which Listeria DB corresponds to the resulting ST types, they can query the **Institut Pasteur** mlst database:

Example: query Listeria ST type here: https://bigsdb.pasteur.fr/cgi-bin/bigsdb/bigsdb.pl?db=pubmlst_listeria_seqdef&page=query

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 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