Apr 17, 2024

Quality control assessment for microbial genomes: GalaxyTrakr MicroRunQC workflow

Forked from Quality control assessment for microbial genomes: GalaxyTrakr MicroRunQC workflow

DOI

dx.doi.org/10.17504/protocols.io.4r3l27oz3g1y/v1

Ruth Timme[1], Sai Laxmi Gubbala Venkata[2], Maria Balkey[1], Robyn Randolph[3], William Wolfgang[4], Errol Strain[3]

[1]US Food and Drug Administration; [2]Bacteriology Laboratory, Wadsworth Center, New York State Department of Health, Albany, New York, USA;
[3]Center for Veterinary Medicine, U.S. Food and Drug Administration, College Park, Maryland, USA; [4]Wadsworth Center NYSDOH

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

Maria

Maria Balkey
US Food and Drug Administration

DOI: dx.doi.org/10.17504/protocols.io.4r3l27oz3g1y/v1

External link: https://galaxytrakr.org

Protocol Citation: Ruth Timme, Sai Laxmi Gubbala Venkata, Maria Balkey, Robyn Randolph, William Wolfgang, Errol Strain 2024. Quality control assessment for microbial genomes: GalaxyTrakr MicroRunQC workflow. **protocols.io** https://dx.doi.org/10.17504/protocols.io.4r3l27oz3g1y/v1

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.

Protocol status: Working

Created: October 26, 2022

Last Modified: April 17, 2024

Protocol Integer ID: 71843

Keywords: WGS, Quality Control, GalaxyTrakr, GenomeTrakr, microbial pathogen survielliance

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. 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.

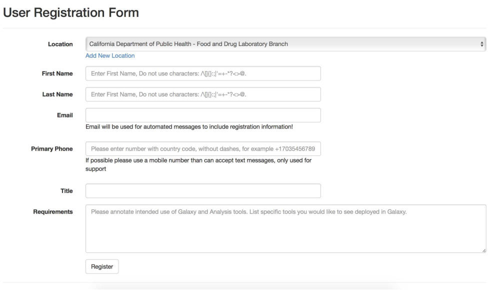
SCOPE: This protocol covers the following tasks:

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

V3: updated with *Cronobacter* thresholds

Account set up

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



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

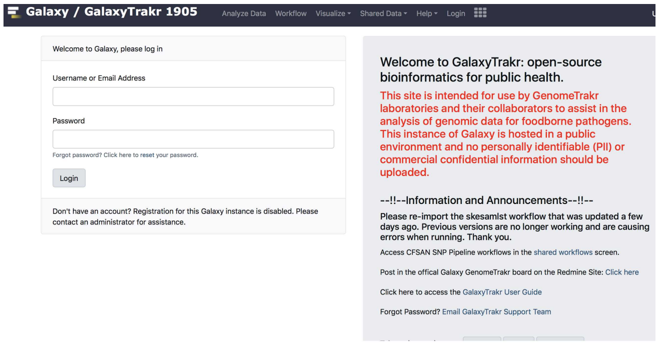
If possible please use a mobile number than can accept text messages, only used for support

Title:

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

Register

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



Welcome to Galaxy, please log in

Username or Email Address

Password

Forgot password? Click here to reset your password.

Login

Don't have an account? Registration for this Galaxy instance is disabled. Please contact an administrator for assistance.

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 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.

Access CFSAN SNP Pipeline workflows in the shared workflows screen.

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

Click here to access the GalaxyTrakr User Guide

Forgot Password? Email GalaxyTrakr Support Team

Create a new history

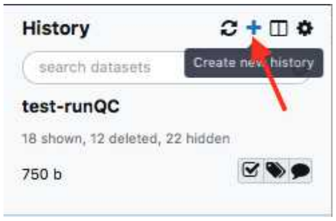
2 **Create a new history.**

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

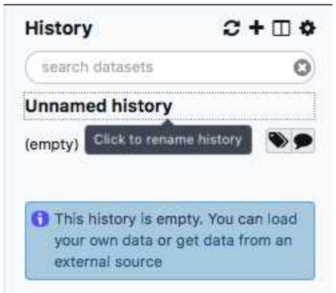
Save your MicroRunQC output here and any other relevant analyses, like serotyping, or AMR detection.

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.

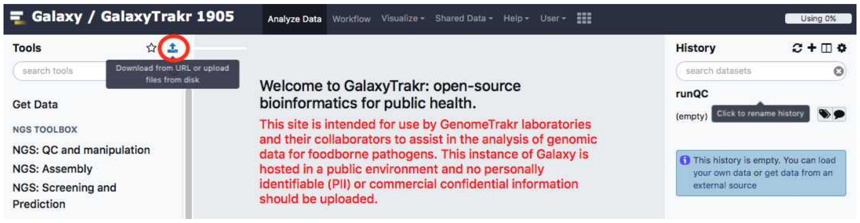


This history is empty. You can load your own data or get data from an external source

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/download icon on the top of the left web page to start an upload process.

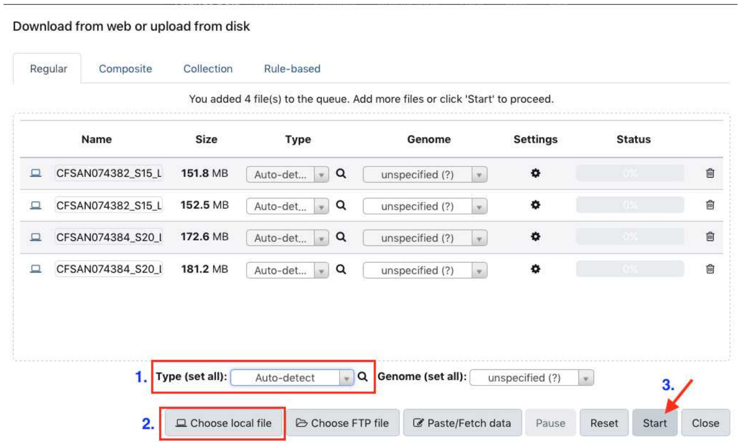


Welcome to GalaxyTrakr: open-source bioinformatics for public health.

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.

This history is empty. You can load your own data or get data from an external source

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).



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

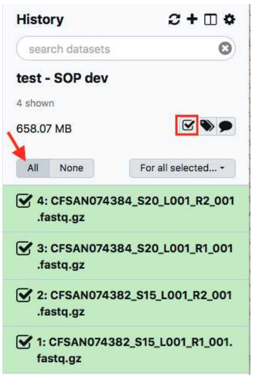
Name	Size	Type	Genome	Settings	Status
CFSAN074382_S15_L	151.8 MB	Auto-det...	unspecified (?)	⚙	0%
CFSAN074382_S15_L	152.5 MB	Auto-det...	unspecified (?)	⚙	0%
CFSAN074384_S20_l	172.6 MB	Auto-det...	unspecified (?)	⚙	0%
CFSAN074384_S20_l	181.2 MB	Auto-det...	unspecified (?)	⚙	0%

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

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

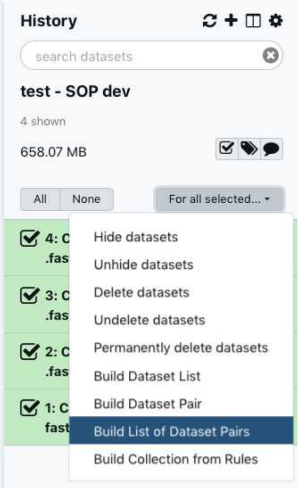
3.3 You have just upload a set of forward and reverse reads. For further analysis these files need to be paired properly so the platform knows which R1 and R2 files go with each sample/isolate. GalaxyTrakr does this by creating a **List of Dataset Pairs.**

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.


History
search datasets
test - SOP dev
4 shown
658.07 MB
All
None
For all selected...
4: CFSAN074384_S20_L001_R2_001.fastq.gz
3: CFSAN074384_S20_L001_R1_001.fastq.gz
2: CFSAN074382_S15_L001_R2_001.fastq.gz
1: CFSAN074382_S15_L001_R1_001.fastq.gz

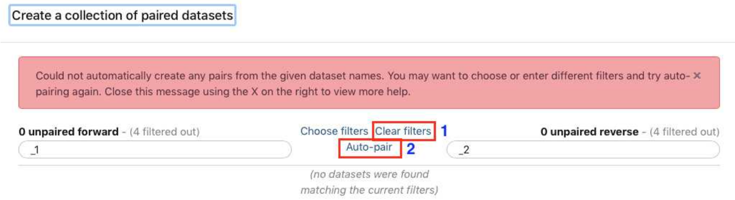
Screenshot of History panal showing recently uploaded files. Note the way the files are named, using R1 and R2 to identify the paired reads. This will be important in the next step. Some naming conventions can be slightly different.

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 (_R1 and _R2 in the example above)

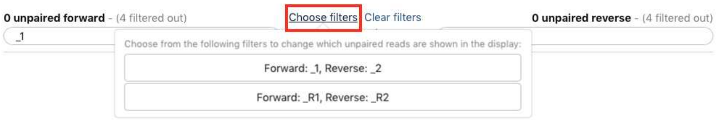
Select **Clear filters**, then click **Auto-pair.**


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)
_1
Choose filters
Clear filters
1
Auto-pair
2
(no datasets were found matching the current filters)
0 unpaired reverse - (4 filtered out)
_2

Alternatively, instead of autopairing you can click "choose filters" and select the appropriate filter for the pairing:



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:

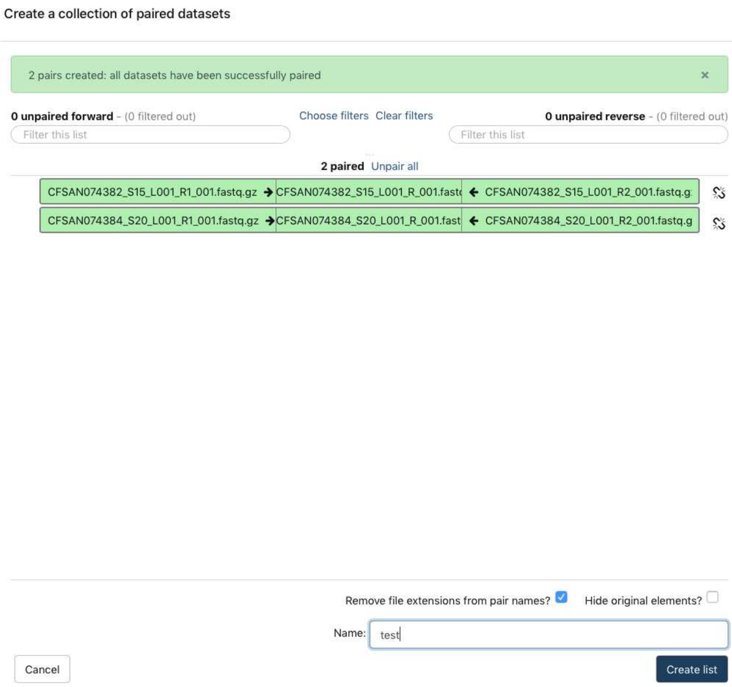
Forward: _1, Reverse: _2

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 0 unpaired reverse - (0 filtered out)

Filter this list Filter this list

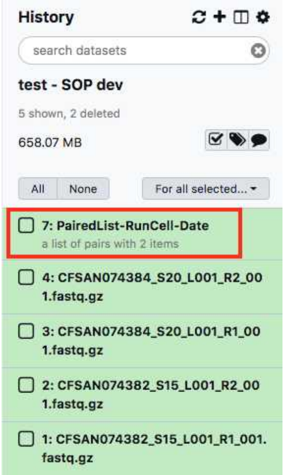
2 paired Unpair all

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.fast	← CFSAN074384_S20_L001_R2_001.fastq.g

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

Name: test

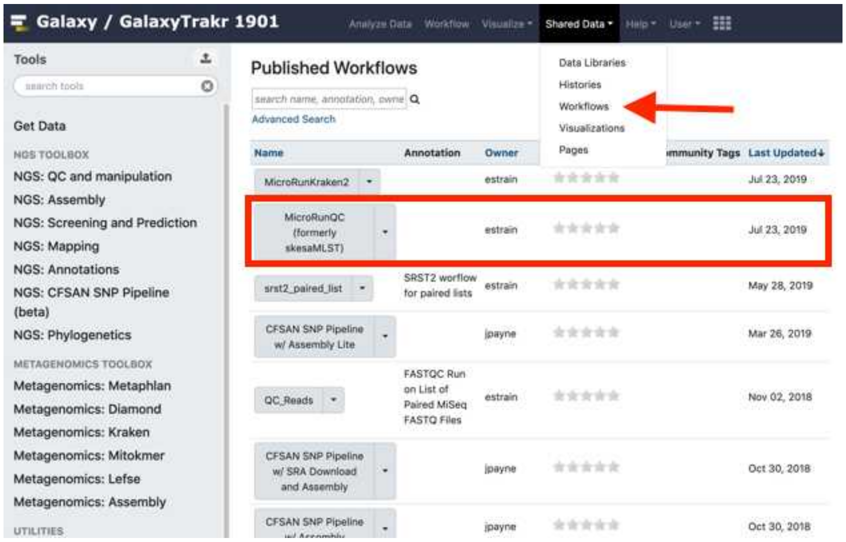
3.7 This paired dataset will now be available for analysis in your history panel. You can run multiple analyses on the same dataset in a history rather than upload the same sequence data to a new history to perform additional analyses. This will help you use your allocated storage space efficiently.


History
search datasets
test - SOP dev
5 shown, 2 deleted
658.07 MB
All
None
For all selected...
7: PairedList-RunCell-Date
a list of pairs with 2 items
4: CFSAN074384_S20_L001_R2_001.fastq.gz
3: CFSAN074384_S20_L001_R1_001.fastq.gz
2: CFSAN074382_S15_L001_R2_001.fastq.gz
1: CFSAN074382_S15_L001_R1_001.fastq.gz

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.

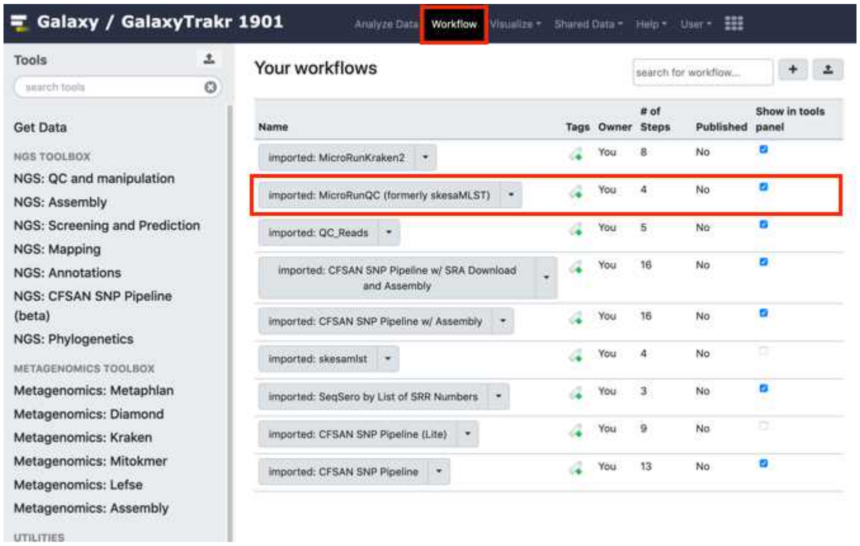
4.1 Navigate to the “Shared Data" drop down menu, choose workflows and from the MicroRunQC drop down menu select import.



Published Workflows

Name	Annotation	Owner		mmunity Tags	Last Updated↓
MicroRunKraken2		estrain	☆☆☆☆☆		Jul 23, 2019
MicroRunQC (formerly skesaMLST)		estrain	☆☆☆☆☆		Jul 23, 2019
srst2_paired_list	SRST2 worflow for paired lists	estrain	☆☆☆☆☆		May 28, 2019
CFSAN SNP Pipeline w/ Assembly Lite		jpayne	☆☆☆☆☆		Mar 26, 2019
QC_Reads	FASTQC Run on List of Paired MiSeq FASTQ Files	estrain	☆☆☆☆☆		Nov 02, 2018
CFSAN SNP Pipeline w/ SRA Download and Assembly		jpayne	☆☆☆☆☆		Oct 30, 2018
CFSAN SNP Pipeline w/ Assembly		jpayne	☆☆☆☆☆		Oct 30, 2018

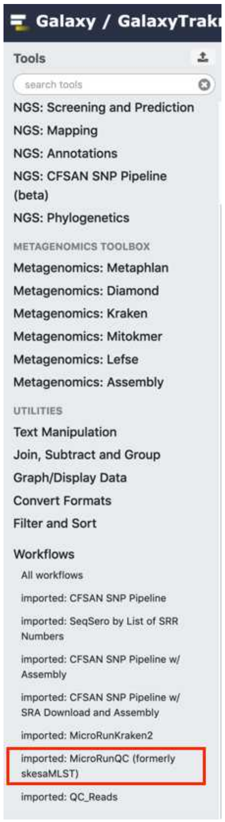
4.2 To see the new workflow in your “Workflows” tools panel on the left, open the Workflow tab and check “show in tools panel” for the workflow of interest.



Your workflows

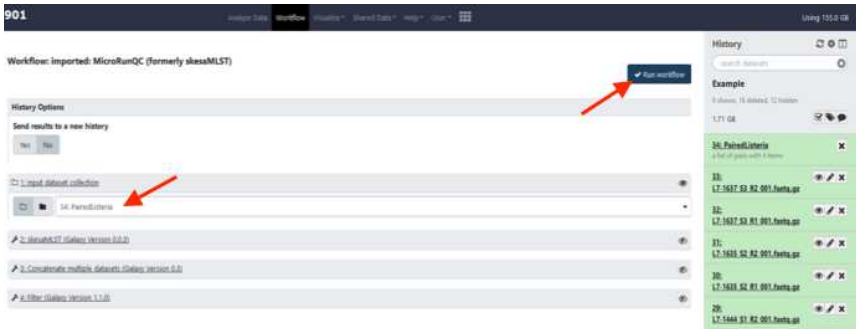
Name	Tags	Owner	# of Steps	Published	Show in tools panel
imported: MicroRunKraken2		You	8	No	☑
imported: MicroRunQC (formerly skesaMLST)		You	4	No	☑
imported: QC_Reads		You	5	No	☑
imported: CFSAN SNP Pipeline w/ SRA Download and Assembly		You	16	No	☑
imported: CFSAN SNP Pipeline w/ Assembly		You	16	No	☑
imported: skesamlst		You	4	No	□
imported: SeqSero by List of SRR Numbers		You	3	No	☑
imported: CFSAN SNP Pipeline (Lite)		You	9	No	□
imported: CFSAN SNP Pipeline		You	13	No	☑

4.3 From the workflow menus select MicroRunQC



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.



Workflow: imported: MicroRunQC (formerly skesaMLST)

History Options

Send results to a new history: Yes No

1: input dataset collection — 34: PairedListeria

2: skesaMLST (Galaxy Version 0.0.2)

3: Concatenate multiple datasets (Galaxy Version 0.0)

4: Filter (Galaxy Version 1.1.0)

History: Example — 1.71 GB

34: PairedListeria

33: L7-1637_S3_R2_001.fastq.gz

32: L7-1637_S3_R1_001.fastq.gz

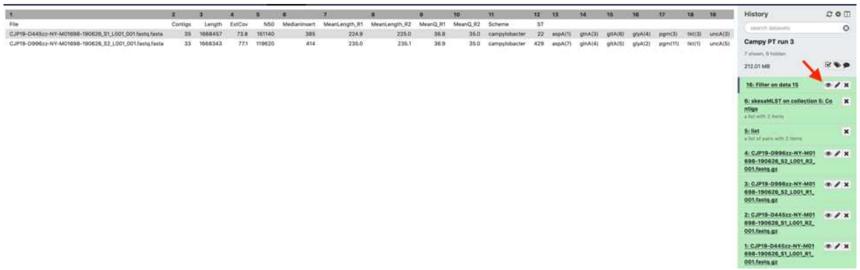
31: L7-1635_S2_R2_001.fastq.gz

30: L7-1635_S2_R1_001.fastq.gz

29: L7-1444_S1_R2_001.fastq.gz

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

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



1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
File	Contigs	Length	EstCov	N50	MedianInsert	MeanLength_R1	MeanLength_R2	MeanQ_R1	MeanQ_R2	Scheme	ST							
CJP19-D445zz-NY-M01698-190626_S1_L001_001.fastq.fasta	35	1668457	73.8	151140	385	224.9	225.0	36.8	36.0	campylobacter	22	aspA(1)	glnA(3)	gltA(6)	glyA(4)	pgm(3)	tkt(3)	uncA(3)
CJP19-D996zz-NY-M01698-190626_S2_L001_001.fastq.fasta	33	1668343	77.1	119620	414	236.0	236.1	36.9	35.0	campylobacter	429	aspA(7)	glnA(4)	gltA(5)	glyA(2)	pgm(11)	tkt(1)	uncA(5)

History: Campy PT run 3 — 7 shown, 9 hidden — 212.01 MB

16: Filter on data 15

6: skesaMLST on collection 5: Contigs — a list with 2 items

5: list — a list of pairs with 2 items

4: CJP19-D996zz-NY-M01698-190626_S2_L001_R2_001.fastq.gz

3: CJP19-D996zz-NY-M01698-190626_S2_L001_R1_001.fastq.gz

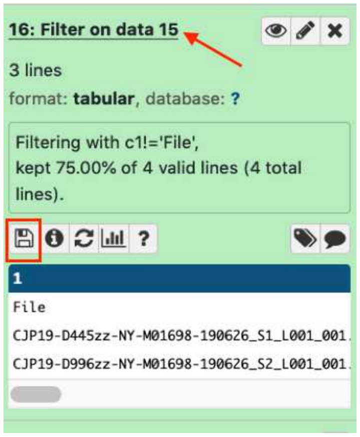
2: CJP19-D445zz-NY-M01698-190626_S1_L001_R2_001.fastq.gz

1: CJP19-D445zz-NY-M01698-190626_S1_L001_R1_001.fastq.gz

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 that can be opened on excel.



16: Filter on data 15

3 lines

format: **tabular**, database: **?**

Filtering with c1!='File',
kept 75.00% of 4 valid lines (4 total lines).

1
File
CJP19-D445zz-NY-M01698-190626_S1_L001_001.
CJP19-D996zz-NY-M01698-190626_S2_L001_001.

5.2 The MicroRunQC output file includes the following metrics:

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.
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 (pubmlst.org) database scheme (e.g. senterica for Salmonella enterica)
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 4 Listeria samples run through the MicroRunQC workflow:

File name	Contigs	Length	EstCov	N50	MedianInsert	MeanLength_R1	MeanLength_R2	MeanQ_R1	MeanQ_R2	Scheme	ST				
FSL-R9-8346	14	2876874	87.8	512255	408	147.6	147.7	33.1	32.7	lmonocytogenes	389	abcZ(52)	bglA(1)	cat(12)	dap
FSL-R9-8348	11	2832172	84.2	1464158	388	147.9	147.9	33.2	32.9	lmonocytogenes	795	abcZ(7)	bglA(10)	cat(18)	dap
FSL-R9-8350	14	2884629	64.2	450082	390	147.2	147.2	33.1	32.7	lmonocytogenes	37	abcZ(5)	bglA(7)	cat(3)	dap
FSL-R9-8352	12	2902520	85.9	1460419	390	148.1	148.2	33.1	32.8	lmonocytogenes	391	abcZ(7)	bglA(6)	cat(62)	dap

Spreadsheet showing example output for 5 *Listeria monocytogenes* samples from a NextSeq sequencing run.

5.4 Quality control threshold guidelines for the GenomeTrakr surveillance network. These are also relevant for NARMS and VetLIRN contributors.

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

A	B	C	D	E	F	G	H
Quality metric	***Salmonella***	***Listeria***	***E. coli***	***Shigella***	***Campylobacter***	***Vibrio para.***	Cronobacter
Average read quality Q score for R1 and R2	>=30	>=30	>=30	>=30	>=30	>=30	>=30
Average coverage	>=30X	>=20X	>=40X	>=40X	>=20X	>=40X	>=20X
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
De novo assembly: no. contigs	<=300	<=300	<=500	<=650	<=300	<=300	<=500