



2

Jan 05, 2022

Wastewater QC workflow in GalaxyTrakr (SSQuAWK) V.2

Jasmine Amirzadegan¹, Tunc Kayikcioglu¹, hugh.rand¹, Ruth Timme², Maria Balkey¹

¹Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration, College Park, Maryland, USA;

²US Food and Drug Administration



dx.doi.org/10.17504/protocols.io.b3icqkaw

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

Jasmine Amirzadegan

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

PURPOSE:

Step-by-step instructions for checking sequence quality for SARS-CoV-2 wastewater samples using **SSQuAWK: SARS - CoV - 2 Sequence Quality Assurance Workflow and Kontrapion**. The SSQuAWK workflow, implemented in a custom Galaxy instance, will produce quality assessments for raw reads (Illumina MiSeq paired-end fastq files).

SCOPE: This protocol covers the following tasks:

1. Set up an account in GalaxyTrakr
2. Create a new history
3. Upload data
4. Execute the SSQuAWK workflow
5. Interpret the results

Version history:

V1: Basic protocol steps with screenshots

V2: Addition of a detailed 12 minute video tutorial

DOI

dx.doi.org/10.17504/protocols.io.b3icqkaw

<https://galaxytrakr.org>

Jasmine Amirzadegan, Tunc Kayikcioglu, hugh.rand , Ruth Timme, Maria Balkey 2022. Wastewater QC workflow in GalaxyTrakr (SSQuAWK). **protocols.io**
<https://dx.doi.org/10.17504/protocols.io.b3icqkaw>
Ruth Timme

WGS, Quality Control, GalaxyTrakr, GenomeTrakr, microbial pathogen surveillance

protocol ,

Jan 05, 2022

Jan 05, 2022

Jan 05, 2022 | Ruth Timme | US Food and Drug Administration

Jan 05, 2022 | Ruth Timme | US Food and Drug Administration

Jan 05, 2022 | Jasmine Amirzadegan

56612

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

Account set up

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 Last 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 +17025450199
If possible please use a mobile number than can accept text messages, only used for support

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

Requirements: Please provide 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>

Galaxy / GalaxyTrakr 1905 Analysis Data Workflow Visualize Shared Data Help Login

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 skesamist workflow that was updated a few days ago. Previous versions are no longer working and are causing errors when running. Thank you.

Access CFSAI 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 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 sequence set with details and date in the history name.

Save your SSQuAWK output here with any other relevant analyses.

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 Create a new history with the "+" symbol in the upper right hand corner. Name your history and press "enter" on your keyboard to save the name.

GalaxyTrakr Analysis Data Workflow Visualize Shared Data Help Login

Tools

- Get Data
- FASTQ tools
- NGS: QC and manipulation
- NGS: Screening and Prediction
- NGS: Mapping
- NGS: Mapping
- NGS: Assembly
- NGS: Phylogenetics
- NGS: CFSAI SNP Pipeline (Beta)
- NGS: Megablast
- NGS: Nanopore
- NGS: NCBI Blast+
- NGS: RNA seq
- NGS: Annotations
- NGS: Virus
- NGS: Ikonu
- NGS: uniguy
- NGS: Seqtk

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 CFSAI 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 GalaxyTrakr User Guide

Click here to access the GalaxyTrakr User Guide

Click here to access the GalaxyTrakr User Guide

Forgot Password? Email GalaxyTrakr Support Team

Take an interactive tour: Galaxy | History | Screenshots

Galaxy is an open platform for supporting data intensive research. Galaxy is developed by The Galaxy Team with the support of many contributors.

The Galaxy Project is supported in part by NIDDK, NCI, The Huck Institute of the University of Wisconsin, The Institute for Genome Sciences and Policy, and Johns Hopkins University.

History

Workflows: 12/20/2018, 11:10:21AM

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

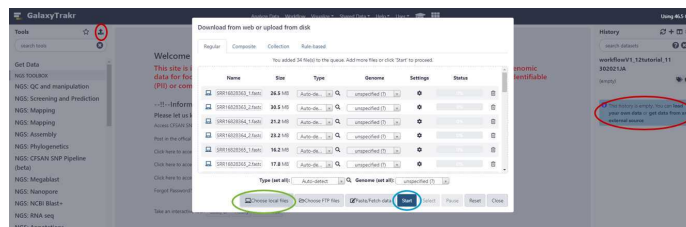
Upload data

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.

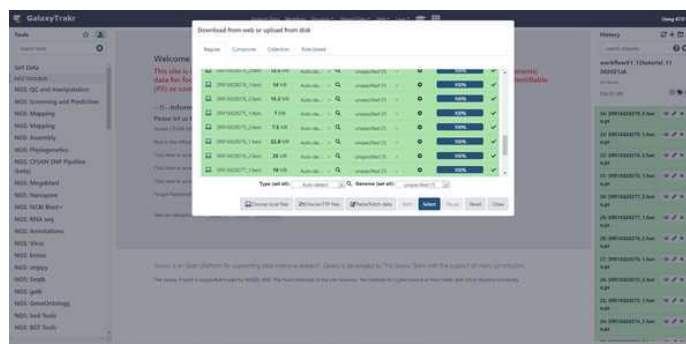
3.1 Upload sequence data to your history, using either of the two options circled in red below.

A window will appear in the middle of your screen. This is where you select your files using the "Choose local files" button at the bottom of the window. The "Choose local files" button is circled in green. These fastq files should be paired (two per sample).

After you've selected your files, press "Start" to initiate your data upload to GalaxyTrakr. The "Start" button is circled in blue.

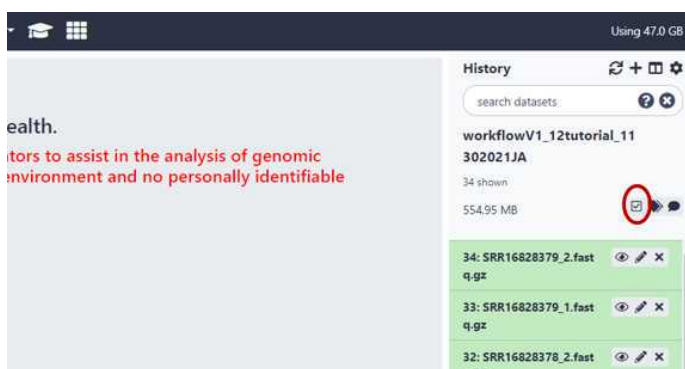


3.2 As the file uploads complete, each row will turn green. If samples are shown with yellow background, then are still uploading.



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. GalaxyTrakr does this by creating a **List of Dataset Pairs**.

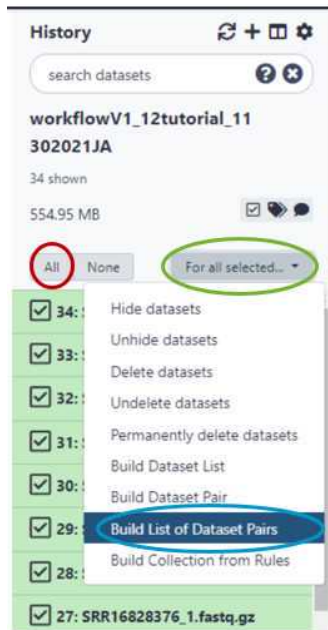
Within your newly created History panel, click the "check box," then select all the files you just uploaded by clicking "All" or by individually selecting the ones you want to pair.



3.4 Check all the files belonging to a pair. In this example, all the files belong to a pair, so I will use the "All" button (circled in red).

Then, use the "For all selected..." dropdown (circled in green), and click on "Build List of Dataset

Pairs" (circled in blue).

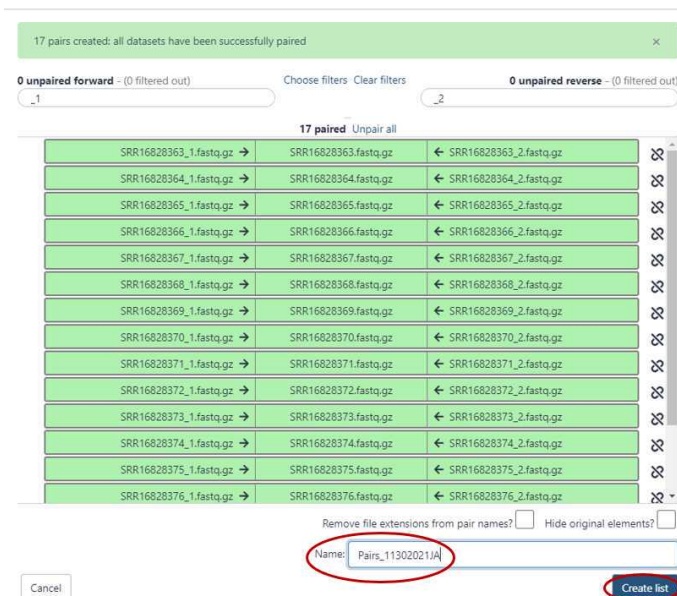


3.5 GalaxyTrakr will automatically pair the files, but it's good to double check.

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

If everything looks good, then choose a name for your pairs (circled red) and "Create List" (also circled red).

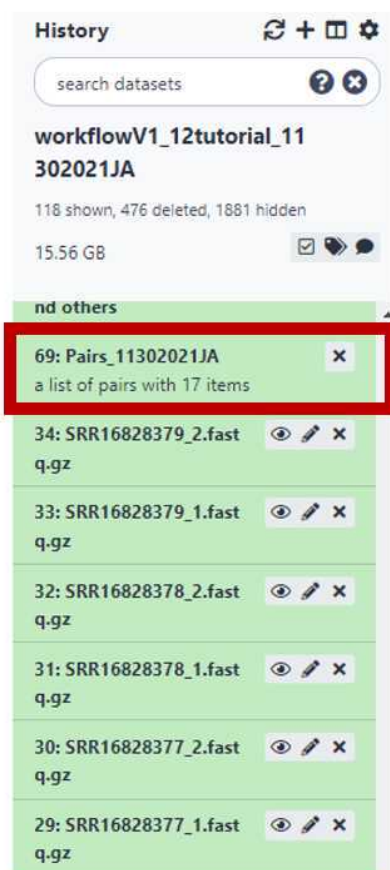
Create a collection of paired datasets



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



- 3.6 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.

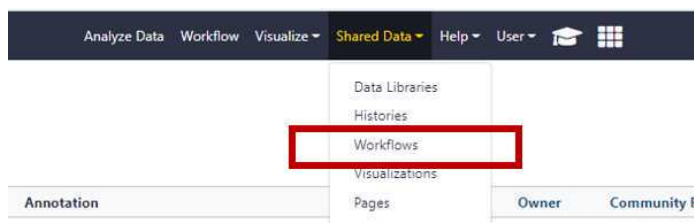


Run the SSQuAWK workflow

- 4 Add the SSQuAWK* workflow to your own "workflows" panel. You only have to do this step once for each new workflow you need.

*SSQuAWK: SARS - CoV - 2 Sequence Quality Assurance Workflow and Kontrapion

- 4.1 Navigate to the "Shared Data" drop down and choose workflows



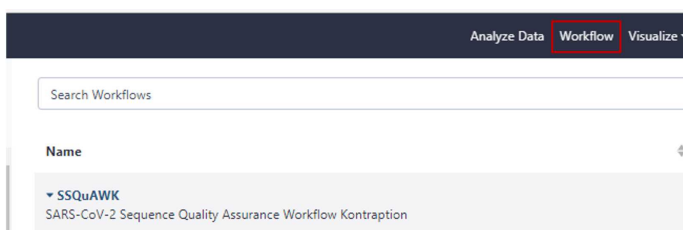
Then, from the SSQuAWK drop down menu, select import.

Published Workflows

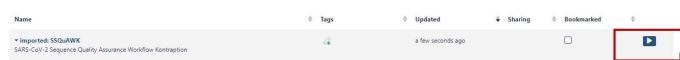
search name, annotation, owner, or Advanced Search

Name	Annotation	Owner	Community Rating	Community Tags	Last Updated:
SSQuAWK	SARS-CoV-2 Sequence Quality Assurance Workflow Kontraption	jasmine_amp	★★★★★	Import Run Workflow	5 seconds ago
NARMS: Unknown or Mixed Run AMR Workflow V2.0	Not bug specific. For mixed MSeq runs or unknown isolates	gmarin	★★★★★	Import Run Workflow	Oct 22, 2021
AMRfinderPlusDT Report WF		gmarin	★★★★★		Oct 22, 2021
NARMS: E. coli AMR Workflow V2.0	E. coli AMR, speciation, and QC	gmarin	★★★★★	Import Run Workflow	Oct 04, 2021

- 4.2 Navigate to the "Workflow" tab in the top ribbon (boxed in red). The workflow will be imported there.

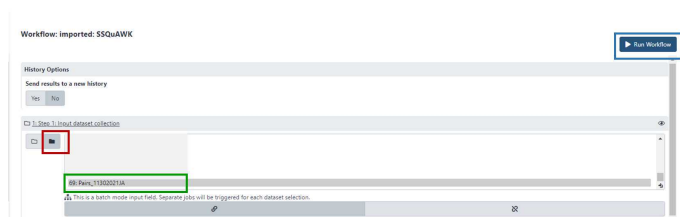


- 4.3 To use the workflow, press the 'play' button (boxed in red) on the right



- 4.4 Select the paired list you created earlier by selecting the folder icon (boxed in red), and then the list of pairs (boxed in green).

Click Run Workflow (boxed in blue).



Running the workflow can take some time depending on the number of samples you are analyzing. Once GalaxyTrakr adds the workflow invocation to the queue, you can choose to 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, the NGSQC_outfile will be green. Click on the "Eye" icon to view the output in the GalaxyTrakr window.

Sample	fwdReads	fwdAvgLen	fwdQual	fwdGC
SRR16828363.fastq	316332	75.5	33.1	
SRR16828364.fastq	229058	75.5	35.7	
SRR16828365.fastq	175990	75.5	34.9	

Interpret the results

5 Download and interpret the results:

- 5.1 Click "NGSQC_outfile" (boxed in red) and then the floppy disc save icon (boxed in blue). The tabular file can be opened in a text reader or converted to a format that can be opened on excel.

217: NGSQC_outfile

4 lines
format: **tabular**, database: ?

1 2 3 4

Sample	fwdReads	fwdAvgLen	fwdGC
SRR16828363.fastq	316332	75.5	33.1
SRR16828364.fastq	229058	75.5	35.7
SRR16828365.fastq	175990	75.5	34.9

- 5.2 The SSQuAWK output file includes the following metrics:

A	B	C
Parameter	Input	Description
Sample	List of Pairs	Sample name from list of pairs
fwdReads	FASTQC	Number of forward reads contributing to the sample pair
fwdAvgLen	FASTQC	Average of all forward read lengths
fwdAvgQ	FASTQC	Average quality of all forward reads
revReads	FASTQC	Number of reverse reads contributing to the sample pair
revAvgLen	FASTQC	Average of all reverse read lengths
revAvgQ	FASTQC	Average quality of all reverse reads
percentHuman	Kraken2	Percentage of reads classified as <i>Homo sapiens</i>
readsHuman	Kraken2	Number of reads classified as <i>Homo sapiens</i>
percentSyntheticSeqs	Kraken2	Percentage of reads classified as non - biological sequences
readsSyntheticSeqs	Kraken2	Number of reads classified as non - biological sequences
percentCovid	Kraken2	Percentage of reads classified as SARS - CoV - 2
readsCovid	Kraken2	Number of reads classified as SARS - CoV - 2

5.3 Example output for 3 pairs run through the SSQuAWK workflow:

A	B	C	D	E	F	G	H	I	J	K	
Sample	fwdReads	fwdAvgLen	fwdAvgQ	revReads	revAvgLen	revAvgQ	percentHuman	readsHuman	percentSyntheticSeqs	readsSyntheticSeqs	perc
SRR16828363.fastq.qz	316332	75.5	33.21	316332	75.5	31.76	0.48	1517	70.88	224206	2
SRR16828364.fastq.qz	229058	75.5	35.71	229058	75.5	34.81	0.38	863	20.92	47920	3
SRR16828365.fastq.qz	175990	75.5	34.9	175990	75.5	33.79	0.5	874	30.04	52862	1

Video Tutorial

6 A more detailed, 12 minute, video version of this protocol: