Wastewater QC workflow in GalaxyTrakr (SSQuAWK) V.1

Jasmine Amirzadegan[1], Tunc Kayikcioglu[1], hugh.rand [1], Ruth Timme[2], Maria Balkey[1]

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

1

Dec 27, 2021

protocol .

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

Ruth Timme
US Food and Drug Administration

PURPOSE:

Step-by-step instructions for checking sequence quality for SARS-CoV-2 wastewater samples using **SSQuAWK: S**ARS - CoV - 2 **S**equence **Qu**ality **A**ssurance **W**orkflow and **K**ontraption. 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

https://galaxytrakr.org

Jasmine Amirzadegan, Tunc Kayikcioglu, hugh.rand , Ruth Timme, Maria Balkey 2021. Wastewater QC workflow in GalaxyTrakr (SSQuAWK). **protocols.io**
https://protocols.io/view/wastewater-qc-workflow-in-galaxytrakr-ssquawk-b2h4qb8w

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

protocol ,

Dec 01, 2021

Dec 27, 2021

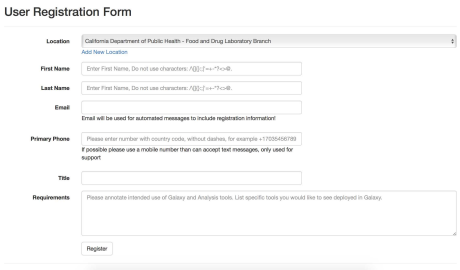
55580

:

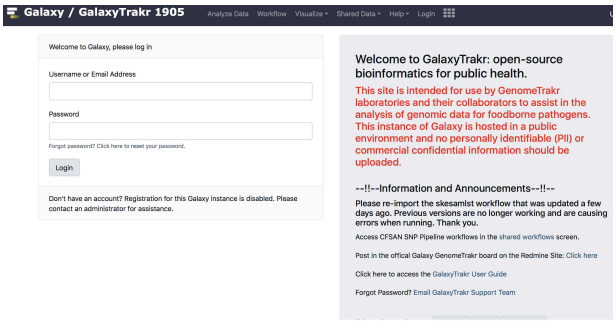
Please note that this protocol is public domain, which supersedes the CC-BY license default used by protocols.io.

Account set up

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



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



Create a new history

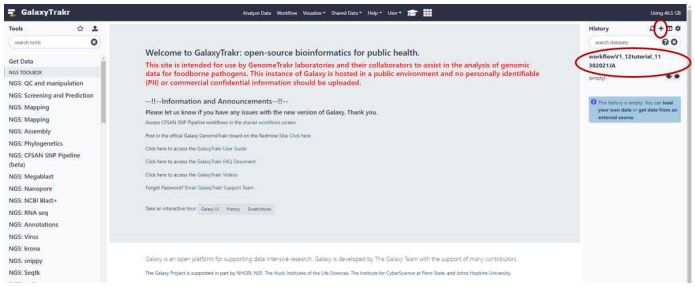
2 **Create a new history.**

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 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.



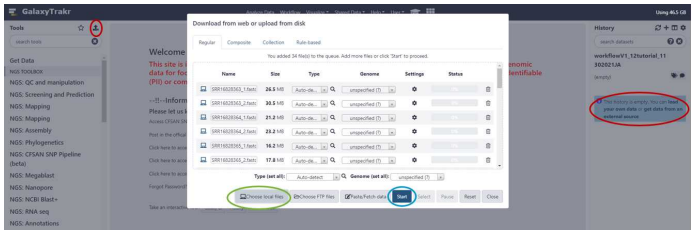
Upload data

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.

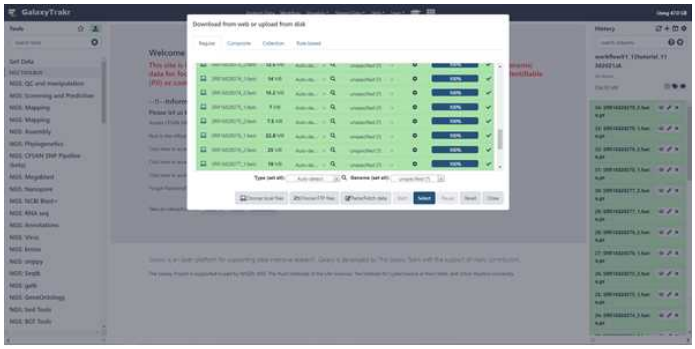
3.1 Upload sequence data to your history, using either of the two options circled in red below.

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.

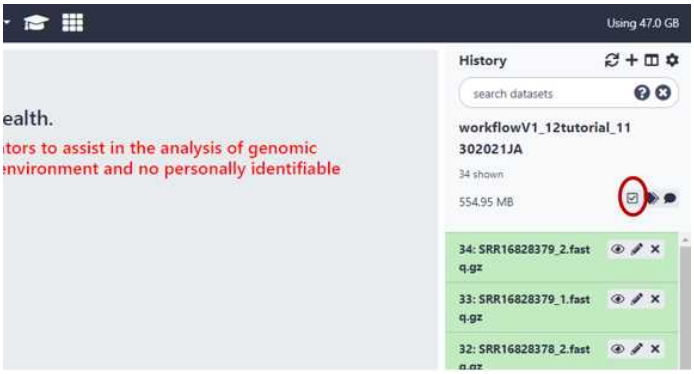


3.2 As the file uploads complete, each row will turn green. If samples are shown with yellow background, then are still uploading.



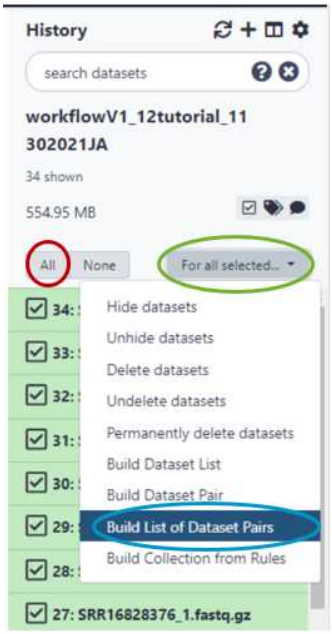
3.3 You have just upload a set of forward and reverse reads. For further analysis these files need to be paired properly so the platform knows which R1 and R2 files go with each sample. GalaxyTrakr does this by creating a **List of Dataset Pairs.**

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).

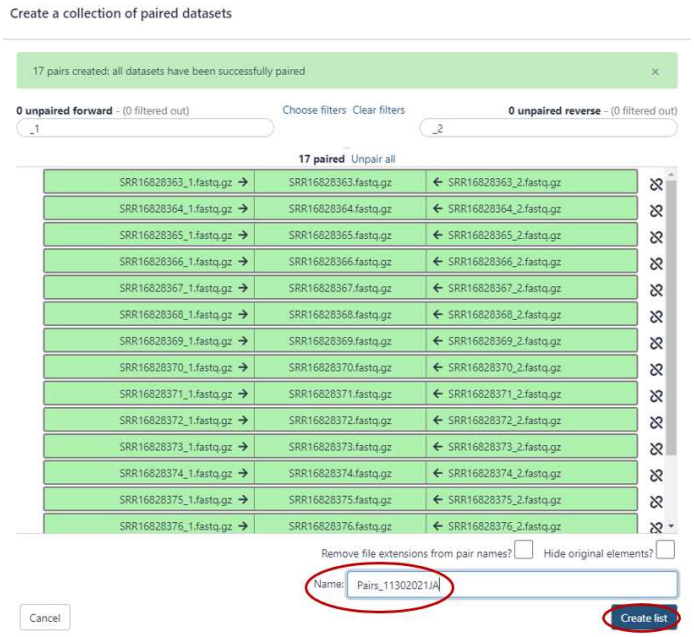
Then, use the "For all selected..." dropdown (circled in green), and click on "Build List of Dataset Pairs" (circled in blue).



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

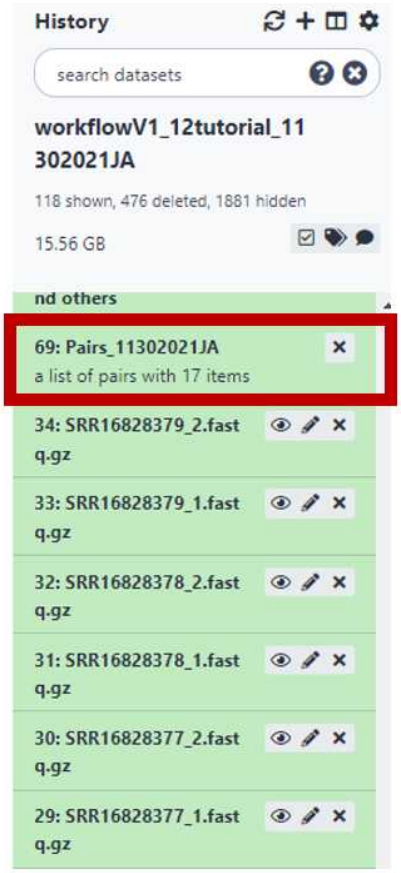
Paired reads will pair in the middle column and turn green.

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



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.

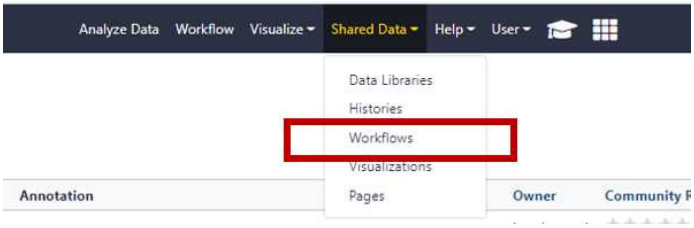


Run the <> workflow

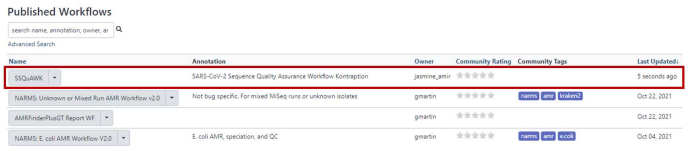
4 **Add the SSQuAWK* workflow to your own "workflows" panel.** You only have to do this step once for each new workflow you need.

***SSQuAWK:** **S**ARS - CoV - 2 **S**equence **Qu**ality **A**ssurance **W**orkflow and **K**ontraption

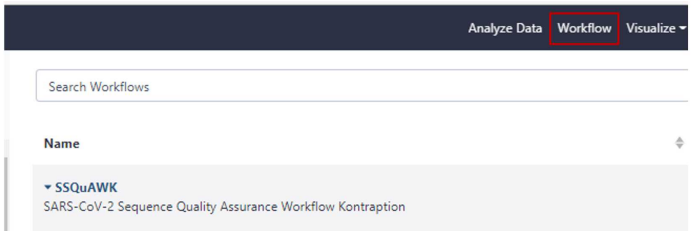
4.1 Navigate to the “Shared Data" drop down and choose workflows



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



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



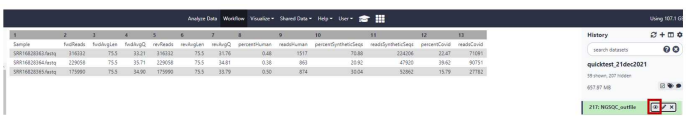
4.3 To use the workflow, press the 'play' button (boxed in red) on the right

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.

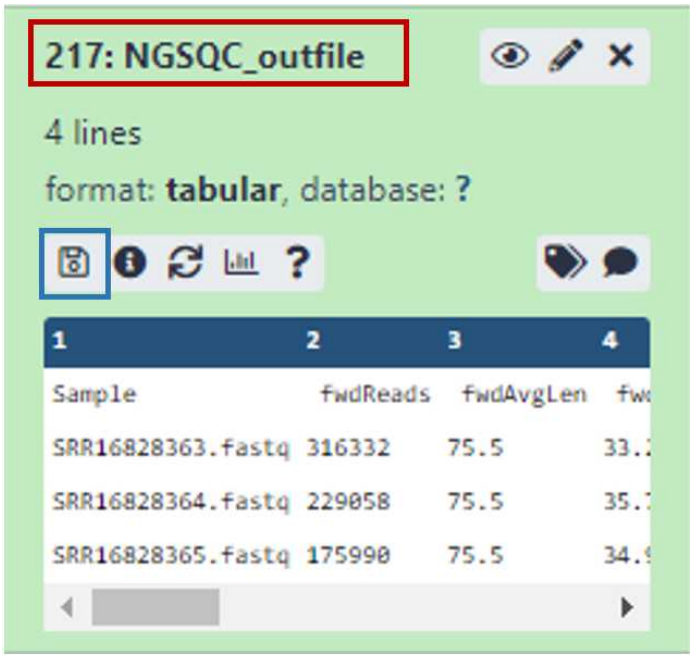
4.5 Upon completion of the pipeline, the NGSQC_outfile will be green. Click on the “Eye” icon to view the output in the GalaxyTrakr window.



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.



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 *Homo sapiens*
readsHuman	Kraken2	Number of reads classified as *Homo sapiens*
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