

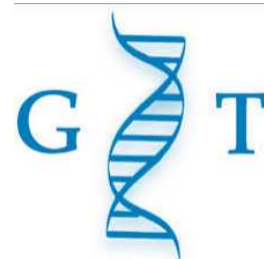
Jan 05, 2022

Version 2

Wastewater QC workflow in GalaxyTrakr (SSQuAWK) V.2

DOI

<https://dx.doi.org/10.17504/protocols.io.b3icqkaw>



Jasmine Amirzadegan¹, tunc.kayikcioglu¹, hugh.rand¹, Ruth E 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

GenomeTrakr

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



Jasmine Amirzadegan

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.b3icqkaw>

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



Protocol Citation: Jasmine Amirzadegan, tunc.kayikcioglu , hugh.rand , Ruth E Timme, Maria Balkey 2022. Wastewater QC workflow in GalaxyTrakr (SSQuAWK). protocols.io <https://dx.doi.org/10.17504/protocols.io.b3icqkaw>Version created by **Ruth Timme**

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: In development

We are still developing and optimizing this protocol

Created: January 05, 2022

Last Modified: January 05, 2022

Protocol Integer ID: 56612

Keywords: WGS, Quality Control, GalaxyTrakr, GenomeTrakr, microbial pathogen survielliance, ssquawk workflow, using ssquawk, sequence quality assurance workflow, ssquawk, galaxytrakr, sequence quality for sar, end fastq file, basic protocol steps with screenshots v2, account in galaxytrakr, wastewater, quality assessments for raw read, checking sequence quality, raw read, basic protocol step, kontraption

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 sequence quality for SARS-CoV-2 wastewater samples using **SSQuAWK: SARS - CoV - 2 Sequence Quality Assurance Workflow** and **Kontraption**. 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



Account set up

1 1. Create a GalaxyTrakr account here:

<https://account.galaxytrakr.org/Account/Register>

User Registration Form

The registration form includes the following fields and instructions:

- Location:** A dropdown menu showing "California Department of Public Health - Food and Drug Laboratory Branch" with a link to "Add New Location".
- First Name:** Text input with placeholder "Enter First Name, Do not use characters: /@[]-?+=*~?<>@".
- Last Name:** Text input with placeholder "Enter First Name, Do not use characters: /@[]-?+=*~?<>@".
- Email:** Text input with a note: "Email will be used for automated messages to include registration information!".
- Primary Phone:** Text input with a note: "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:** Text input.
- Requirements:** Text area with a note: "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 login page features a navigation bar at the top with the title "Galaxy / GalaxyTrakr 1905" and links for "Analyze Data", "Workflow", "Visualize", "Shared Data", "Help", and "Login".

The main content area is split into two columns:

- Left Column (Login Form):**
 - Header: "Welcome to Galaxy, please log in"
 - Field: "Username or Email Address" with a text input.
 - Field: "Password" with a text input.
 - 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 Column (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

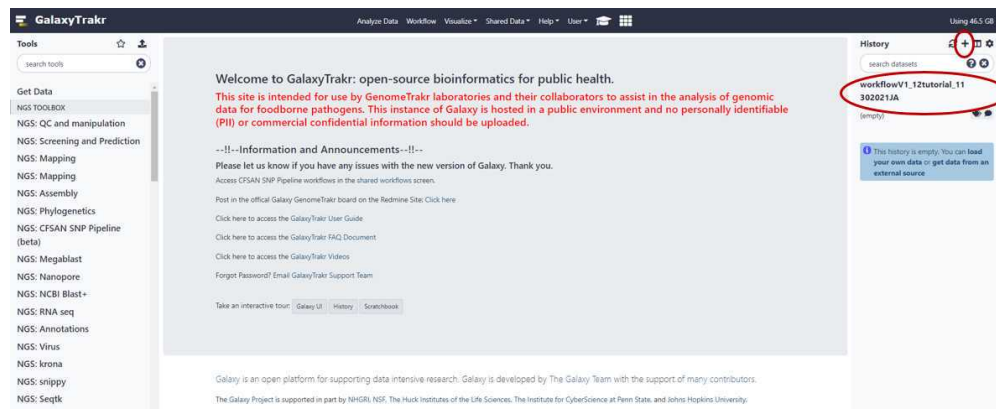
2 Create a new history.

We recommend creating a new history for each new MiSeq 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.



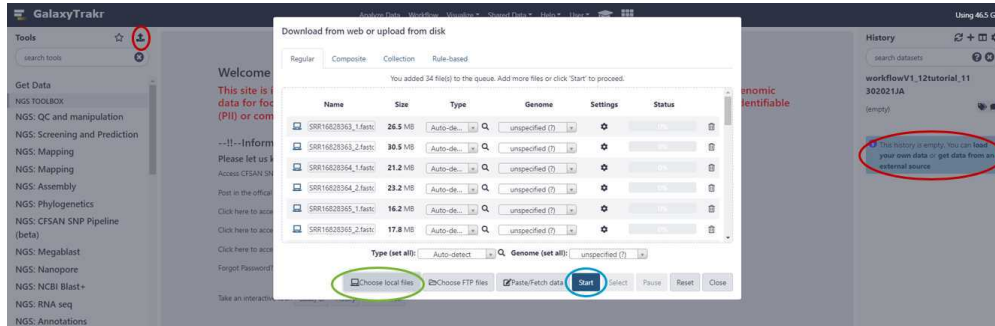
Upload data

- 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.

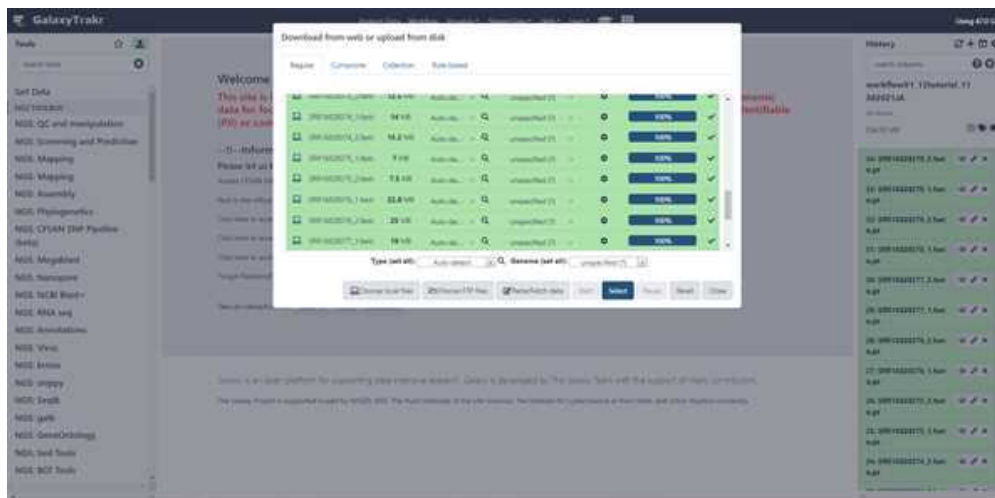
A window will appear in the middle of your screen. This is where you select your files using the "Choose local files" button at the bottom of the window. The "Choose local

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.

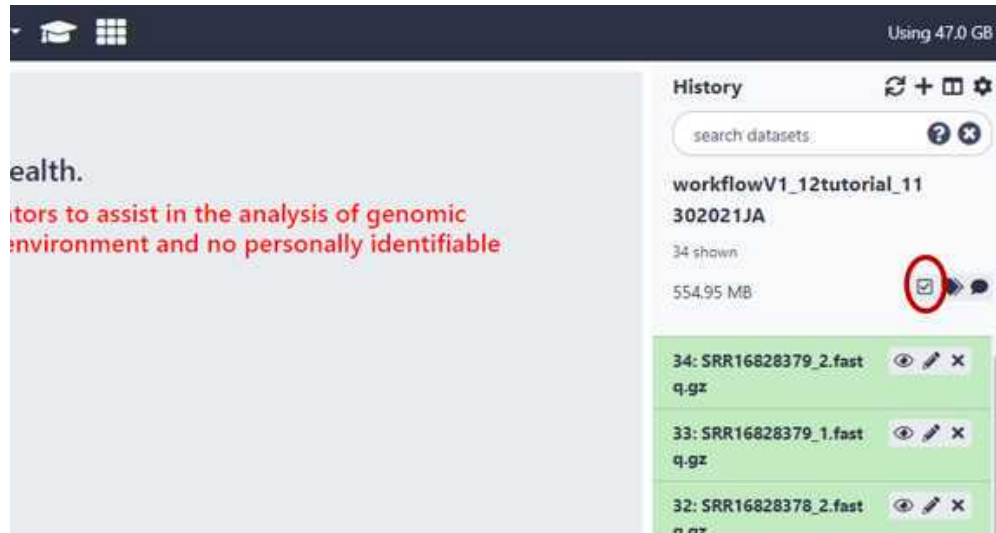


3.2 As the file uploads complete, each row will turn green. If samples are shown with yellow background, then are still uploading.



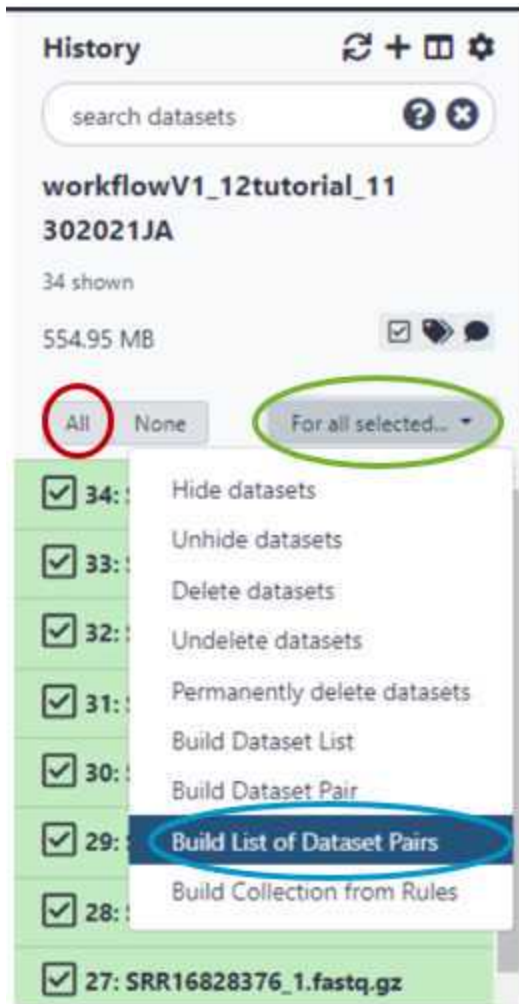
3.3 You have just upload a set of forward and reverse reads. For further analysis these files need to be paired properly so the platform knows which R1 and R2 files go with each sample. GalaxyTrakr does this by creating a **List of Dataset Pairs**.

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

17 pairs created: all datasets have been successfully paired

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

_1 _2

17 paired Unpair all

SRR16828363_1.fastq.gz	→	SRR16828363.fastq.gz	←	SRR16828363_2.fastq.gz	⌗
SRR16828364_1.fastq.gz	→	SRR16828364.fastq.gz	←	SRR16828364_2.fastq.gz	⌗
SRR16828365_1.fastq.gz	→	SRR16828365.fastq.gz	←	SRR16828365_2.fastq.gz	⌗
SRR16828366_1.fastq.gz	→	SRR16828366.fastq.gz	←	SRR16828366_2.fastq.gz	⌗
SRR16828367_1.fastq.gz	→	SRR16828367.fastq.gz	←	SRR16828367_2.fastq.gz	⌗
SRR16828368_1.fastq.gz	→	SRR16828368.fastq.gz	←	SRR16828368_2.fastq.gz	⌗
SRR16828369_1.fastq.gz	→	SRR16828369.fastq.gz	←	SRR16828369_2.fastq.gz	⌗
SRR16828370_1.fastq.gz	→	SRR16828370.fastq.gz	←	SRR16828370_2.fastq.gz	⌗
SRR16828371_1.fastq.gz	→	SRR16828371.fastq.gz	←	SRR16828371_2.fastq.gz	⌗
SRR16828372_1.fastq.gz	→	SRR16828372.fastq.gz	←	SRR16828372_2.fastq.gz	⌗
SRR16828373_1.fastq.gz	→	SRR16828373.fastq.gz	←	SRR16828373_2.fastq.gz	⌗
SRR16828374_1.fastq.gz	→	SRR16828374.fastq.gz	←	SRR16828374_2.fastq.gz	⌗
SRR16828375_1.fastq.gz	→	SRR16828375.fastq.gz	←	SRR16828375_2.fastq.gz	⌗
SRR16828376_1.fastq.gz	→	SRR16828376.fastq.gz	←	SRR16828376_2.fastq.gz	⌗

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

Name: Pairs_11302021JA

Cancel Create list

Alternatively, instead of auto-pairing 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 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

workflowV1_12tutorial_11

302021JA

118 shown, 476 deleted, 1881 hidden

15.56 GB

and others

69: Pairs_11302021JA

a list of pairs with 17 items

34: SRR16828379_2.fast

q.gz

33: SRR16828379_1.fast

q.gz

32: SRR16828378_2.fast

q.gz

31: SRR16828378_1.fast

q.gz

30: SRR16828377_2.fast

q.gz

29: SRR16828377_1.fast

q.gz

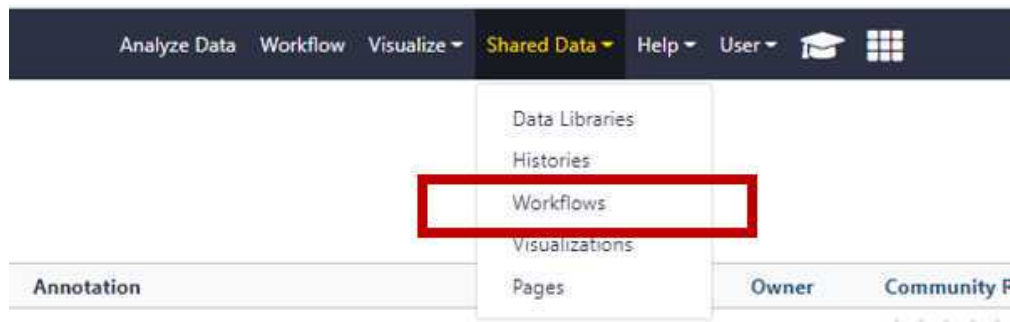
Run the SSQuAWK workflow



- 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



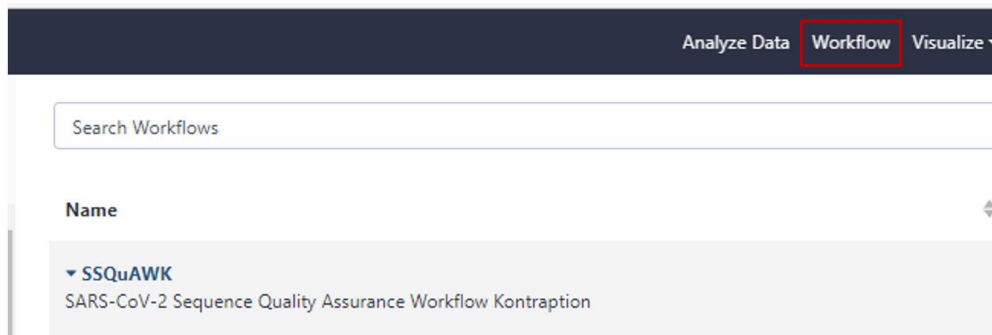
Then, from the SSQuAWK drop down menu, select import.

Published Workflows

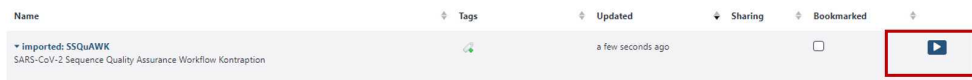
search name, annotation, owner, or
Advanced Search

Name	Annotation	Owner	Community Rating	Community Tags	Last Updated
SSQuAWK	SARS-CoV-2 Sequence Quality Assurance Workflow Kontrapion	jasmine_amir	★★★★★		5 seconds ago
NARMS: Unknown or Mixed Run AMR Workflow v2.0	Not bug specific. For mixed MiSeq runs or unknown isolates	gmartin	★★★★★	narms amr kraken2	Oct 22, 2021
AMRfinderPlusGT Report WF		gmartin	★★★★★		Oct 22, 2021
NARMS: E. coli AMR Workflow V2.0	E. coli AMR, speciation, and QC	gmartin	★★★★★	narms amr ecoli	Oct 04, 2021

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

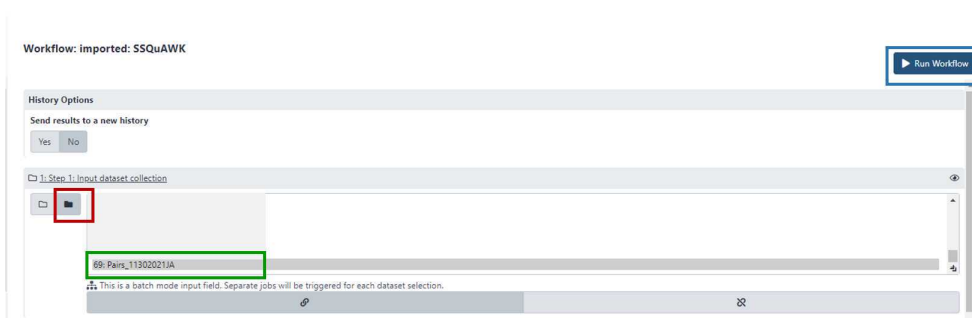


4.3 To use the workflow, press the 'play' button (boxed in red) on the right



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.

The screenshot shows the GalaxyTrakr interface. On the left is a table with 13 columns. The first column is 'Sample'. The next three columns are 'fastq' files. The next three columns are 'rev' files. The next three columns are 'percent' values. The last three columns are 'reads' values. On the right is a 'History' panel showing a search for 'quicktest_21dec2021' and a list of jobs. The job '217: NGSQC_outfile' is highlighted in green and has a red box around its 'Eye' icon.

1	2	3	4	5	6	7	8	9	10	11	12	13
Sample	fastq1	fastq2	fastq3	rev1	rev2	rev3	percentHuman	readsHuman	percentSyntheticSeqs	readsSyntheticSeqs	percentCovid	readsCovid
SR16828363.fastq	316332	75.5	33.21	316332	75.5	31.76	0.48	1517	70.88	224206	22.47	71091
SR16828364.fastq	229058	75.5	35.71	229058	75.5	34.81	0.38	863	20.92	47920	39.62	90751
SR16828365.fastq	175990	75.5	34.90	175990	75.5	33.79	0.50	874	30.04	52862	15.79	27782

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.

1	2	3	4
Sample	fwdReads	fwdAvgLen	fwdAvgQ
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

A	B	C
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	L	M
Sample	fw dR eads	fw d Avg Len	fw dA vg Q	rev Re ads	rev Av gL en	rev Avg Q	per cen tHu man	rea ds Hu man	perc entS ynth eticS eqs	read sSyn theti cSeq s	per cen tCo vid	rea ds Co vid
SRR16828363.fastq.qz	316332	75.5	33.21	316332	75.5	31.76	0.48	1517	70.88	224206	22.47	71091
SRR16828364.fastq.qz	229058	75.5	35.71	229058	75.5	34.81	0.38	863	20.92	47920	39.62	90751



	A	B	C	D	E	F	G	H	I	J	K	L	M
	SRR1 6828 365.f astq. qz	175 99 0	75. 5	34. 9	175 99 0	75. 5	33. 79	0.5	87 4	30.0 4	5286 2	15. 79	27 78 2

Video Tutorial

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



SSQUAWK

SARS – CoV – 2 Surveillance Quality Assurance Workflow and Kontraption

Center for Food Safety and Applied Nutrition

Jasmine Amirzadegan, Tunc Kayikcioglu, Hugh Rand, Ruth E. Timme, and Maria Balkey