Mar 31, 2022 Version 4

Wastewater QC workflow in GalaxyTrakr (SSQuAWK3) V.4

DOI
https://dx.doi.org/10.17504/protocols.io.kxygxzk5dv8j/v4

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

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

Jasmine Amirzadegan

OPEN ACCESS

DOI: https://dx.doi.org/10.17504/protocols.io.kxygxzk5dv8j/v4

External link: https://galaxytrakr.org

Protocol Citation: Jasmine Amirzadegan, Tunc Kayikcioglu, hugh.rand , Ruth Timme, Maria Balkey 2022. Wastewater QC workflow in GalaxyTrakr (SSQuAWK3). **protocols.io https://dx.doi.org/10.17504/protocols.io.kxygxzk5dv8j/v4**Version created by **Jasmine Amirzadegan**

Protocol status: In development
We are still developing and optimizing this protocol

Created: March 31, 2022

Last Modified: March 31, 2022

Protocol Integer ID: 60115

Keywords: WGS, Quality Control, GalaxyTrakr, GenomeTrakr, microbial pathogen survielliance, wastewater qc workflow in galaxytrakr, ssquawk3 workflow, new genome mapping metrics protocol v4 ssquawk version, underlying galaxytrakr step, wastewater qc workflow, using ssquawk3, underlying galaxytrakr job, galaxytrakr step, galaxytrakr job, ssquawk version, ssquawk3, sequence quality assurance workflow, ssquawk, galaxytrakr, sequence quality for sar, basic protocol steps with screenshots protocol v2, fastq file, account in galaxytrakr, end fastq file, checking sequence quality, quality assessments for raw read, wastewater sample, screenshots protocol v2, wastewater, qc placeholder columns for sra metadata template, raw read, workflow, detailed 12 minute video tutorial protocol v3, cfsan, basic protocol step

Disclaimer

Abstract

PURPOSE:

Step-by-step instructions for checking sequence quality for SARS-CoV-2 wastewater samples using **SSQuAWK3:** **S**ARS - CoV - 2 **S**equence **Qu**ality **A**ssurance **W**orkflow and **K**ontraption, version **3**. The SSQuAWK3 workflow, implemented in CFSAN's custom Galaxy instance (GalaxyTrakr) will produce quality assessments for raw reads (Illumina MiSeq paired-end fastq files).

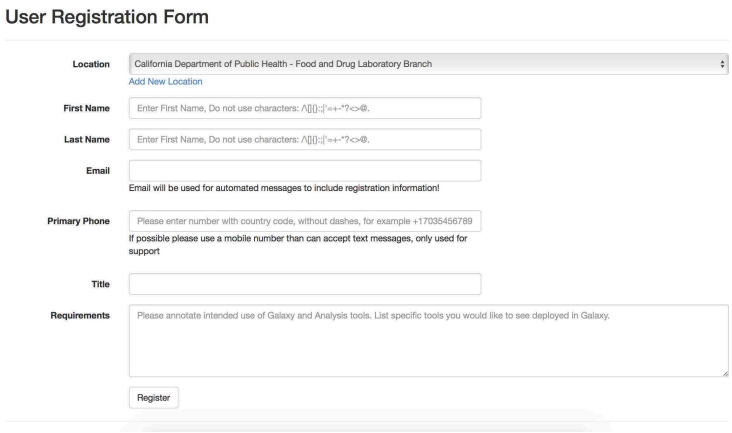
SCOPE: This protocol covers the following tasks:

1. Set up an account in GalaxyTrakr
2. Create a new history
3. Upload data and reference files
4. Execute the SSQuAWK3 workflow
5. Interpret the results

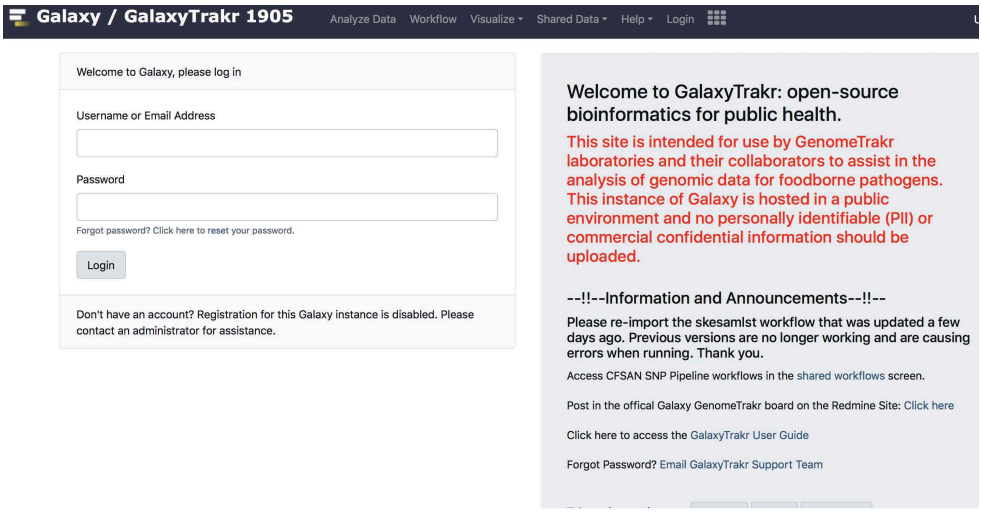
Protocol and SSQuAWK workflow version history:
Protocol V1, SSQuAWK version 1: Basic protocol steps with screenshots
Protocol V2, SSQuAWK version 1: Addition of a detailed 12 minute video tutorial
Protocol V3, SSQuAWK version 2: Addition of 5 new genome mapping metrics
Protocol V4 SSQuAWK version 3: Metrics now reported with fewer softwares, fewer underlying GalaxyTrakr jobs, and about 50% fewer underlying GalaxyTrakr steps. Cleaner output table formats now include QC placeholder columns for SRA metadata template.

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

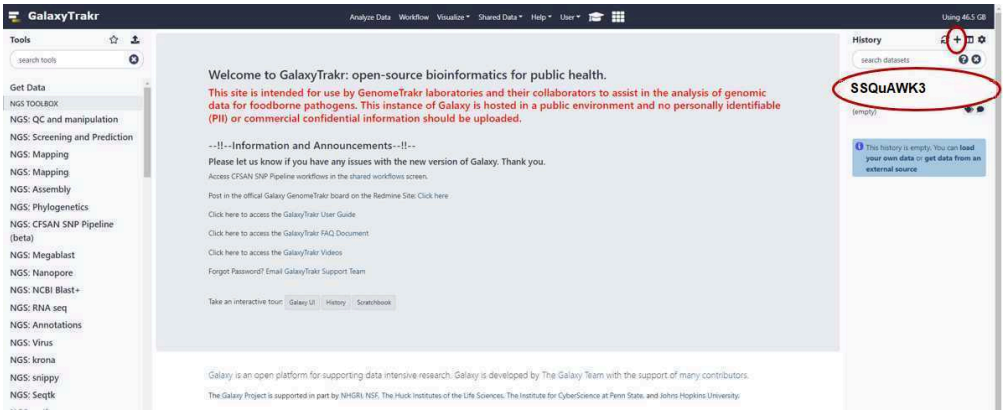
2 **Create a new history.**

We recommend creating a new history for each new MiSeq sequence set with details and date in the history name.

Save your SSQuAWK2 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 sequence data

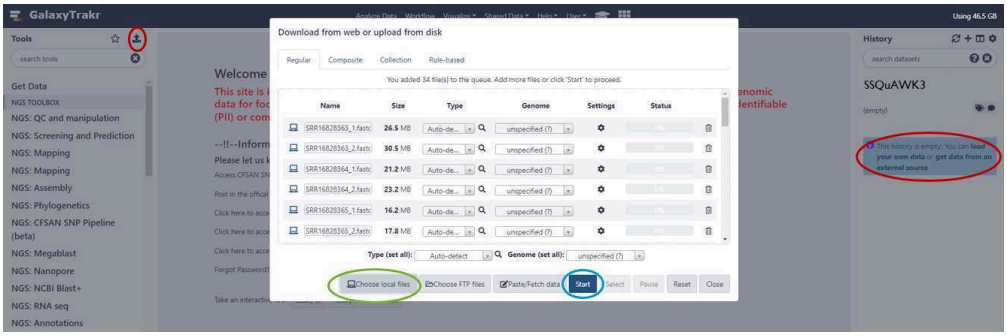
3 **This section will describe the process for uploading raw fastq files into your active History panel.** After the files have been uploaded they will stay in your account until they are deleted.

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

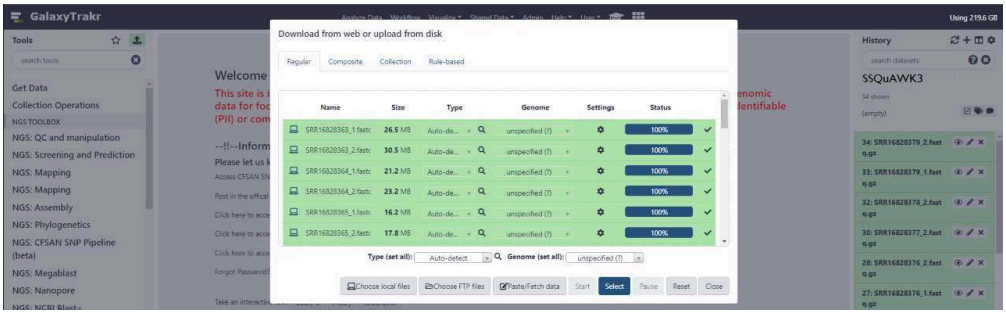
A window will appear in the middle of your screen. This is where you select your files using the "Choose local files" button at the bottom of the window. The "Choose local

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.

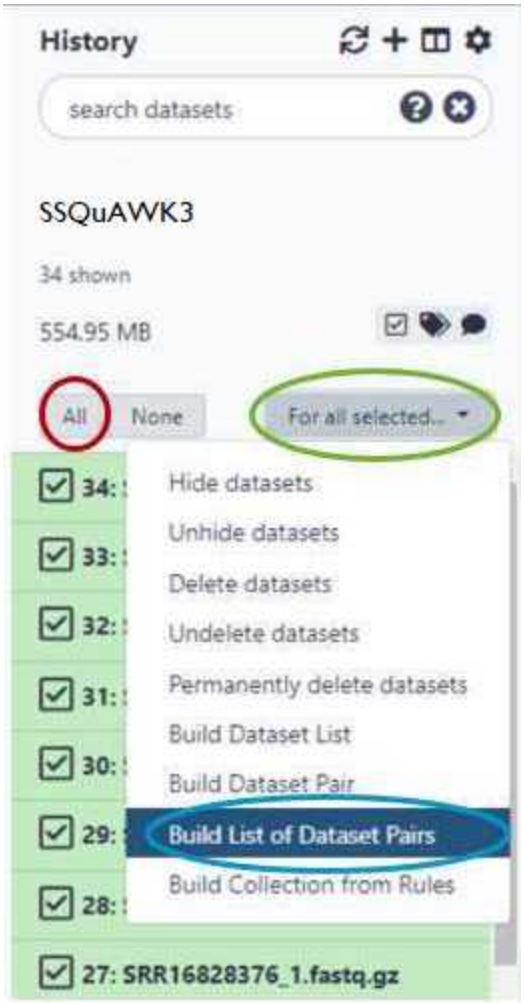


3.3 You have just upload a set of forward and reverse reads. For further analysis these files need to be paired properly so the platform knows which R1 and R2 files go with each sample. GalaxyTrakr does this by creating a **List of Dataset Pairs.**

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).

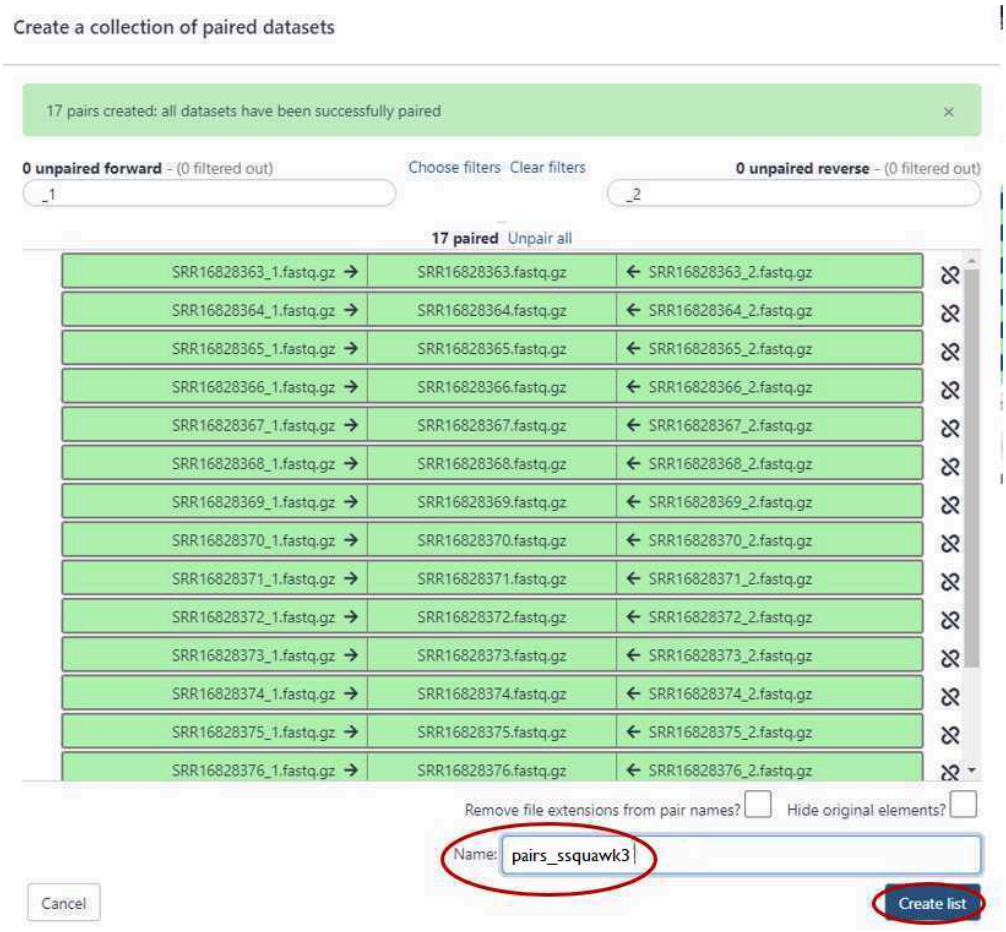
Then, use the "For all selected..." dropdown (circled in green), and click on "Build List of Dataset Pairs" (circled in blue).



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.

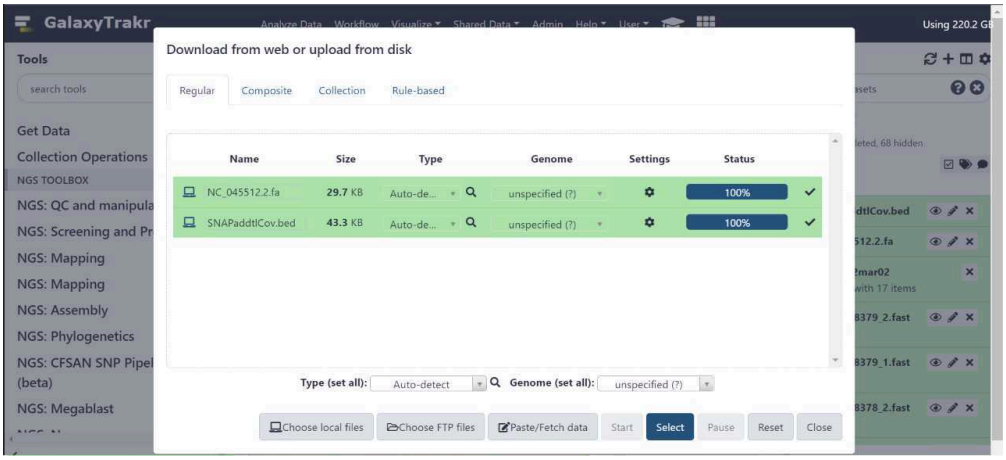
If everything looks good, then choose a name for your pairs (circled red) and "Create List" (also circled red).



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.

4 To the existing history, also upload (1) the **provided reference.fasta file** and (1) a primer.bed file.



4.1 **SSQuAWK2 is only compatible with the 22903 nt reference genome file obtained from NCBI 'NC_045512.2'. It is provided here for your convenience:** NC_045512.2.fa

4.2 **The primer.bed file should correspond to the SARS - CoV - 2 enrichment primer panel kit used.**

QIAseqDIRECT.bed QIAseq Direct kit

SNAPStd.bed SNAP standard kit

SNAPaddtlCov.bed SNAP additional coverage kit

varskipShort.bed NEB VarSkip Short (version 1) kit

VSSv2.primer.bed NEB VarSkip Short (version 2) kit

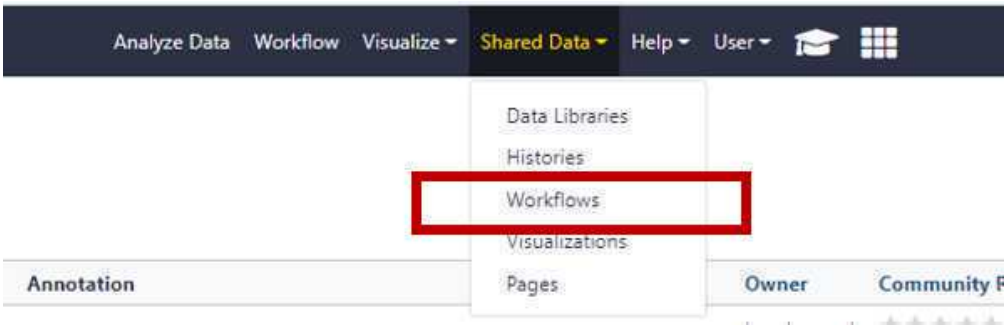
ARTICv4.bed ARTIC v4 primer schemes

Run the SSQuAWK workflow

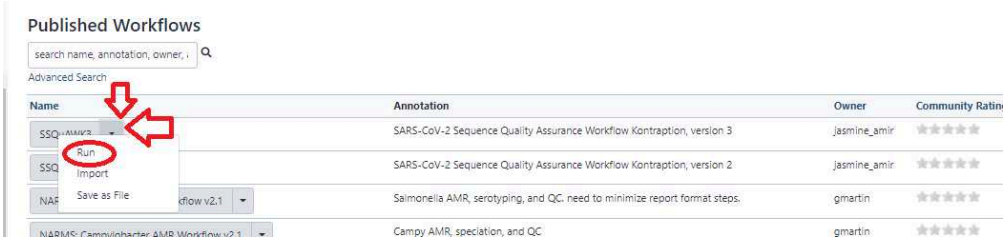
5 **Access the SSQuAWK3* workflow with the "workflows" panel.**

***SSQuAWK3: S**ARS - CoV - 2 **S**equence **Qu**ality **A**ssurance **W**orkflow **K**ontraption, version **3**

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



Then, from the SSQuaWK3 drop down menu, select "Run".



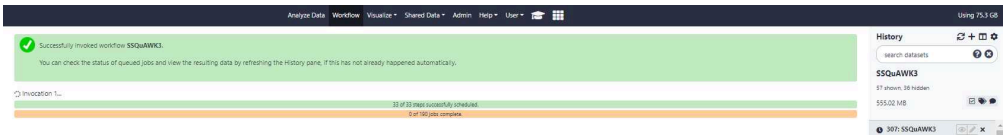
5.2 Select the paired list you created earlier by selecting the folder icon (boxed in red), and then the list of pairs (boxed in green).

Boxed in gold: Select the reference fasta file from your history.

Boxed in blue: Select the bed file from your history.

Click Run Workflow (boxed in purple).

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.



Run the SSQuAWK workflow

5.3 Upon completion of the pipeline, the output file for **SSQuAWK3** will be green. Click on the “Eye” icon to view in GalaxyTrakr window.

Interpret the results

6 **Download and interpret the results:**

6.1 Click the output file text for "**SSQuAWK3"** (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 in Excel.

6.2 The SSQuAWK3 output file includes the following metrics:

	A	B	C
	Parameter	***Input***	***Description***
	Sample	List of Pairs	Sample name from list of pairs
	0xGenomeCov	Bowtie2, samtools, ivar_trim	Percentage of nucleotides that do not cover the genome at all (zero times)
	<10xGenomeCov	Bowtie2, samtools, ivar_trim	Percentage of nucleotides that barely cover the genome (less than 10 times)
	nReads	Bowtie2	Total number of reads
	avgLen	Bowtie2, samtools	Average read length

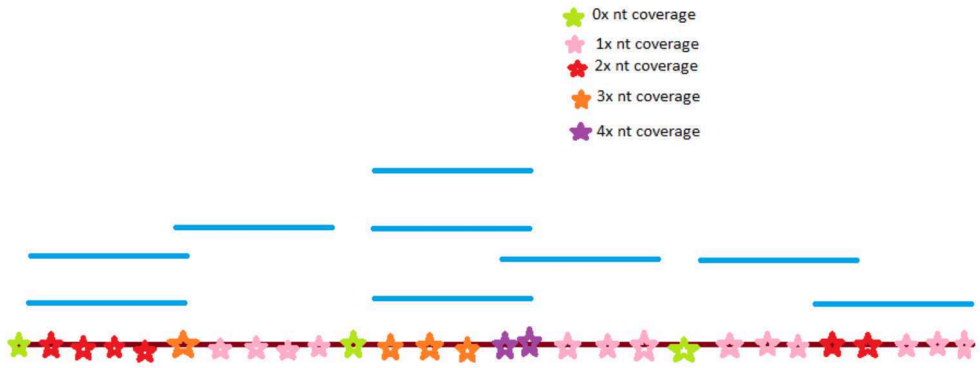
	A	B	C
	avgLenPassFilt	Bowtie2, samtools, ivar_trim	Average read length after iVar_trim filtering*
	avgQual	Bowtie2, samtools	Average read quality
	avgQualPassFilt	Bowtie2, samtools, ivar_trim	Average read length after iVar_trim filtering*
	avgCovPassQual	Bowtie2, samtools, ivar_trim	Average number and percentage of nts from sequence reads that map to the genome
	readsAlign	Bowtie2, samtools	Number and percentage of reads that aligned to the reference sequence.
	readsAlignPassFilt	Bowtie2, samtools, ivar_trim	Number and percentage of reads that aligned to the reference sequence after iVar_trim filtering*.
	humanReads	Kraken2	Number and percentage of reads classified as *Homo sapiens*
	SARS-CoV-2Reads	Kraken2	Number and percentage of reads classified as SARS - CoV - 2
	syntheticSeqsReads	Kraken2	Number and percentage of reads classified as non - biological sequences
	quality_control_method_name	SSQuAWK	Name of the method or pipline used to evaluate sequence quality
	quality_control_method_version	3.0	Version number of the quality control pipeline or method used
	quality_control_determination		Result of the quality control accessment. Blank if pass/fail thresholds have not been estabolished or "sequence flagged for potential quality control issues" if relevant.
	quality_control_issues		More information for sequences that have a QC flag issue

*** The iVar_trim filter parameters: minReadLen = 30, minQual_slidingWindow = 20, and slidingWindow = 4 nt.**

6.3 What is nucleotide coverage?! Let's look at 2 simple pictures

In the figure above, let the burgundy line represent the entire reference genome.

The blue lines are the reads, as sequenced nucleotides.



In the figure above, each star, drawn on the burgundy line (reference genome) is a **nucleotide position**.

There are 28 stars, so we will say our genome is 28 nucleotides long.

We can use coverage to determine the quality of our sequences (blue lines).

The lime green stars along the genome represent 0X coverage, because we did not sequence any reads with **nucleotides positions covering that reference nucleotide position.** There are no blue lines that we sequenced there!

There are 3 nucleotide positions with 0x coverage. The total genome is 28 nucleotides long.

*percent_nt0Xcov = (nucleotidePositions0Xcov / genomeLength) * 100*

*percent_nt0Xcov = (3 / 28) *100*

percent_nt0xcov = 10.71%

In most ideal scenarios, higher coverage indicates better sequence quality.

For example, 100x coverage is better than 10x coverage.

Since we want **higher coverage**, percent_nt0Xcov and percent_ntLess10Xcov are ideally **lower percentages.**

0x coverage and 10x coverage indicate "no coverage" and "poor coverage", respectively.

Generally, we expect avgReadCov in 100's or 1000's*

If **percent_nt0Xcov** is a higher percentage, say 50%*, that means half of the genome was not covered by our sequences. The quality is not ideal.

** These values are not official threshold and only used for illustrative purposes.*

Threshold guidance is 'in progress', and planned to be announced after further analyses.

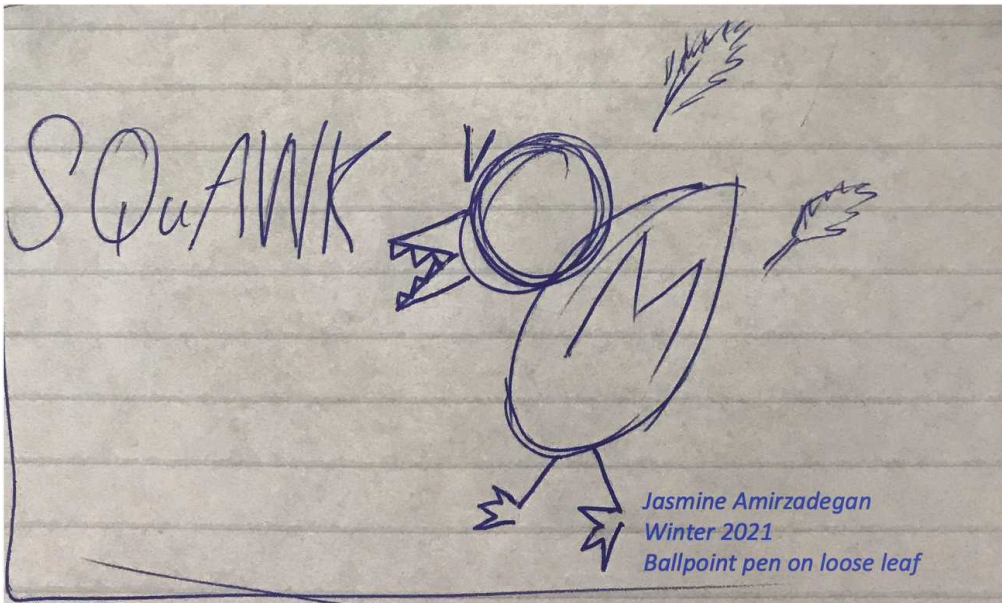
6.4 Example output for the first 3 pairs run through the SSQuAWK3 workflow:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
	Sample	0xGenomeCov	<10xGenomeCov	nReads	avgLen	avgLenPassFilt	avgQual	avgQualPassFilt	avgCovPassQual	readsAlign	readsAlignPassFilt	humanReads	SARS-CoV-2Reads	syntheticSeqsReads	quality_control_method_name	quality_control_method_version	quality_control_determination	quality_control_determination
	SRR16828363.fastq.gz	107nt (0%)	31nt (0%)	6326664	151	151	37.8	37.9	6888X	13863637 (21%)	13632727 (21%)	1517reads (0.48%)	71091reads (22.47%)	224206reads (70.88%)	SSQuAWK	3.0		
	SRR16828364.fastq.gz	76nt (0%)	31nt (0%)	4588116	151	151	37.8	37.9	8900X	17991913 (39%)	17634848 (38%)	863reads (0.38%)	90751reads (39.62%)	479220reads (20.92%)	SSQuAWK	3.0		

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
	SRR16828365.fastq.gz	76nt (0%)	77nt (0%)	3519800	151	151	37.8	37.9	272X	549288 (15%)	539588 (15%)	874reads (0.50%)	27782reads (15.79%)	52862reads (30.04%)	SSQuAWK	3.0		

Video Tutorial

7 Thanks for using SSQuAWK!



8 **New to GalaxyTrakr? Check out this detailed, 12 minute video overview of the SSQuAWK (version 1) protocol before trying SSQuAWK3.**

Video edit:
*"SSQuawk allows users to check the sequence quality of SARS-CoV-2 wastewater samples in **CFSAN's custom Galaxy instance, called GalaxyTrakr.** This generates a single report file from raw Illumina MiSeq paired-end fastq file inputs."*