



Version 9

Oct 05, 2022

Wastewater QC workflow in GalaxyTrakr (SSQuAWK4) V.9

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

In Development

Share

dx.doi.org/10.17504/protocols.io.kxygzk5dv8j/v9

GenomeTrakr

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

Jasmine Amirzadegan

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 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 SSQuAWK workflow
5. Interpret the results

Protocol and SSQuAWK workflow version history:

- **Protocol V9 SSQuAWK version 4: Protocol version 9 includes some additional primer bed files. All SSQuAWK4 protocol steps listed in this guide are applicable to SSQuAWK4.0.2.**
- *Protocol V8 SSQuAWK version 4: Protocol version 8 has minor text corrections. The SSQuAWK version 4 workflow required minor edits on an updated backend system of GalaxyTrakr, and thus is now labeled as "SSQuAWK4.0.2" in GalaxyTrakr. All SSQuAWK4 protocol steps listed in this guide are applicable to SSQuAWK4.0.2.*
- *Protocol V7 SSQuAWK version 4: Protocol now includes a QC determination guidance table. The SSQuAWK version 4 workflow required minor bug edits on the backend, thus is now labeled as "SSQuAWK4.0.1" in GalaxyTrakr. All SSQuAWK4 protocol steps listed in this guide are applicable to SSQuAWK4.0.1.*
- *Protocol V6 SSQuAWK version 4: Best practice guidance on fastq.gz file uploads and new QC metric.*
- *Protocol V5 SSQuAWK version 3: Previous protocol version had broken links for FASTA and BED files, this version fixes the links.*
- *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.*
- *Protocol V3, SSQuAWK version 2: Addition of 5 new genome mapping metrics*
- *Protocol V2, SSQuAWK version 1: Addition of a detailed 12 minute video tutorial*
- *Protocol V1, SSQuAWK version 1: Basic protocol steps with screenshots*

Participation was supported by the:

- Research Participation Program at the U.S. Food and Drug Administration administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy and the U.S. Food and Drug Administration.
- Joint Institute for Food Safety and Applied Nutrition (JIFSAN), University of Maryland by the Food and Drug Administration (FDA) of the U.S. Department of Health and Human Services (HHS).
- The American Rescue Plan Act of 2021, Congress provided temporary funding for FDA to develop the capacity to sequence SARS-CoV-2 RNA from wastewater samples and to conduct a sampling and sequencing project through 2022.

DOI

dx.doi.org/10.17504/protocols.io.kxygzk5dv8j/v9

EXTERNAL LINK

<https://galaxytrakr.org>

Jasmine Amirzadegan, Tunc Kayikcioglu, hugh.rand, Ruth Timme, Maria Balkey 2022. Wastewater QC workflow in GalaxyTrakr (SSQuAWK4). **protocols.io**
<https://dx.doi.org/10.17504/protocols.io.kxygxzk5dv8j/v9>
 Version created by Jasmine Amirzadegan

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

_____ 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

Oct 05, 2022

Oct 05. 2022

70879


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

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	<input type="text"/>
	<small>Enter First Name. Do not use characters /[](){}~+=-!*<>@.</small>
Last Name	<input type="text"/>
	<small>Enter Last Name. Do not use characters /[](){}~+=-!*<>@.</small>
Email	<input type="text"/>
	<small>Email will be used for automated messages to include registration information!</small>
Primary Phone	<input type="text"/>
	<small>Please enter number with country code, without dashes, for example +77096809788 if possible please use a mobile number than can accept text messages, only used for support</small>
Title	<input type="text"/>
Requirements	<small>Please annotate intended use of Galaxy and Analysis tools. List specific tools you would like to see deployed in Galaxy</small>
	<input type="button" value="Register"/>

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

 Galaxy / GalaxyTrakr 2015

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 keasnikam 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 official Galaxy GenomeTrakr board on the Redmine site: Click here

Click here to access the GalaxyTrakr User Guide

Forgot Password? Email GalaxyTrakr Support Team

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

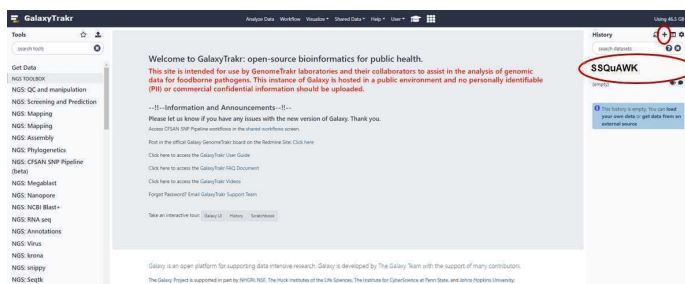
be purged/deleted so as not to occupy the limited storage space in your account.

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.

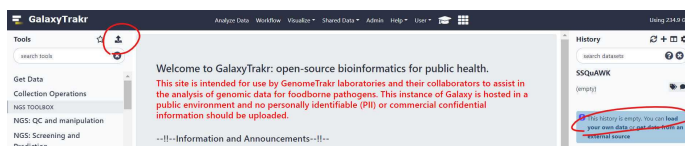
- 2.1 Create a new history with the "+" symbol in the upper right hand corner. Name your history and press "enter" on your keyboard to save the name.



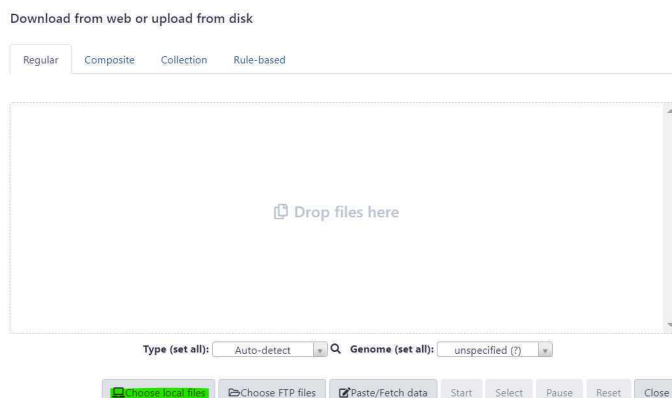
Upload Sequence Data

- 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 highlighted in green.



- 3.2 ⚠

Before initiating the file upload, double check that the file "Type" is appropriately set.

Do not use the "Auto - Detect" option.

Sequence data uploaded using the "auto - detect" option may be subject to file corruption.

This can result in various downstream analysis issues, including empty output metrics and job errors.

Download from web or upload from disk

Regular Composite Collection Rule-based

You added 34 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status
16828363_1.fastq.gz	26.5 MB	Auto-de...	unspecified (?)		0%
SRR16828363_2.fastq	30.5 MB	Auto-de...	unspecified (?)		0%
SRR16828364_1.fastq	21.2 MB	Auto-de...	unspecified (?)		0%
SRR16828364_2.fastq	23.2 MB	Auto-de...	unspecified (?)		0%
SRR16828365_1.fastq	16.2 MB	Auto-de...	unspecified (?)		0%
SRR16828365_2.fastq	17.8 MB	Auto-de...	unspecified (?)		0%

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

Choose local files Choose FTP files Paste/Fetch data Start Select Pause Reset Close

Instead, use the "Type (set all)" dropdown to select the correct file type.

Download from web or upload from disk

Regular Composite Collection Rule-based

You added 34 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status
16828363_1	26.5 MB	fastq.gz	unspecified (?)		0%
SRR16828363_2.fastq	30.5 MB	fastq.gz	unspecified (?)		0%
SRR16828364_1.fastq	21.2 MB	fastq.gz	unspecified (?)		0%
SRR16828364_2.fastq	23.2 MB	fastq.gz	unspecified (?)		0%
SRR16828365_1.fastq	16.2 MB	fastq.gz	unspecified (?)		0%
SRR16828365_2.fastq	17.8 MB	fastq.gz	unspecified (?)		0%

Type (set all): fastq.gz Genome (set all): unspecified (?)

Choose local files Choose FTP files Paste/Fetch data Start Select Pause Reset Close

Once the file type is set, press "Start" to initiate your data upload to GalaxyTrakr. The "Start" button is circled in green.

Download from web or upload from disk

Regular Composite Collection Rule-based

You added 34 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status
SRR16828363_1.fastq	26.5 MB	fastq.gz	unspecified (?)		0%
SRR16828363_2.fastq	30.5 MB	fastq.gz	unspecified (?)		0%
SRR16828364_1.fastq	21.2 MB	fastq.gz	unspecified (?)		0%
SRR16828364_2.fastq	23.2 MB	fastq.gz	unspecified (?)		0%
SRR16828365_1.fastq	16.2 MB	fastq.gz	unspecified (?)		0%
SRR16828365_2.fastq	17.8 MB	fastq.gz	unspecified (?)		0%

Type (set all): fastq.gz Genome (set all): unspecified (?)

Choose local files Choose FTP files Paste/Fetch data Start Select Pause Reset Close

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

3.6 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)

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_ssquawk

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)

Choose from the following filters to change which unpaired reads are shown in the display:

Forward: _1, Reverse: _2

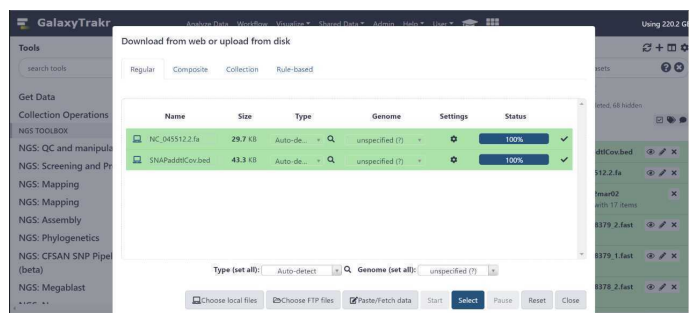
Forward: _R1, Reverse: _R2

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.

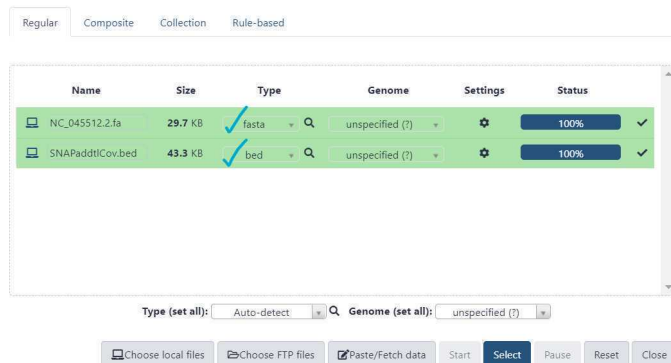


Upload reference data

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



Download from web or upload from disk



- 4.1 **SSQuAWK 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.**

QIAseq Direct:  [QIAseqDIRECT.bed](#)

QIAseq Direct Boosted:  [QIAseqDIRECT_booster.bed](#)

SNAP standard kit:  [SNAPStd.bed](#)

SNAP additional coverage kit:  [SNAPaddtlCov.bed](#)

NEB VarSkip Short, version 1a:  [VSSv1a.bed](#)

NEB VarSkip Short, version 2a:  [VSSv2a.bed](#)

NEB VarSkip Short, version 2b:  [VSSv2b.bed](#)

ARTIC v4 primer schemes:  [ARTICv4.bed](#)

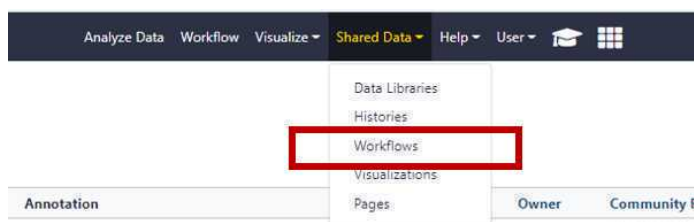
ARTIC v4.1 primer schemes:  [ARTICv4_1.bed](#)

Run the SSQuAWK workflow

5 Access the SSQuAWK4.0.2* workflow with the "workflows" panel.

***SSQuAWK4.0.1: SARS - CoV - 2 Sequence Quality Assurance Workflow Kontrapion, version 4.0.2**

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



Then, from the SSQuAWK4.0.1 drop down menu, select "Run".

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, version 4	jasmine_amir	★★★★★		5 seconds ago
NARMS: Unknown or Mixed Run AMR Workflow v2.0	Not bug specific. For mixed MSeq runs or unknown isolates	gmartin	★★★★★	Name Rate Details	Oct 22, 2021
AMRfinderPhoDT Report WF		gmartin	★★★★★		Oct 22, 2021
NARMS: E-coli AMR Workflow V2.0	E-coli AMR, speciation, and QC	gmartin	★★★★★	Name Rate Details	Oct 04, 2021

(circled in blue). The tabular file can be opened in a text reader or converted to a format that can be opened in Excel.

History ↺ + ▢ ⚙

search datasets ? ×

SSQuAWK

39 shown, 5 deleted, 269 hidden

1.22 GB ☑ 🔍 💬

309: SSQuAWK4 👁 ✎ ×

18 lines

format: **tabular**, database: ?

📄 ? ↺ ↻ 🔍 💬

1	2	3
Sample	0xGenomeCov	<10xGenomeCov
SRR16828363.fastq.gz	107nt (0%)	138nt (0%)
SRR16828364.fastq.gz	76nt (0%)	107nt (0%)
SRR16828365.fastq.gz	76nt (0%)	153nt (0%)
SRR16828366.fastq.gz	443nt (1%)	4962nt (1%)

71: SNAPaddtlCov.bed 👁 ✎ ×

70: NC_045512.2.fa 👁 ✎ ×

69: pairs_ssquawk ×

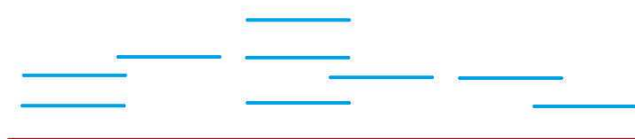
a list of pairs with 17 items

6.2 The SSQuAWK4.0.1 output file includes the following metrics:

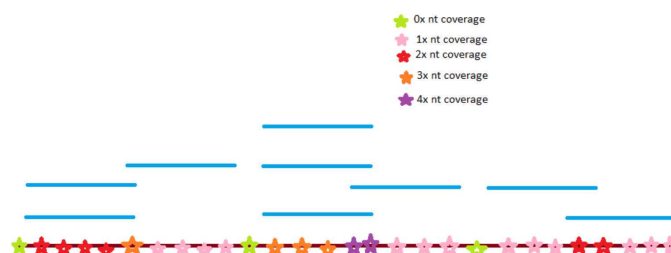
A	B	C
<i>Parameter</i>	<i>Tool</i>	<i>Description</i>
Sample	List of Pairs	Sample name from list of pairs
0xGenomeCov	Bowtie2, samtools, ivar_trim	Percentage of nucleotides that do not cover the genome at all (zero times)
<10xGenomeCov	Bowtie2, samtools, ivar_trim	Percentage of nucleotides that barely cover the genome (less than 10 times)
nReads	Bowtie2	Total number of reads
avgLen	Bowtie2, samtools	Average read length
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*.
SNR	Bowtie2, ivar_trim, Python3 and Pandas	SNR is "Signal to Noise Ratio". When the sequence dataset contains one paired sequence set containing "negativeControl" in its name, SNR will be calculated. Otherwise, the SNR metric will return "NA". The SNR calculation is as follows, for each sequence file pair: $\text{readsAlignPassFilt} / \text{readsAlignPassFilt_negativeControl}$
humanReads	Kraken2	Number and percentage of reads classified as <i>Homo sapiens</i>
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 pipeline used to evaluate sequence quality
quality_control_method_version	4.0.2	Version number of the quality control pipeline or method used
quality_control_determination		Result of the quality control assessment. Blank if pass/fail thresholds have not been established 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.

$$\text{percent_nt0Xcov} = (\text{nucleotidePositions0Xcov} / \text{genomeLength}) * 100$$

$$\text{percent_nt0Xcov} = (3 / 28) * 100$$

$$\text{percent_nt0Xcov} = 10.71\%$$

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.*

6.4 Example output for the first 3 pairs run through the SSQuAWK4.0.2 workflow:

A	B	C	D	E	F	G	H	I	J	K
Sample	0xGenomeCov	<10xGenomeCov	nReads	avgLen	avgLenPassFilt	avgQual	avgQualPassFilt	avgCovPassQual	readsAlign	readsAlignPassFilt
SRR16828363.fastq.gz	107nt (0%)	138nt (0%)	632664	151	151	37.8	37.9	688X	138637 (21%)	136327 (21%)
SRR16828364.fastq.gz	76nt (0%)	107nt (0%)	458116	151	151	37.8	37.9	890X	179913 (39%)	176348 (38%)
SRR16828365.fastq.gz	76nt (0%)	153nt (0%)	351980	151	151	37.8	37.9	272X	54928 (15%)	53958 (15%)

6.5 QC metric guidance for QC attributes on SRA metadata

A	B	C	D	E	F	G
QC bin	Subjective definition	% Genome uncovered (10X)	Average coverage	Other observations	SRA submission	FDA CFSAN Dashboard
A	No QC issues evident	~5%	~1000X	Majority of reads are SARS-CoV-2	"quality_control_determination" = no quality control issues identified	Included
B	Minor QC issues	6% - 40%	~100X		"quality_control_determination" = minor quality control issues identified	Included
C	Insufficient coverage	40% - 95%	< 100X	Insufficient data mapped for confidence	"quality_control_determination" = sequence flagged for potential quality control issues	Excluded
F	Significant QC and/or study design issues	>95%	< 10X	Suspected contamination (SNR low), low sequence quality, other process errors identified	Do not submit	N/A

Video Tutorial

7 Thanks for using SSQuAWK!

