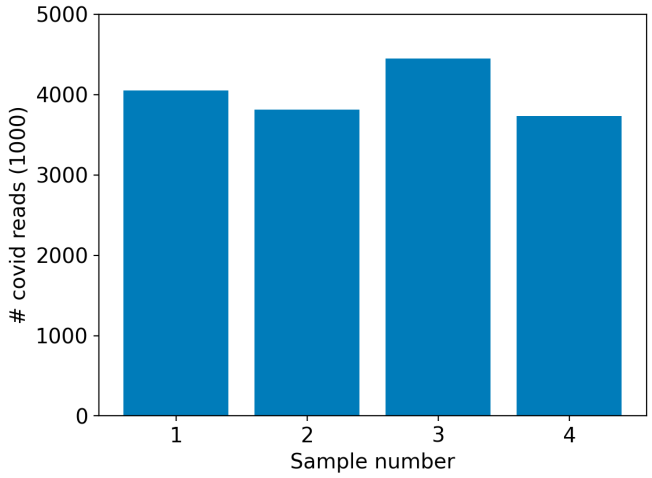
U.S. FOOD & DRUG ADMINISTRATION

CFSAN/OAO
BIOSTATISTICS AND BIOINFORMATICS STAFF

WASTEWATER SARS-COV2 ANALYSIS REPORT

Summary

Sample#	Sample name	Total #reads	Reads aligned PF*	Genomic coordinates 0X	Genomic coordinates <10X
1	SRR22214907	4232286	4053932 (95%)	213nt (0%)	214nt (0%)
2	SRR22214908	3971518	3815695 (96%)	478nt (1%)	538nt (1%)
3	SRR22214909	4625222	4453080 (96%)	151nt (0%)	172nt (0%)
4	SRR22214910	3914096	3731782 (95%)	223nt (0%)	333nt (1%)



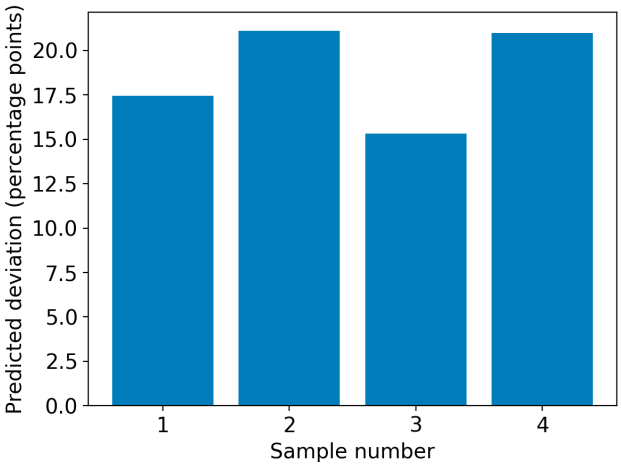
*Quantity of raw reads that align to the reference sequence and pass filter, i.e. the read length after adaptor trimming ≥30 and minimum read quality ≥20 within a sliding window of width 4. SNR refers to the ratio of SC2-mapping reads aligned that pass filter in the sample vs. that in the auto-detected negative control samples (if any). The dashed line represents the baseline level of covid reads detected from the negative control or their average if multiple negative controls we included.

QC-bot (Experimental)

QC category	Subjective definition	Objective metrics
A	No QC issues evident	0x coordinates <1% 10x coordinates <5% average coverage > 1000X average quality score >35 for Illumina, >15 if ONT, >70 if PacBio HiFi most abundant taxon is coronovirinae
B	Some QC issues, but accurate variant calling possible	0x coordinates <20% 10X coordinates < 40% >80% of diverse SNPs covered average coverage > 100X average quality score >35 for Illumina >15 if ONT, >70 if PacBio HiFi
C	Some QC issues, and accurate variant calling impossible	0x coordinates <99% 10X coordinates <95%

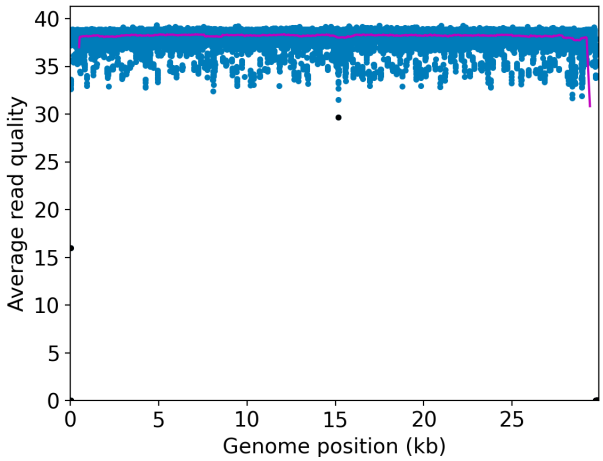
F	Significant QC/study design issues	Contamination (SNR<50) No/negligible coverage (< 1X) Biological/technical replicates' results are irreconcileable.

Sample Number	Suggested category	Suggested QC flags
1	A	None
2	B/C	low_coverage_breadth
3	A	None
4	A	None

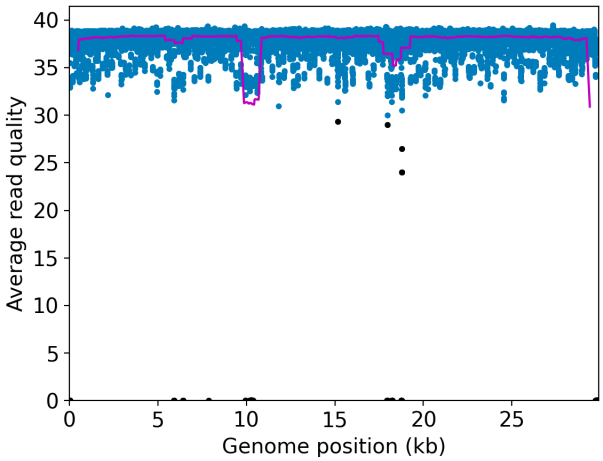


Machine-learning based prediction of the SC2 variant calling accuracy of Freyja of this dataset. The model is a random forest trained on FDA/CFSAN's experimental wastewater WGS data obtained in January 2022 and aims to assess the impact of the potential coverage gaps on the variant abundance estimates. The plotted values represent the predicted deviation of the omicron percentage points from the value that would have been obtained if the coverage was near-complete.

SRR22214907

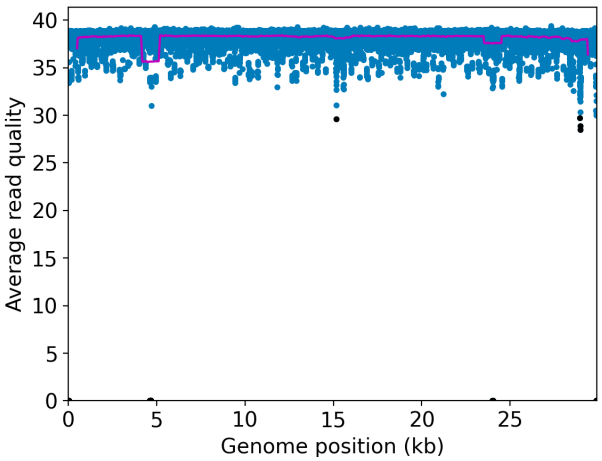


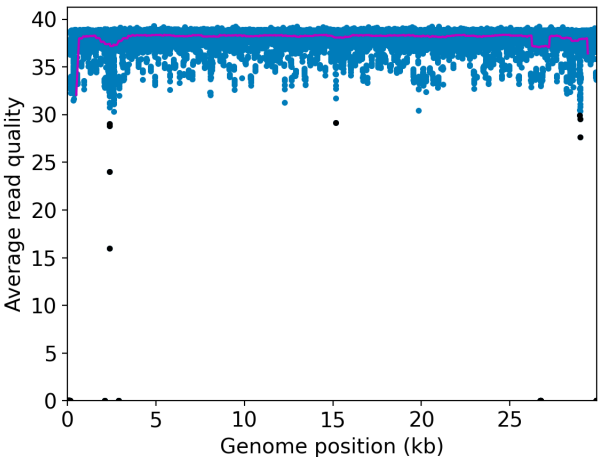
SRR22214908



SRR22214909

SRR22214910


Average read quality
Genome position (kb)
0
5
10
15
20
25
30
35
40


Average read quality
Genome position (kb)
0
5
10
15
20
25
30
35
40