U.S. FOOD & DRUG ADMINISTRATION

CFSAN/OAO
BIOSTATISTICS AND BIOINFORMATICS STAFF

WASTEWATER SARS-COV2 ANALYSIS REPORT

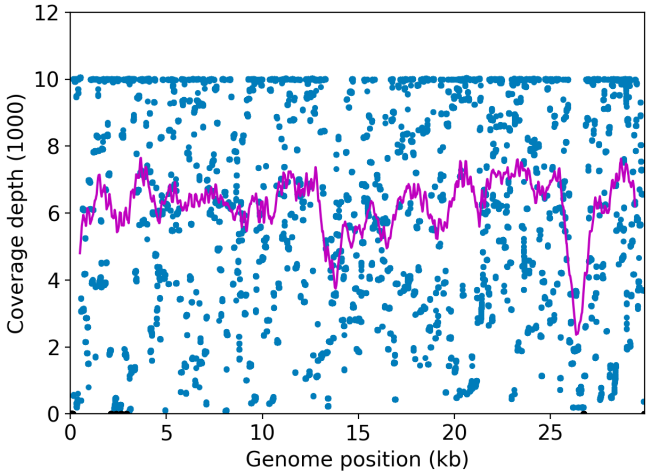
Sample name:	SRR22214910
Date generated:	2022-12-06, 11:59:30 EST
Timestamp of C-WAP version used:	Mon Dec 5 13:25:44 2022 -0500
Executed by:	Jasmine Amirzadegan (Jasmine.Amirzadegan@fda.hhs.gov)
Executed on:	172.20.44.224 (aka n224.raven.cfsan)

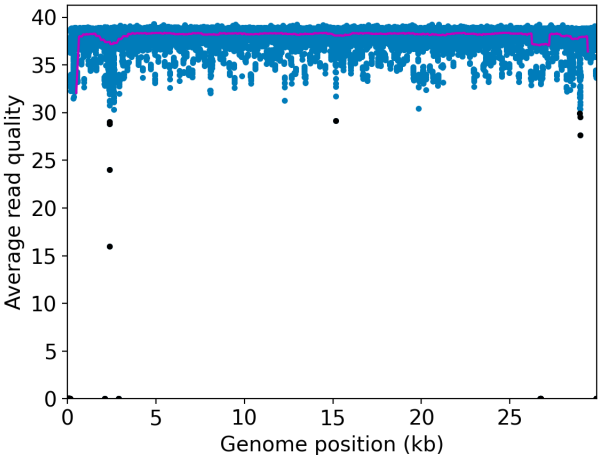
Sequencing summary

Sequencing chemistry:	AMPLICON with Illumina MiSeq
Source site:	USA: Alabama (missing,?)
Sampling date:	2022-10-25
Collected by:	FDA Center for Food Safety and Applied Nutrition
Sequenced by:	Missing
Total number of reads:	3914096
Reads aligned:	3759638 (96%)
Average read quality:	38.1
Average read length:	149
Reads passing filter:	3731782 (95%)
Average read quality passing filter:	38.2
Average read length passing filter:	149
Average coverage passing filter:	18594X

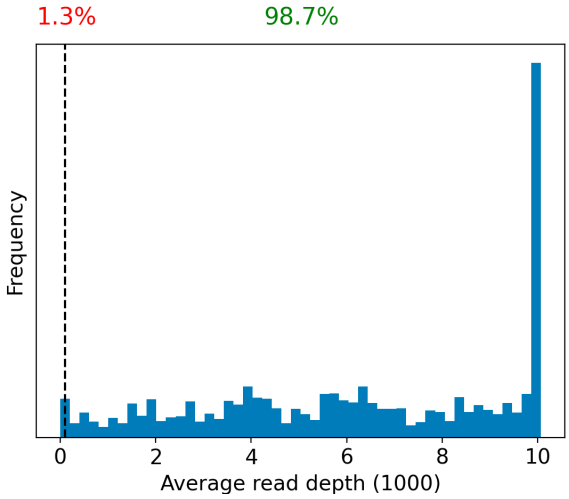
A read passes filter if the read length after adaptor trimming ≥ 30 and minimum read quality ≥ 20 within a sliding window of width 4.

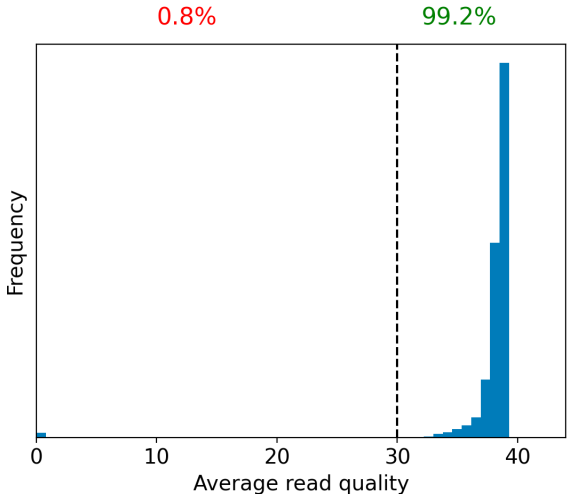
Overall sequence characteristics

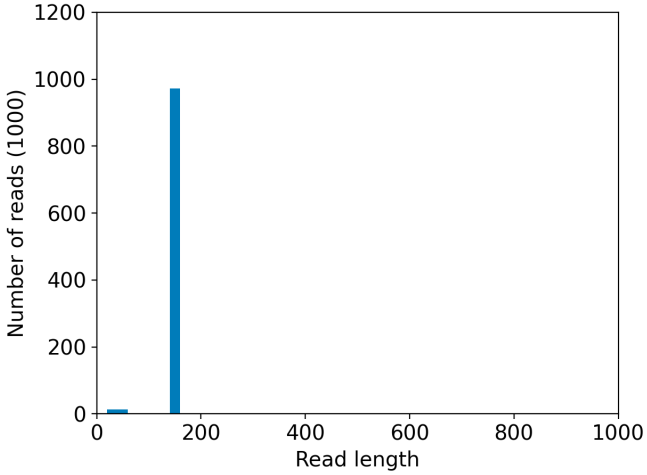


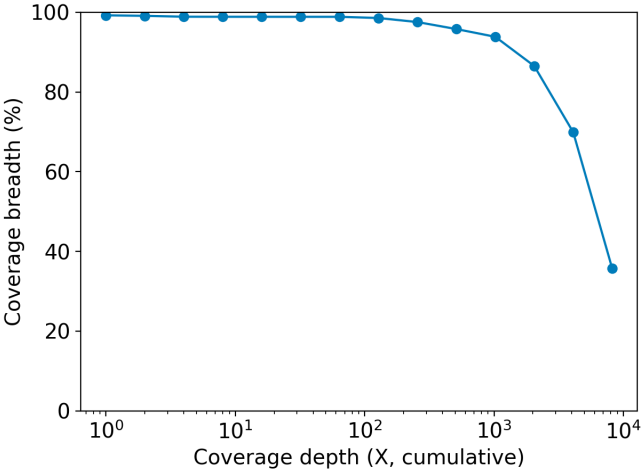


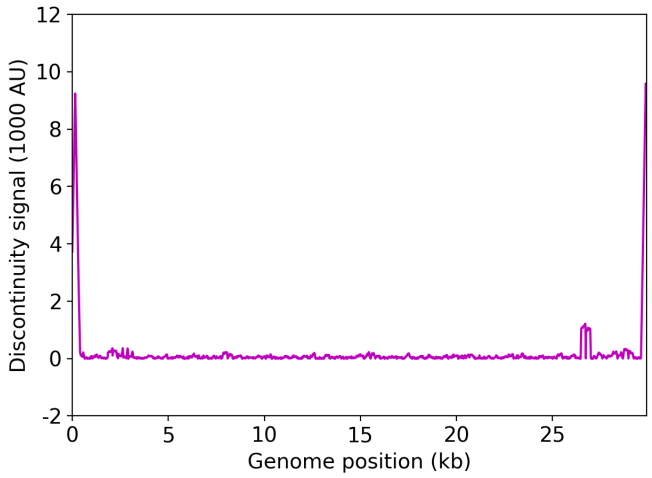
NOTE: The red shaded areas marked with a (*) are not covered by the design of the library preparation kit and hence excluded from analyses. Magenta curves represent moving average with a window width of 1kb.





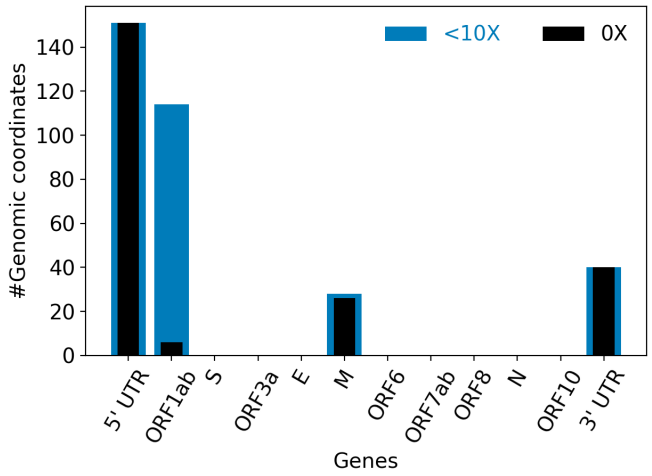


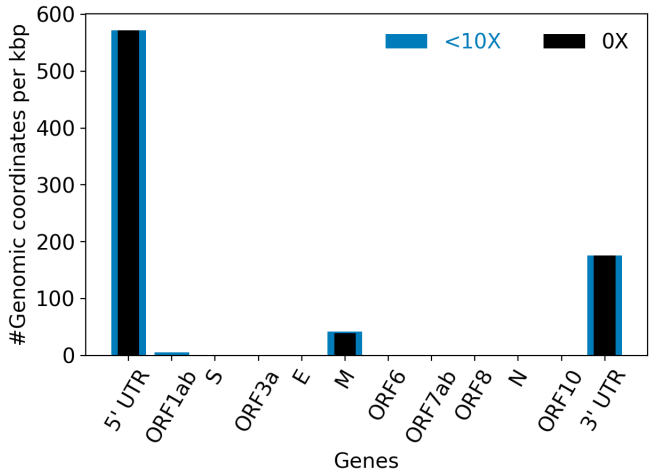




	Uncovered coordinates (0X)	Poorly covered coordinates (<10X)
# Inaccessible genomic coordinates by kit design:	-1nt (0%)	-1nt (0%)
All genomic coordinates:	223nt (0%)	333nt (1%)
Common SNPs:	0nt (0%)	0nt (0%)
Diverse SNPs:	9nt (4%)	9nt (4%)
Rare SNPs:	36nt (3%)	36nt (3%)

SNPs refer to the polymorphic sites currently in circulation that were detected out of recent GISAID entries. The sites that differ from the SC2 reference sequence are denoted as "common" if [90%, 100%] of the submissions carry this mutation, whereas those that are prevalent in [0%,10%] of the submissions are grouped under the "rare" category. The population is still diverse at the mutation sites that are observed in (10%,90%) of the entries and these coordinates are grouped under the "diverse" category.

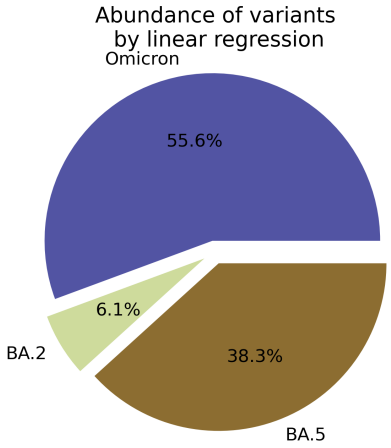


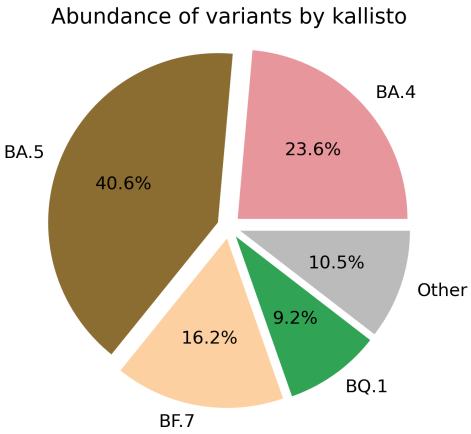


Hits to SARS-Cov2 genome (kraken2):	1895228 reads (96.84%)
Hits to human genome (kraken2):	345 reads (0.02%)
Hits to synthetic sequences (kraken2, taxid 28384):	50 reads (0.00%)
Most abundant organisms (kraken2, family level):	Coronaviridae (96.84%) Geobacteraceae (0.08%) Akkermansiaceae (0.04%)

	Desulfovibrionaceae (0.02%) Enterobacteriaceae (0.02%)

Detected variants (Experimental)

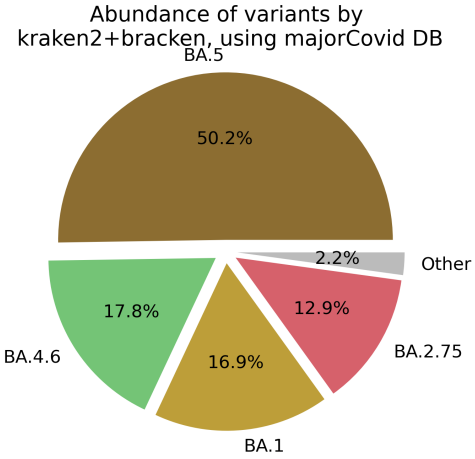


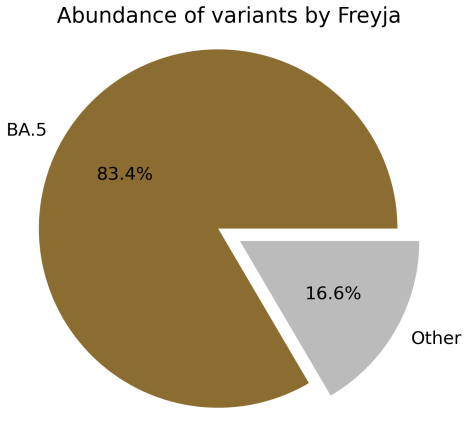


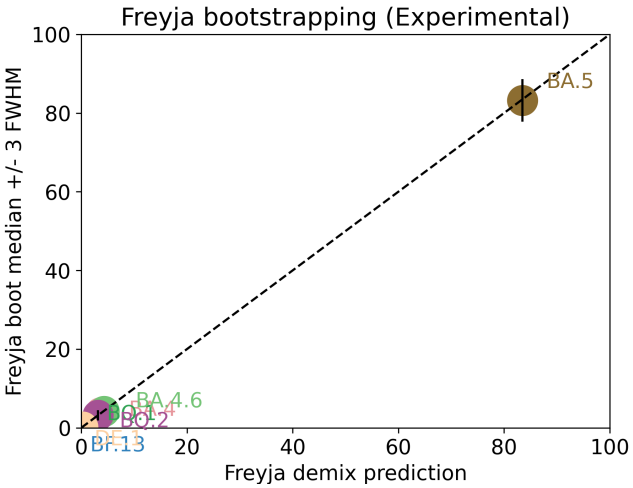
Based on deconvolution, B.1.1.529 is estimated to constitute 55.67% of the viral particles and hence is the most abundant variant in the sample. The R^2 for the linear regression was 0.57. Variants that were detected less than 5% were grouped under "Other"

Based on the consensus sequence of the observed reads, the "ensemble-averaged sequence" most closely resembles the BA.5.2 lineage. If this is a sample consisting of a single source of pathogens or an overwhelming majority of the different sources are infected with the same variant, the sample is dominated by this variant.

Based on mapping individual reads to the variant consensus sequences in the reference database, kallisto predicts that the sample is dominated by BA.5 lineage. Accuracy of this measure is expected to improve if the input data consists of long reads as opposed to convolution.







Under the assumption that the presence of a variant requires the detection of all respective mutations of the variant, the characteric mutations which support the presence of the respective variant are indicated in the respective column of the table. Numbers show the number of mutations detected, if any, and the number of mutations expected to be present based on the variant definitions.

VOC	B.1.617.2	BA.1	BA.2	BA.3	BA.4	BA.5
Characteristic mutations detected	(3 of 13) S:G142D S:L452R S:T478K	(2 of 26) NUC:C25000T NUC:C25584T	(20 of 31) N:S413R NUC:A20055G NUC:A9424G NUC:C10198T NUC:C12880T NUC:C15714T NUC:C25000T NUC:C25584T NUC:C4321T NUC:G10447A ORF1AB:G1307S ORF1AB:L3027F ORF1AB:S135R ORF1AB:T3090I ORF1AB:T842I S:D405N S:R408S S:S371F S:T19I S:T376A	(9 of 21) N:S413R NUC:C12880T NUC:C15714T NUC:G10447A ORF1AB:G1307S ORF1AB:S135R ORF1AB:T3090I S:D405N S:S371F	(23 of 31) N:P151S N:S413R NUC:A20055G NUC:C10198T NUC:C12880T NUC:C15714T NUC:C25000T NUC:C25584T NUC:C4321T NUC:G10447A NUC:G12160A NUC:G27788T ORF1AB:G1307S ORF1AB:S135R ORF1AB:T3090I ORF1AB:T842I S:D405N S:F486V S:L452R S:S371F S:T19I S:T376A S:V213G	(22 of 28) M:D3N N:S413R NUC:A20055G NUC:C10198T NUC:C12880T NUC:C15714T NUC:C25000T NUC:C25584T NUC:C4321T NUC:G10447A NUC:G12160A ORF1AB:G1307S ORF1AB:S135R ORF1AB:T3090I ORF1AB:T842I S:D405N S:F486V S:L452R S:S371F S:T19I S:T376A S:V213G

Jaccard Index is a measure of similarity between two sets A and B, reaching the maximum value of 1 if A=B and minimum value of 0 if A ∩ B = {}. In the c(d) representation below, c represents the Jaccard index of the set of mutations that were experimentally detected for this sample as listed above, whereas d refers to the ideal value of the Jaccard index expected from complete genome coverage without any sequencing errors.

	B.1.617.2	BA.1	BA.2	BA.3	BA.4	BA.5
B.1.617.2	1.00 (1.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.04 (0.02)	0.04 (0.03)
BA.1	0.00 (0.00)	1.00 (1.00)	0.10 (0.10)	0.00 (0.21)	0.09 (0.08)	0.09 (0.08)
BA.2	0.00 (0.00)	0.10 (0.10)	1.00 (1.00)	0.45 (0.33)	0.65 (0.63)	0.68 (0.59)
BA.3	0.00 (0.00)	0.00 (0.21)	0.45 (0.33)	1.00 (1.00)	0.39 (0.30)	0.41 (0.29)
BA.4	0.04 (0.02)	0.09 (0.08)	0.65 (0.63)	0.39 (0.30)	1.00 (1.00)	0.88 (0.84)

BA.5	0.04 (0.03)	0.09 (0.08)	0.68 (0.59)	0.41 (0.29)	0.88 (0.84)	1.00 (1.00)

Detected mutations

Excluded from this pdf version due to file size limitations.