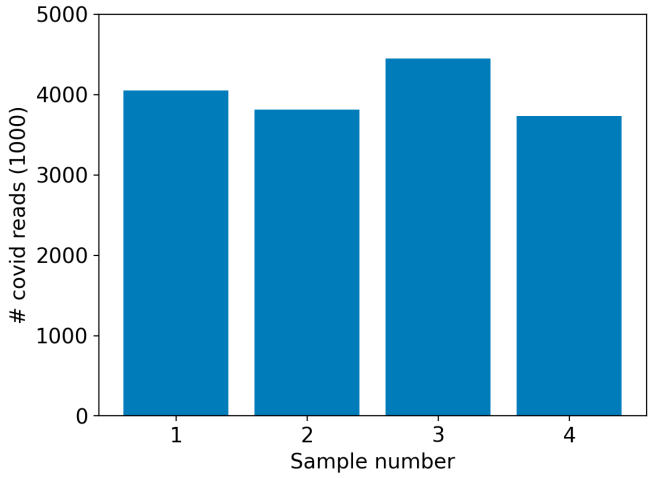
U.S. FOOD & DRUG ADMINISTRATION

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%)



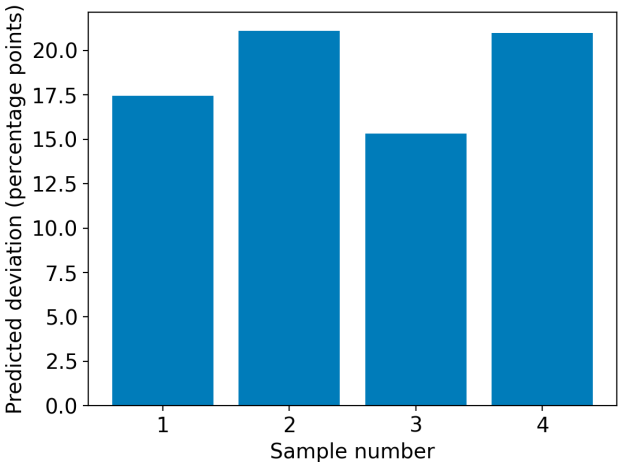
*Quantity of raw reads that align to the reference sequence and pass filter, i.e. the read length after adaptor trimming ≥30 and minimum read quality ≥20 within a sliding window of width 4. SNR refers to the ratio of SC2-mapping reads aligned that pass filter in the sample vs. that in the auto-detected 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%

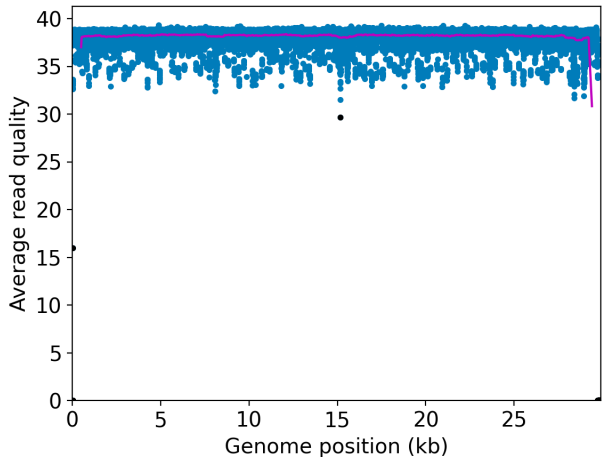
F	Significant QC/study design issues	Contamination (SNR<50) No/negligible coverage (< 1X) Biological/technical replicates' results are irreconcileable.

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

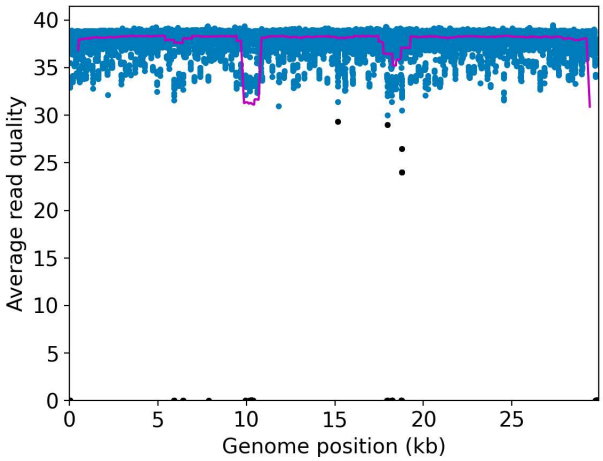


Machine-learning based prediction of the SC2 variant calling accuracy of Freyja of this dataset. The model is a random forest trained on FDA/CFSAN's experimental wastewater WGS data obtained in January 2022 and aims to assess the impact of the potential coverage gaps on the variant abundance estimates. 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

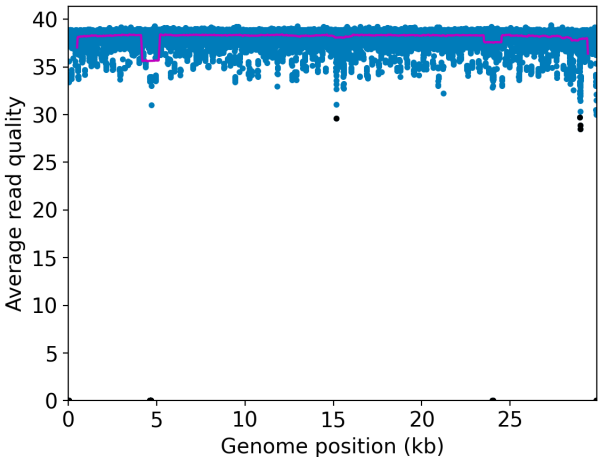


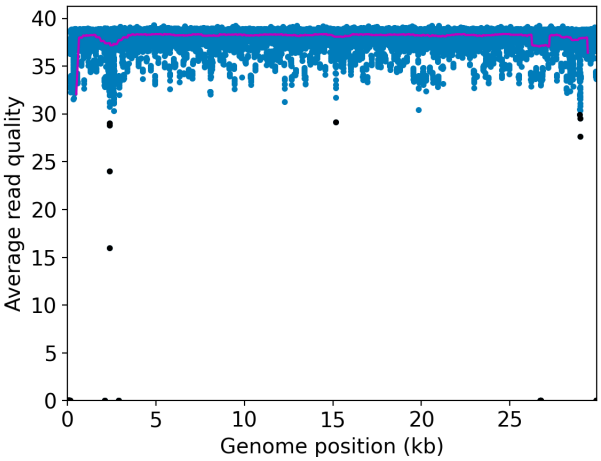
SRR22214908



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

U.S. FOOD & DRUG ADMINISTRATION

CFSAN/OAO
BIOSTATISTICS AND BIOINFORMATICS STAFF

WASTEWATER SARS-COV2 ANALYSIS REPORT

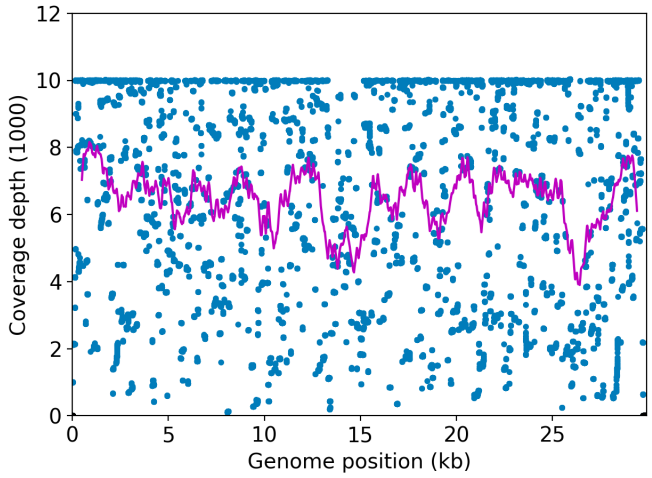
Sample name:	SRR22214907
Date generated:	2023-06-29, 17:43:10 EDT
Timestamp of C-WAP version used:	Thu Jun 29 16:12:34 2023 -0400
Executed by:	Jasmine Amirzadegan (Jasmine.Amirzadegan@fda.hhs.gov)
Executed on:	172.20.44.122 (aka n122.raven.cfsan)

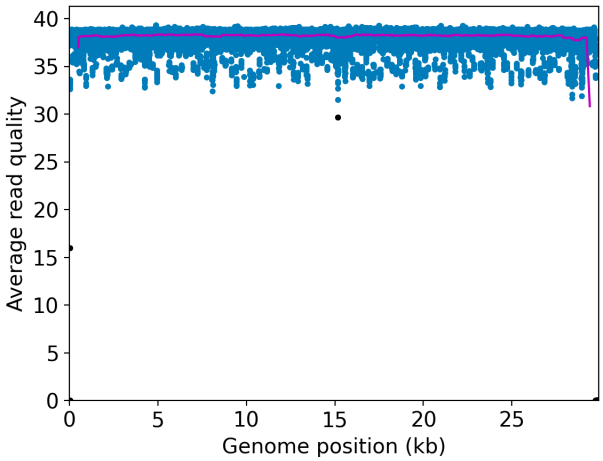
Sequencing summary

Sequencing chemistry:	AMPLICON with Illumina MiSeq
Source site:	USA: Mississippi (missing,?)
Sampling date:	2022-10-25
Collected by:	FDA Center for Food Safety and Applied Nutrition
Sequenced by:	Missing
Total number of reads:	4232286
Reads aligned:	4092919 (96%)
Average read quality:	38.1
Average read length:	149
Reads passing filter:	4053932 (95%)
Average read quality passing filter:	38.2
Average read length passing filter:	149
Average coverage passing filter:	20199X

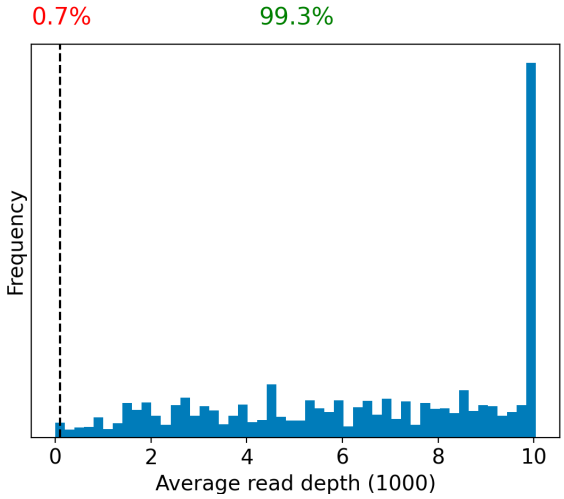
A read passes filter if the read length after adaptor trimming ≥ 30 and minimum read quality ≥ 20 within a sliding window of width 4.

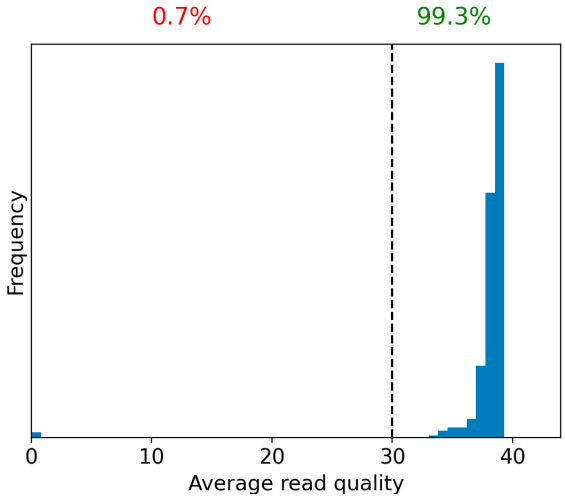
Overall sequence characteristics

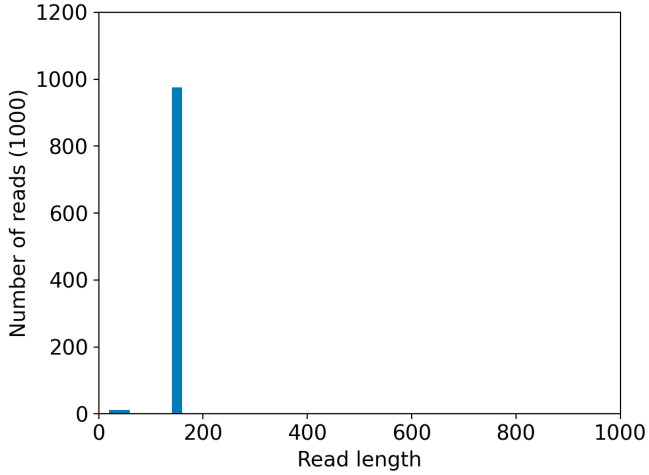


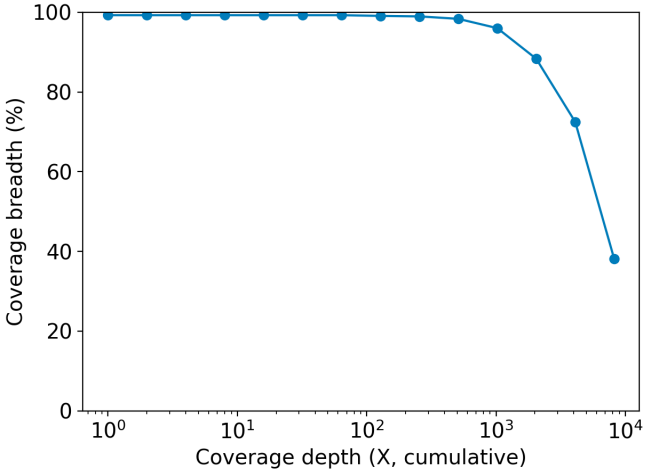


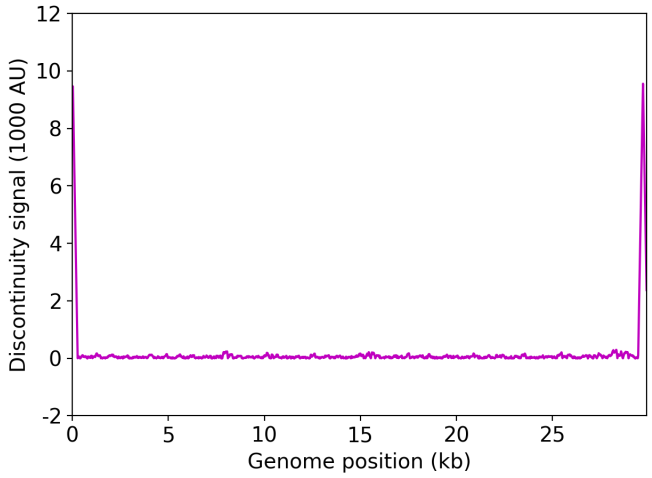
NOTE: The red shaded areas marked with a (*) are not covered by the design of the library preparation kit and hence excluded from analyses. Magenta curves represent moving average with a window width of 1kb.





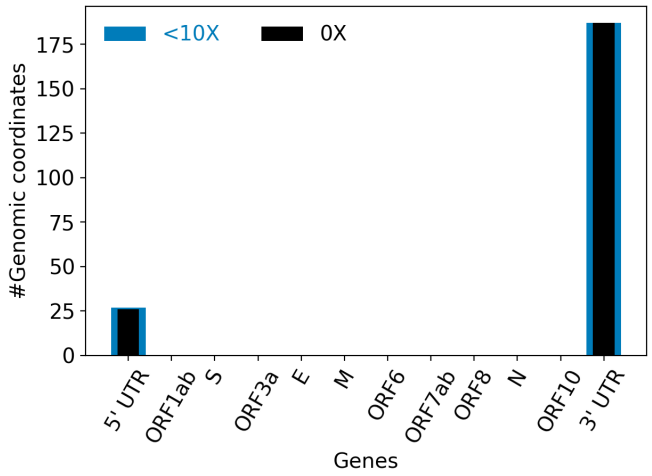


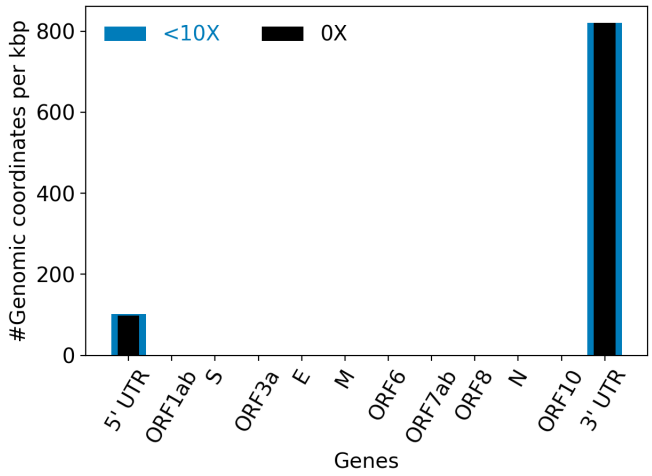




	Uncovered coordinates (0X)	Poorly covered coordinates (<10X)
# Inaccessible genomic coordinates by kit design:	-1nt (0%)	-1nt (0%)
All genomic coordinates:	213nt (0%)	214nt (0%)
Common SNPs:	0nt (0%)	0nt (0%)
Diverse SNPs:	55nt (24%)	55nt (24%)
Rare SNPs:	17nt (1%)	17nt (1%)

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.

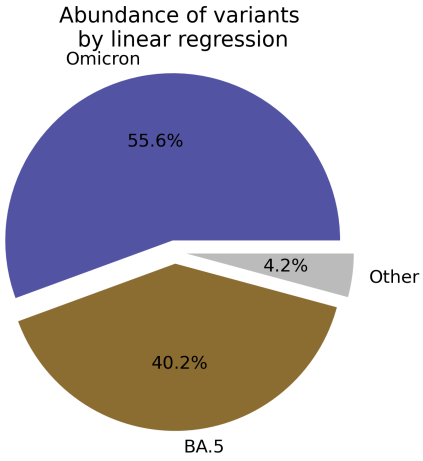


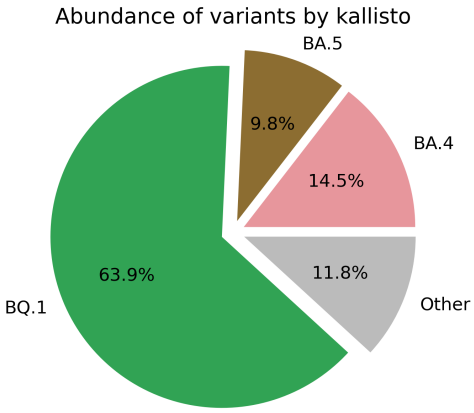


Hits to SARS-Cov2 genome (kraken2):	2065794 reads (97.62%)
Hits to human genome (kraken2):	357 reads (0.02%)
Hits to synthetic sequences (kraken2, taxid 28384):	88 reads (0.00%)
Most abundant organisms (kraken2, family level):	Coronaviridae (97.62%) Geobacteraceae (0.10%) Bacteroidaceae (0.04%)

	Desulfovibrionaceae (0.02%) Staphylococcaceae (0.02%)

Detected variants (Experimental)

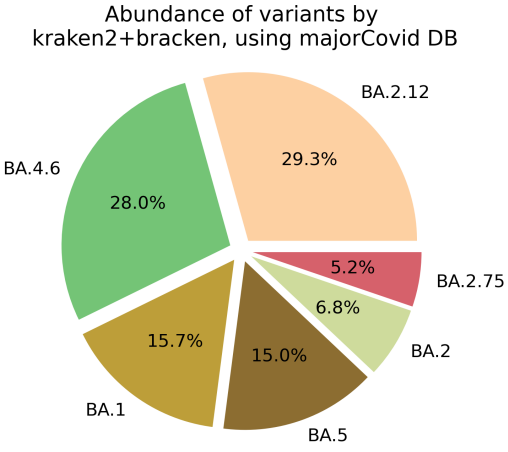


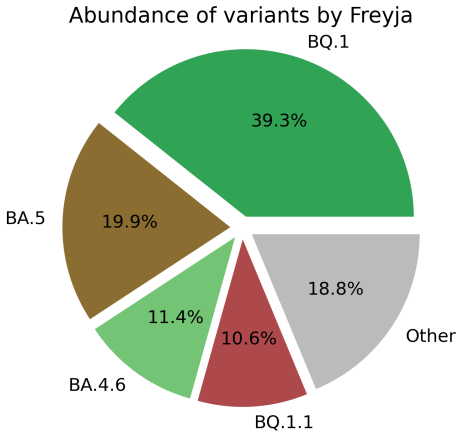


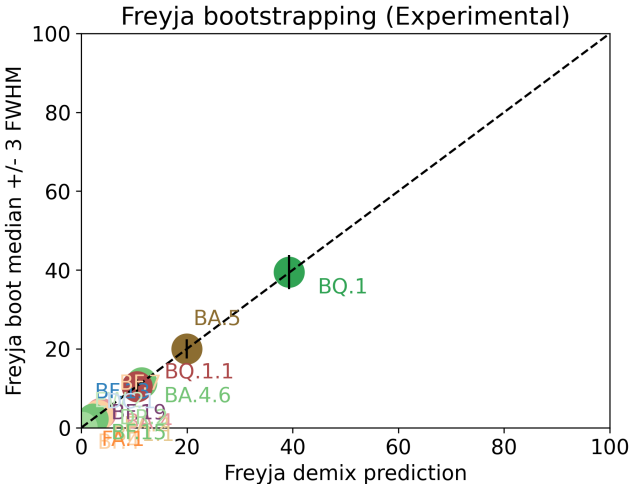
Based on deconvolution, B.1.1.529 is estimated to constitute 55.62% of the viral particles and hence is the most abundant variant in the sample. The R^2 for the linear regression was 0.58. Variants that were detected less than 5% were grouped under "Other"

Based on the consensus sequence of the observed reads, the "ensemble-averaged sequence" most closely resembles the BA.5.3 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 BQ.1 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	(21 of 31) N:S413R NUC:A20055G NUC:A9424G NUC:C10198T NUC:C12880T NUC:C15714T NUC:C25000T NUC:C25584T NUC:C26858T NUC:C4321T NUC:G10447A ORF1AB:G1307S ORF1AB:L3027F ORF1AB:S135R ORF1AB:T3090I ORF1AB:T842I S:D405N S:R408S S:S371F S:T19I S:T376A	(10 of 21) N:S413R NUC:C12880T NUC:C15714T NUC:C26858T NUC:G10447A ORF1AB:G1307S ORF1AB:S135R ORF1AB:T3090I S:D405N S:S371F	(24 of 31) N:P151S N:S413R NUC:A20055G NUC:C10198T NUC:C12880T NUC:C15714T NUC:C25000T NUC:C25584T NUC:C26858T 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)	0.08 (0.08)	0.09 (0.08)
BA.2	0.00 (0.00)	0.10 (0.10)	1.00 (1.00)	0.48 (0.33)	0.67 (0.63)	0.65 (0.59)
BA.3	0.00 (0.00)	0.00 (0.21)	0.48 (0.33)	1.00 (1.00)	0.42 (0.30)	0.39 (0.29)

BA.4	0.04 (0.02)	0.08 (0.08)	0.67 (0.63)	0.42 (0.30)	1.00 (1.00)	0.84 (0.84)
BA.5	0.04 (0.03)	0.09 (0.08)	0.65 (0.59)	0.39 (0.29)	0.84 (0.84)	1.00 (1.00)

Detected mutations

Excluded from this pdf version due to file size limitations.

U.S. FOOD & DRUG ADMINISTRATION

CFSAN/OAO
BIOSTATISTICS AND BIOINFORMATICS STAFF

WASTEWATER SARS-COV2 ANALYSIS REPORT

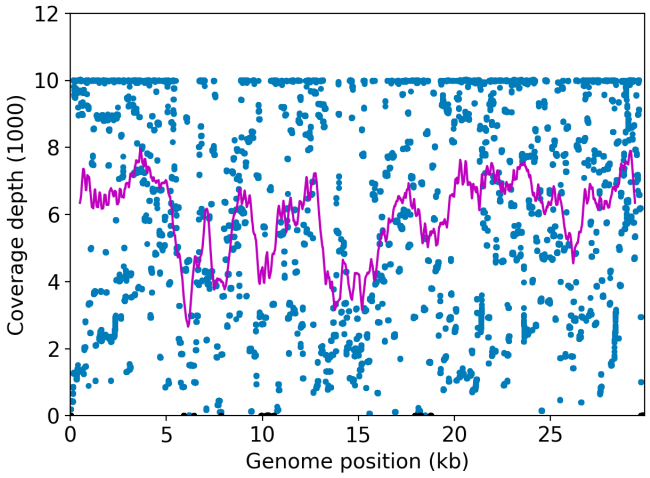
Sample name:	SRR22214908
Date generated:	2023-06-29, 17:34:06 EDT
Timestamp of C-WAP version used:	Thu Jun 29 16:12:34 2023 -0400
Executed by:	Jasmine Amirzadegan (Jasmine.Amirzadegan@fda.hhs.gov)
Executed on:	172.20.44.158 (aka n158.raven.cfsan)

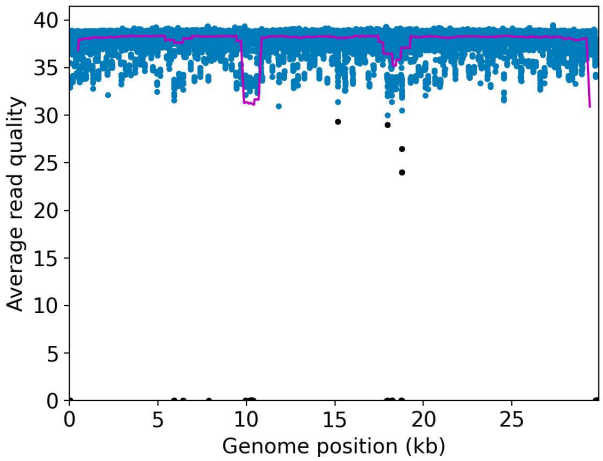
Sequencing summary

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:	3971518
Reads aligned:	3853242 (97%)
Average read quality:	38.1
Average read length:	149
Reads passing filter:	3815695 (96%)
Average read quality passing filter:	38.2
Average read length passing filter:	149
Average coverage passing filter:	19012X

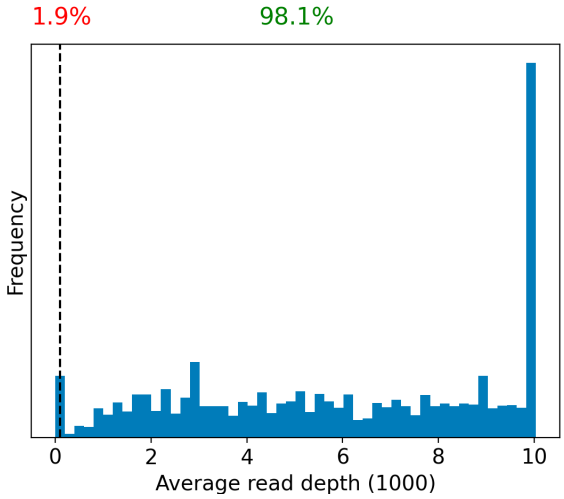
A read passes filter if the read length after adaptor trimming ≥ 30 and minimum read quality ≥ 20 within a sliding window of width 4.

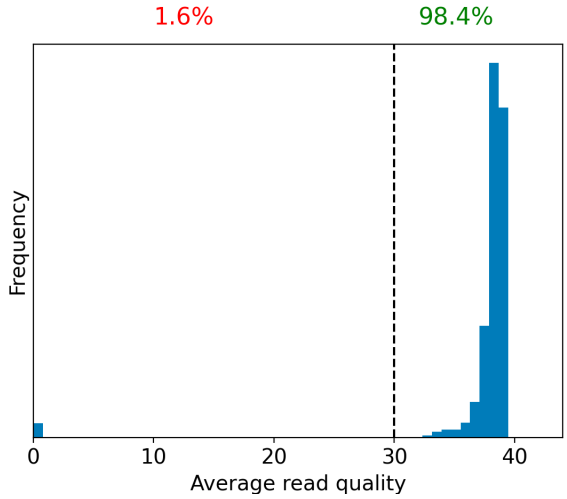
Overall sequence characteristics

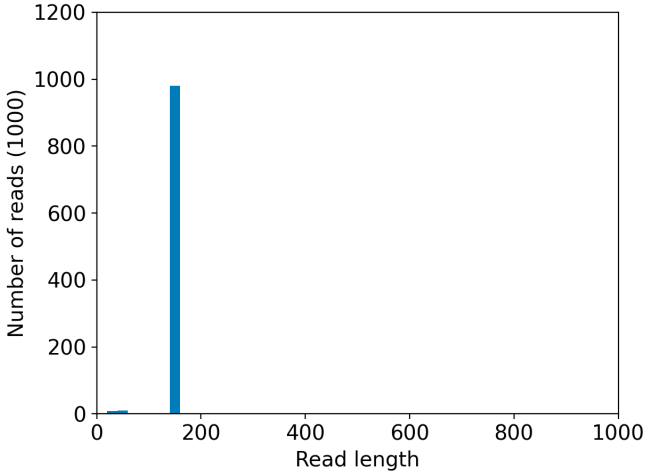


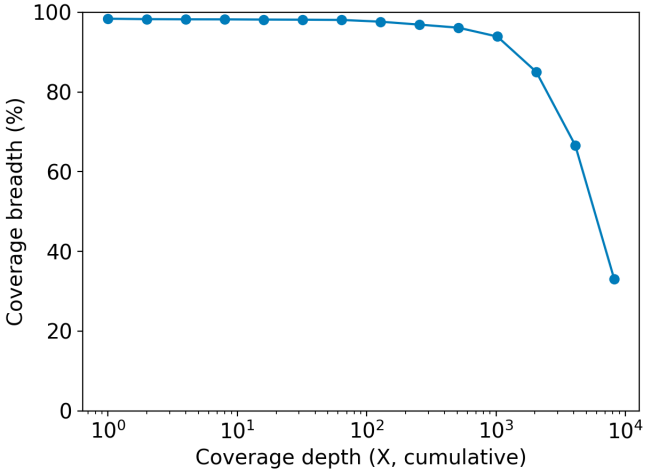


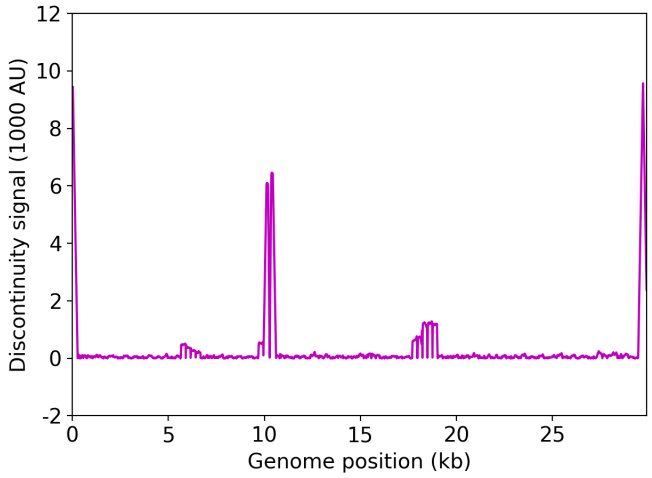
NOTE: The red shaded areas marked with a (*) are not covered by the design of the library preparation kit and hence excluded from analyses. Magenta curves represent moving average with a window width of 1kb.





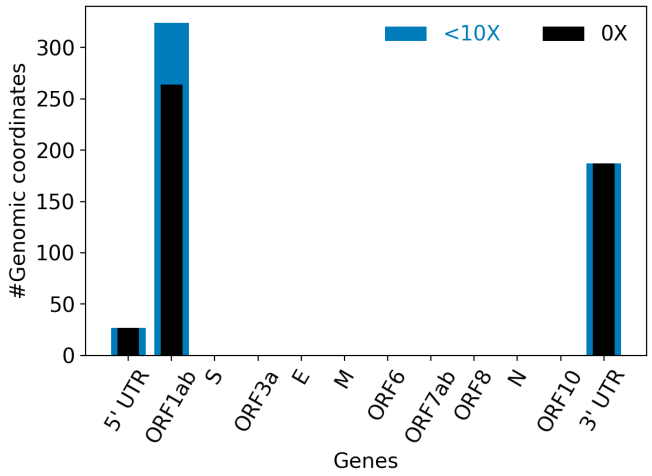


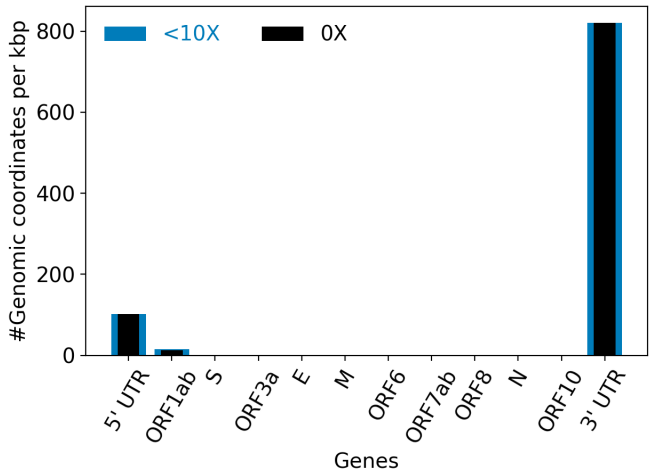




	Uncovered coordinates (0X)	Poorly covered coordinates (<10X)
# Inaccessible genomic coordinates by kit design:	-1nt (0%)	-1nt (0%)
All genomic coordinates:	478nt (1%)	538nt (1%)
Common SNPs:	0nt (0%)	0nt (0%)
Diverse SNPs:	56nt (24%)	56nt (24%)
Rare SNPs:	17nt (1%)	17nt (1%)

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.

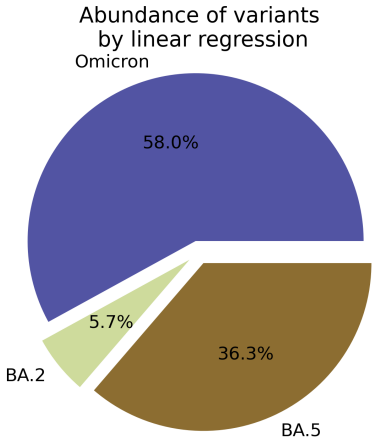


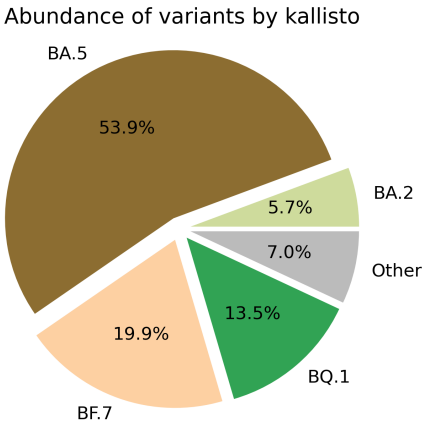


Hits to SARS-Cov2 genome (kraken2):	1939011 reads (97.65%)
Hits to human genome (kraken2):	449 reads (0.02%)
Hits to synthetic sequences (kraken2, taxid 28384):	43 reads (0.00%)
Most abundant organisms (kraken2, family level):	Coronaviridae (97.65%) Aeromonadaceae (0.14%) Enterobacteriaceae (0.05%)

	Geobacteraceae (0.05%) Hominidae (0.02%)

Detected variants (Experimental)

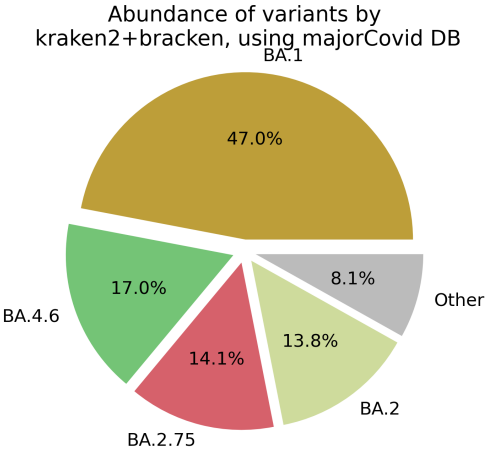


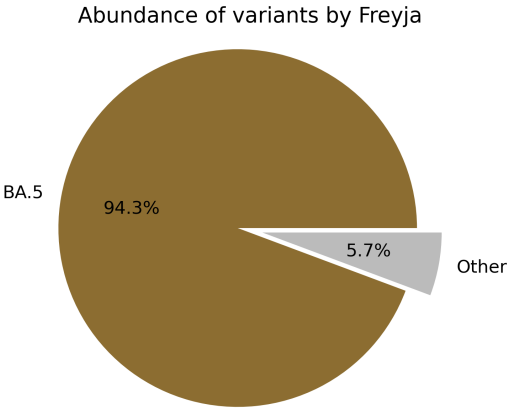


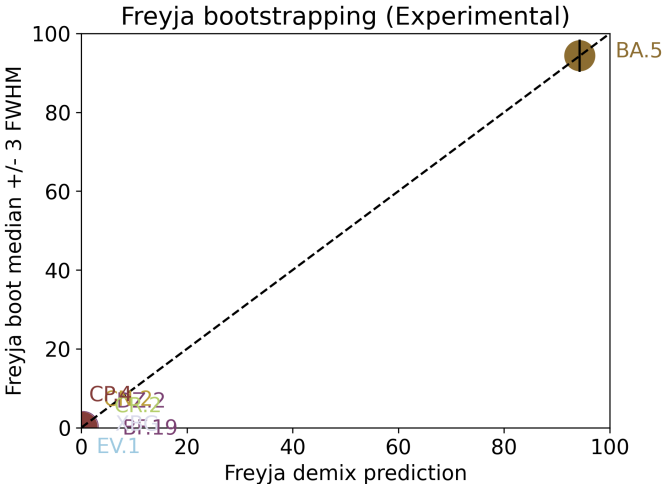
Based on deconvolution, B.1.1.529 is estimated to constitute 57.98% of the viral particles and hence is the most abundant variant in the sample. The R^2 for the linear regression was 0.56. Variants that were detected less than 5% were grouped under "Other"

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	(19 of 31) N:S413R NUC:A20055G NUC:A9424G 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	(20 of 31) N:S413R NUC:A20055G 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	(21 of 28) M:D3N N:S413R NUC:A20055G 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)	0.05 (0.02)	0.04 (0.03)
BA.1	0.00 (0.00)	1.00 (1.00)	0.11 (0.10)	0.00 (0.21)	0.10 (0.08)	0.10 (0.08)
BA.2	0.00 (0.00)	0.11 (0.10)	1.00 (1.00)	0.47 (0.33)	0.70 (0.63)	0.67 (0.59)
BA.3	0.00 (0.00)	0.00 (0.21)	0.47 (0.33)	1.00 (1.00)	0.45 (0.30)	0.43 (0.29)
BA.4	0.05 (0.02)	0.10 (0.08)	0.70 (0.63)	0.45 (0.30)	1.00 (1.00)	0.95 (0.84)
BA.5	0.04 (0.03)	0.10 (0.08)	0.67 (0.59)	0.43 (0.29)	0.95 (0.84)	1.00 (1.00)

Detected mutations

Excluded from this pdf version due to file size limitations.

U.S. FOOD & DRUG ADMINISTRATION

CFSAN/OAO
BIOSTATISTICS AND BIOINFORMATICS STAFF

WASTEWATER SARS-COV2 ANALYSIS REPORT

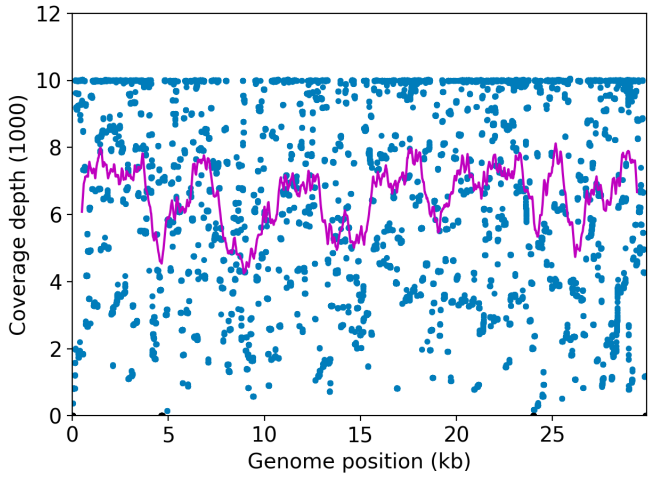
Sample name:	SRR22214909
Date generated:	2023-06-29, 17:36:10 EDT
Timestamp of C-WAP version used:	Thu Jun 29 16:12:34 2023 -0400
Executed by:	Jasmine Amirzadegan (Jasmine.Amirzadegan@fda.hhs.gov)
Executed on:	172.20.44.158 (aka n158.raven.cfsan)

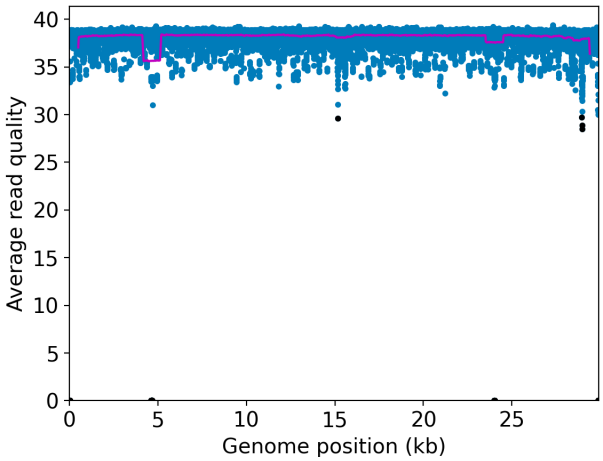
Sequencing summary

Sequencing chemistry:	AMPLICON with Illumina MiSeq
Source site:	USA: Mississippi (missing,?)
Sampling date:	2022-10-25
Collected by:	FDA Center for Food Safety and Applied Nutrition
Sequenced by:	Missing
Total number of reads:	4625222
Reads aligned:	4483541 (96%)
Average read quality:	38.1
Average read length:	149
Reads passing filter:	4453080 (96%)
Average read quality passing filter:	38.2
Average read length passing filter:	149
Average coverage passing filter:	22188X

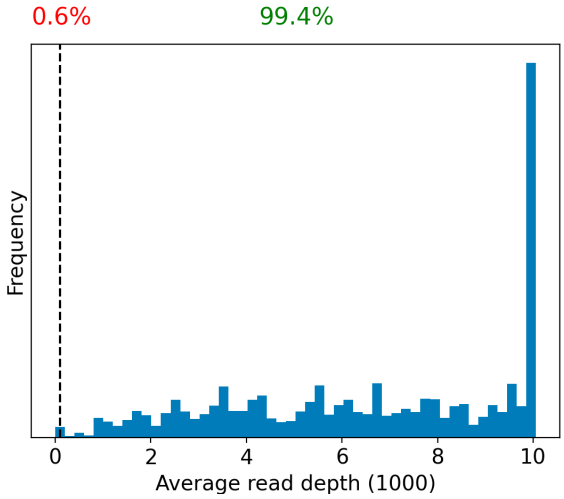
A read passes filter if the read length after adaptor trimming ≥ 30 and minimum read quality ≥ 20 within a sliding window of width 4.

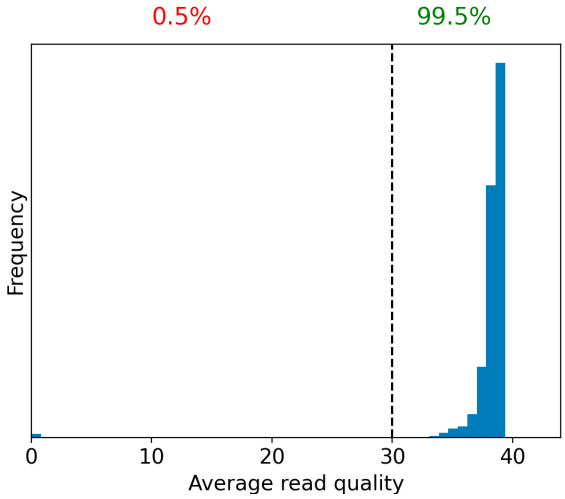
Overall sequence characteristics

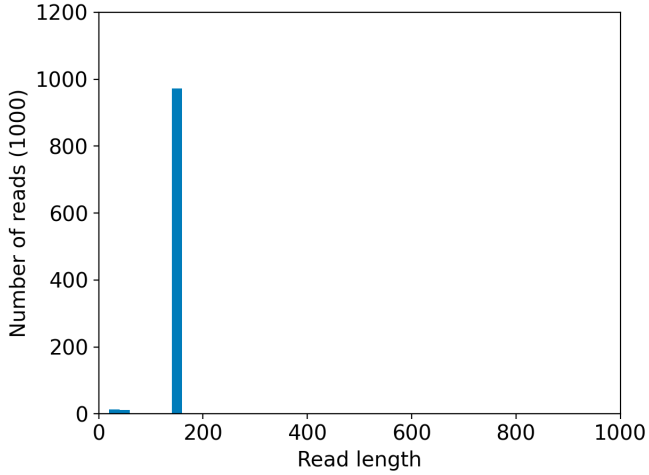


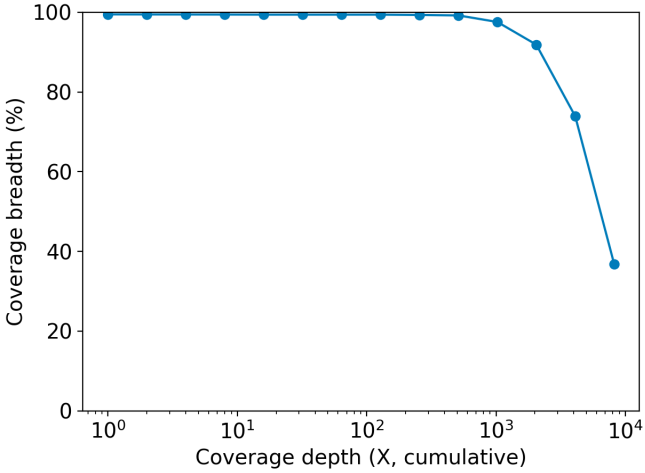


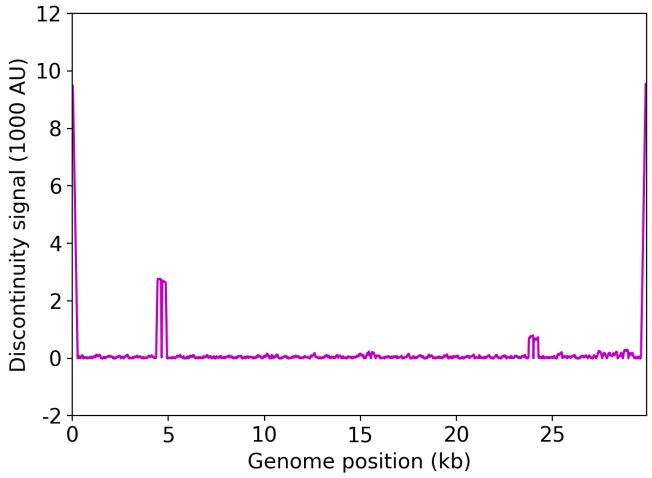
NOTE: The red shaded areas marked with a (*) are not covered by the design of the library preparation kit and hence excluded from analyses. Magenta curves represent moving average with a window width of 1kb.





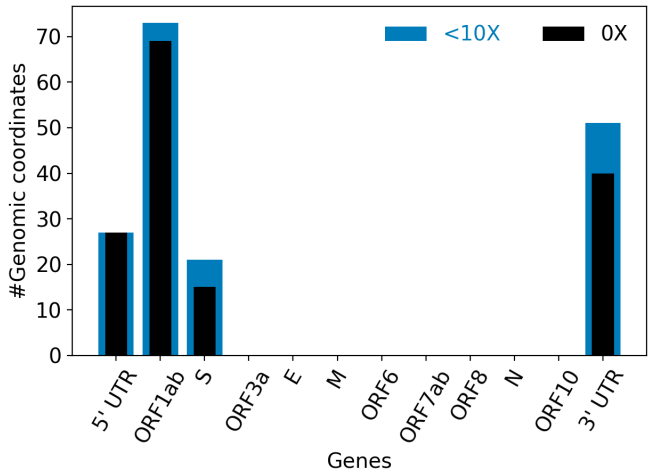


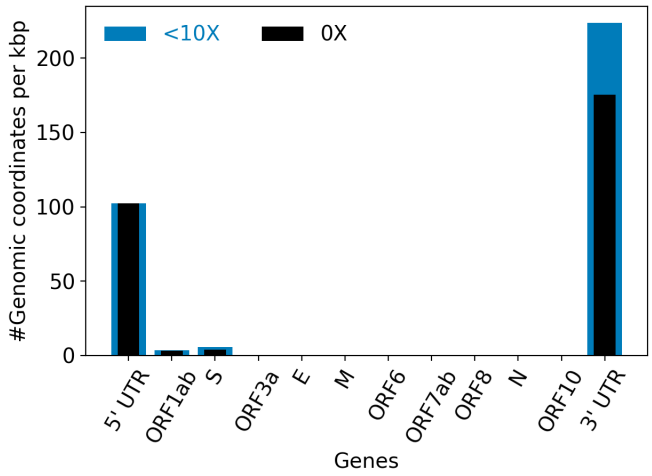




	Uncovered coordinates (0X)	Poorly covered coordinates (<10X)
# Inaccessible genomic coordinates by kit design:	-1nt (0%)	-1nt (0%)
All genomic coordinates:	151nt (0%)	172nt (0%)
Common SNPs:	0nt (0%)	0nt (0%)
Diverse SNPs:	8nt (3%)	19nt (8%)
Rare SNPs:	1nt (0%)	1nt (0%)

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.

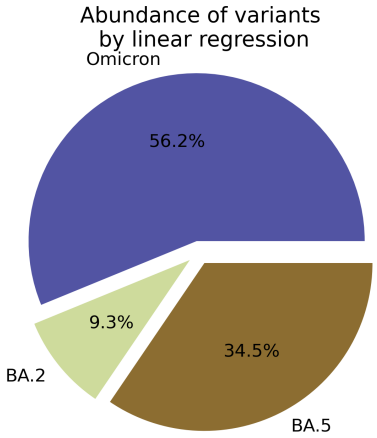


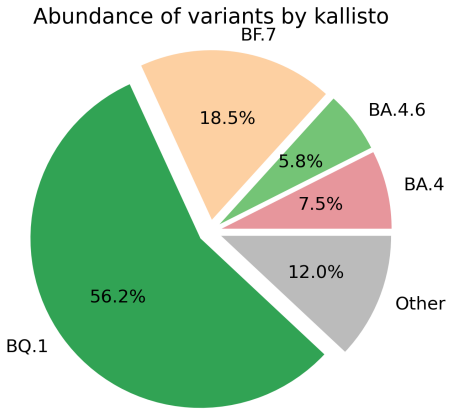


Hits to SARS-Cov2 genome (kraken2):	2259920 reads (97.72%)
Hits to human genome (kraken2):	487 reads (0.02%)
Hits to synthetic sequences (kraken2, taxid 28384):	2 reads (0.00%)
Most abundant organisms (kraken2, family level):	Coronaviridae (97.72%) Bacteroidaceae (0.05%) Burkholderiaceae (0.02%)

	Hominidae (0.02%) Flavobacteriaceae (0.02%)

Detected variants (Experimental)

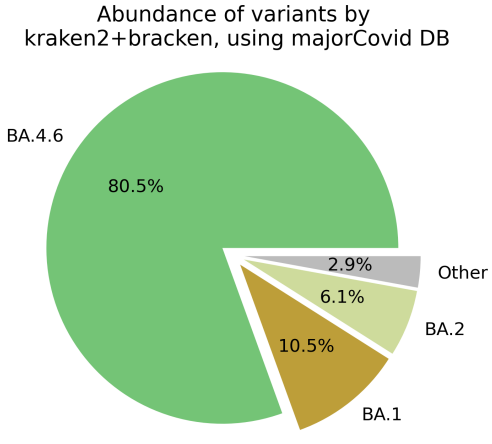


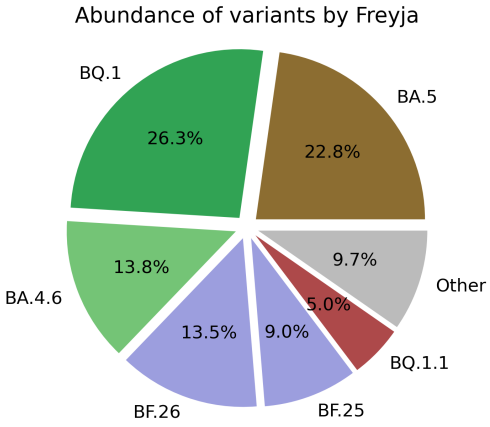


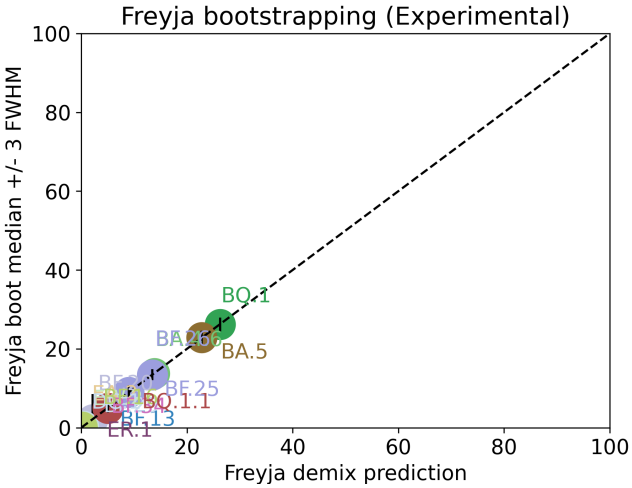
Based on deconvolution, B.1.1.529 is estimated to constitute 56.25% of the viral particles and hence is the most abundant variant in the sample. The R^2 for the linear regression was 0.57. Variants that were detected less than 5% were grouped under "Other"

Based on the consensus sequence of the observed reads, the "ensemble-averaged sequence" most closely resembles the BA.5 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 BQ.1 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	(21 of 31) N:S413R NUC:A20055G NUC:A9424G NUC:C10198T NUC:C12880T NUC:C15714T NUC:C25000T NUC:C25584T NUC:C26858T NUC:C4321T NUC:G10447A ORF1AB:G1307S ORF1AB:L3027F ORF1AB:S135R ORF1AB:T3090I ORF1AB:T842I S:D405N S:R408S S:S371F S:T19I S:T376A	(10 of 21) N:S413R NUC:C12880T NUC:C15714T NUC:C26858T NUC:G10447A ORF1AB:G1307S ORF1AB:S135R ORF1AB:T3090I S:D405N S:S371F	(24 of 31) N:P151S N:S413R NUC:A20055G NUC:C10198T NUC:C12880T NUC:C15714T NUC:C25000T NUC:C25584T NUC:C26858T 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)	0.08 (0.08)	0.09 (0.08)
BA.2	0.00 (0.00)	0.10 (0.10)	1.00 (1.00)	0.48 (0.33)	0.67 (0.63)	0.65 (0.59)
BA.3	0.00 (0.00)	0.00 (0.21)	0.48 (0.33)	1.00 (1.00)	0.42 (0.30)	0.39 (0.29)

BA.4	0.04 (0.02)	0.08 (0.08)	0.67 (0.63)	0.42 (0.30)	1.00 (1.00)	0.84 (0.84)
BA.5	0.04 (0.03)	0.09 (0.08)	0.65 (0.59)	0.39 (0.29)	0.84 (0.84)	1.00 (1.00)

Detected mutations

Excluded from this pdf version due to file size limitations.

U.S. FOOD & DRUG ADMINISTRATION

CFSAN/OAO
BIOSTATISTICS AND BIOINFORMATICS STAFF

WASTEWATER SARS-COV2 ANALYSIS REPORT

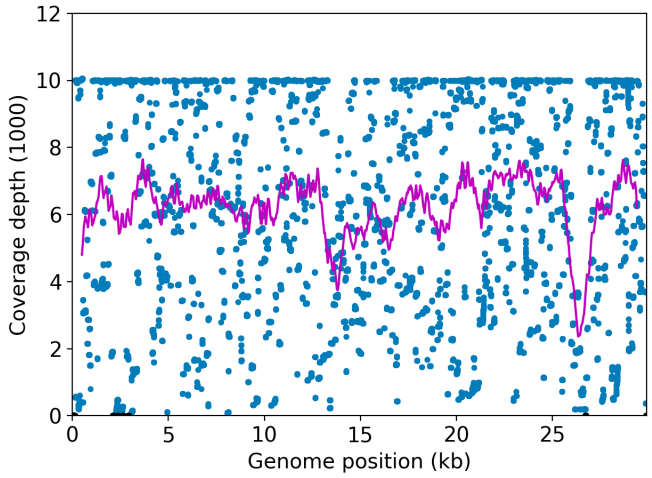
Sample name:	SRR22214910
Date generated:	2023-06-29, 17:47:46 EDT
Timestamp of C-WAP version used:	Thu Jun 29 16:12:34 2023 -0400
Executed by:	Jasmine Amirzadegan (Jasmine.Amirzadegan@fda.hhs.gov)
Executed on:	172.20.44.122 (aka n122.raven.cfsan)

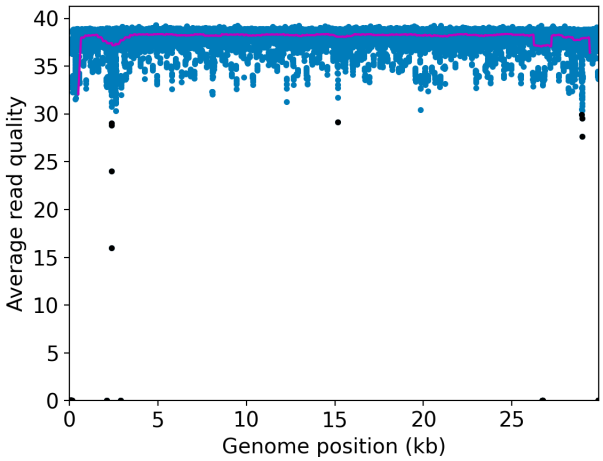
Sequencing summary

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

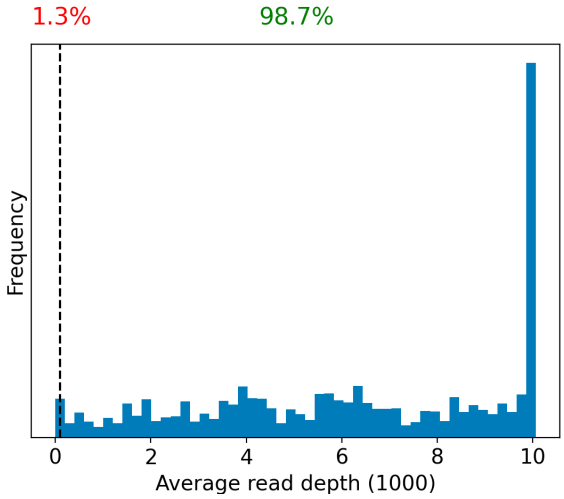
A read passes filter if the read length after adaptor trimming ≥ 30 and minimum read quality ≥ 20 within a sliding window of width 4.

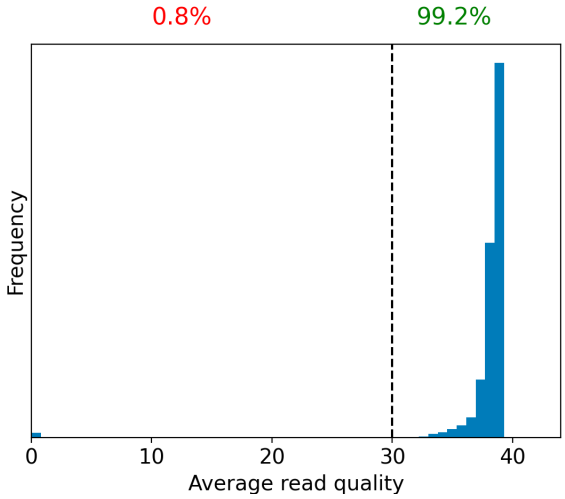
Overall sequence characteristics

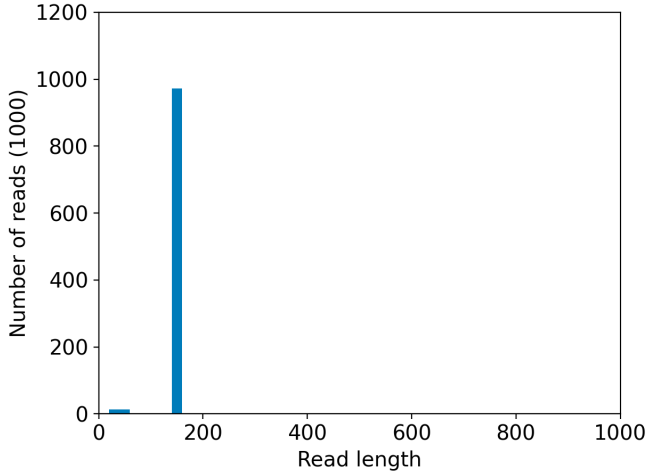


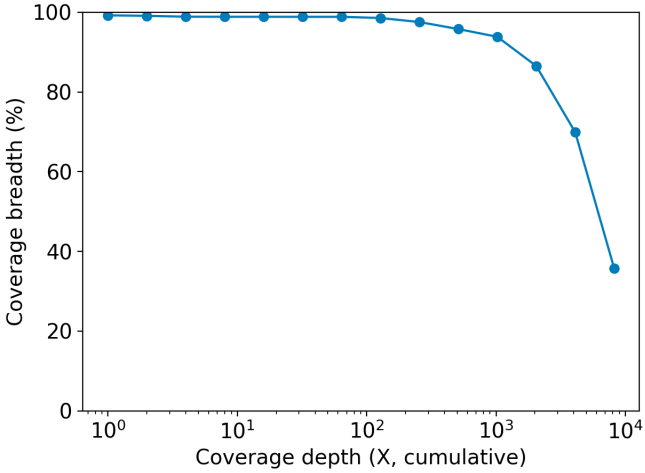


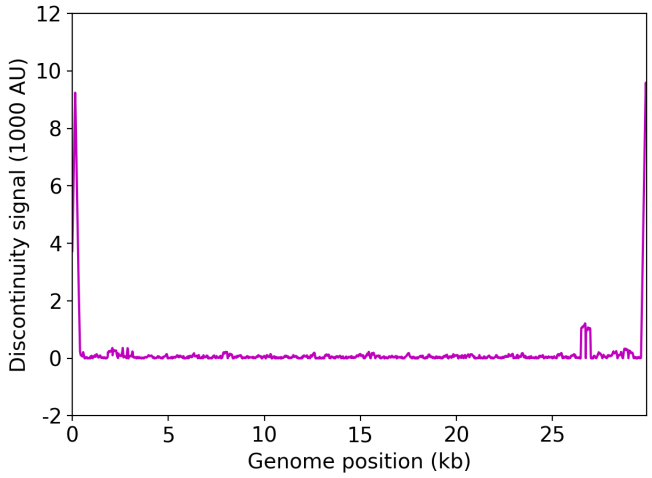
NOTE: The red shaded areas marked with a (*) are not covered by the design of the library preparation kit and hence excluded from analyses. Magenta curves represent moving average with a window width of 1kb.





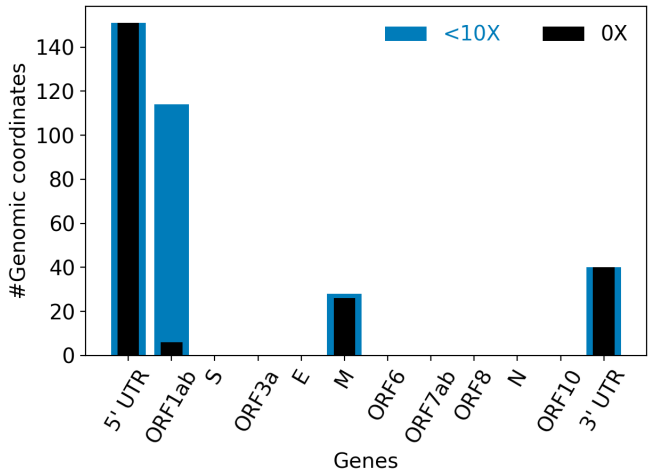


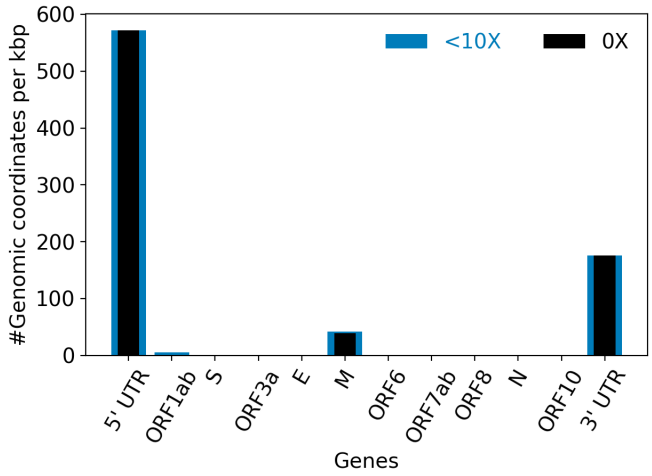




	Uncovered coordinates (0X)	Poorly covered coordinates (<10X)
# Inaccessible genomic coordinates by kit design:	-1nt (0%)	-1nt (0%)
All genomic coordinates:	223nt (0%)	333nt (1%)
Common SNPs:	0nt (0%)	0nt (0%)
Diverse SNPs:	9nt (4%)	9nt (4%)
Rare SNPs:	36nt (3%)	36nt (3%)

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.

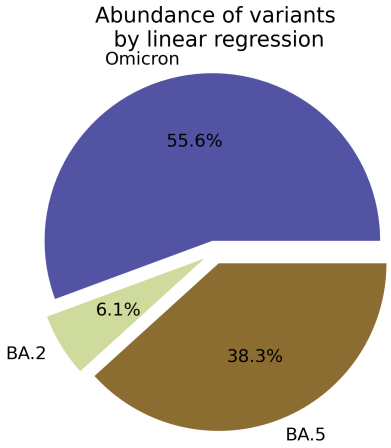


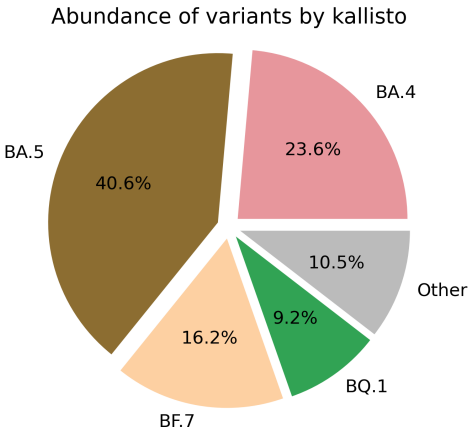


Hits to SARS-Cov2 genome (kraken2):	1895226 reads (96.84%)
Hits to human genome (kraken2):	345 reads (0.02%)
Hits to synthetic sequences (kraken2, taxid 28384):	50 reads (0.00%)
Most abundant organisms (kraken2, family level):	Coronaviridae (96.84%) Bacillaceae (0.32%) Akkermansiaceae (0.04%)

	Desulfovibrionaceae (0.02%) Enterobacteriaceae (0.02%)

Detected variants (Experimental)

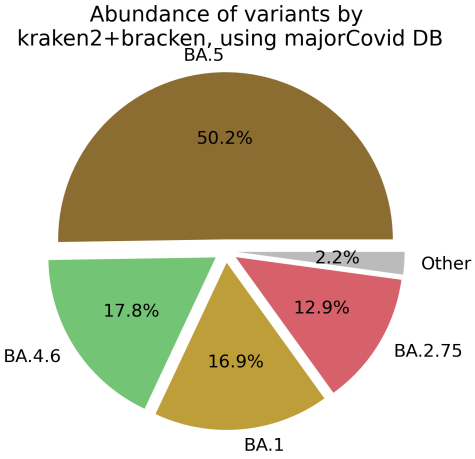


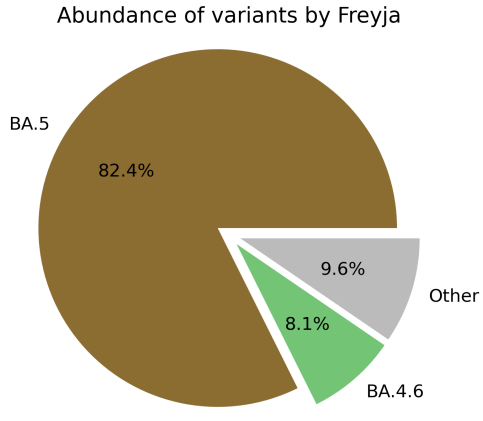


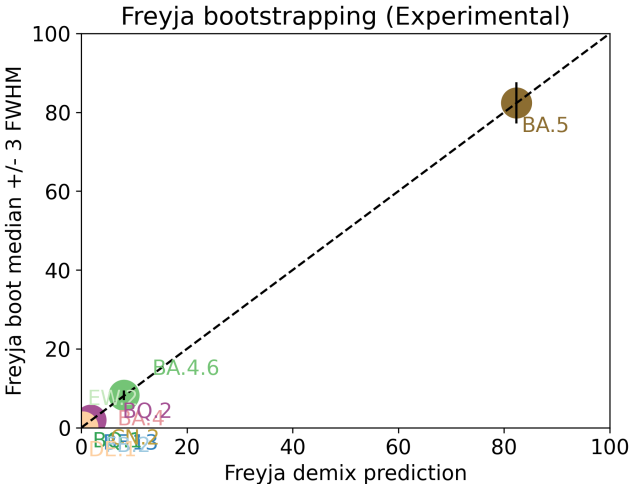
Based on deconvolution, B.1.1.529 is estimated to constitute 55.67% of the viral particles and hence is the most abundant variant in the sample. The R^2 for the linear regression was 0.57. Variants that were detected less than 5% were grouped under "Other"

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.