U.S. FOOD & DRUG ADMINISTRATION

CFSAN/OAO
BIOSTATISTICS AND BIOINFORMATICS STAFF

WASTEWATER SARS-COV2 ANALYSIS REPORT

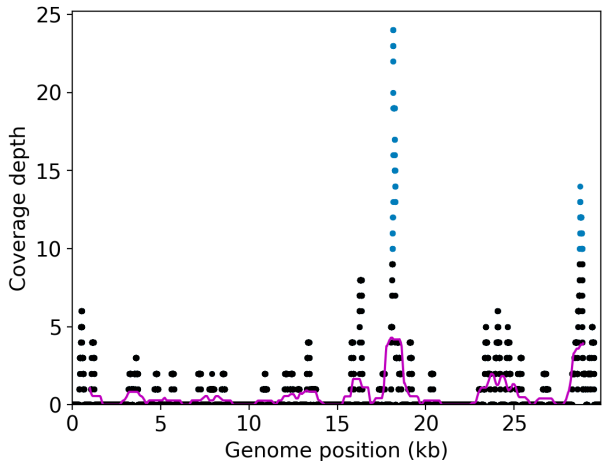
Sample name:	SRR16828014
Date generated:	2022-03-15, 14:20:14 EDT
Executed by:	Tunc Kayikcioglu (Tunc.Kayikcioglu@fda.hhs.gov)
Executed on:	172.20.44.145 (aka n145.raven.cfsan)

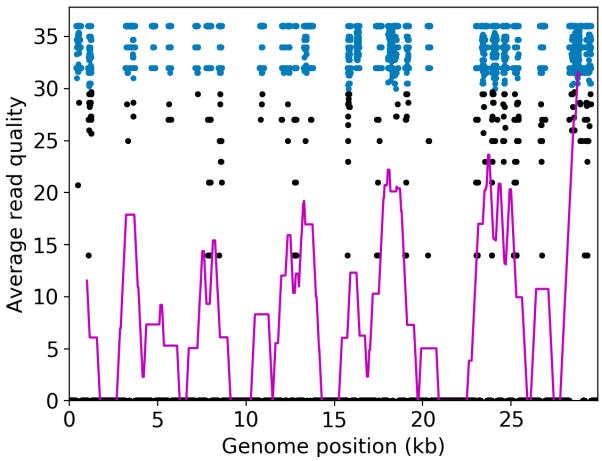
Sequencing summary

Sequencing chemistry:	AMPLICON with NextSeq 500
Source site:	USA: Washington (?,?)
Sampling date:	2021-01-04
Collected by:	Aquavitas
Sequenced by:	Missing
Total number of reads:	4096
Reads aligned:	157 (3%)
Average read quality:	34.4
Average read length:	141
Reads passing filter:	154 (3%)
Average read quality passing filter:	34.5
Average read length passing filter:	141
Average coverage passing filter:	0X

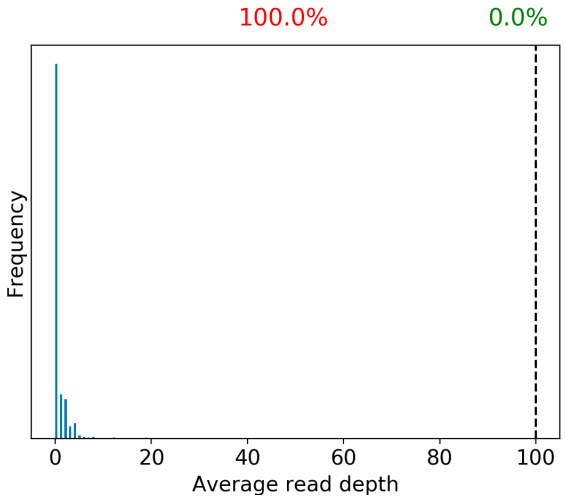
A read passes filter if the read length after adaptor trimming ≥ 30 and minimum read quality ≥ 20 within a sliding window of width 4.

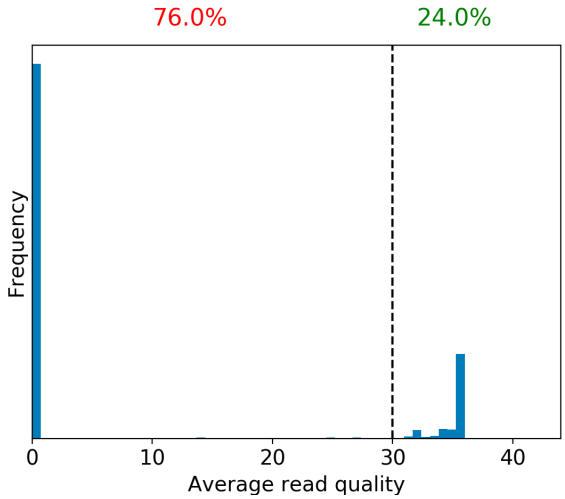
Overall sequence characteristics

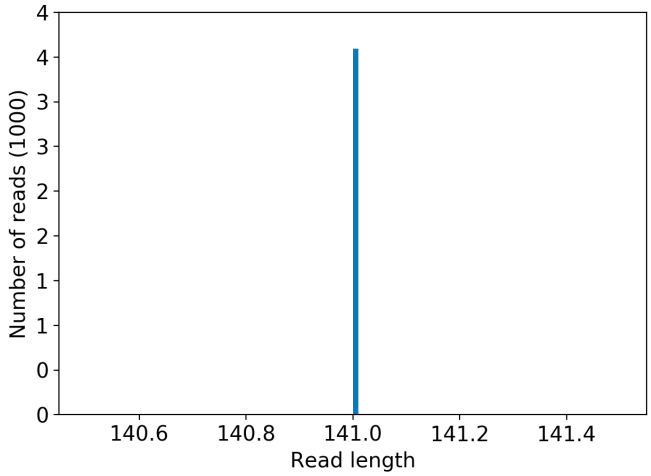


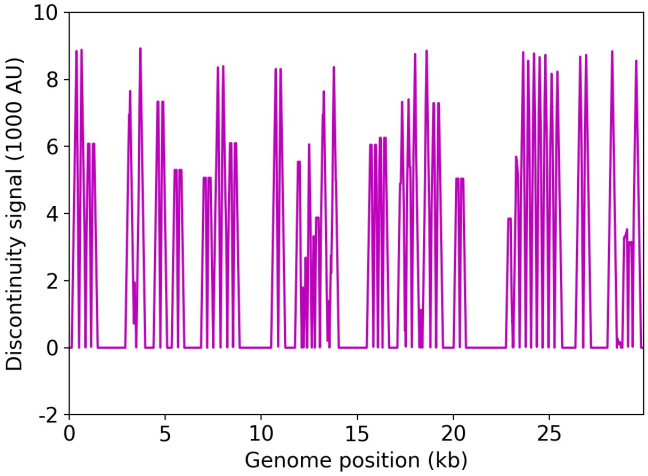


NOTE: The red shaded areas marked with a (*) are not covered by the design of the library preparation kit and hence excluded from analyses. Magenta curves represent moving average with a window width of 1kb.









WARNING: The sequence coverage is very low (0X)

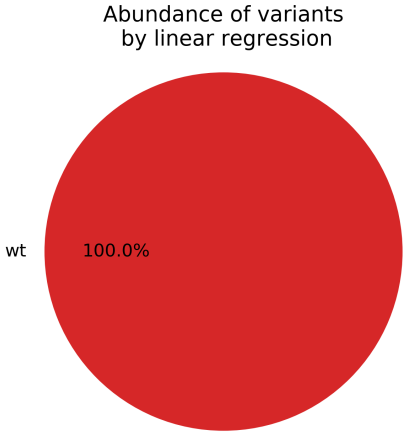
	Uncovered coordinates (0X)	Poorly covered coordinates (<10X)
# Inaccessible genomic coordinates by kit design:	-1nt (0%)	-1nt (0%)

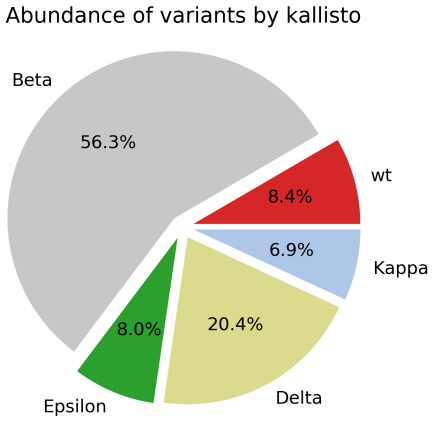
All genomic coordinates:	22376nt (74%)	29615nt (99%)
Common SNPs:	16nt (50%)	31nt (96%)
Diverse SNPs:	491nt (97%)	505nt (100%)
Rare SNPs:	2315nt (92%)	2500nt (100%)

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):	282 reads (13.77%)
Hits to human genome (kraken2):	302 reads (14.75%)
Hits to synthetic sequences (kraken2, taxid 28384):	0 reads (0.00%)

Detected variants (Experimental)

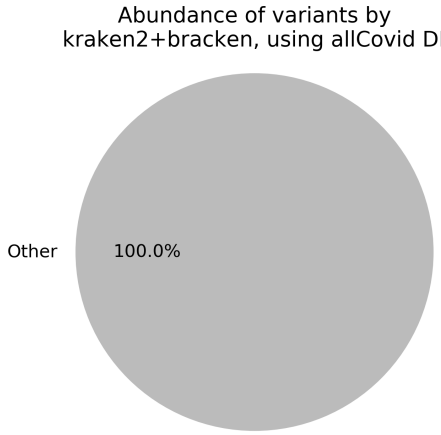


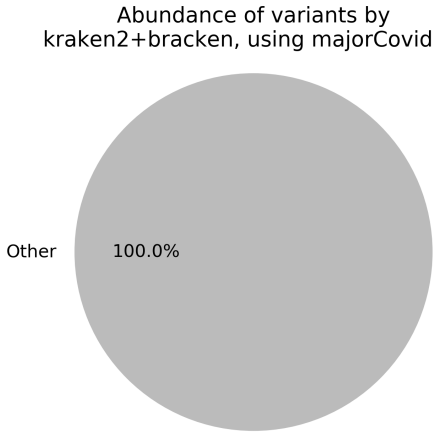


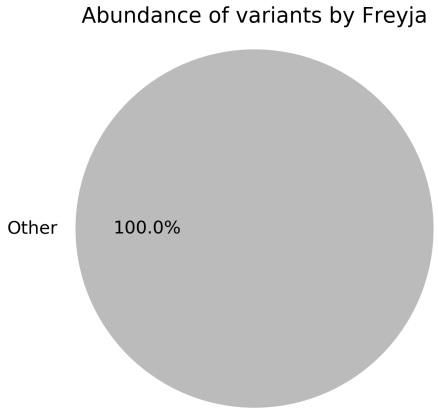
Based on deconvolution, wt is estimated to constitute 100.00% of the viral particles and hence is the most abundant variant in the sample. The R^2 for the linear regression was 1.00. Variants that were detected less than 5% were grouped under "Other"

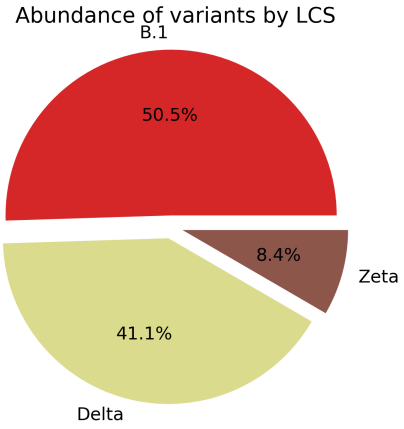
Based on the consensus sequence of the observed reads, the "ensemble-averaged sequence" most closely resembles the Unknown lineage. If this is a sample consisting of a single source of pathogens or an overwhelming majority of the different sources are infected with the same variant, the sample is dominated 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 Beta 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	AY.4	AY.4.2	B.1.617.2	BA.1	BA.2	BA.3
Characteristic mutations detected	(0 of 19)	(0 of 3)	(0 of 13)	(0 of 23)	(0 of 28)	(0 of 19)

Detected mutations

Only genomic coordinates with at least 10X coverage were considered.

Position	Ref. base	Alt. base	Alt. freq	p-value	Mutation name	Compatible lineages
18268	G	T	0.364	5.49E-04	ORF1AB:E1601*	None found
18270	A	C	0.500	1.55E-03	ORF1AB:E1601D	None found
18271	G	T	0.385	1.62E-04	ORF1AB:E1602*	None found