# COVID-19 subject S-210128-01783

2021-05-05

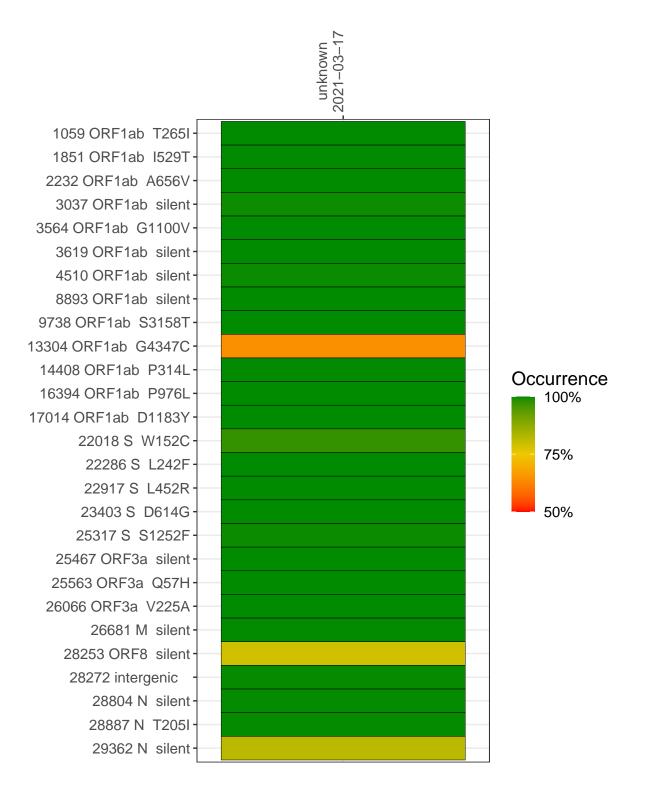
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software tool (Rambaut et al 2020) for genomes with > 90% sequence coverage.

Table 1. Sample summary.

Experiment	Туре	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP1056-1	single experiment	NA	unknown	2021-03-17	6.48	NA	92.1%	91.7%

#### Variants shared across samples

The heat map below shows how variants (reference genome /home/everett/projects/SARS-CoV-2-Philadelphia/Wuhan-Hu-1) are shared across subject samples where the percent variance is colored. Variants are called if a variant position is covered by 5 or more reads, the alternative base is found in > 50% of read pairs and the variant yields a PHRED score > 20. Gray tiles denote positions where the variant was not the major variant or no variants were found. The relative base compositions of each experiment used to calculate tiles are shown in the following plot where the total number of position reads are shown atop of each plot.



#### unknown 2021-03-17

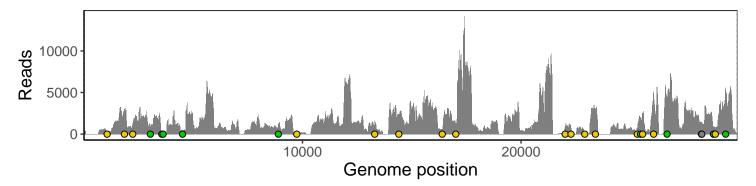
	2021-03-17
1059 ORF1ab T265I	212
1851 ORF1ab I529T	2377
2232 ORF1ab A656V	1298
3037 ORF1ab silent	870
3564 ORF1ab G1100V	268
3619 ORF1ab silent	150
4510 ORF1ab silent	427
8893 ORF1ab silent	1033
9738 ORF1ab S3158T	103
13304 ORF1ab G4347C	1362
14408 ORF1ab P314L	2103
16394 ORF1ab P976L	606
17014 ORF1ab D1183Y	1213
22018 S W152C	863
22286 S L242F	708
22917 S L452R	45
23403 S D614G	2936
25317 S S1252F	757
25467 ORF3a silent	1926
25563 ORF3a Q57H	2477
26066 ORF3a V225A	4688
26681 M silent	3760
28253 ORF8 silent	1439
28272 intergenic	1352
28804 N silent	897
28887 N T205I	776
29362 N silent	5113
	1056-1
	105



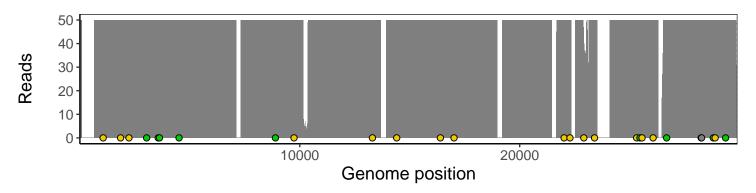
### Analyses of individual experiments and composite results

#### $VSP1056\text{-}1 \mid 2021\text{-}03\text{-}17 \mid unknown \mid S\text{-}210128\text{-}01783 \mid genomes \mid single \ experiment$

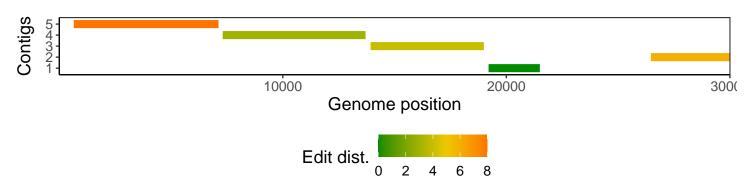
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



Excerpt from plot above focusing on reads coverage from 0 to 50 NT.



The longest five assembled contigs are shown below colored by their edit distance to the reference genome.



## Software environment

Software/R package	Version
R	3.4.0
bwa	0.7.17-r1198-dirty
samtools	1.10 Using htslib 1.10
bcftools	1.10.2-34-g1a12af0-dirty Using htslib 1.10.2-57-gf58a6f3
pangolin	2.3.8
genbankr	1.4.0
optparse	1.6.0
forcats	0.3.0
stringr	1.4.0
dplyr	0.8.1
purrr	0.2.5
readr	1.1.1
tidyr	0.8.1
tibble	2.1.2
ggplot2	3.0.0
tidyverse	1.2.1
ShortRead	1.34.2
${\it Genomic Alignments}$	1.12.2
SummarizedExperiment	1.6.5
DelayedArray	0.2.7
matrixStats	0.54.0
Biobase	2.36.2
Rsamtools	1.28.0
GenomicRanges	1.28.6
$\operatorname{GenomeInfoDb}$	1.12.3
Biostrings	2.44.2
XVector	0.16.0
IRanges	2.10.5
S4Vectors	0.14.7
BiocParallel	1.10.1
BiocGenerics	0.22.1