# COVID-19 subject S-210226-01457

2021-05-05

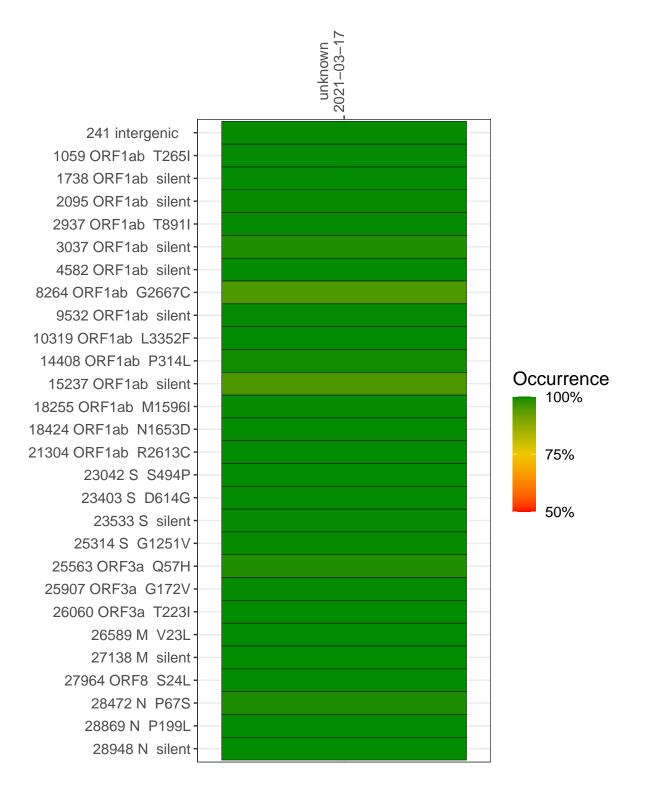
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP1060-1	single experiment	NA	unknown	2021-03-17	29.84	B.1.2	99.9%	99.8%

#### 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

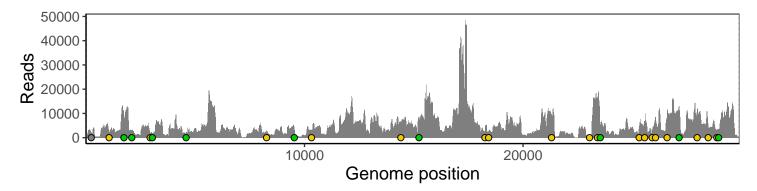
	2021-03-17
241 intergenic	1979
1059 ORF1ab T265I	2335
1738 ORF1ab silent	7816
2095 ORF1ab silent	2682
2937 ORF1ab T891I	1847
3037 ORF1ab silent	1621
4582 ORF1ab silent	1413
8264 ORF1ab G2667C	2508
9532 ORF1ab silent	4146
10319 ORF1ab L3352F	3173
14408 ORF1ab P314L	5888
15237 ORF1ab silent	7901
18255 ORF1ab M1596l	2730
18424 ORF1ab N1653D	2572
21304 ORF1ab R2613C	9039
23042 S S494P	82
23403 S D614G	16188
23533 S silent	4870
25314 S G1251V	4790
25563 ORF3a Q57H	6700
25907 ORF3a G172V	2006
26060 ORF3a T223I	8249
26589 M V23L	10083
27138 M silent	12987
27964 ORF8 S24L	11321
28472 N P67S	5652
28869 N P199L	2991
28948 N silent	3539
	0-1
	7060-1



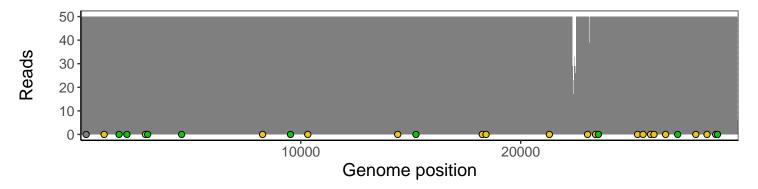
## Analyses of individual experiments and composite results

## $VSP1060\text{-}1 \mid 2021\text{-}03\text{-}17 \mid unknown \mid S\text{-}210226\text{-}01457 \mid genomes \mid single \ experiment$

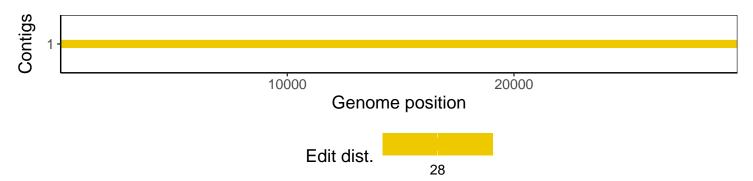
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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