COVID-19 subject S-210122-02276

2021-04-17

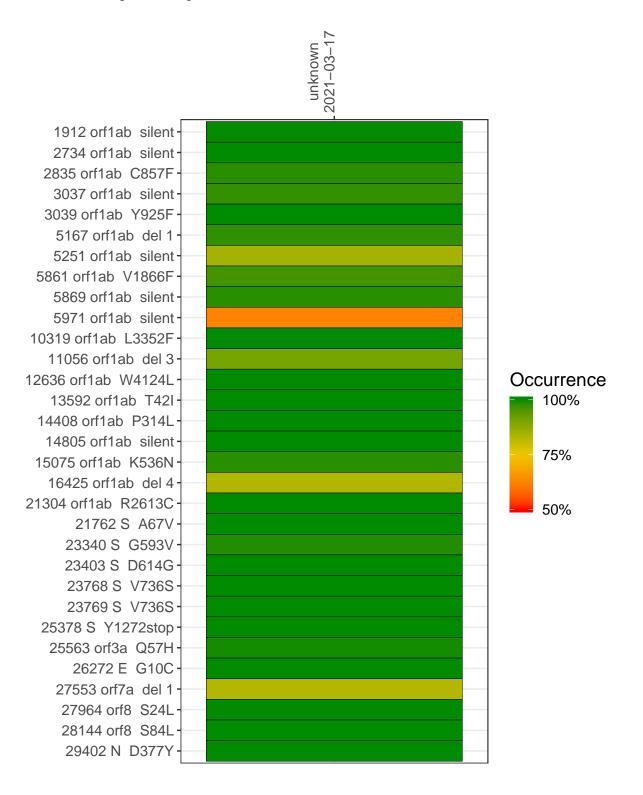
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP1051-1	single experiment	NA	unknown	2021-03-17	1.96	NA	57.2%	55.8%

Variants shared across samples

The heat map below shows how variants (reference genome /home/everett/projects/SARS-CoV-2-Philadelphia/USA-WA1-2020) 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

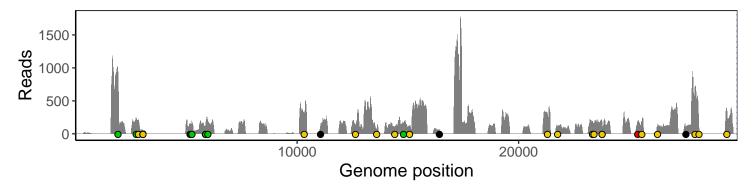
	2021-03-17
1912 orf1ab silent	921
2734 orf1ab silent	220
2835 orf1ab C857F	188
3037 orf1ab silent	44
3039 orf1ab Y925F	43
5167 orf1ab del 1	148
5251 orf1ab silent	191
5861 orf1ab V1866F	108
5869 orf1ab silent	127
5971 orf1ab silent	186
10319 orf1ab L3352F	309
11056 orf1ab del 3	117
12636 orf1ab W4124L	126
13592 orf1ab T42I	67
14408 orf1ab P314L	93
14805 orf1ab silent	161
15075 orf1ab K536N	134
16425 orf1ab del 4	65
21304 orf1ab R2613C	282
21762 S A67V	101
23340 S G593V	183
23403 S D614G	165
23768 S V736S	176
23769 S V736S	176
25378 S Y1272stop	181
25563 orf3a Q57H	196
26272 E G10C	118
27553 orf7a del 1	105
27964 orf8 S24L	705
28144 orf8 S84L	104
29402 N D377Y	397
	1051–1
	105



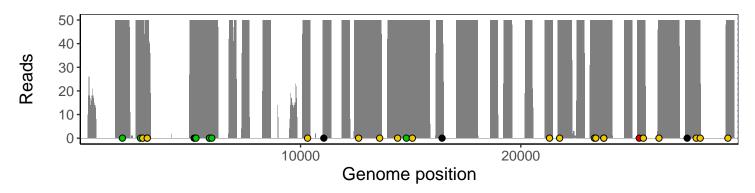
Analyses of individual experiments and composite results

$VSP1051-1 \ | \ 2021-03-17 \ | \ unknown \ | \ S-210122-02276 \ | \ genomes \ | \ single \ experiment$

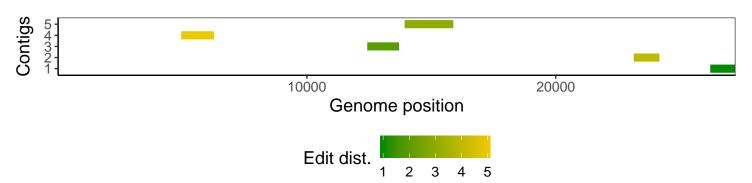
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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