COVID-19 subject UPHS-0617

2021-06-23

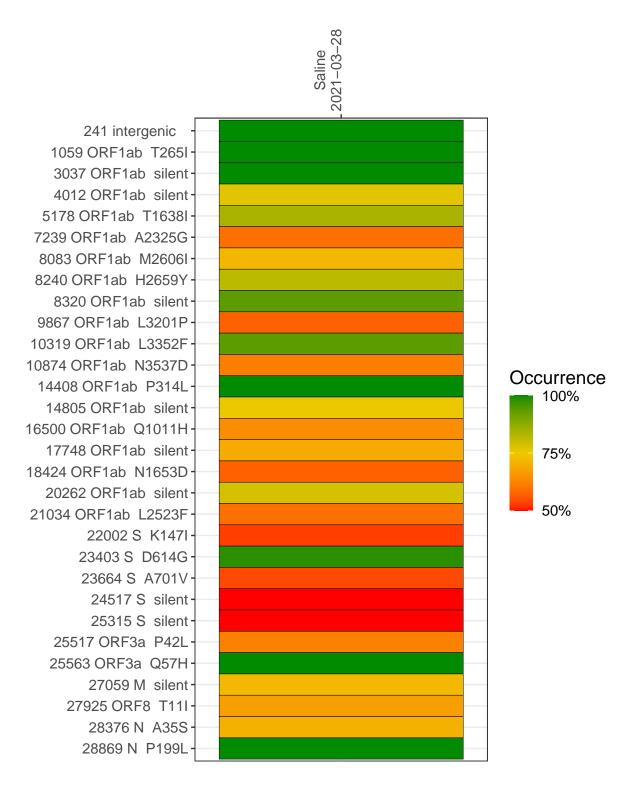
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP1802-1	single experiment	NA	Saline	2021-03-28	22.50	B.1.2	99.7%	99.6%

Variants shared across samples

The heat map below shows how variants (reference genome /home/common/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.



Saline 2021-03-28

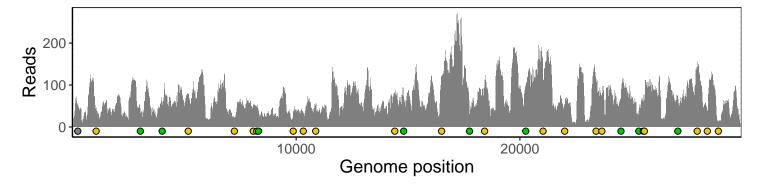
	2021-03-28
241 intergenic	43
1059 ORF1ab T265I	41
3037 ORF1ab silent	39
4012 ORF1ab silent	56
5178 ORF1ab T1638I	69
7239 ORF1ab A2325G	22
8083 ORF1ab M2606I	49
8240 ORF1ab H2659Y	44
8320 ORF1ab silent	31
9867 ORF1ab L3201P	21
10319 ORF1ab L3352F	31
10874 ORF1ab N3537D	44
14408 ORF1ab P314L	71
14805 ORF1ab silent	69
16500 ORF1ab Q1011H	111
17748 ORF1ab silent	26
18424 ORF1ab N1653D	100
20262 ORF1ab silent	88
21034 ORF1ab L2523F	108
22002 S K147I	49
23403 S D614G	114
23664 S A701V	72
24517 S silent	66
25315 S silent	34
25517 ORF3a P42L	55
25563 ORF3a Q57H	71
27059 M silent	74
27925 ORF8 T11I	152
28376 N A35S	37
28869 N P199L	12
	2–7
	VSP1802-1
	/SP



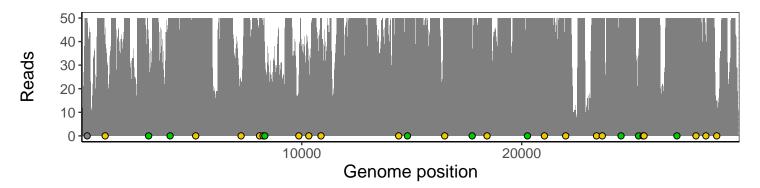
Analyses of individual experiments and composite results

$VSP1802\text{-}1 \mid 2021\text{-}03\text{-}28 \mid Saline \mid UPHS\text{-}0617 \mid genomes \mid single \ experiment$

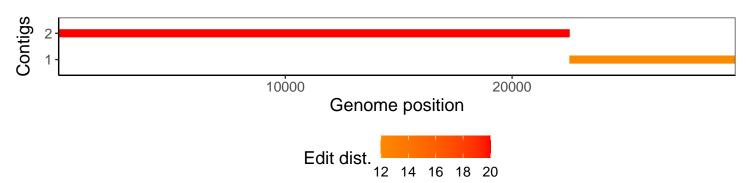
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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	3.1.3
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	3.3.3
tidyverse	1.2.1
ShortRead	1.34.2
GenomicAlignments	1.12.2
${\bf Summarized Experiment}$	1.6.5
DelayedArray	0.2.7
matrixStats	0.54.0
Biobase	2.36.2
Rsamtools	1.28.0
GenomicRanges	1.28.6
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