COVID-19 subject UPHS-1644

2021-06-23

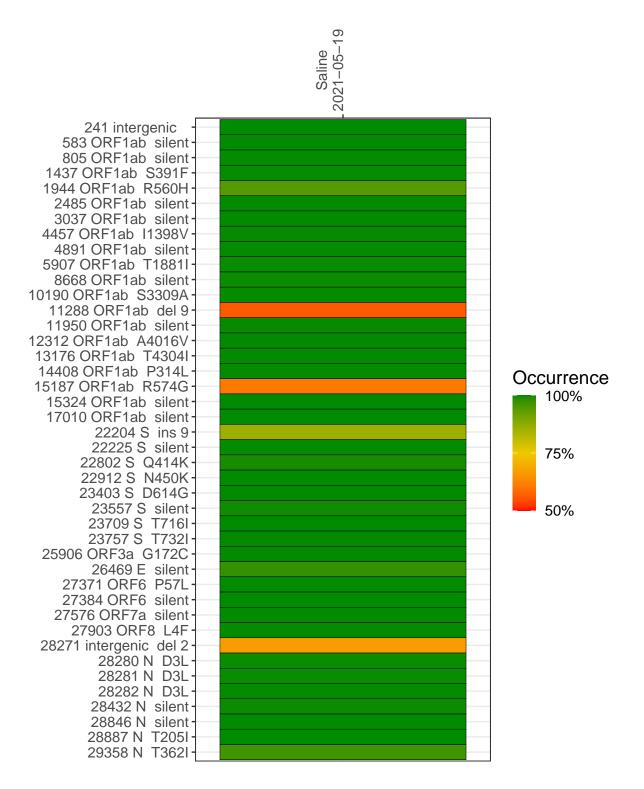
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software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

Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP2945-1	single experiment	NA	Saline	2021-05-19	29.80	B.1.214.2	99.8%	99.8%

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 2021-05-19

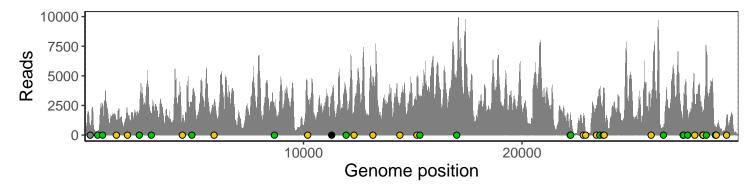
	2021-05-19
241 intergenic	696
583 ORF1ab silent	373
805 ORF1ab silent	2401
1437 ORF1ab S391F	1266
1944 ORF1ab R560H	1733
2485 ORF1ab silent	1697
3037 ORF1ab silent	1617
4457 ORF1ab I1398V	1403
4891 ORF1ab silent	2306
5907 ORF1ab T1881I	2377
8668 ORF1ab silent	2334
10190 ORF1ab S3309A	3334
11288 ORF1ab del 9	1756
11950 ORF1ab silent	3292
12312 ORF1ab A4016V	1937
13176 ORF1ab T4304I	4849
14408 ORF1ab P314L	2107
15187 ORF1ab R574G	4881
15324 ORF1ab silent	3110
17010 ORF1ab silent	4870
22204 S ins 9	1154
22225 S silent	1793
22802 S Q414K	1591
22912 S N450K	173
23403 S D614G	2485
23557 S silent	3746
23709 S T716I	2986
23757 S T732I	3966
25906 ORF3a G172C	2254
26469 E silent	646
27371 ORF6 P57L	1627
27384 ORF6 silent	2129
27576 ORF7a silent	1887
27903 ORF8 L4F	2686
28271 intergenic del 2	1648
28280 N D3L	1084
28281 N D3L	1084
28282 N D3L	1171
28432 N silent	6803
28846 N silent	775
28887 N T205I	631
29358 N T362I	1512
	<u> </u>
	345.
	\sim



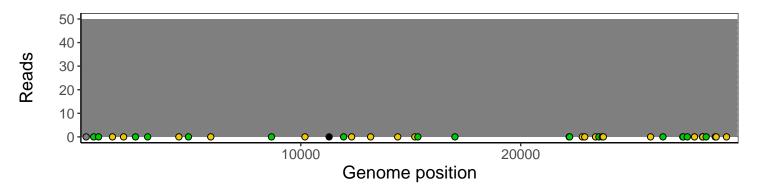
Analyses of individual experiments and composite results

$VSP2945\text{-}1 \mid 2021\text{-}05\text{-}19 \mid Saline \mid UPHS\text{-}1644 \mid genomes \mid single \ experiment$

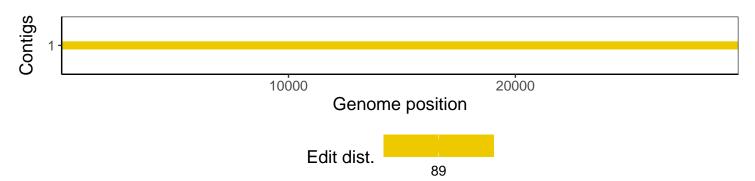
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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