COVID-19 subject UPHS-0093

2021-06-23

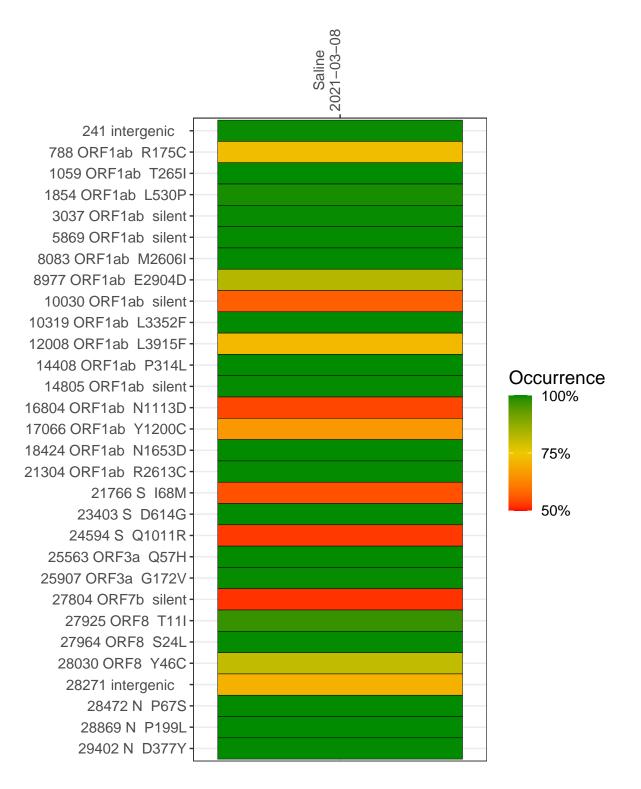
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP1024-1	single experiment	NA	Saline	2021-03-08	29.86	B.1.2	99.7%	99.7%

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 2021-03-08

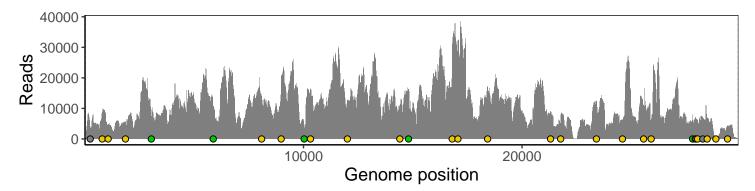
	2021 00 00
241 intergenic	3658
788 ORF1ab R175C	8481
1059 ORF1ab T265I	4581
1854 ORF1ab L530P	5409
3037 ORF1ab silent	7615
5869 ORF1ab silent	9618
8083 ORF1ab M2606I	13515
8977 ORF1ab E2904D	19517
10030 ORF1ab silent	4559
10319 ORF1ab L3352F	15054
12008 ORF1ab L3915F	12527
14408 ORF1ab P314L	11006
14805 ORF1ab silent	16259
16804 ORF1ab N1113D	30810
17066 ORF1ab Y1200C	31717
18424 ORF1ab N1653D	15970
21304 ORF1ab R2613C	8849
21766 S 168M	7453
23403 S D614G	12299
24594 S Q1011R	6211
25563 ORF3a Q57H	10812
25907 ORF3a G172V	5019
27804 ORF7b silent	7756
27925 ORF8 T11I	6908
27964 ORF8 S24L	6612
28030 ORF8 Y46C	6140
28271 intergenic	7070
28472 N P67S	6869
28869 N P199L	948
29402 N D377Y	4207
	4- 1-



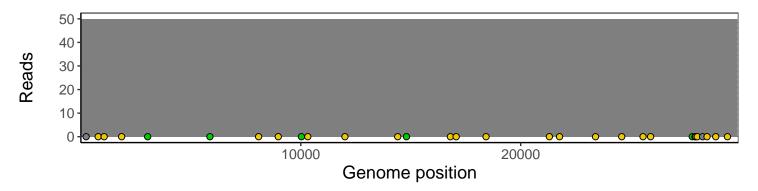
Analyses of individual experiments and composite results

$VSP1024\text{-}1 \mid 2021\text{-}03\text{-}08 \mid Saline \mid UPHS\text{-}0093 \mid genomes \mid single \ experiment$

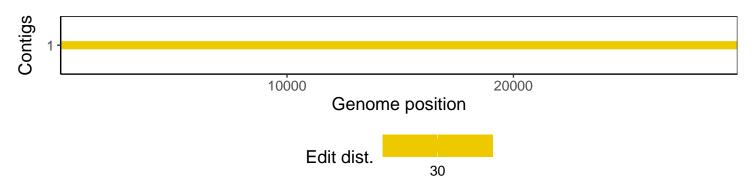
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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