COVID-19 subject UPHS-0088

2021-05-05

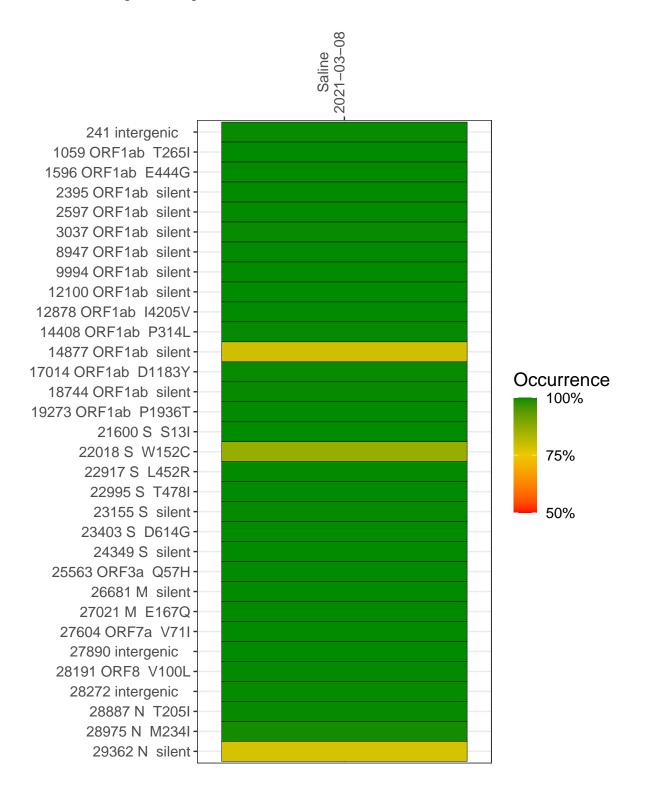
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP1020-1	single experiment	NA	Saline	2021-03-08	29.99	B.1.429	99.8%	99.7%

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.



Saline 2021-03-08

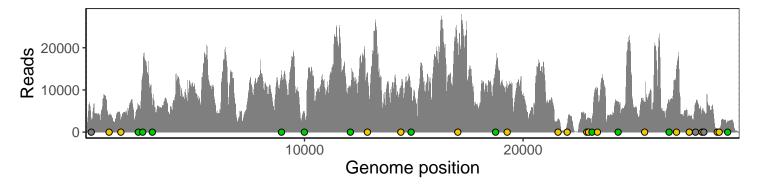
	202: 00 00
241 intergenic	3062
1059 ORF1ab T265I	4004
1596 ORF1ab E444G	2549
2395 ORF1ab silent	6525
2597 ORF1ab silent	12438
3037 ORF1ab silent	6634
8947 ORF1ab silent	13724
9994 ORF1ab silent	4620
12100 ORF1ab silent	9612
12878 ORF1ab I4205V	13910
14408 ORF1ab P314L	10836
14877 ORF1ab silent	10294
17014 ORF1ab D1183Y	19022
18744 ORF1ab silent	15116
19273 ORF1ab P1936T	10899
21600 S S13I	3188
22018 S W152C	2075
22917 S L452R	2718
22995 S T478I	3269
23155 S silent	3264
23403 S D614G	10388
24349 S silent	6752
25563 ORF3a Q57H	7185
26681 M silent	6507
27021 M E167Q	14929
27604 ORF7a V71I	4159
27890 intergenic	6714
28191 ORF8 V100L	5218
28272 intergenic	4914
28887 N T205I	692
28975 N M234I	570
29362 N silent	2253
	7



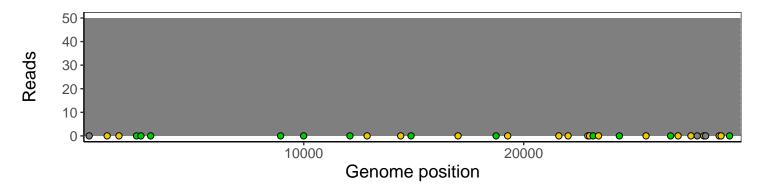
Analyses of individual experiments and composite results

$VSP1020\text{-}1 \mid 2021\text{-}03\text{-}08 \mid Saline \mid UPHS\text{-}0088 \mid genomes \mid single \ experiment$

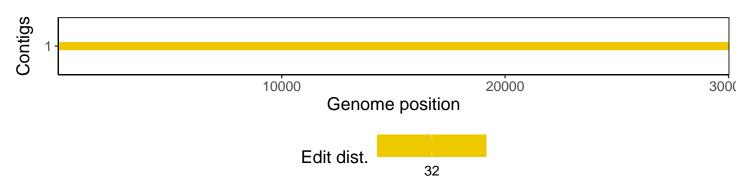
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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