COVID-19 subject UPHS-0338

2021-05-05

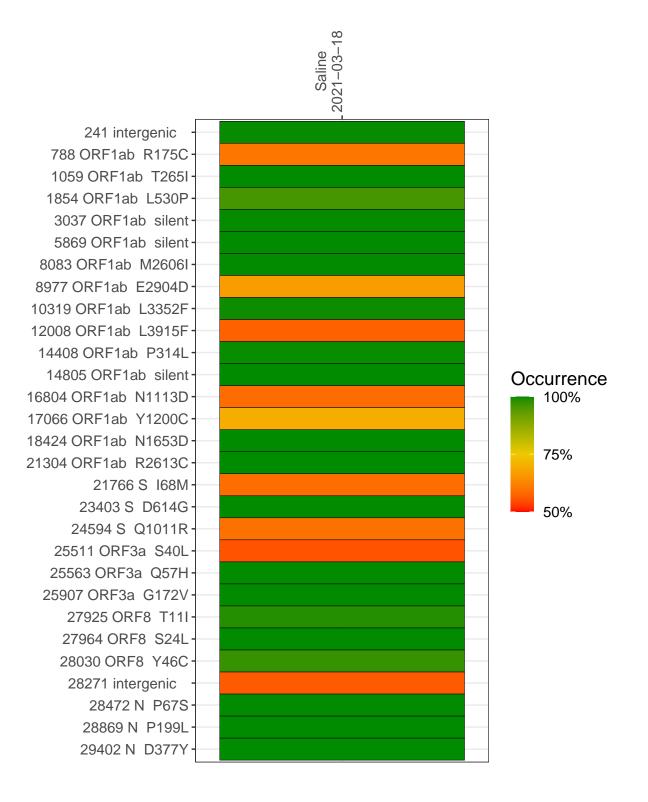
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP1383-1	single experiment	NA	Saline	2021-03-18	29.85	B.1.2	99.7%	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 2021-03-18

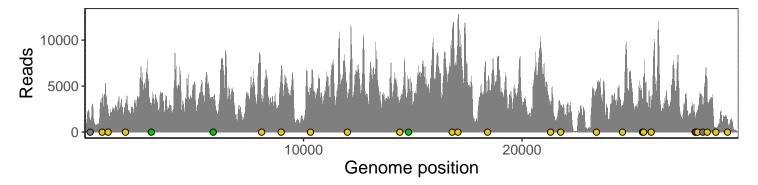
	2021-03-10
241 intergenic	1237
788 ORF1ab R175C	3452
1059 ORF1ab T265I	2206
1854 ORF1ab L530P	2961
3037 ORF1ab silent	2606
5869 ORF1ab silent	3326
8083 ORF1ab M2606I	3361
8977 ORF1ab E2904D	6780
10319 ORF1ab L3352F	4680
12008 ORF1ab L3915F	4056
14408 ORF1ab P314L	2679
14805 ORF1ab silent	5686
16804 ORF1ab N1113D	11095
17066 ORF1ab Y1200C	12124
18424 ORF1ab N1653D	4161
21304 ORF1ab R2613C	2895
21766 S 168M	2939
23403 S D614G	4568
24594 S Q1011R	2774
25511 ORF3a S40L	3507
25563 ORF3a Q57H	5901
25907 ORF3a G172V	2528
27925 ORF8 T11I	1531
27964 ORF8 S24L	1836
28030 ORF8 Y46C	2292
28271 intergenic	3009
28472 N P67S	5693
28869 N P199L	446
29402 N D377Y	1384
	1383–1
	138



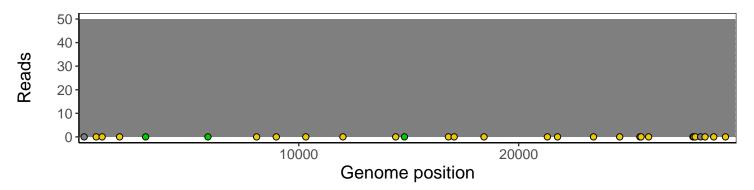
Analyses of individual experiments and composite results

VSP1383-1 | 2021-03-18 | Saline | UPHS-0338 | genomes | single experiment

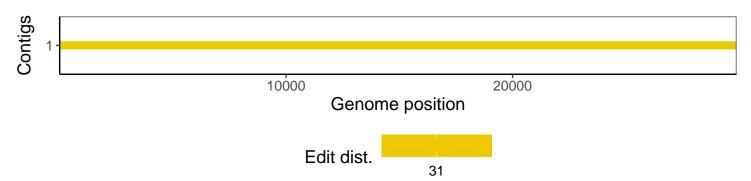
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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