COVID-19 subject UPHS-0312

2021-04-17

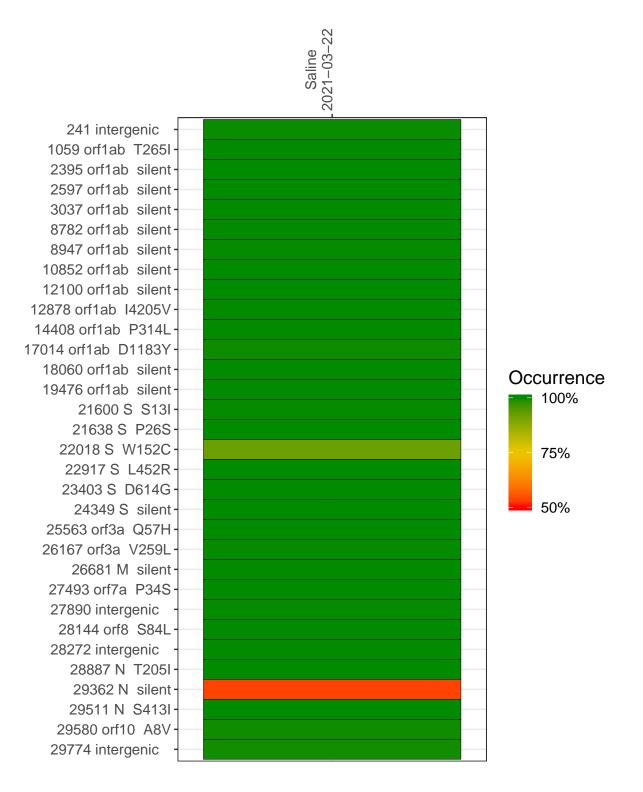
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP1357-1	single experiment	NA	Saline	2021-03-22	29.91	B.1.429	99.9%	99.8%

Variants shared across samples

The heat map below shows how variants (reference genome /home/everett/projects/SARS-CoV-2-Philadelphia/USA-WA1-2020) 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 2021-03-22

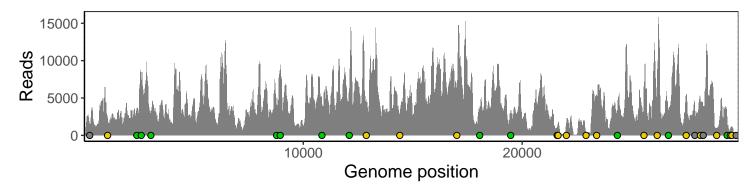
	2021-03-22
241 intergenic	1462
1059 orf1ab T265I	2083
2395 orf1ab silent	3603
2597 orf1ab silent	6043
3037 orf1ab silent	2728
8782 orf1ab silent	3758
8947 orf1ab silent	9166
10852 orf1ab silent	4167
12100 orf1ab silent	5690
12878 orf1ab I4205V	4133
14408 orf1ab P314L	2407
17014 orf1ab D1183Y	5841
18060 orf1ab silent	3335
19476 orf1ab silent	2718
21600 S S13I	1117
21638 S P26S	1470
22018 S W152C	1580
22917 S L452R	357
23403 S D614G	5309
24349 S silent	2795
25563 orf3a Q57H	6368
26167 orf3a V259L	5532
26681 M silent	2994
27493 orf7a P34S	4808
27890 intergenic	3362
28144 orf8 S84L	4511
28272 intergenic	3757
28887 N T205I	895
29362 N silent	1951
29511 N S413I	891
29580 orf10 A8V	1105
29774 intergenic	204
	. 1



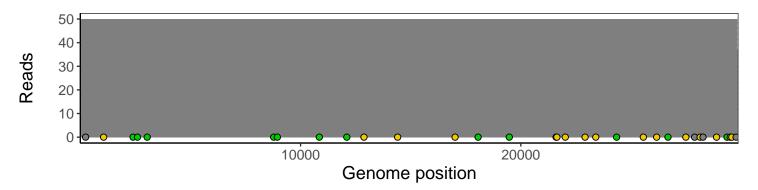
Analyses of individual experiments and composite results

VSP1357-1 | 2021-03-22 | Saline | UPHS-0312 | genomes | single experiment

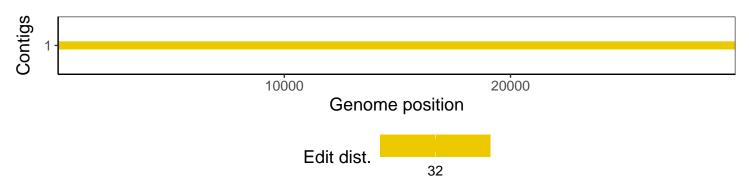
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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