COVID-19 subject UPHS-0096

2021-04-17

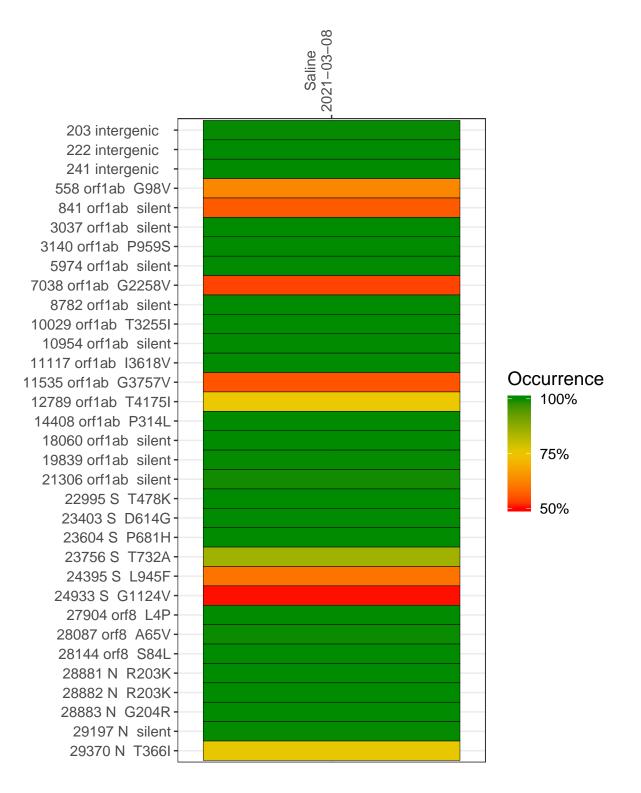
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software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

Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP1027-1	single experiment	NA	Saline	2021-03-08	26.09	B.1.1.519	99.0%	98.9%

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 2021-03-08

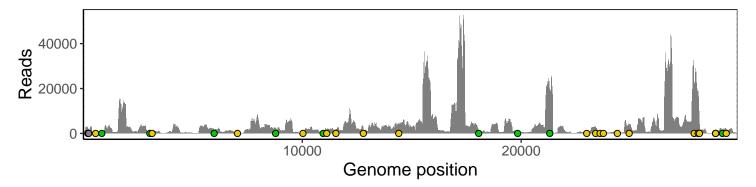
	2021 00 00
203 intergenic	2197
222 intergenic	2081
241 intergenic	1577
558 orf1ab G98V	314
841 orf1ab silent	267
3037 orf1ab silent	369
3140 orf1ab P959S	350
5974 orf1ab silent	1341
7038 orf1ab G2258V	323
8782 orf1ab silent	2084
10029 orf1ab T3255I	251
10954 orf1ab silent	738
11117 orf1ab I3618V	1667
11535 orf1ab G3757V	1805
12789 orf1ab T4175I	2072
14408 orf1ab P314L	3848
18060 orf1ab silent	72 6
19839 orf1ab silent	1167
21306 orf1ab silent	18527
22995 S T478K	27
23403 S D614G	2904
23604 S P681H	1312
23756 S T732A	1175
24395 S L945F	695
24933 S G1124V	3748
27904 orf8 L4P	18671
28087 orf8 A65V	16791
28144 orf8 S84L	4586
28881 N R203K	25
28882 N R203K	25
28883 N G204R	25
29197 N silent	1527
29370 N T366I	1682
	$\overline{}$



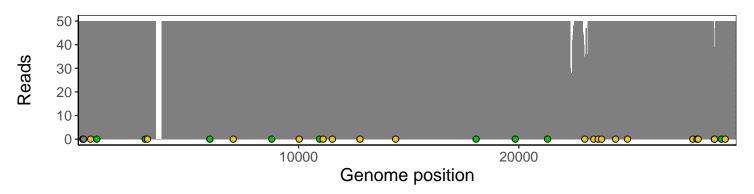
Analyses of individual experiments and composite results

$VSP1027\text{-}1 \mid 2021\text{-}03\text{-}08 \mid Saline \mid UPHS\text{-}0096 \mid genomes \mid single \ experiment$

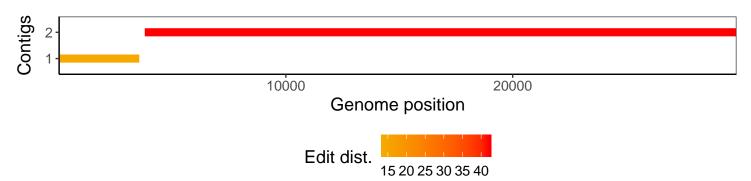
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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