COVID-19 subject UPHS-0076

2021-03-25

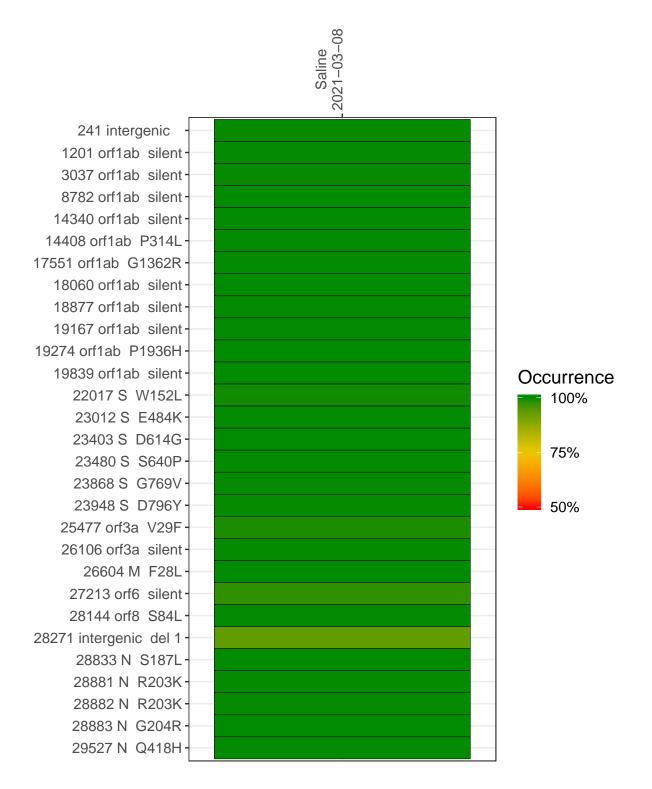
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP1008-1	single experiment	NA	Saline	2021-03-08	29.87	R.1	99.9%	99.8%

Variants shared across samples

The heat map below shows how variants (reference genome 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

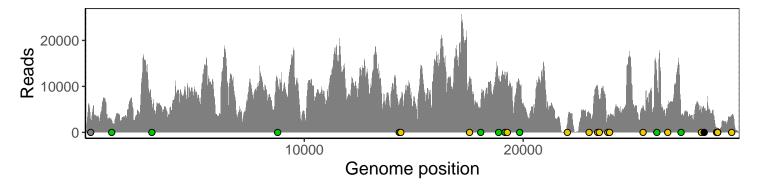
	2021-03-08
241 intergenic	2703
1201 orf1ab silent	2552
3037 orf1ab silent	5563
8782 orf1ab silent	11006
14340 orf1ab silent	6059
14408 orf1ab P314L	7331
17551 orf1ab G1362R	11559
18060 orf1ab silent	6529
18877 orf1ab silent	11266
19167 orf1ab silent	12978
19274 orf1ab P1936H	11758
19839 orf1ab silent	10162
22017 S W152L	1761
23012 S E484K	3989
23403 S D614G	10310
23480 S S640P	7188
23868 S G769V	3676
23948 S D796Y	3198
25477 orf3a V29F	5670
26106 orf3a silent	11844
26604 M F28L	5606
27213 orf6 silent	3229
28144 orf8 S84L	4420
28271 intergenic del 1	5096
28833 N S187L	1073
28881 N R203K	752
28882 N R203K	749
28883 N G204R	752
29527 N Q418H	3439
	3–1



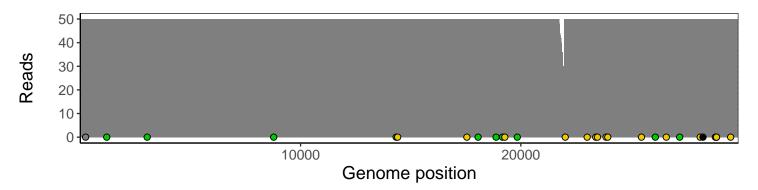
Analyses of individual experiments and composite results

$VSP1008\text{-}1 \mid 2021\text{-}03\text{-}08 \mid Saline \mid UPHS\text{-}0076 \mid genomes \mid single \ experiment$

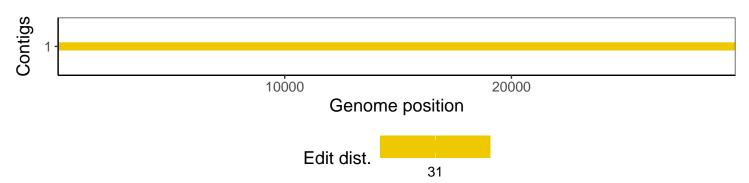
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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	2.3.3
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	3.0.0
tidyverse	1.2.1
ShortRead	1.34.2
$\operatorname{GenomicAlignments}$	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