COVID-19 subject UPHS-0083

2021-03-25

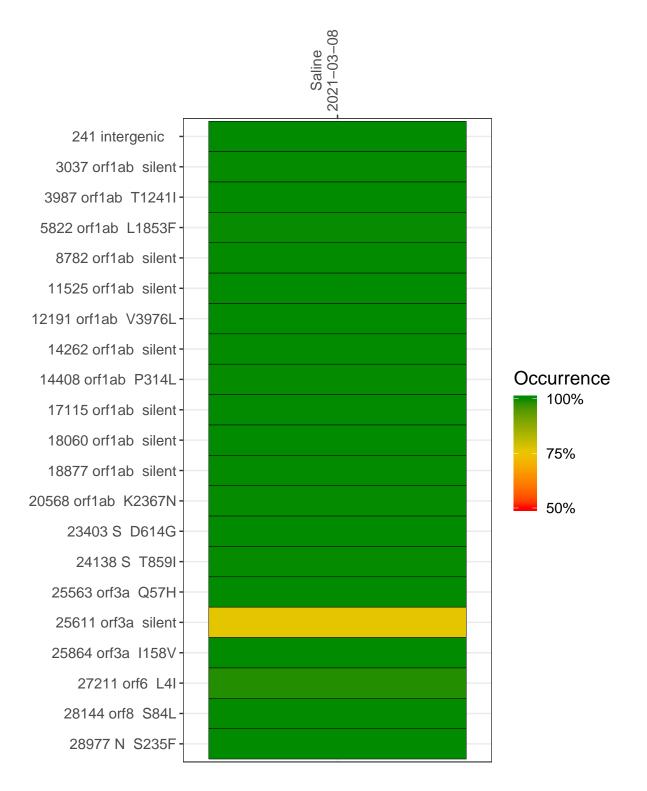
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP1015-1	single experiment	NA	Saline	2021-03-08	29.88	B.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

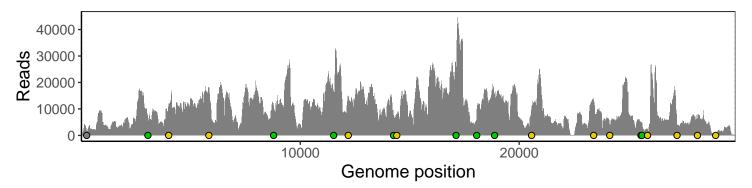
241 intergenic	1722
3037 orf1ab silent	5824
3987 orf1ab T1241I	11604
5822 orf1ab L1853F	17913
8782 orf1ab silent	10717
11525 orf1ab silent	15826
12191 orf1ab V3976L	14292
14262 orf1ab silent	7300
14408 orf1ab P314L	7910
17115 orf1ab silent	24297
18060 orf1ab silent	9152
18877 orf1ab silent	15087
20568 orf1ab K2367N	9859
20568 orf1ab K2367N	9859
20568 orf1ab K2367N 23403 S D614G	9859 13742
20568 orf1ab K2367N 23403 S D614G 24138 S T859I	9859 13742 4306
20568 orf1ab K2367N 23403 S D614G 24138 S T859I 25563 orf3a Q57H	9859 13742 4306 5108
20568 orf1ab K2367N 23403 S D614G 24138 S T859I 25563 orf3a Q57H 25611 orf3a silent	9859 13742 4306 5108
20568 orf1ab K2367N 23403 S D614G 24138 S T859I 25563 orf3a Q57H 25611 orf3a silent 25864 orf3a I158V	9859 13742 4306 5108 4845 2430
20568 orf1ab K2367N 23403 S D614G 24138 S T859I 25563 orf3a Q57H 25611 orf3a silent 25864 orf3a I158V 27211 orf6 L4I	9859 13742 4306 5108 4845 2430 3508



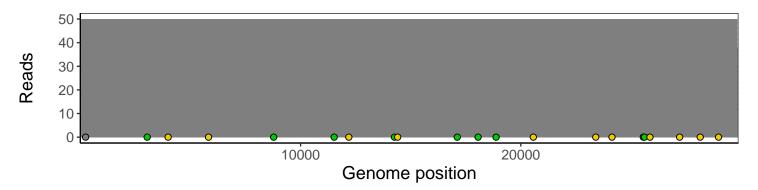
Analyses of individual experiments and composite results

$VSP1015\text{-}1 \mid 2021\text{-}03\text{-}08 \mid Saline \mid UPHS\text{-}0083 \mid genomes \mid single \ experiment$

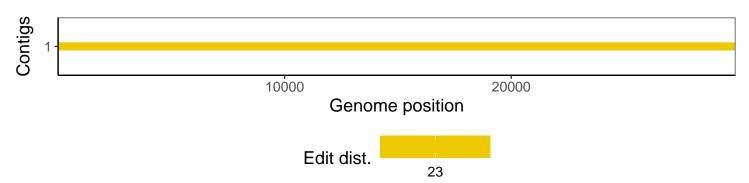
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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