COVID-19 subject UPHS-0868

2021-05-21

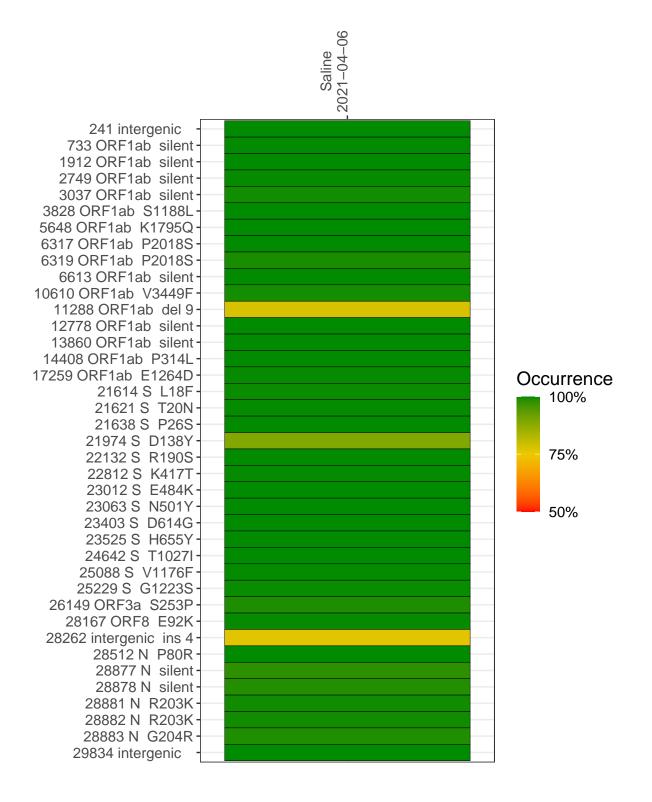
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP2082-2	single experiment	NA	Saline	2021-04-06	29.81	P.1	99.8%	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 2021-04-06

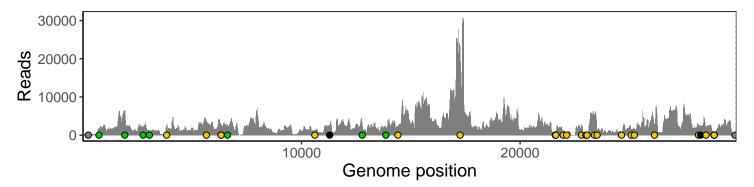
	2021-04-00
241 intergenic	312
733 ORF1ab silent	1754
1912 ORF1ab silent	5927
2749 ORF1ab silent	2084
3037 ORF1 b silent	1045
3828 ORF1ab S1188L	2530
5648 ORF1ab K1795Q	3322
6317 ORF1ab P2018S	2624
6319 ORF1ab P2018S	2532
6613 ORF1ab silent	2641
10610 ORF1ab V3449F	1212
11288 ORF1ab del 9	1585
12778 ORF1ab silent	3053
13860 ORF1ab silent	1813
14408 ORF1ab P314L	2120
17259 ORF1ab E1264D	20668
21614 S L18F	971
21621 S T20N	953
21638 S P26S	1004
21974 S D138Y	925
22132 S R190S	746
22812 S K417T	2496
23012 S E484K	22
23063 S N501Y	32
23403 S D614G	5286
23525 S H655Y	1122
24642 S T1027I	911
25088 S V1176F	1610
25229 S G1223S	1628
26149 ORF3a S253P	2788
28167 ORF8 E92K	2151
28262 intergenic ins 4	1816
28512 N P80R	2153
28877 N silent	473
28878 N silent	469
28881 N R203K	467
28882 N R203K	467
28883 N G204R	480
29834 intergenic	70
	-2
	282
	VSP2082-2
	S >

No data

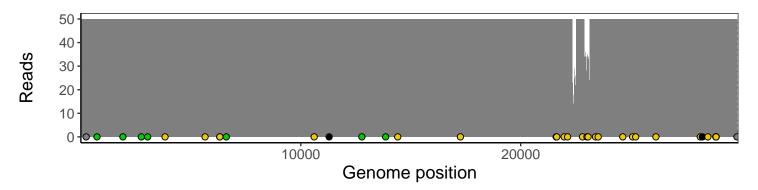
Analyses of individual experiments and composite results

$VSP2082-2\mid 2021-04-06\mid Saline\mid UPHS-0868\mid genomes\mid single\ experiment$

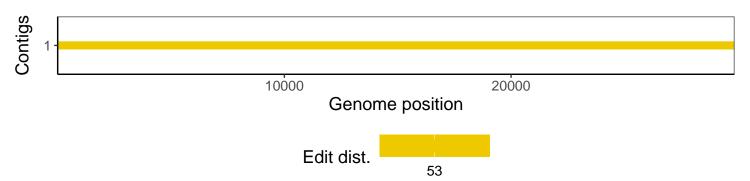
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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	3.3.3
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