COVID-19 subject UPHS-1391

2021-06-23

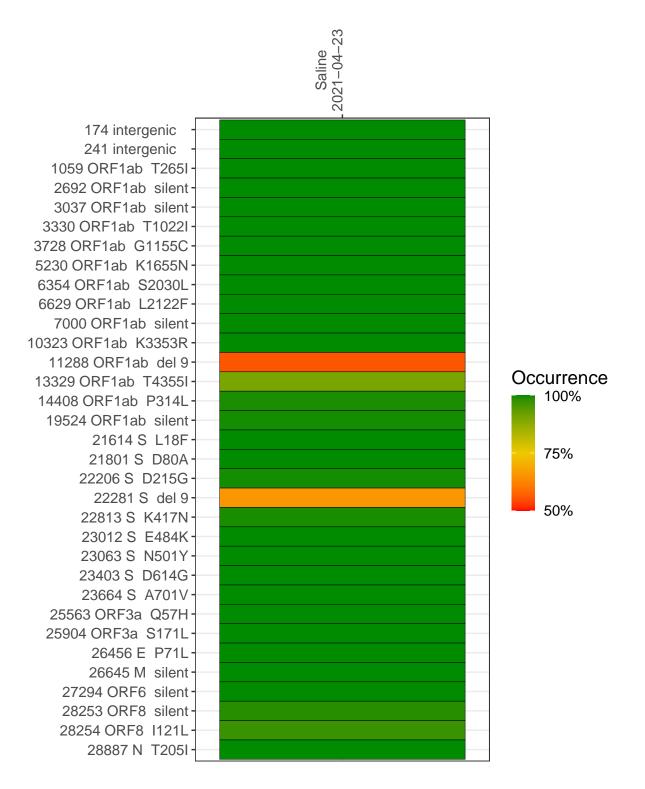
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP2646-1	single experiment	NA	Saline	2021-04-23	22.27	B.1.351	99.6%	99.0%

Variants shared across samples

The heat map below shows how variants (reference genome /home/common/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.



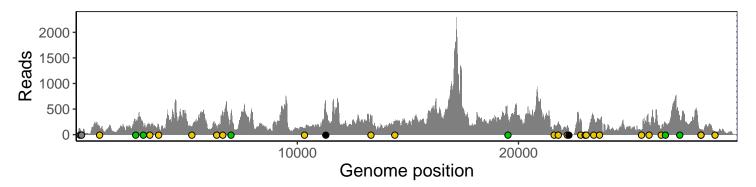
Saline 2021-04-23

	2021-04-23
174 intergenic	115
241 intergenic	56
1059 ORF1ab T265I	150
2692 ORF1ab silent	242
3037 ORF1ab silent	157
3330 ORF1ab T1022I	178
3728 ORF1ab G1155C	188
5230 ORF1ab K1655N	146
6354 ORF1ab S2030L	164
6629 ORF1ab L2122F	352
7000 ORF1ab silent	419
10323 ORF1ab K3353R	130
11288 ORF1ab del 9	335
13329 ORF1ab T4355I	326
14408 ORF1ab P314L	236
19524 ORF1ab silent	343
21614 S L18F	161
21801 S D80A	249
22206 S D215G	166
22281 S del 9	52
22813 S K417N	281
23012 S E484K	28
23063 S N501Y	48
23403 S D614G	359
23664 S A701V	270
25563 ORF3a Q57H	99
25904 ORF3a S171L	243
26456 E P71L	107
26645 M silent	221
27294 ORF6 silent	339
28253 ORF8 silent	69
28254 ORF8 I121L	71
28887 N T205I	57
	7
	VSP2646-1
	SPS
	Š

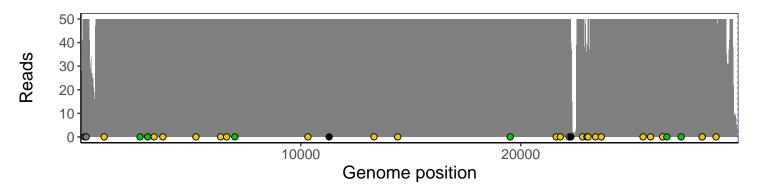
Analyses of individual experiments and composite results

$VSP2646\text{-}1 \mid 2021\text{-}04\text{-}23 \mid Saline \mid UPHS\text{-}1391 \mid genomes \mid single \ experiment$

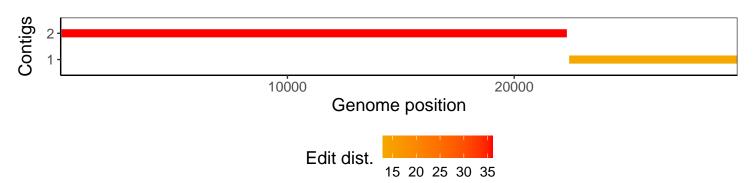
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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	3.1.3				
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	3.3.3				
tidyverse	1.2.1				
ShortRead	1.34.2				
GenomicAlignments	1.12.2				
${\bf Summarized Experiment}$	1.6.5				
DelayedArray	0.2.7				
matrixStats	0.54.0				
Biobase	2.36.2				
Rsamtools	1.28.0				
GenomicRanges	1.28.6				
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				