COVID-19 subject UPHS-1609

2021-06-23

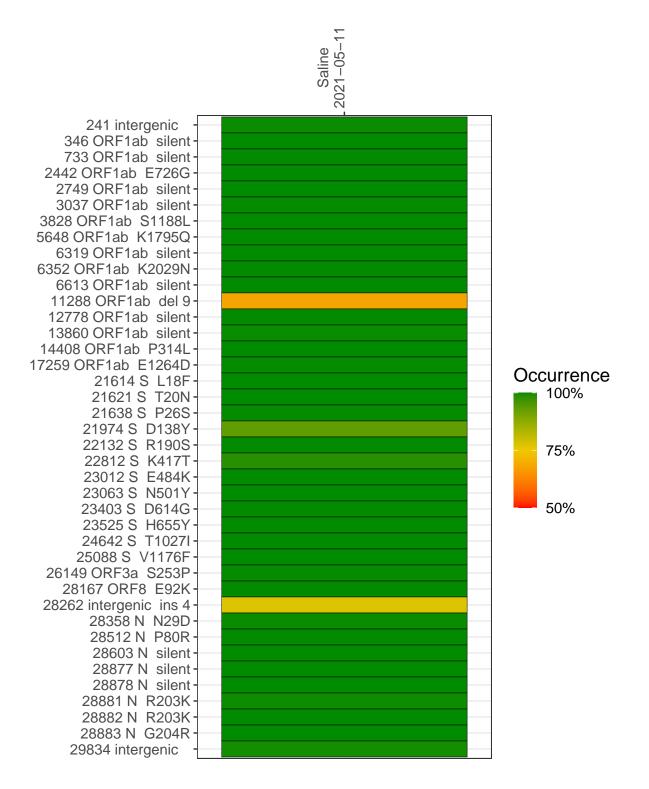
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP2910-1	single experiment	NA	Saline	2021-05-11	29.84	P.1	99.8%	99.7%

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.



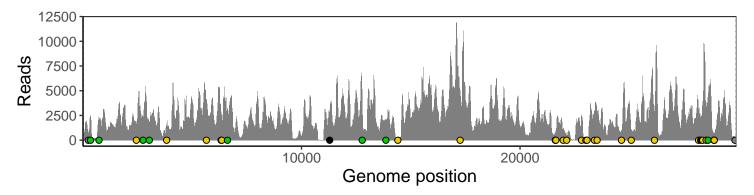
Saline 2021-05-11

	2021-05-11
241 intergenic	743
346 ORF1ab silent	2440
733 ORF1ab silent	2134
2442 ORF1ab E726G	1673
2749 ORF1ab silent	2276
3037 ORF1ab silent	1655
3828 ORF1ab S1188L	1581
5648 ORF1ab K1795Q	3735
6319 ORF1ab silent	2802
6352 ORF1ab K2029N	2225
6613 ORF1ab silent	3180
11288 ORF1ab del 9	1806
12778 ORF1ab silent	4337
13860 ORF1ab silent	1771
14408 ORF1ab P314L	169
17259 ORF1ab E1264D	4456
21614 S L18F	725
21621 S T20N	705
21638 S P26S	938
21974 S D138Y	947
22132 S R190S	504
22812 S K417T	2349
23012 S E484K	24
23063 S N501Y	57
23403 S D614G	2942
23525 S H655Y	2595
24642 S T1027I	1033
25088 S V1176F	1098
26149 ORF3a S253P	2823
28167 ORF8 E92K	3101
28262 intergenic ins 4	1735
28358 N N29D	3770
28512 N P80R	3639
28603 N silent	4326
28877 N silent	327
28878 N silent	321
28881 N R203K	321
28882 N R203K	321
28883 N G204R	329
29834 intergenic	169
	7
	VSP2910-1
	526
	/SF
	>

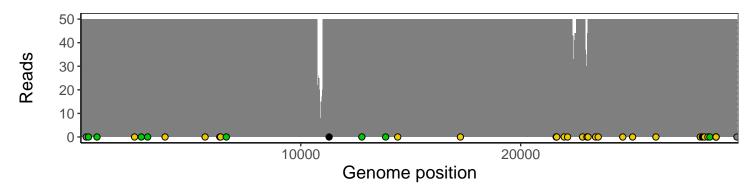
Analyses of individual experiments and composite results

VSP2910-1 | 2021-05-11 | Saline | UPHS-1609 | genomes | single experiment

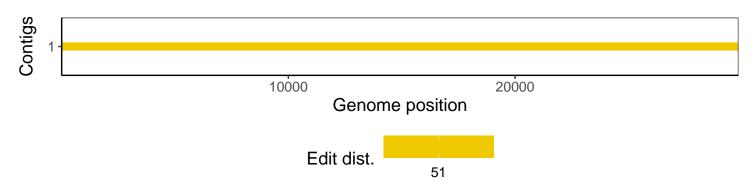
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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				