COVID-19 subject HUP Q-0213

2021-05-05

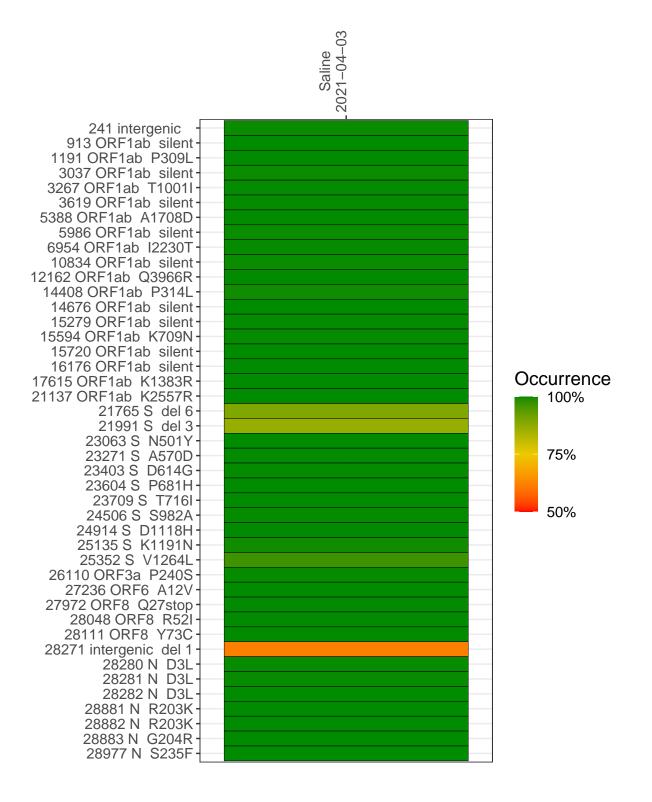
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP1775-1	single experiment	NA	Saline	2021-04-03	29.92	B.1.1.7	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 2021-04-03

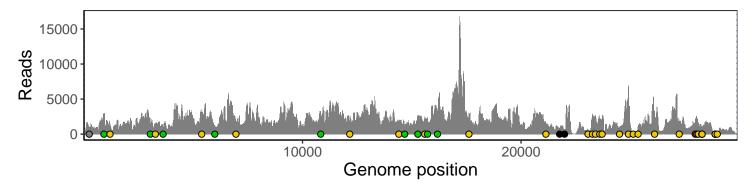
	2021-04-03
241 intergenic	906
913 ORF1ab silent	2825
1191 ORF1ab P309L	1369
3037 ORF1ab silent	1560
3267 ORF1ab T1001I	1627
3619 ORF1ab silent	998
5388 ORF1ab A1708D	3020
5986 ORF1ab silent	1269
6954 ORF1ab I2230T	1356
10834 ORF1ab silent	2405
12162 ORF1ab Q3966R	2178
14408 ORF1ab P314L	1713
14676 ORF1ab silent	980
15279 ORF1ab silent	2280
15594 ORF1ab K709N	2185
15720 ORF1ab silent	2232
16176 ORF1ab silent	3243
17615 ORF1ab K1383R	2802
21137 ORF1ab K2557R	1484
21765 S del 6	722
21991 S del 3	507
23063 S N501Y	656
23271 S A570D	2634
23403 S D614G	2645
23604 S P681H	2086
23709 S T716I	2147
24506 S S982A	997
24914 S D1118H	6953
25135 S K1191N	275
25352 S V1264L	1018
26110 ORF3a P240S	3257
27236 ORF6 A12V	2037
27972 ORF8 Q27stop	2141
28048 ORF8 R52I	2189
28111 ORF8 Y73C	1558
28271 intergenic del 1	954
28280 N D3L	574
28281 N D3L	574
28282 N D3L	632
28881 N R203K	204
28882 N R203K	204
28883 N G204R	204
28977 N S235F	273
20077 14 02001	
	VSP1775-1
	<u> </u>
	<u>6</u>
	>



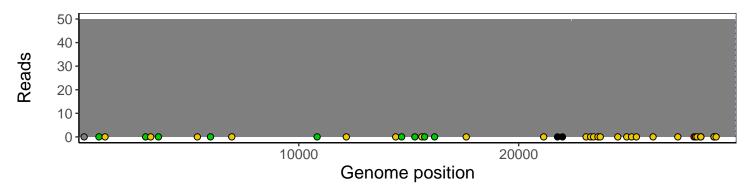
Analyses of individual experiments and composite results

VSP1775-1 | 2021-04-03 | Saline | HUP Q-0213 | genomes | single experiment

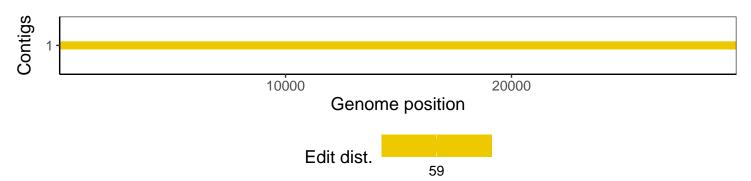
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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	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