COVID-19 subject HUP Q-0119

2021-05-05

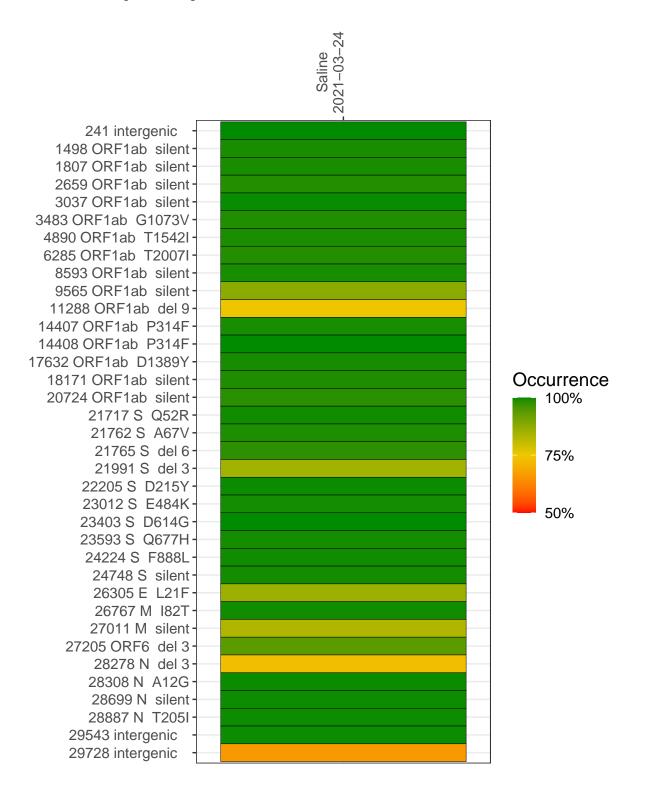
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP1460-1	single experiment	NA	Saline	2021-03-24	29.76	B.1.525	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.



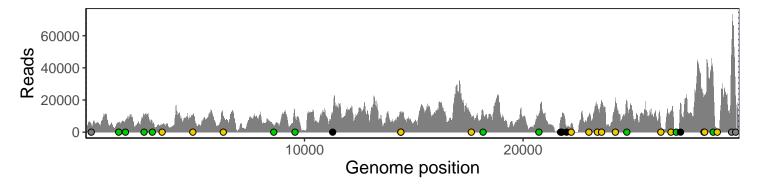
Saline 2021-03-24

	2021 00 21
241 intergenic	4313
1498 ORF1ab silent	5834
1807 ORF1ab silent	7218
2659 ORF1ab silent	11077
3037 ORF1ab silent	5382
3483 ORF1ab G1073V	3889
4890 ORF1ab T1542I	9276
6285 ORF1ab T2007I	8670
8593 ORF1ab silent	4102
9565 ORF1ab silent	10142
11288 ORF1ab del 9	7607
14407 ORF1ab P314F	9897
14408 ORF1ab P314F	10034
17632 ORF1ab D1389Y	10321
18171 ORF1ab silent	12292
20724 ORF1ab silent	11064
21717 S Q52R	7515
21762 S A67V	6056
21765 S del 6	5687
21991 S del 3	2953
22205 S D215Y	5107
23012 S E484K	8768
23403 S D614G	16288
23593 S Q677H	19319
24224 S F888L	17894
24748 S silent	18226
26305 E L21F	2940
26767 M 182T	12222
27011 M silent	637
27205 ORF6 del 3	16601
28278 N del 3	16149
28308 N A12G	22685
28699 N silent	32371
28887 N T205I	2341
29543 intergenic	52956
29728 intergenic	335
	7-
	460-1
	-

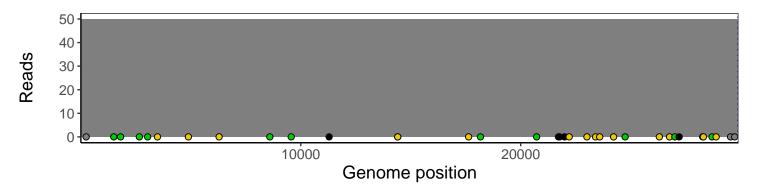
Analyses of individual experiments and composite results

VSP1460-1 | 2021-03-24 | Saline | HUP Q-0119 | genomes | single experiment

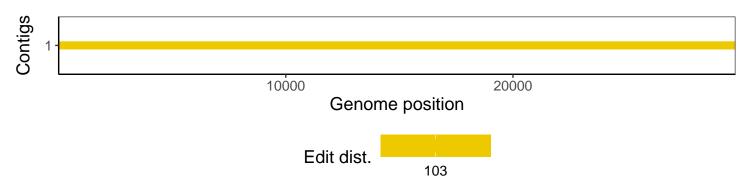
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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