# COVID-19 subject UK strain SARS-CoV-2

2021-06-23

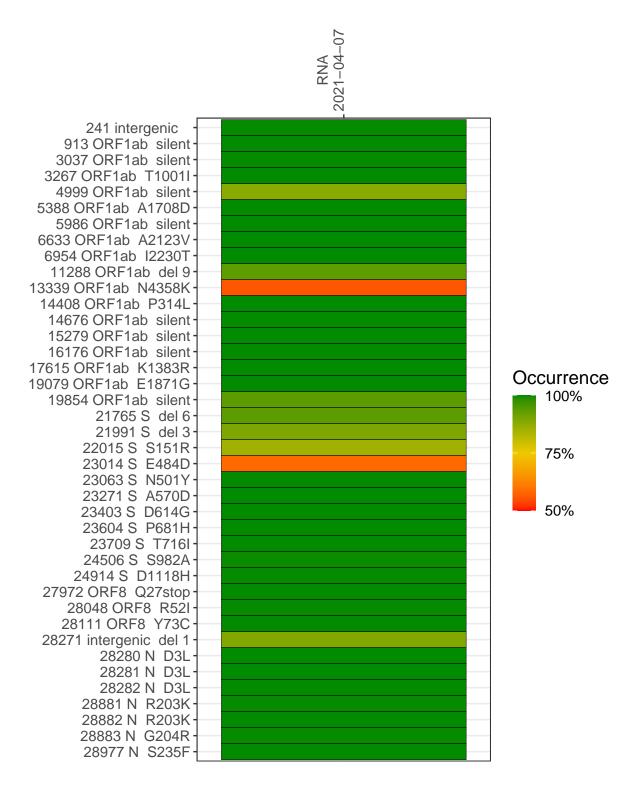
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP1975-1	single experiment	NA	RNA	2021-04-07	29.86	B.1.1.7	99.9%	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.



#### RNA 2021-04-07

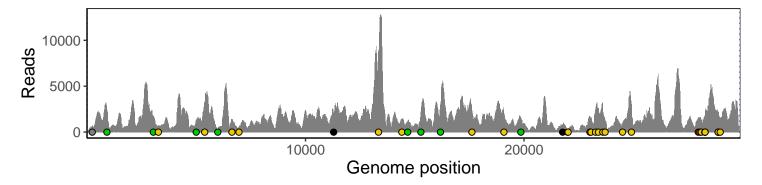
	2021-04-07
241 intergenic	647
913 ORF1ab silent	2465
3037 ORF1ab silent	1631
3267 ORF1ab T1001I	699
4999 ORF1ab silent	1360
5388 ORF1ab A1708D	3227
5986 ORF1ab silent	352
6633 ORF1ab A2123V	1367
6954 ORF1ab I2230T	529
11288 ORF1ab del 9	1960
13339 ORF1ab N4358K	8749
14408 ORF1ab P314L	1342
14676 ORF1ab silent	1170
15279 ORF1ab silent	2165
16176 ORF1ab silent	2974
17615 ORF1ab K1383R	2480
19079 ORF1ab E1871G	2064
19854 ORF1ab silent	1477
21765 S del 6	722
21991 S del 3	307
22015 S S151R	307
23014 S E484D	1744
23063 S N501Y	1679
23271 S A570D	2485
23403 S D614G	2591
23604 S P681H	3075
23709 S T716I	1310
24506 S S982A	944
24914 S D1118H	3221
27972 ORF8 Q27stop	1503
28048 ORF8 R52I	1563
28111 ORF8 Y73C	1401
28271 intergenic del 1	1750
28280 N D3L	1522
28281 N D3L	1522
28282 N D3L	1554
28881 N R203K	1919
28882 N R203K	1908
28883 N G204R	1919
28977 N S235F	1601
	7
	.576



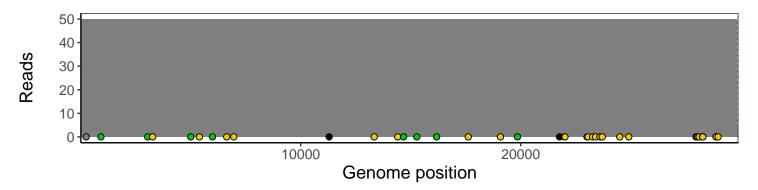
### Analyses of individual experiments and composite results

#### $VSP1975\text{-}1 \mid 2021\text{-}04\text{-}07 \mid RNA \mid UK \text{ strain SARS-CoV-2} \mid genomes \mid single \text{ experiment}$

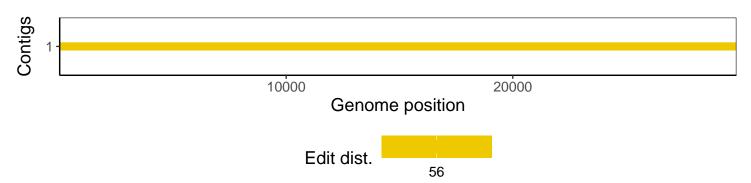
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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				