COVID-19 subject SARS_CoV_248

2021-06-29

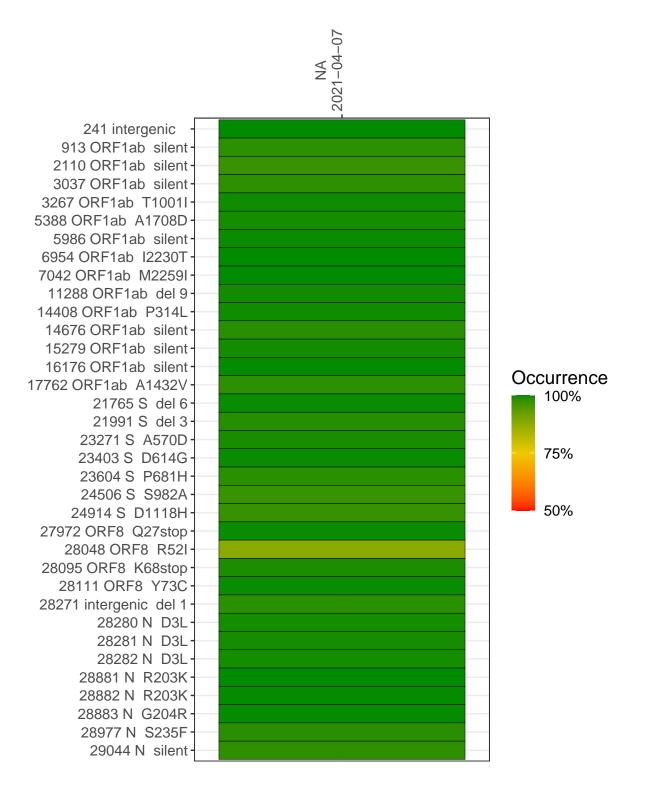
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP3069-1	single experiment	NA	NA	2021-04-07	29.82	B.1.1.7	99.8%	99.6%

Variants shared across samples

The heat map below shows how variants (reference genome /home/common/SARS-CoV-2-Philadelphia/NC_0455) 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.



NA 2021-04-07

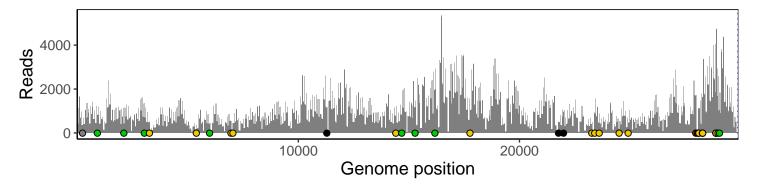
	2021-04-07
241 intergenic	342
913 ORF1ab silent	829
2110 ORF1ab silent	419
3037 ORF1ab silent	711
3267 ORF1ab T1001I	473
5388 ORF1ab A1708D	161
5986 ORF1ab silent	409
6954 ORF1ab I2230T	550
7042 ORF1ab M2259I	433
11288 ORF1ab del 9	1154
14408 ORF1ab P314L	774
14676 ORF1ab silent	839
15279 ORF1ab silent	1154
16176 ORF1ab silent	1927
17762 ORF1ab A1432V	1032
21765 S del 6	478
21991 S del 3	949
23271 S A570D	397
23403 S D614G	876
23604 S P681H	460
24506 S S982A	180
24914 S D1118H	470
27972 ORF8 Q27stop	1690
28048 ORF8 R52I	92
28095 ORF8 K68stop	1113
28111 ORF8 Y73C	1103
28271 intergenic del 1	1634
28280 N D3L	1617
28281 N D3L	1617
28282 N D3L	1617
28881 N R203K	1755
28882 N R203K	1755
28883 N G204R	1755
28977 N S235F	1672
29044 N silent	1026
	VSP3069-1
	P3C
	•



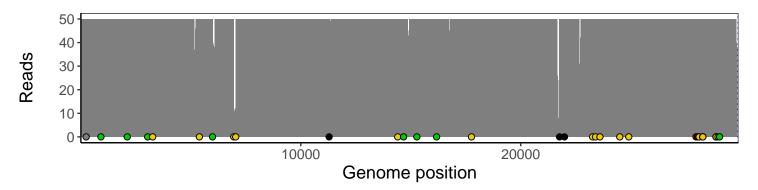
Analyses of individual experiments and composite results

$VSP3069\text{-}1 \mid 2021\text{-}04\text{-}07 \mid NA \mid SARS_CoV_248 \mid genomes \mid single \; experiment$

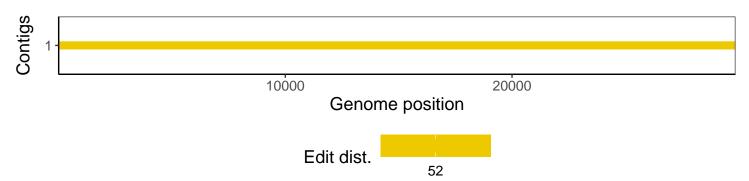
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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