COVID-19 subject MPCluster2-Seq13

2021-03-04

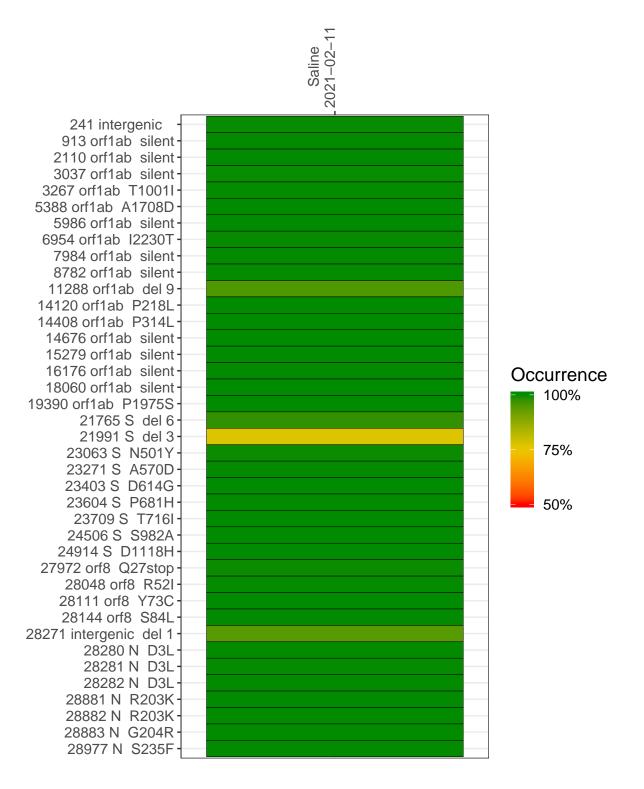
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP0805-1	single experiment	NA	Saline	2021-02-11	28.36	B.1.1.7	99.8%	99.8%

Variants shared across samples

The heat map below shows how variants (reference genome USA-WA1-2020) 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 2021–02–11

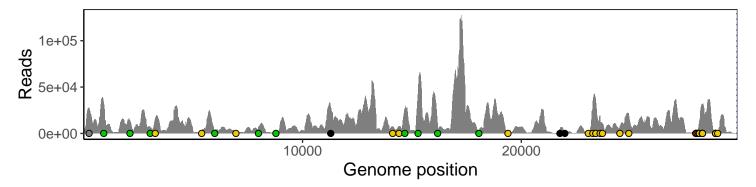
	2021-02-11
241 intergenic	26476
913 orf1ab silent	32606
2110 orf1ab silent	13938
3037 orf1ab silent	6107
3267 orf1ab T1001I	16714
5388 orf1ab A1708D	2962
5986 orf1ab silent	3981
6954 orf1ab I2230T	2647
7984 orf1ab silent	6823
8782 orf1ab silent	3020
11288 orf1ab del 9	26469
14120 orf1ab P218L	7443
14408 orf1ab P314L	7778
14676 orf1ab silent	26444
15279 orf1ab silent	44977
16176 orf1ab silent	8671
18060 orf1ab silent	2679
19390 orf1ab P1975S	1738
21765 S del 6	4578
21991 S del 3	1299
23063 S N501Y	1901
23271 S A570D	29838
23403 S D614G	36485
23604 S P681H	19208
23709 S T716I	10385
24506 S S982A	7423
24914 S D1118H	20886
27972 orf8 Q27stop	3975
28048 orf8 R52I	3353
28111 orf8 Y73C	9111
28144 orf8 S84L	13597
28271 intergenic del 1	31325
28280 N D3L	29515
28281 N D3L	29515
28282 N D3L	29566
28881 N R203K	3255
28882 N R203K	3254
28883 N G204R	3254
28977 N S235F	1218
	7



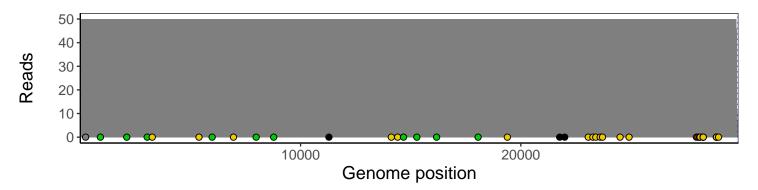
Analyses of individual experiments and composite results

$VSP0805\text{-}1 \mid 2021\text{-}02\text{-}11 \mid Saline \mid MPCluster2\text{-}Seq13 \mid genomes \mid single \ experiment$

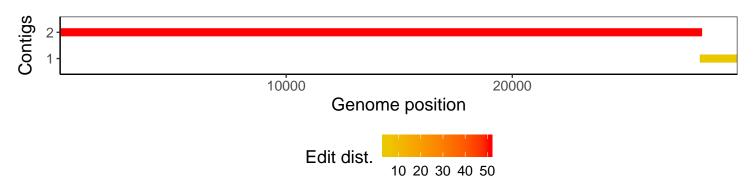
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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	2.3.3
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	3.0.0
tidyverse	1.2.1
ShortRead	1.34.2
$\operatorname{GenomicAlignments}$	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