

COVID-19 subject 505

2021-05-21

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.

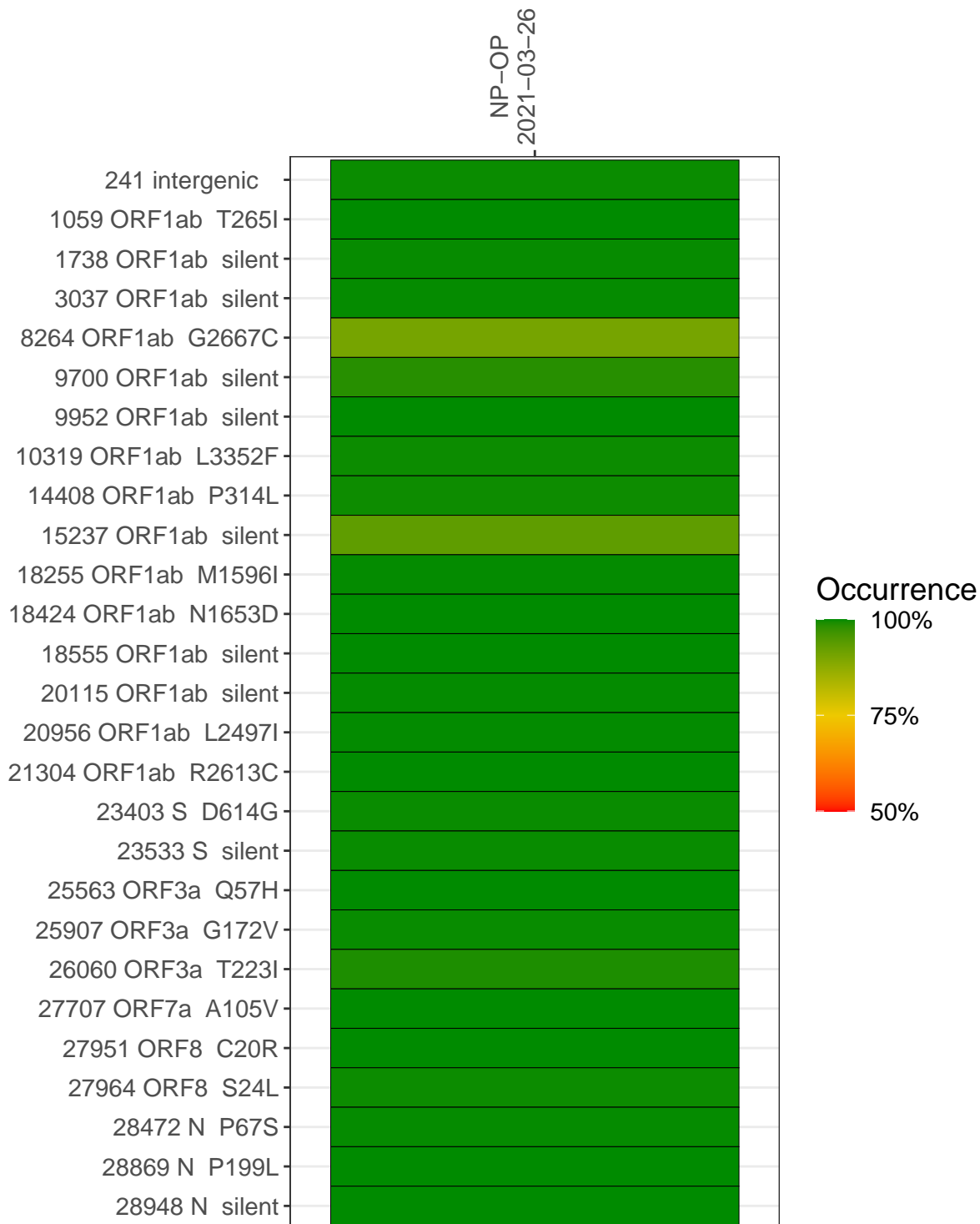
Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software tool (Rambaut et al 2020) for genomes with $> 90\%$ sequence coverage.

Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (≥ 5 reads)
VSP1435-1	single experiment	NA	NP-OP	2021-03-26	29.83	B.1.2	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.



NP-OP
2021-03-26

241 intergenic	430
1059 ORF1ab T265I	900
1738 ORF1ab silent	694
3037 ORF1ab silent	838
8264 ORF1ab G2667C	597
9700 ORF1ab silent	472
9952 ORF1ab silent	710
10319 ORF1ab L3352F	1648
14408 ORF1ab P314L	915
15237 ORF1ab silent	1019
18255 ORF1ab M1596I	1083
18424 ORF1ab N1653D	946
18555 ORF1ab silent	1038
20115 ORF1ab silent	941
20956 ORF1ab L2497I	1496
21304 ORF1ab R2613C	623
23403 S D614G	1429
23533 S silent	1001
25563 ORF3a Q57H	744
25907 ORF3a G172V	531
26060 ORF3a T223I	2365
27707 ORF7a A105V	379
27951 ORF8 C20R	724
27964 ORF8 S24L	1039
28472 N P67S	863
28869 N P199L	169
28948 N silent	209

Base change

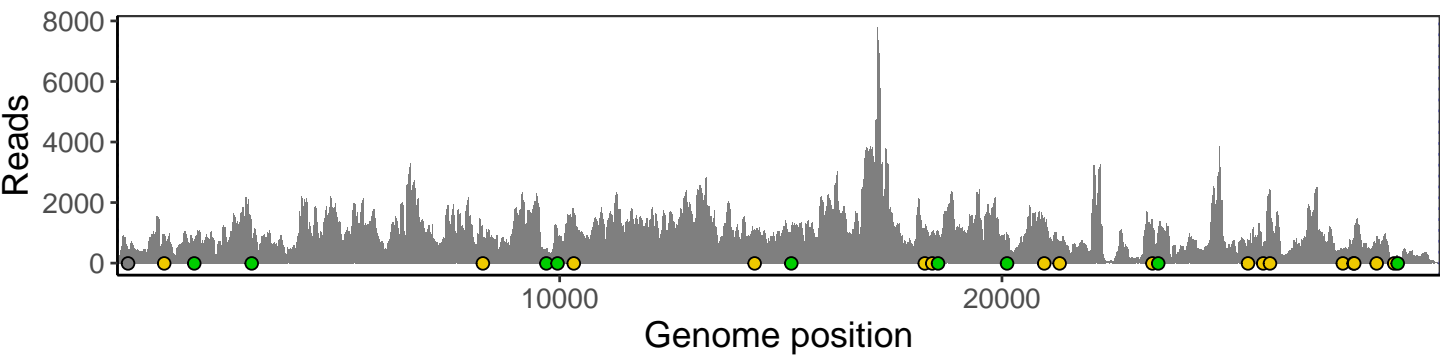


VSP1435-1

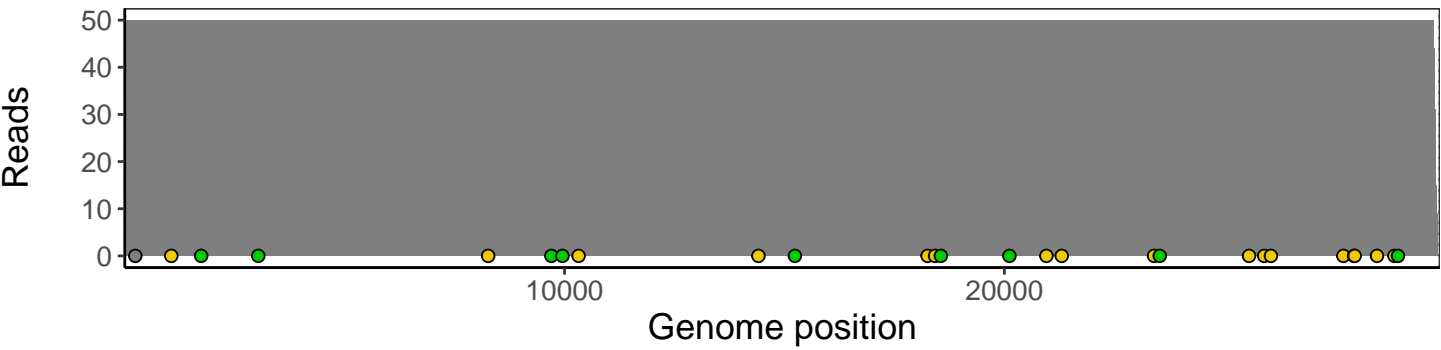
Analyses of individual experiments and composite results

VSP1435-1 | 2021-03-26 | NP-OP | 505no | genomes | single experiment

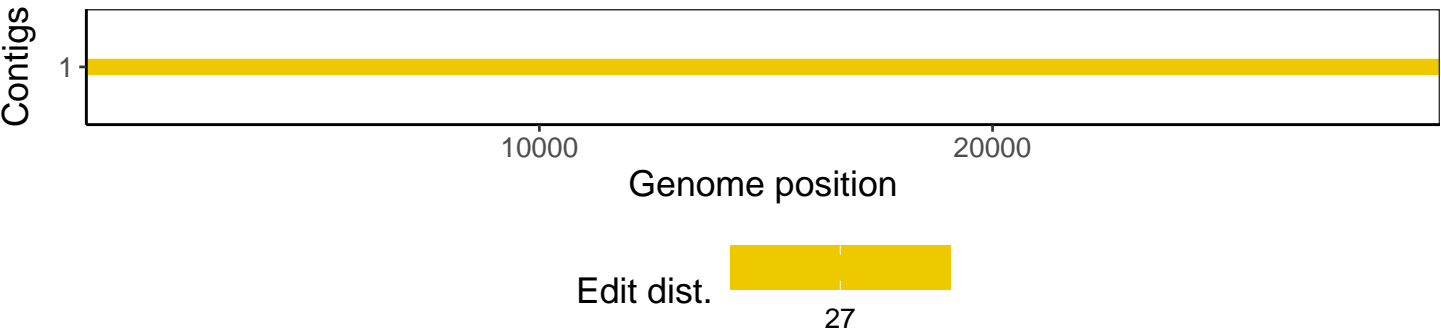
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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 htlib 1.10
bcftools	1.10.2-34-g1a12af0-dirty Using htlib 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
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
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