COVID-19 subject UPHS-0469

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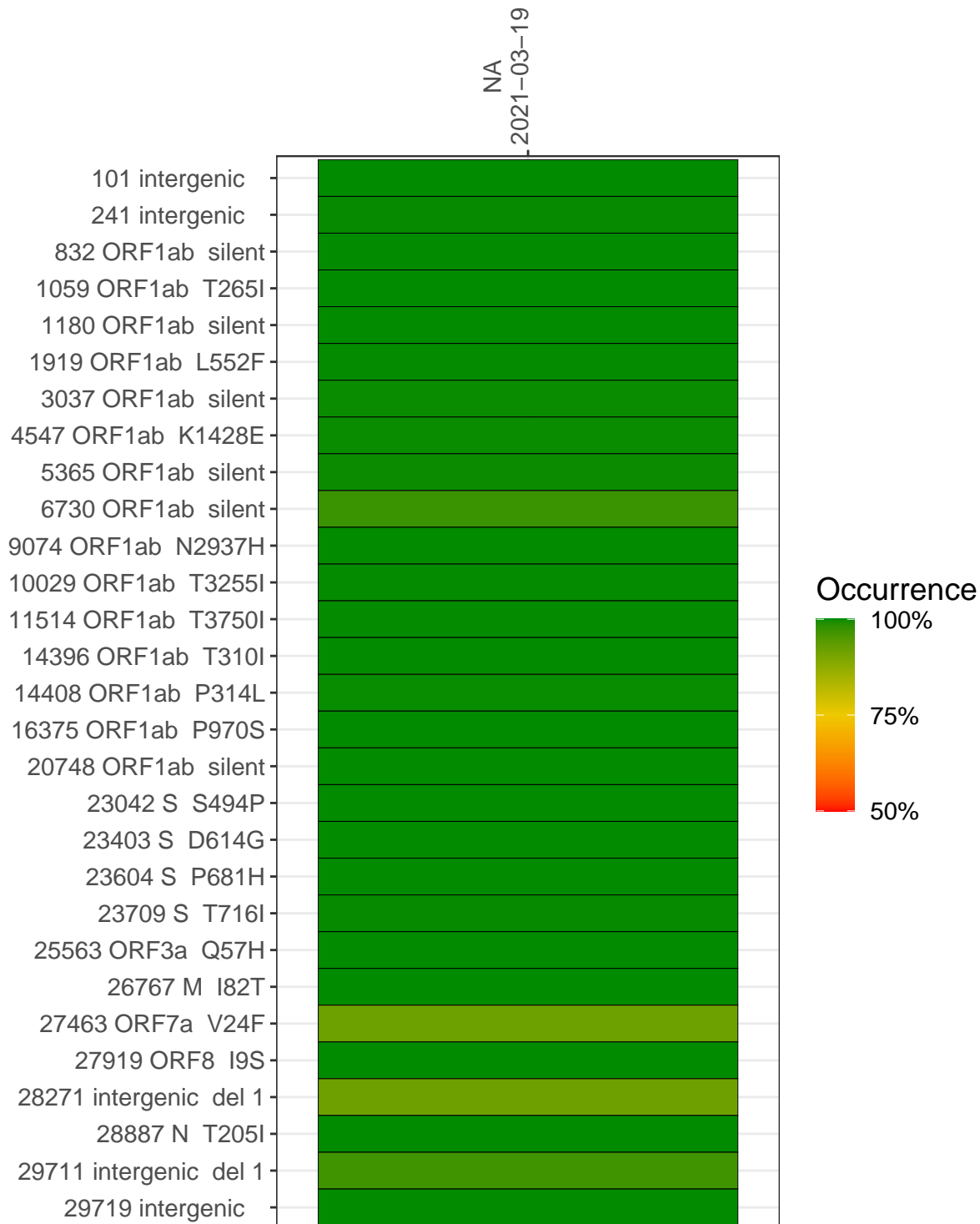 sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software tool (Rambaut et al 2020) for genomes with > 90% sequence coverage.
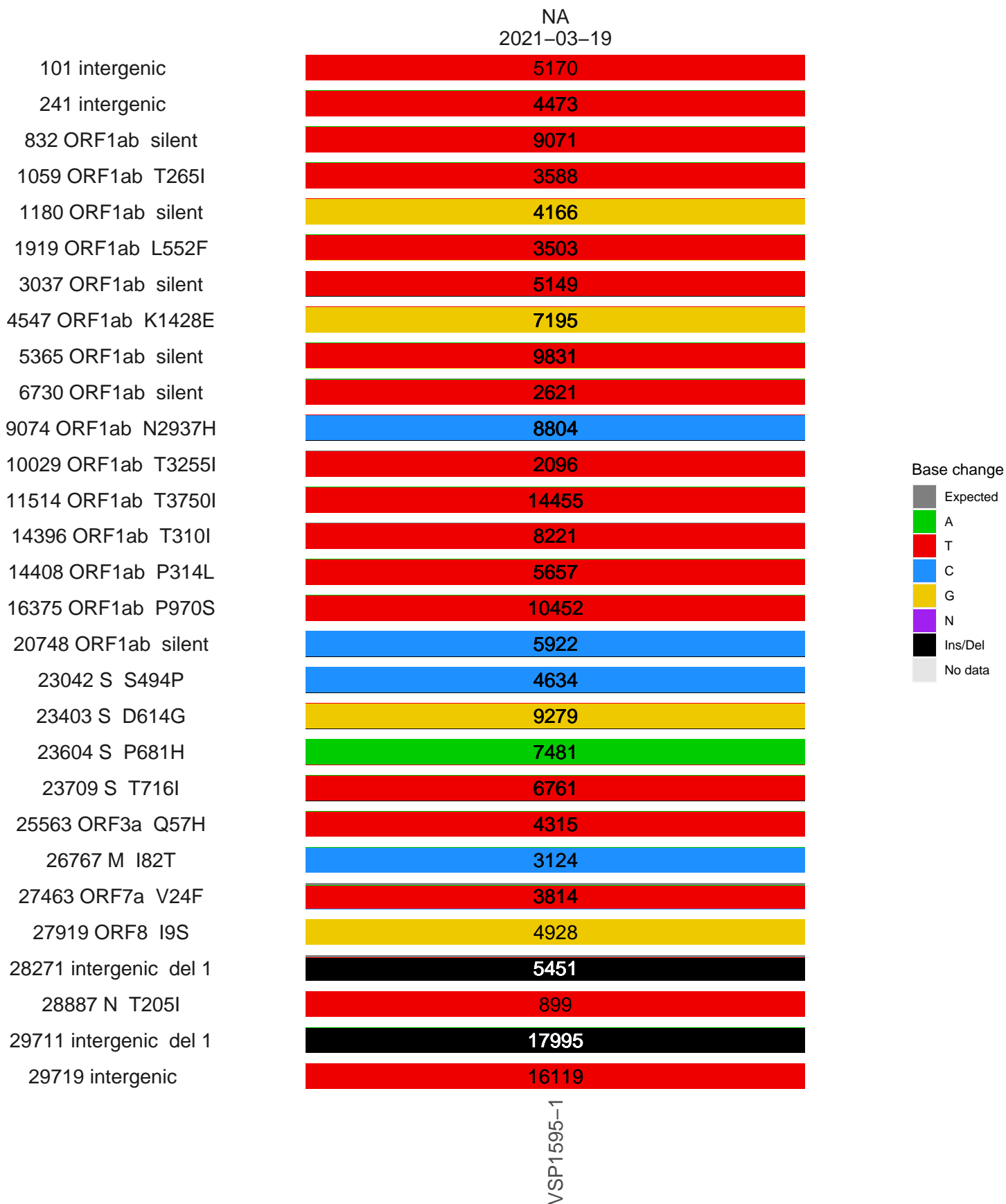
Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP1595-1	single experiment	NA	NA	2021-03-19	29.91	B.1.575	99.8%	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.

NA
2021−03−19

Position	VSP1595−1
101 intergenic	5170
241 intergenic	4473
832 ORF1ab silent	9071
1059 ORF1ab T265I	3588
1180 ORF1ab silent	4166
1919 ORF1ab L552F	3503
3037 ORF1ab silent	5149
4547 ORF1ab K1428E	7195
5365 ORF1ab silent	9831
6730 ORF1ab silent	2621
9074 ORF1ab N2937H	8804
10029 ORF1ab T3255I	2096
11514 ORF1ab T3750I	14455
14396 ORF1ab T310I	8221
14408 ORF1ab P314L	5657
16375 ORF1ab P970S	10452
20748 ORF1ab silent	5922
23042 S S494P	4634
23403 S D614G	9279
23604 S P681H	7481
23709 S T716I	6761
25563 ORF3a Q57H	4315
26767 M I82T	3124
27463 ORF7a V24F	3814
27919 ORF8 I9S	4928
28271 intergenic del 1	5451
28887 N T205I	899
29711 intergenic del 1	17995
29719 intergenic	16119

Base change
- Expected
- A
- T
- C
- G
- N
- Ins/Del
- No data

3

Analyses of individual experiments and composite results

**VSP1595-1 | 2021-03-19 | NA | UPHS-0469 | 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.
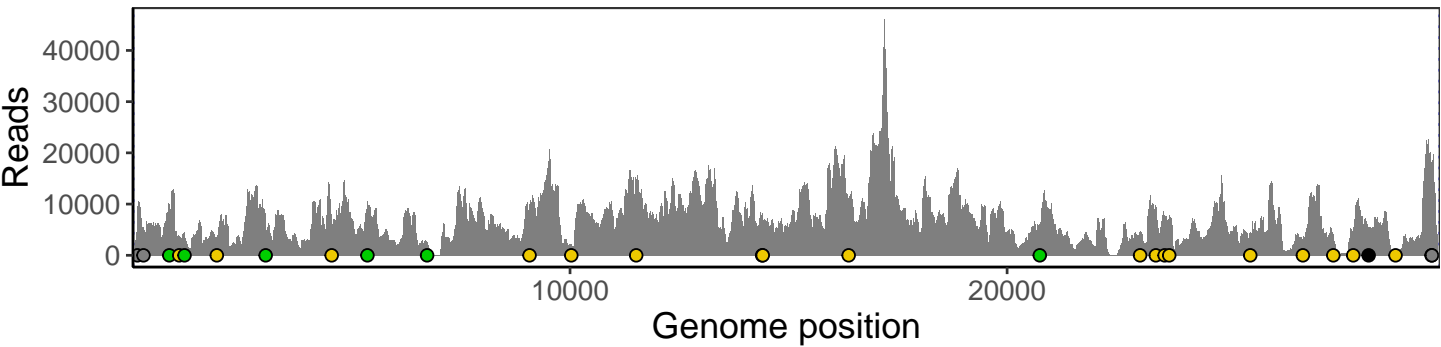
Excerpt from plot above focusing on reads coverage from 0 to 50 NT.
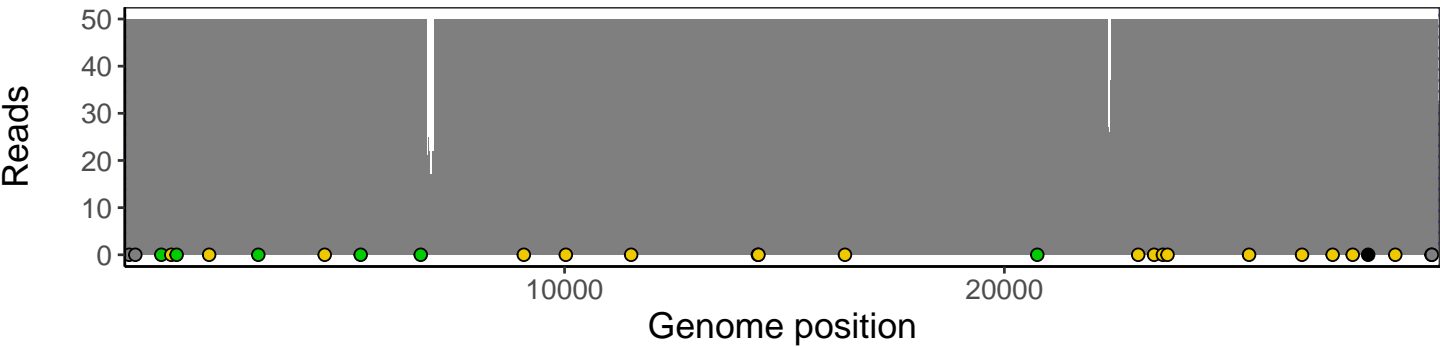
The longest five assembled contigs are shown below colored by their edit distance to the reference genome.

4
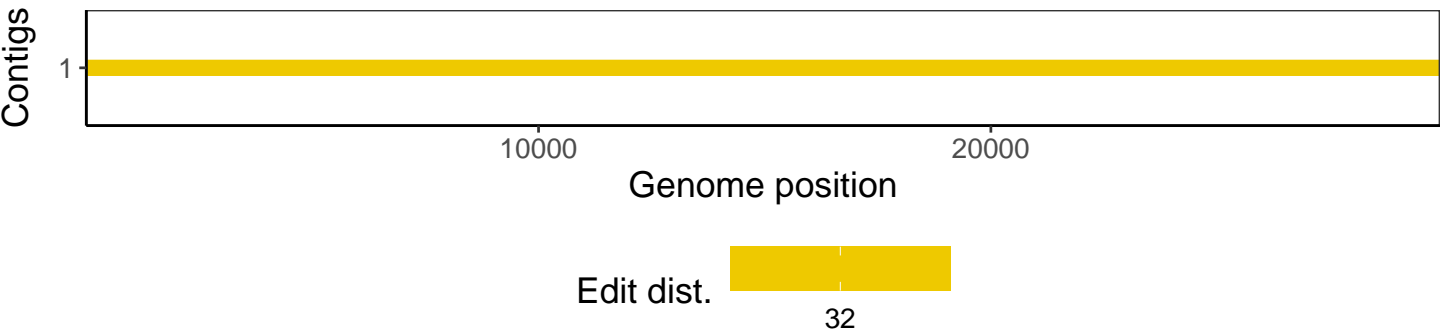
Software environment

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
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