COVID-19 subject UPHS-0416

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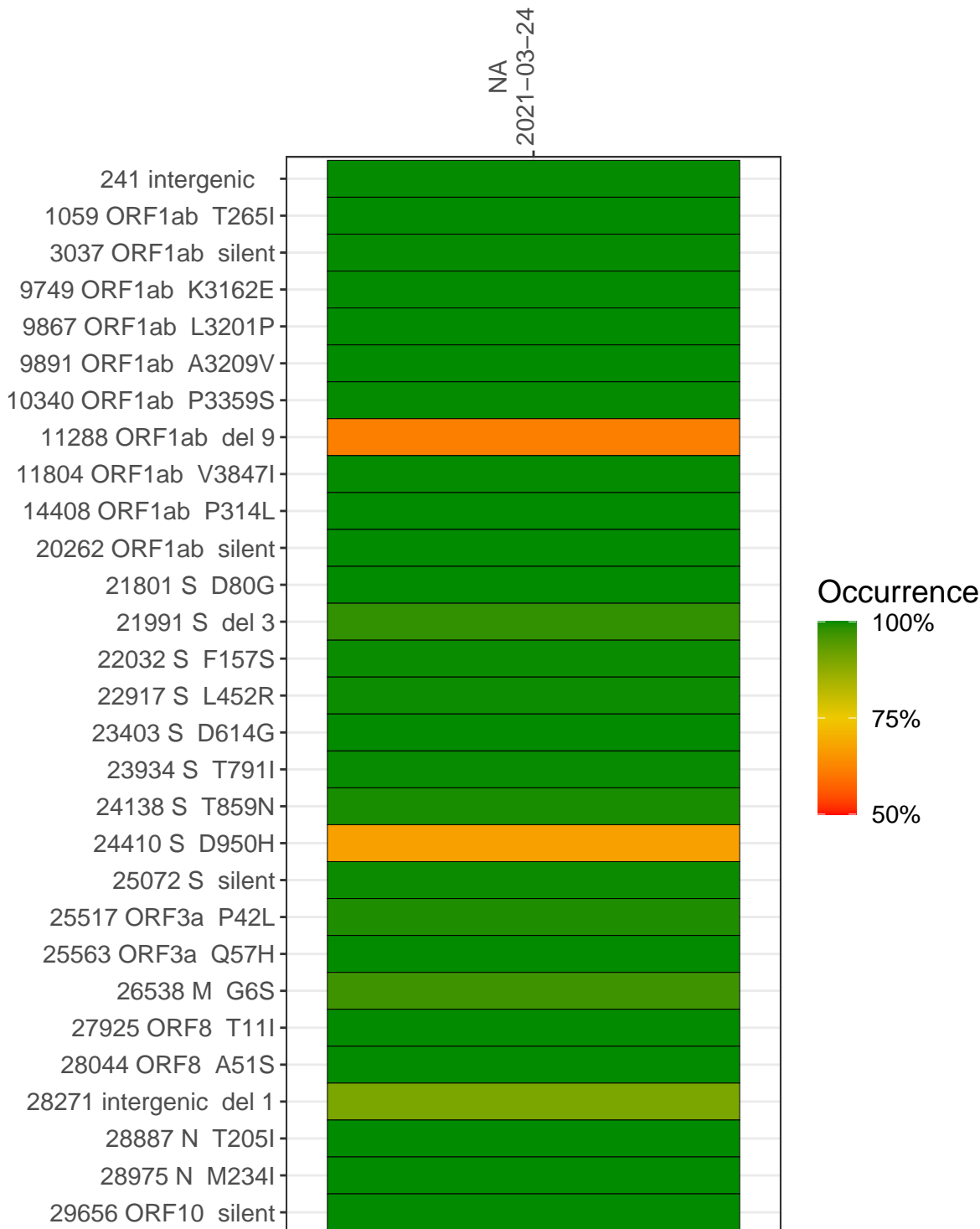 sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software tool (Rambaut et al 2020) for genomes with > 90% sequence coverage.
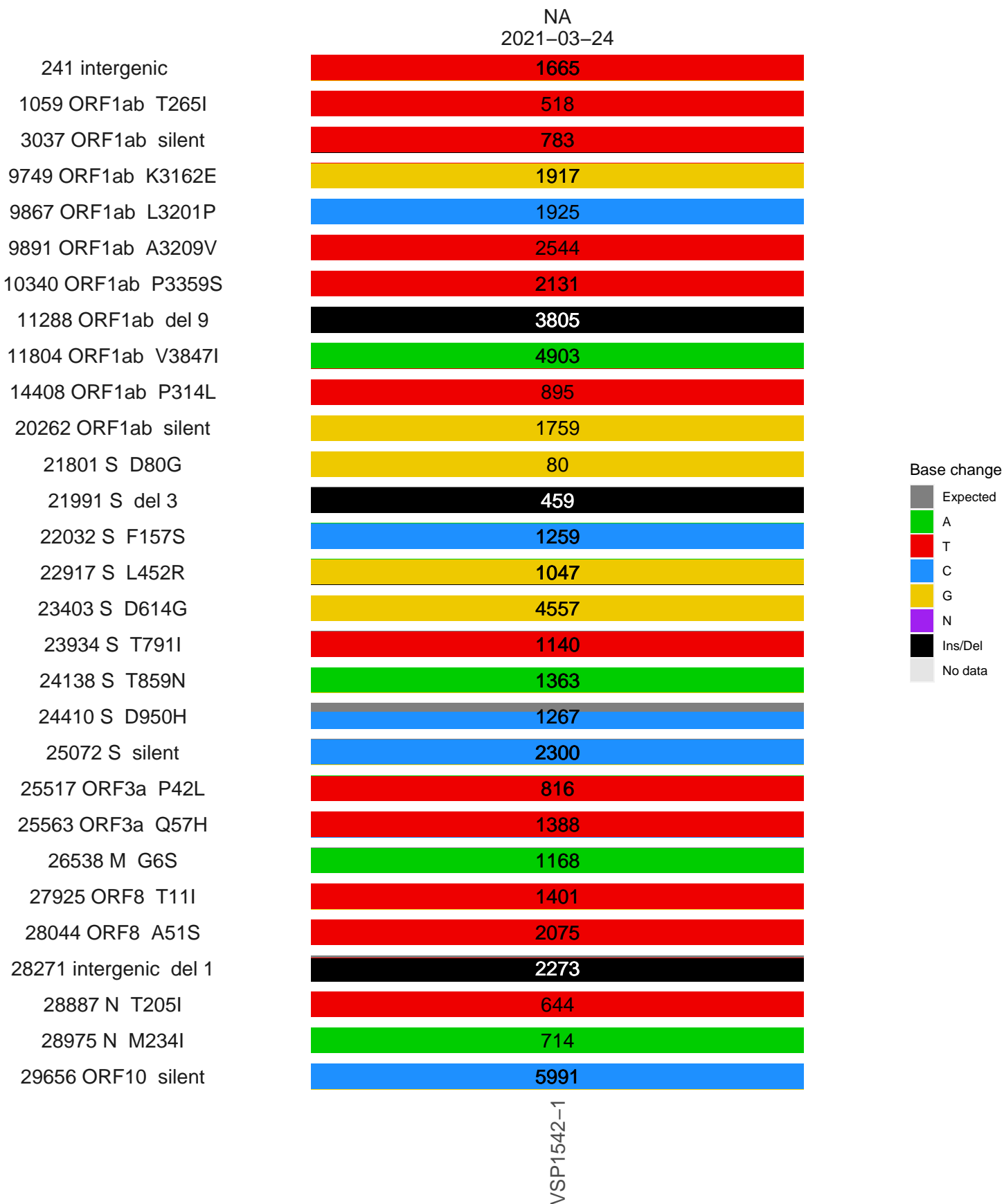
Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP1542-1	single experiment	NA	NA	2021-03-24	22.29	B.1.526	99.8%	99.1%

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
2021–03–24

Position	Value
241 intergenic	1665
1059 ORF1ab T265I	518
3037 ORF1ab silent	783
9749 ORF1ab K3162E	1917
9867 ORF1ab L3201P	1925
9891 ORF1ab A3209V	2544
10340 ORF1ab P3359S	2131
11288 ORF1ab del 9	3805
11804 ORF1ab V3847I	4903
14408 ORF1ab P314L	895
20262 ORF1ab silent	1759
21801 S D80G	80
21991 S del 3	459
22032 S F157S	1259
22917 S L452R	1047
23403 S D614G	4557
23934 S T791I	1140
24138 S T859N	1363
24410 S D950H	1267
25072 S silent	2300
25517 ORF3a P42L	816
25563 ORF3a Q57H	1388
26538 M G6S	1168
27925 ORF8 T11I	1401
28044 ORF8 A51S	2075
28271 intergenic del 1	2273
28887 N T205I	644
28975 N M234I	714
29656 ORF10 silent	5991

VSP1542–1

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

3

Analyses of individual experiments and composite results

**VSP1542-1 | 2021-03-24 | NA | UPHS-0416 | 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.
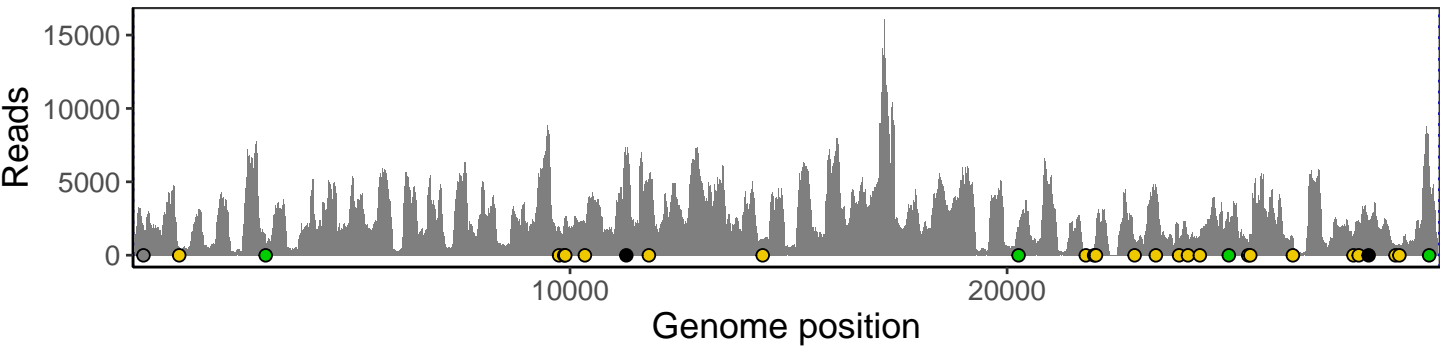
Excerpt from plot above focusing on reads coverage from 0 to 50 NT.
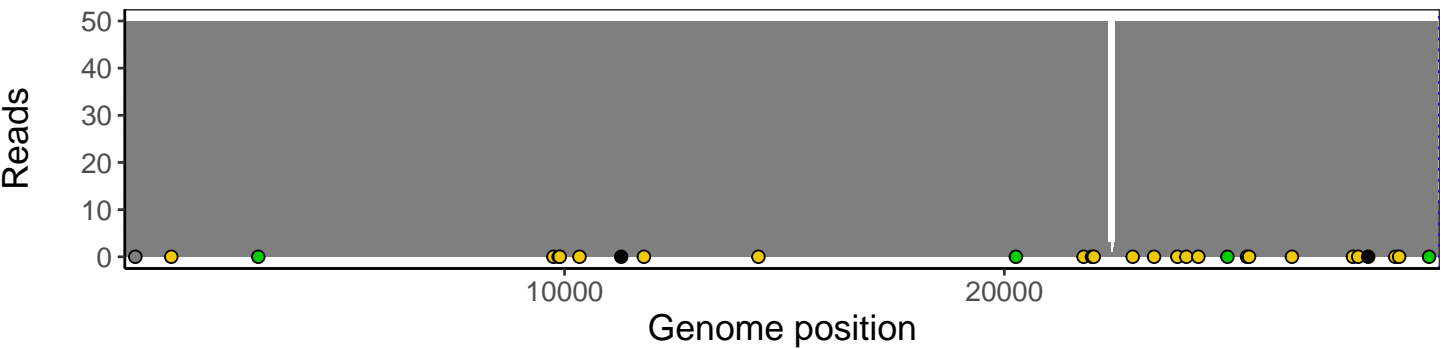
The longest five assembled contigs are shown below colored by their edit distance to the reference genome.
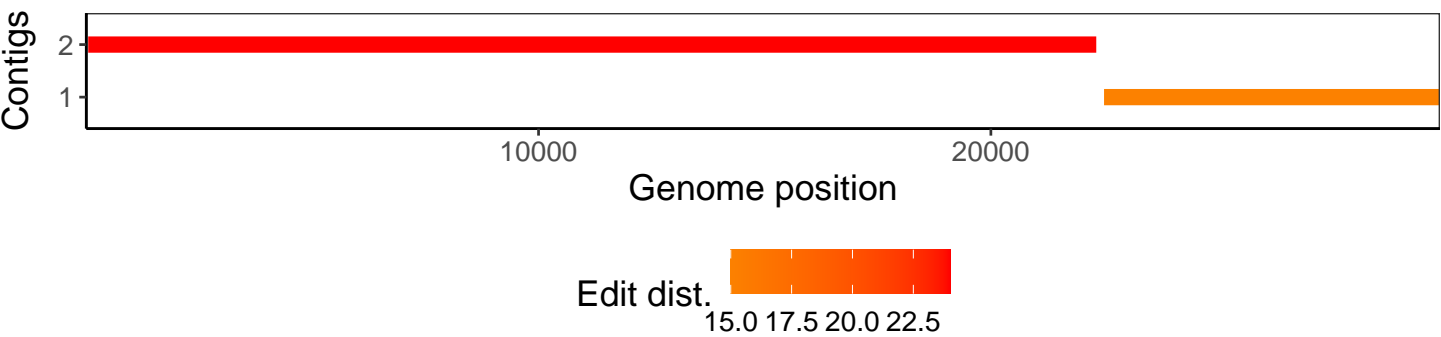
Software environment

Software/R package	Version
R	3.4.0
bwa	0.7.17-r1198-dirty
samtools	1.10 Using 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