COVID-19 subject UPHS-0574

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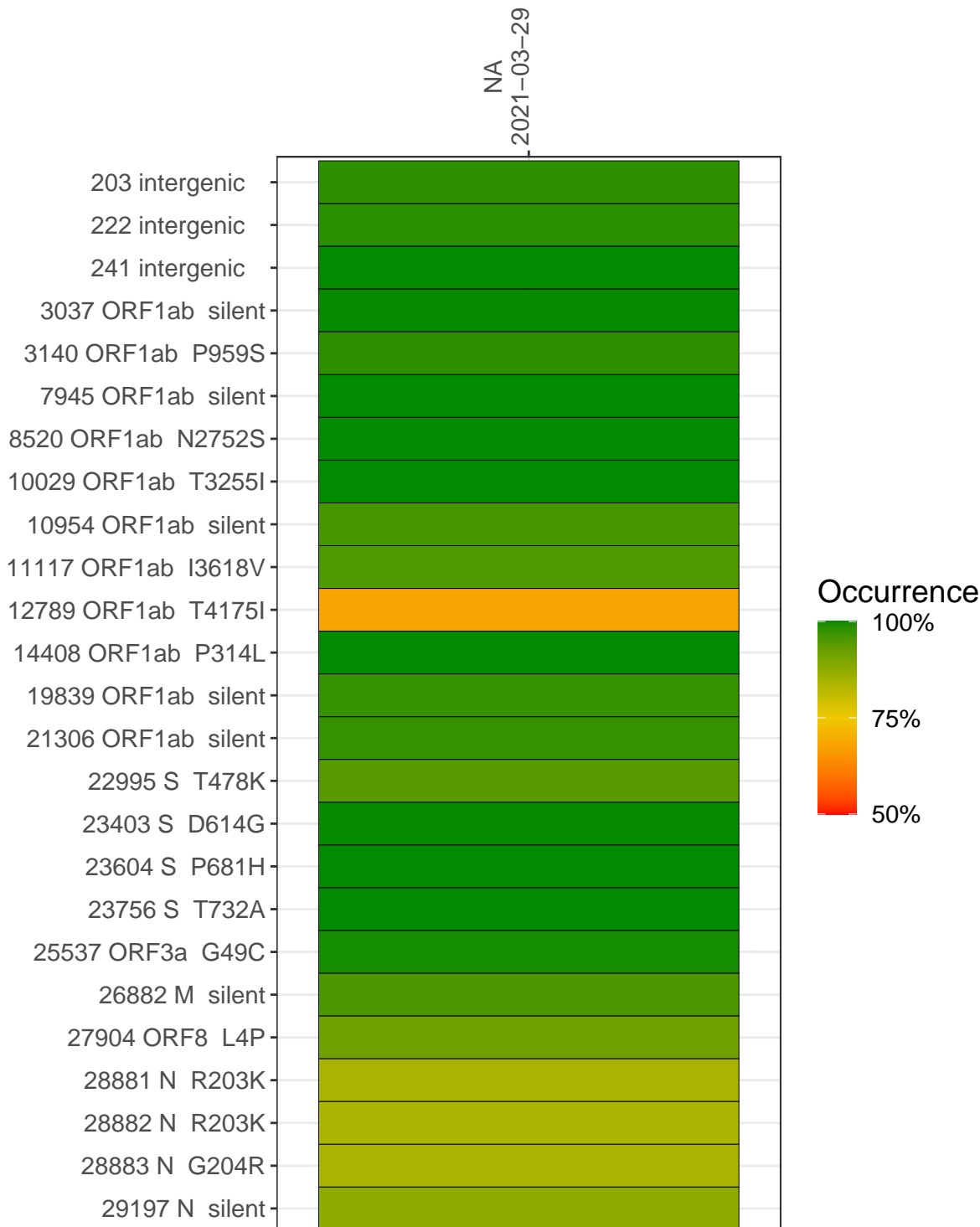 sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software tool (Rambaut et al 2020) for genomes with > 90% sequence coverage.
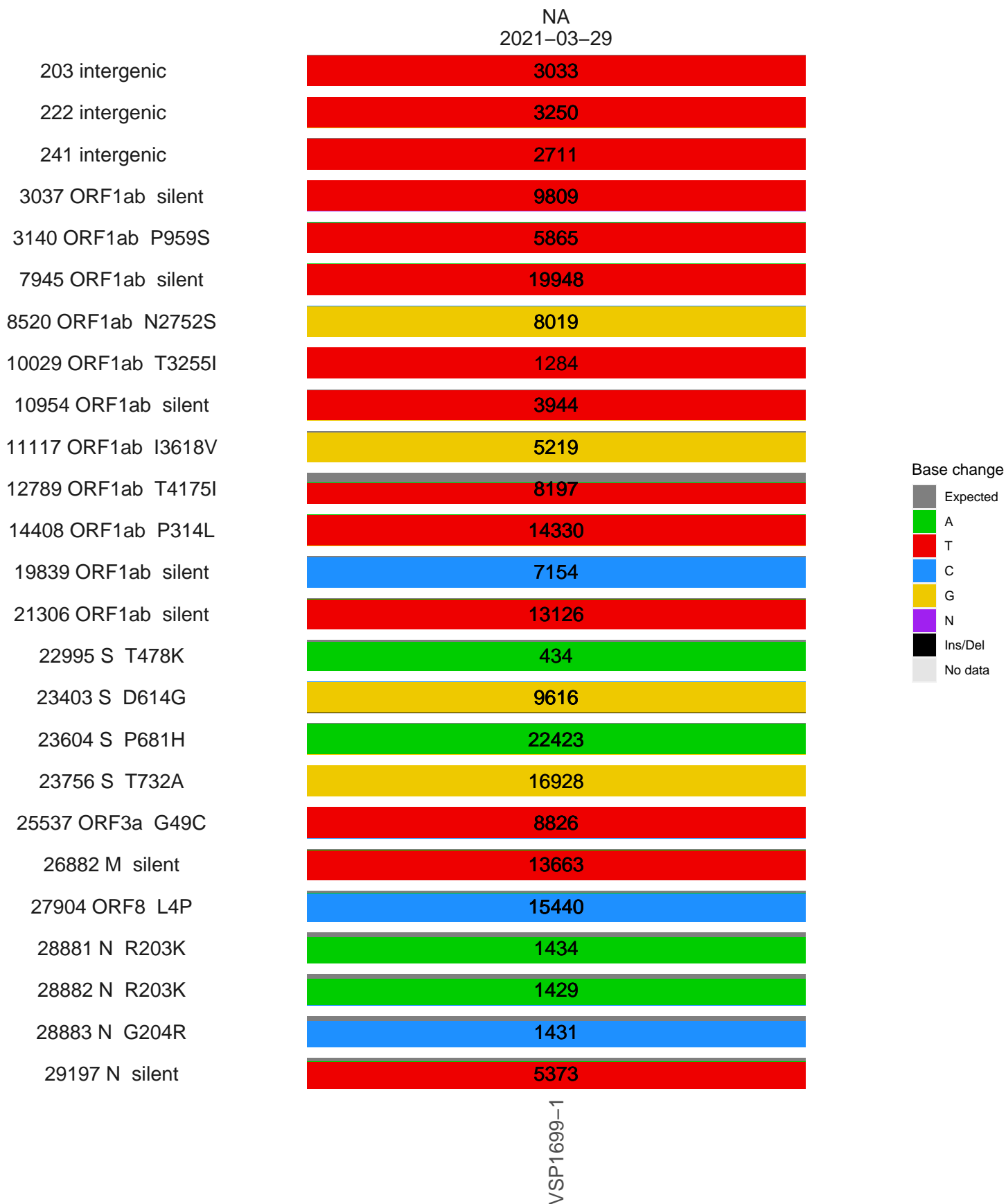
Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP1699-1	single experiment	NA	NA	2021-03-29	29.84	B.1.1.519	99.9%	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
2021-03-29

Position	Base change	VSP1699-1
203 intergenic	T	3033
222 intergenic	T	3250
241 intergenic	T	2711
3037 ORF1ab silent	T	9809
3140 ORF1ab P959S	T	5865
7945 ORF1ab silent	T	19948
8520 ORF1ab N2752S	G	8019
10029 ORF1ab T3255I	T	1284
10954 ORF1ab silent	T	3944
11117 ORF1ab I3618V	G	5219
12789 ORF1ab T4175I	T	8197
14408 ORF1ab P314L	T	14330
19839 ORF1ab silent	C	7154
21306 ORF1ab silent	T	13126
22995 S T478K	A	434
23403 S D614G	G	9616
23604 S P681H	A	22423
23756 S T732A	G	16928
25537 ORF3a G49C	T	8826
26882 M silent	T	13663
27904 ORF8 L4P	C	15440
28881 N R203K	A	1434
28882 N R203K	A	1429
28883 N G204R	C	1431
29197 N silent	T	5373

Base change
- Expected (grey)
- A (green)
- T (red)
- C (blue)
- G (yellow)
- N (purple)
- Ins/Del (black)
- No data (light grey)

3

Analyses of individual experiments and composite results

## VSP1699-1 | 2021-03-29 | NA | UPHS-0574 | 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.
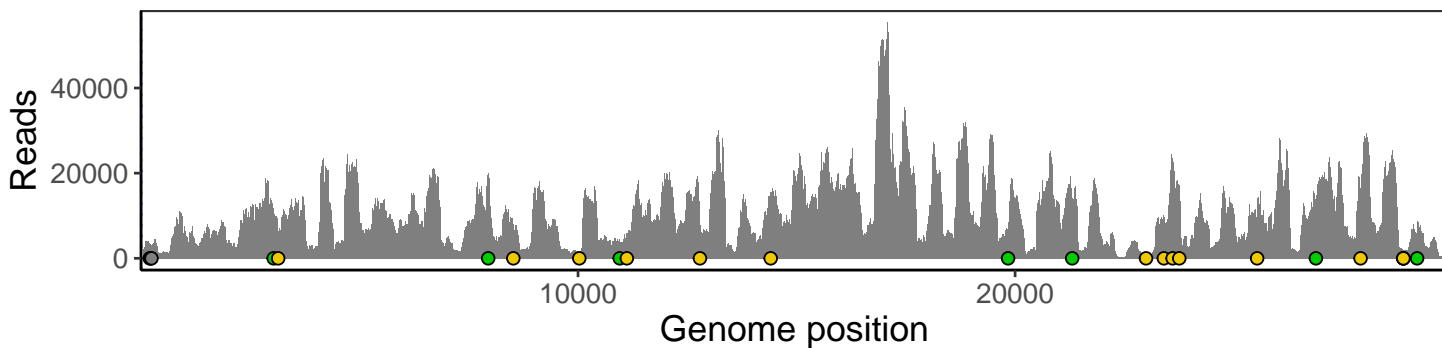
Excerpt from plot above focusing on reads coverage from 0 to 50 NT.
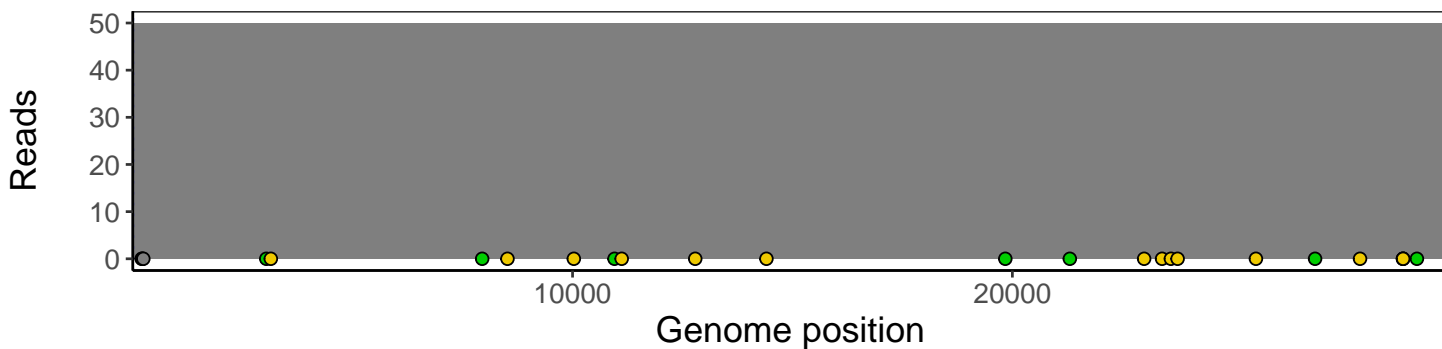
The longest five assembled contigs are shown below colored by their edit distance to the reference genome.
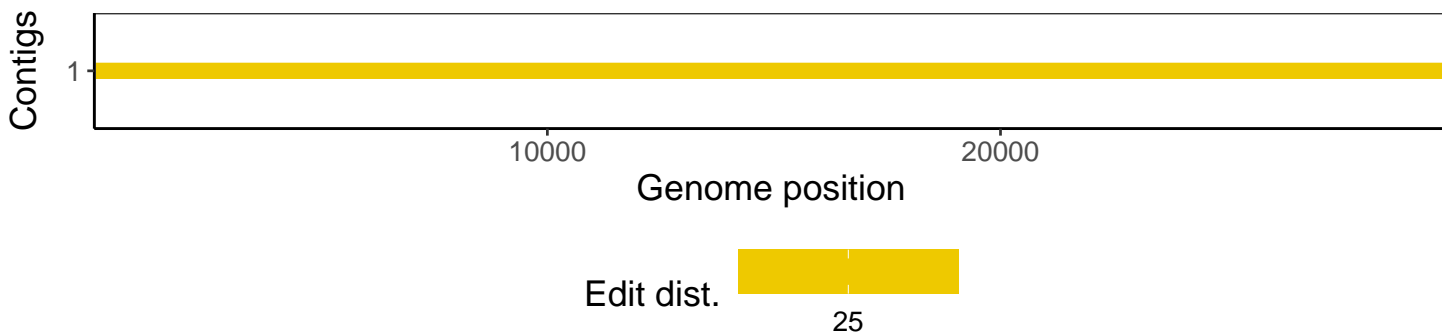
Software environment

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