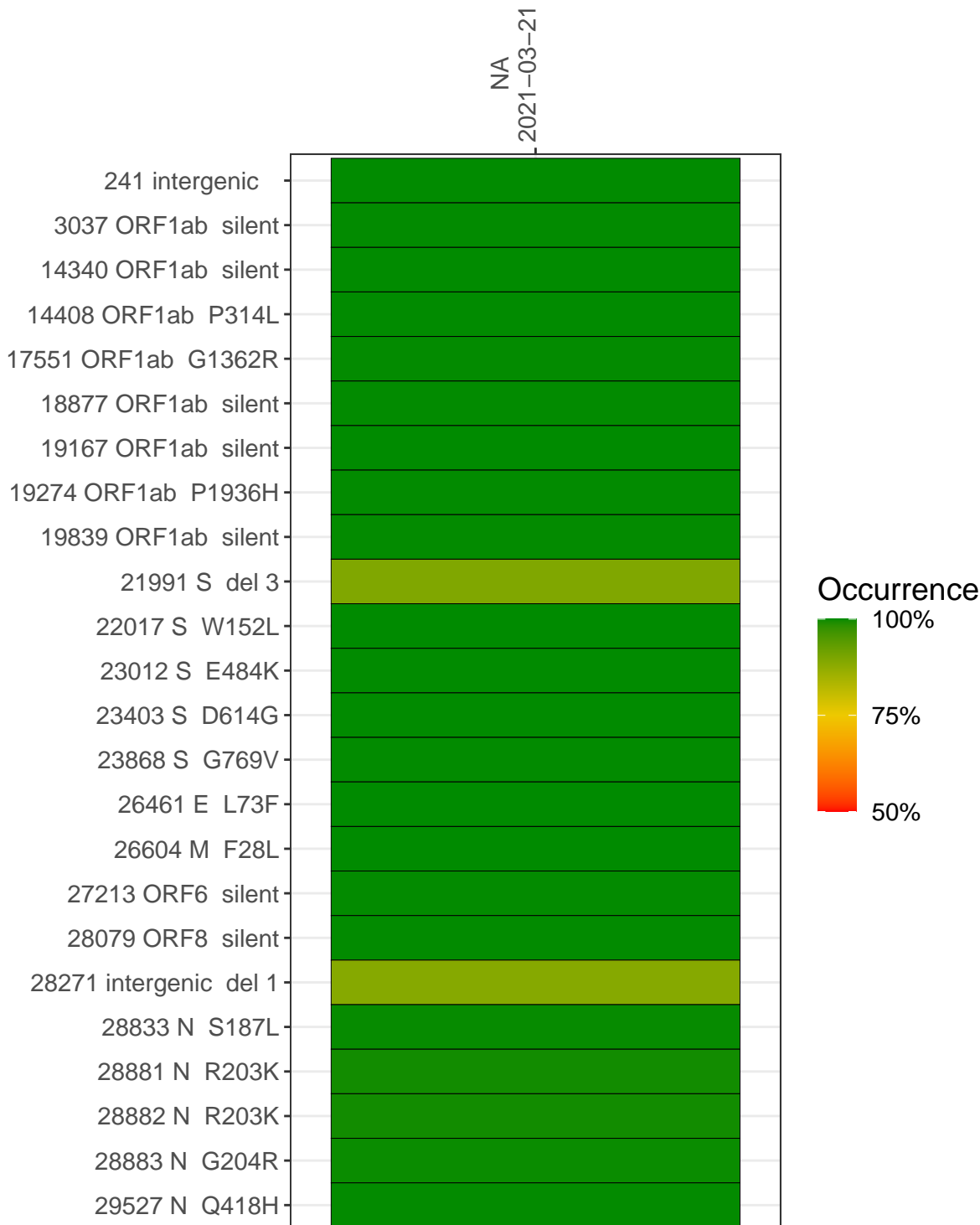# COVID-19 subject UPHS-0498

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
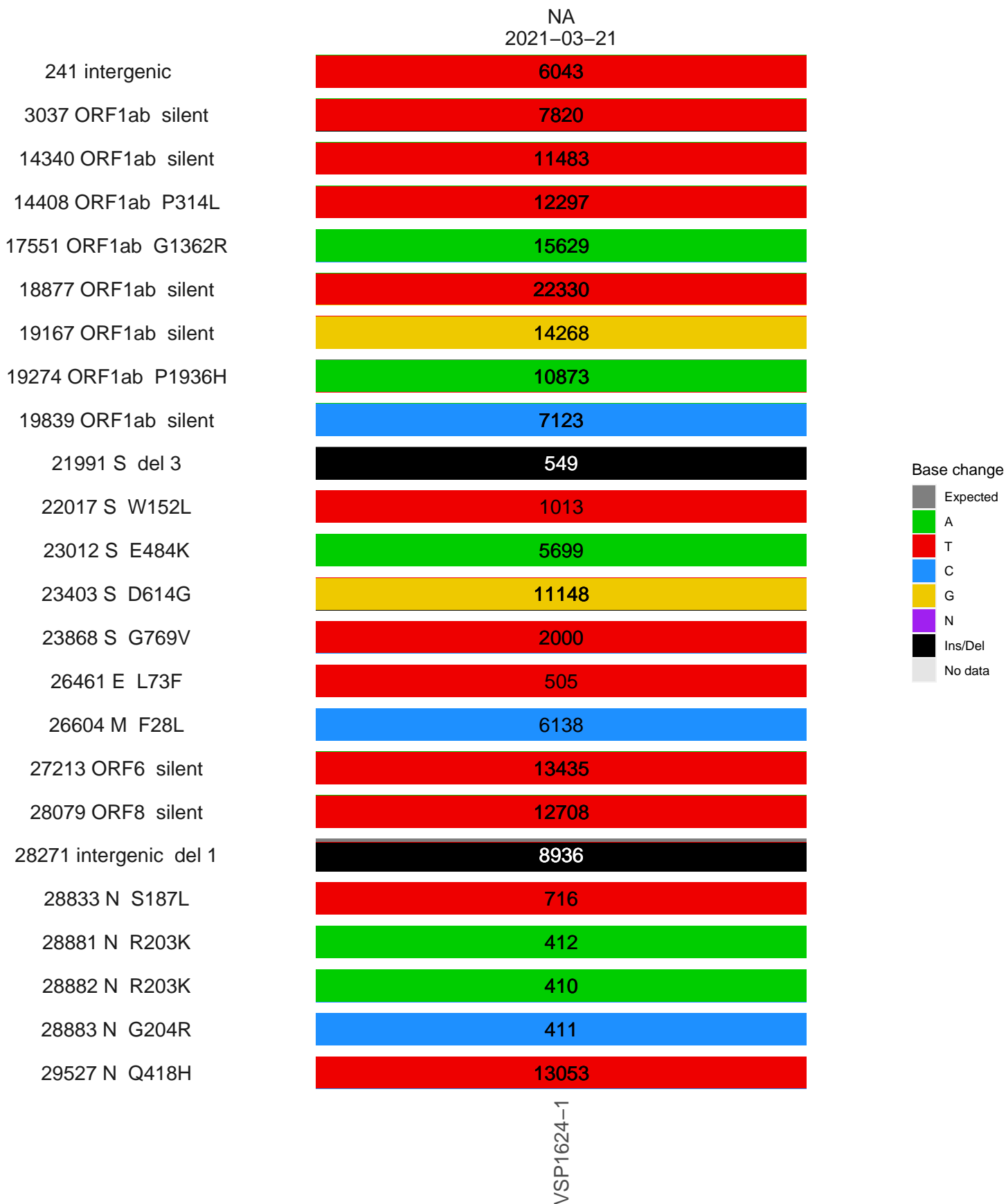software tool (Rambaut et al 2020) for genomes with > 90% sequence coverage.

Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP1624-1	single experiment	NA	NA	2021-03-21	21.94	R.1	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
2021-03-21

Position / Gene / Mutation	VSP1624-1	Base change
241 intergenic	6043	T
3037 ORF1ab silent	7820	T
14340 ORF1ab silent	11483	T
14408 ORF1ab P314L	12297	T
17551 ORF1ab G1362R	15629	A
18877 ORF1ab silent	22330	T
19167 ORF1ab silent	14268	G
19274 ORF1ab P1936H	10873	A
19839 ORF1ab silent	7123	C
21991 S del 3	549	Ins/Del
22017 S W152L	1013	T
23012 S E484K	5699	A
23403 S D614G	11148	G
23868 S G769V	2000	T
26461 E L73F	505	T
26604 M F28L	6138	C
27213 ORF6 silent	13435	T
28079 ORF8 silent	12708	T
28271 intergenic del 1	8936	Ins/Del
28833 N S187L	716	T
28881 N R203K	412	A
28882 N R203K	410	A
28883 N G204R	411	C
29527 N Q418H	13053	T

Base change

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

Analyses of individual experiments and composite results

## VSP1624-1 | 2021-03-21 | NA | UPHS-0498 | 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.
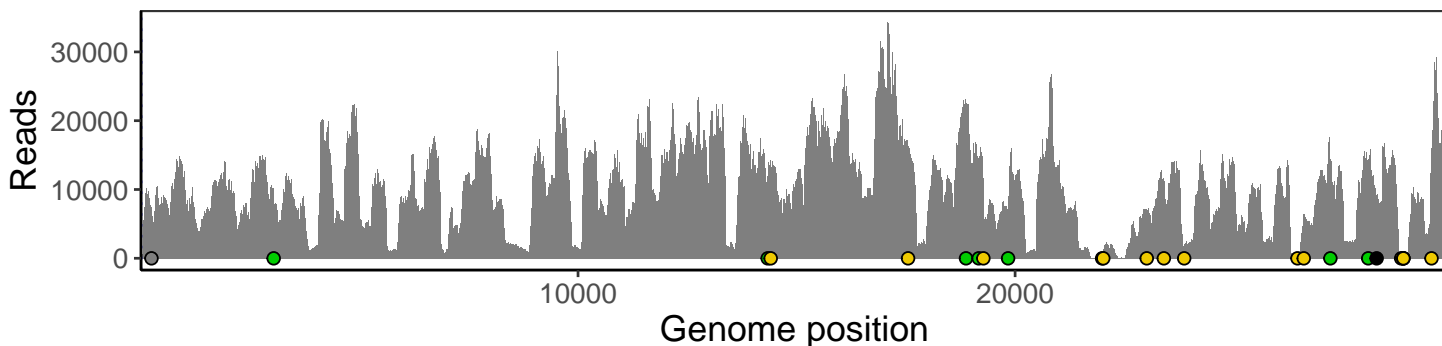
Excerpt from plot above focusing on reads coverage from 0 to 50 NT.
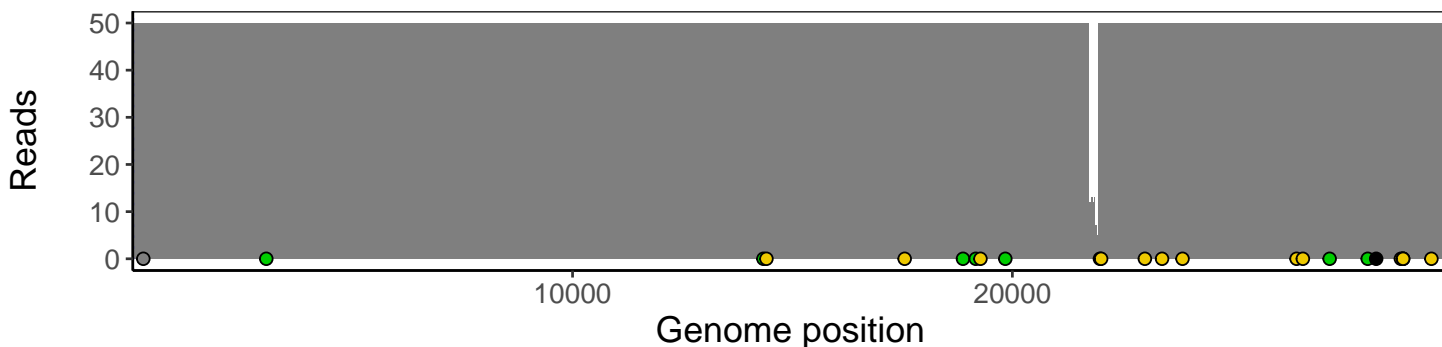
The longest five assembled contigs are shown below colored by their edit distance to the reference genome.
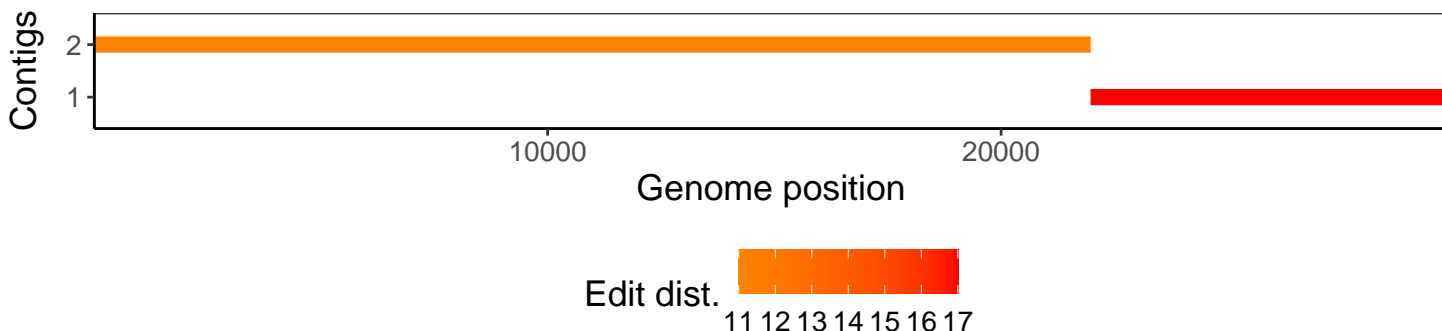
Software environment

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