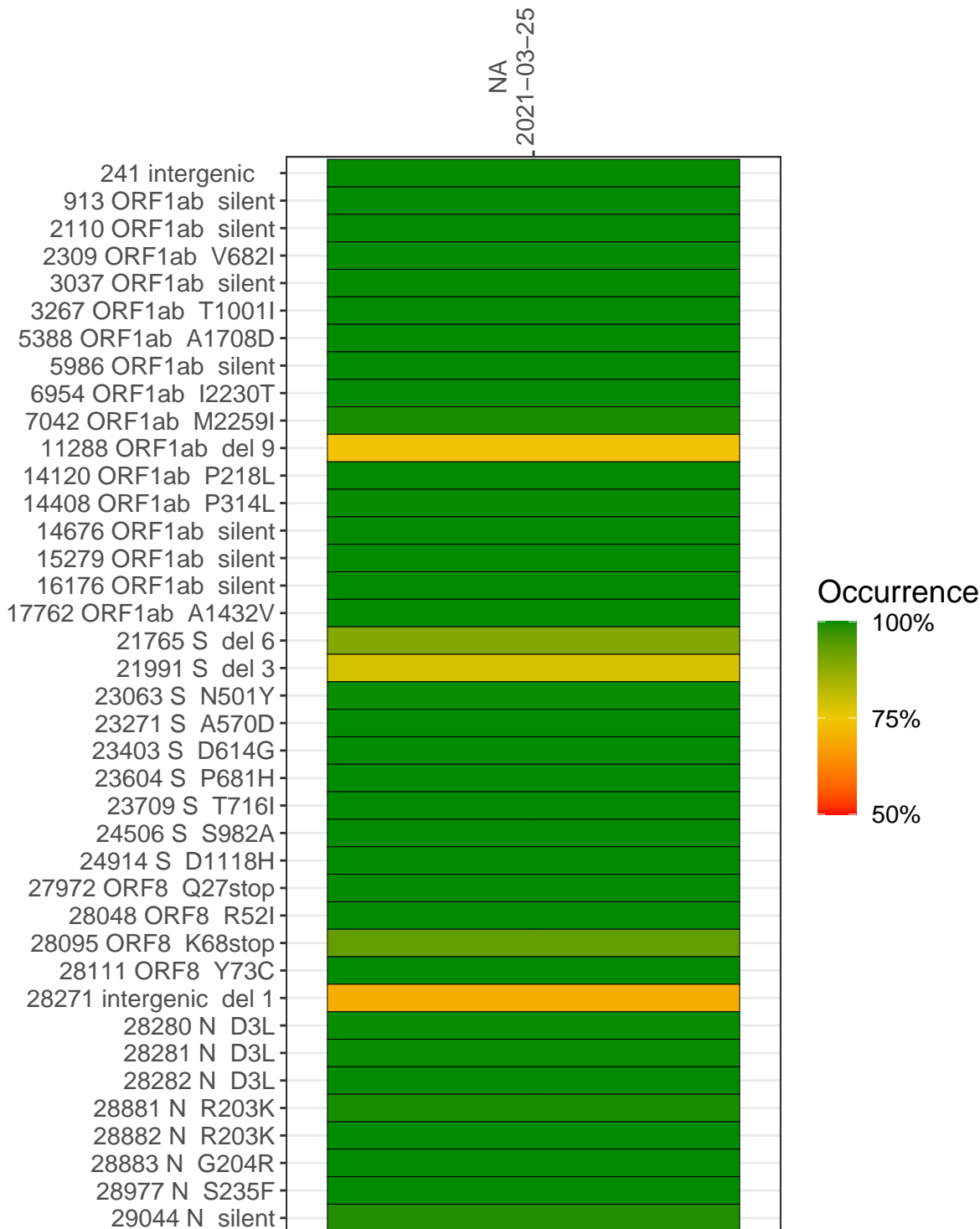# COVID-19 subject UPHS-0520

2021-06-03

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.
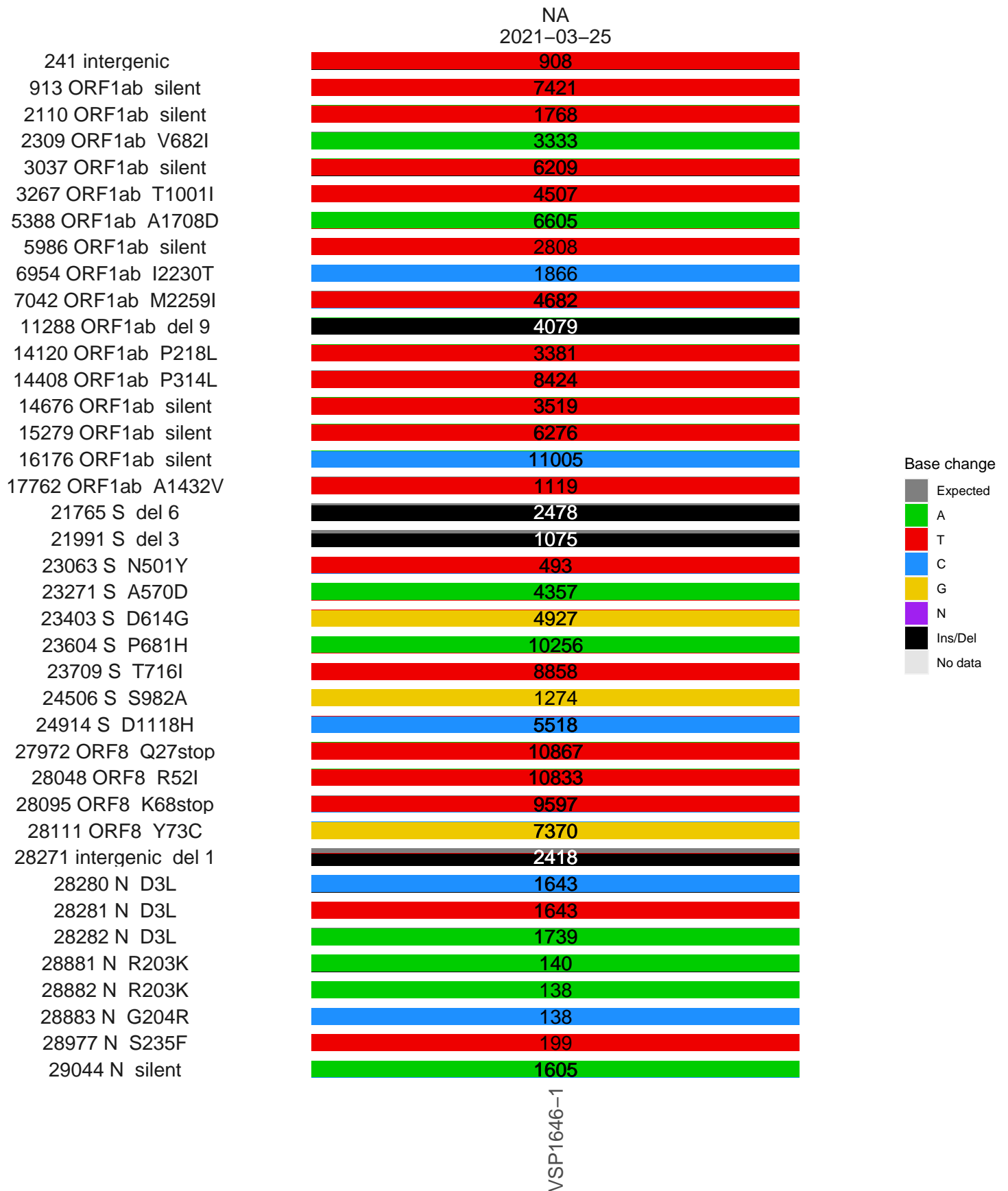
Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP1646-1	single experiment	NA	NA	2021-03-25	29.83	B.1.1.7	99.9%	99.8%

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
2021−03−25

Position/Annotation	Value	Base
241 intergenic	908	T
913 ORF1ab silent	7421	T
2110 ORF1ab silent	1768	T
2309 ORF1ab V682I	3333	A
3037 ORF1ab silent	6209	T
3267 ORF1ab T1001I	4507	T
5388 ORF1ab A1708D	6605	A
5986 ORF1ab silent	2808	T
6954 ORF1ab I2230T	1866	C
7042 ORF1ab M2259I	4682	T
11288 ORF1ab del 9	4079	Ins/Del
14120 ORF1ab P218L	3381	T
14408 ORF1ab P314L	8424	T
14676 ORF1ab silent	3519	T
15279 ORF1ab silent	6276	T
16176 ORF1ab silent	11005	C
17762 ORF1ab A1432V	1119	T
21765 S del 6	2478	Ins/Del
21991 S del 3	1075	Ins/Del
23063 S N501Y	493	T
23271 S A570D	4357	A
23403 S D614G	4927	G
23604 S P681H	10256	A
23709 S T716I	8858	T
24506 S S982A	1274	G
24914 S D1118H	5518	C
27972 ORF8 Q27stop	10867	T
28048 ORF8 R52I	10833	T
28095 ORF8 K68stop	9597	T
28111 ORF8 Y73C	7370	G
28271 intergenic del 1	2418	Ins/Del
28280 N D3L	1643	C
28281 N D3L	1643	T
28282 N D3L	1739	A
28881 N R203K	140	A
28882 N R203K	138	A
28883 N G204R	138	C
28977 N S235F	199	T
29044 N silent	1605	A

VSP1646−1

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

**VSP1646-1 | 2021-03-25 | NA | UPHS-0520 | 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.
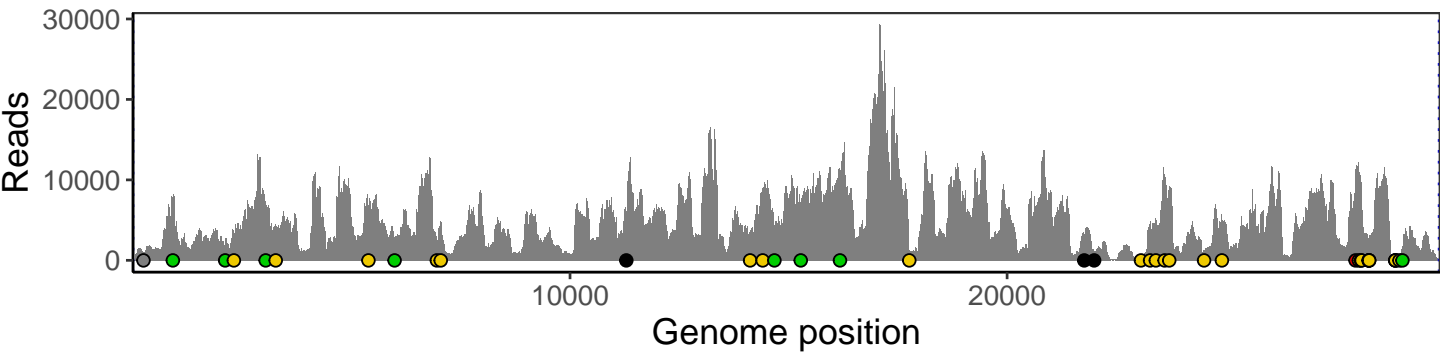
Excerpt from plot above focusing on reads coverage from 0 to 50 NT.
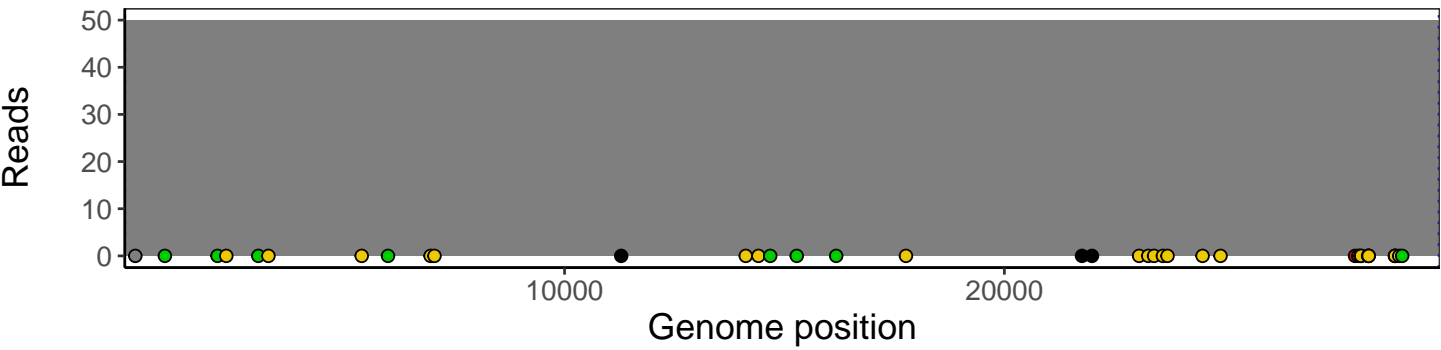
The longest five assembled contigs are shown below colored by their edit distance to the reference genome.

4
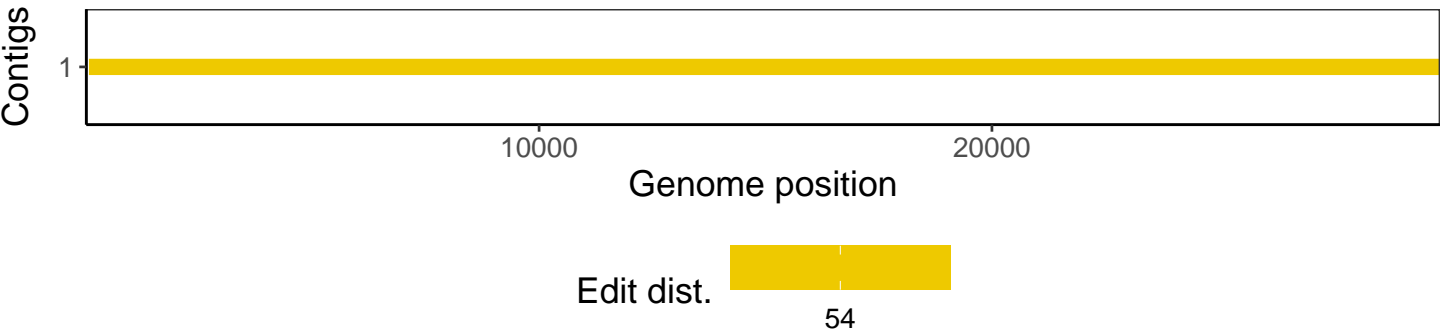
Software environment

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	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