COVID-19 subject UPHS-0860

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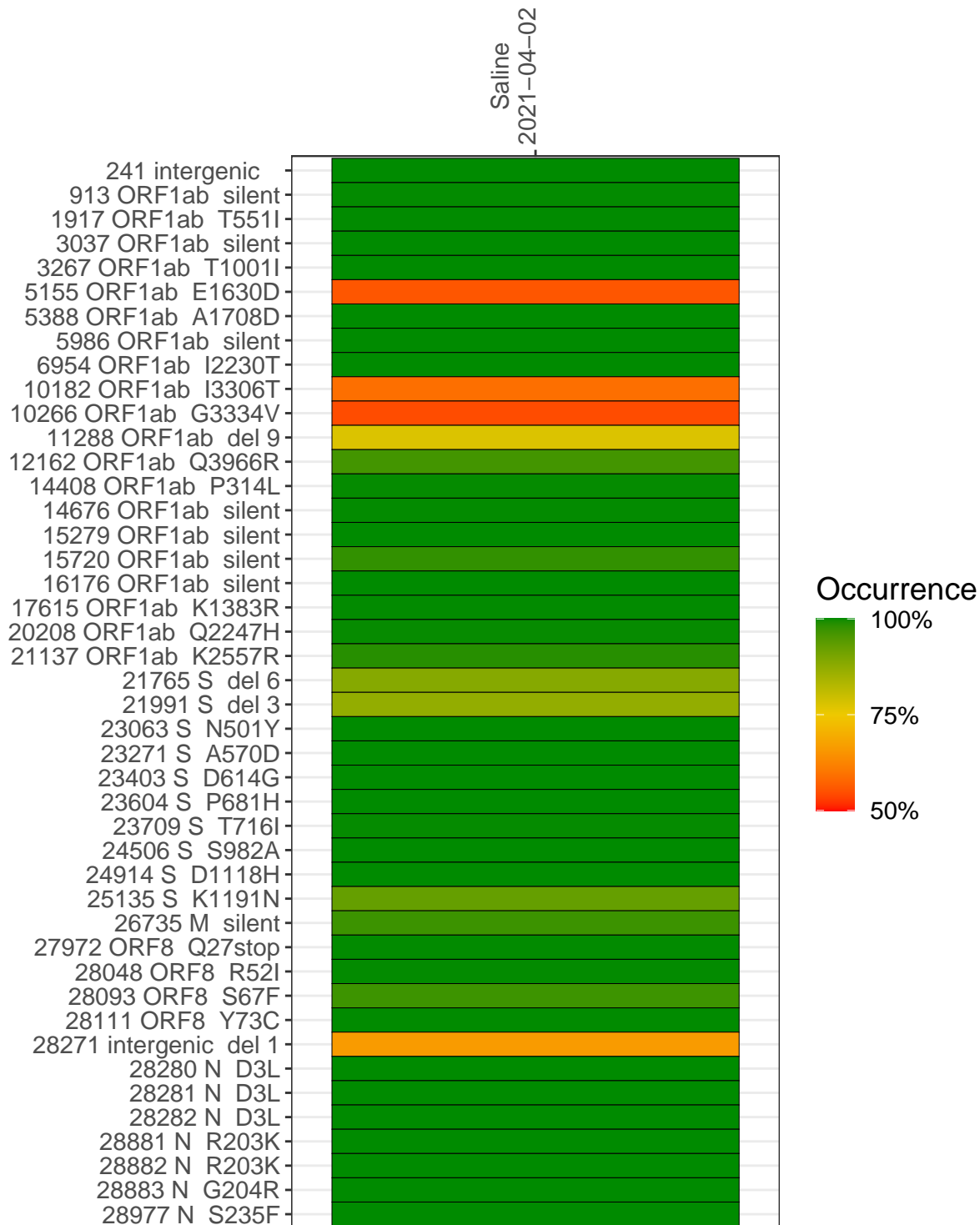 sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software tool (Rambaut et al 2020) for genomes with > 90% sequence coverage.
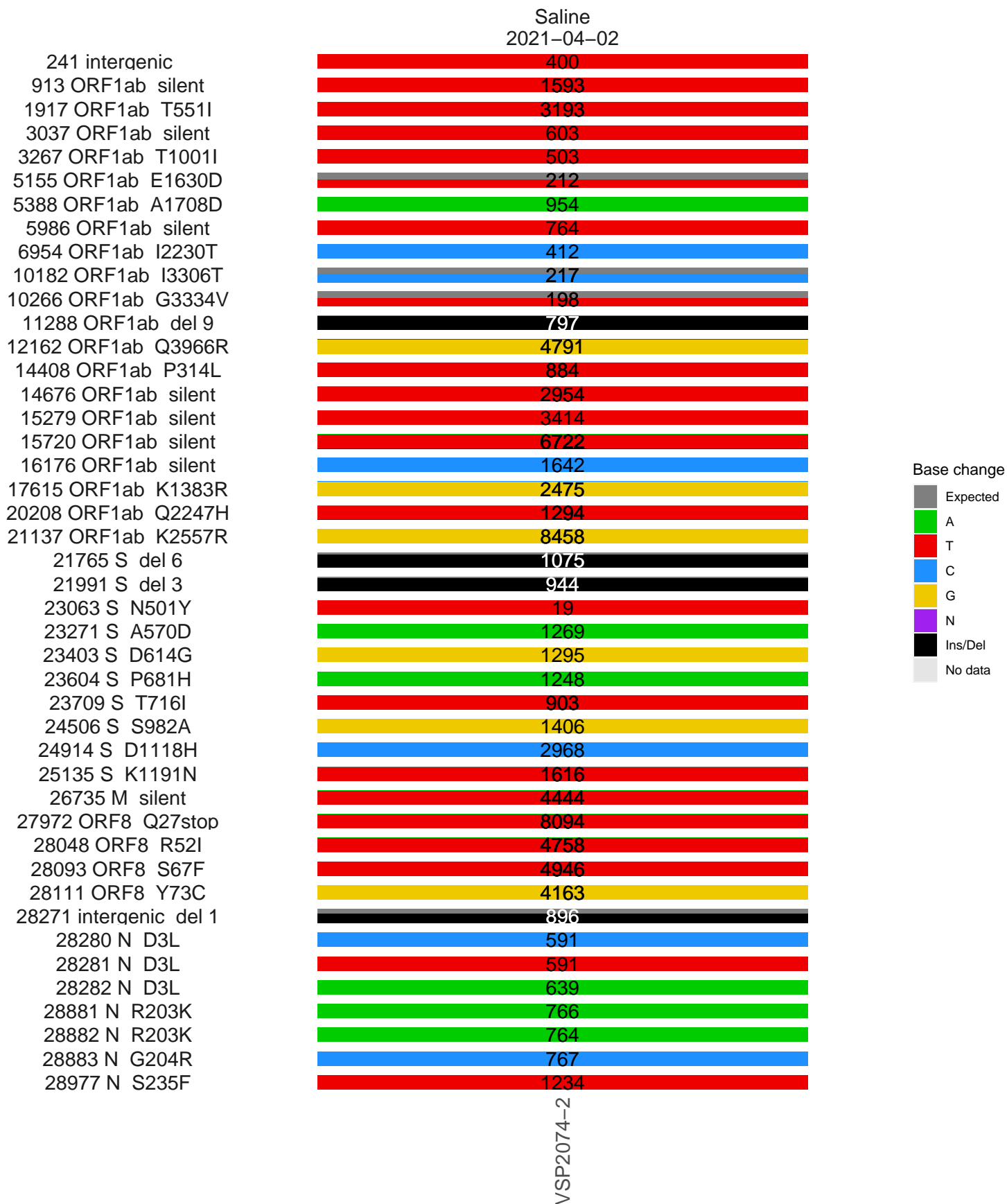
Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP2074-2	single experiment	NA	Saline	2021-04-02	29.65	B.1.1.7	99.3%	99.2%

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.

Saline
2021−04−02

Position	VSP2074-2
241 intergenic	400
913 ORF1ab silent	1593
1917 ORF1ab T551I	3193
3037 ORF1ab silent	603
3267 ORF1ab T1001I	503
5155 ORF1ab E1630D	212
5388 ORF1ab A1708D	954
5986 ORF1ab silent	764
6954 ORF1ab I2230T	412
10182 ORF1ab I3306T	217
10266 ORF1ab G3334V	198
11288 ORF1ab del 9	797
12162 ORF1ab Q3966R	4791
14408 ORF1ab P314L	884
14676 ORF1ab silent	2954
15279 ORF1ab silent	3414
15720 ORF1ab silent	6722
16176 ORF1ab silent	1642
17615 ORF1ab K1383R	2475
20208 ORF1ab Q2247H	1294
21137 ORF1ab K2557R	8458
21765 S del 6	1075
21991 S del 3	944
23063 S N501Y	19
23271 S A570D	1269
23403 S D614G	1295
23604 S P681H	1248
23709 S T716I	903
24506 S S982A	1406
24914 S D1118H	2968
25135 S K1191N	1616
26735 M silent	4444
27972 ORF8 Q27stop	8094
28048 ORF8 R52I	4758
28093 ORF8 S67F	4946
28111 ORF8 Y73C	4163
28271 intergenic del 1	896
28280 N D3L	591
28281 N D3L	591
28282 N D3L	639
28881 N R203K	766
28882 N R203K	764
28883 N G204R	767
28977 N S235F	1234

Base change

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

VSP2074−2

3

Analyses of individual experiments and composite results

## VSP2074-2 | 2021-04-02 | Saline | UPHS-0860 | 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.
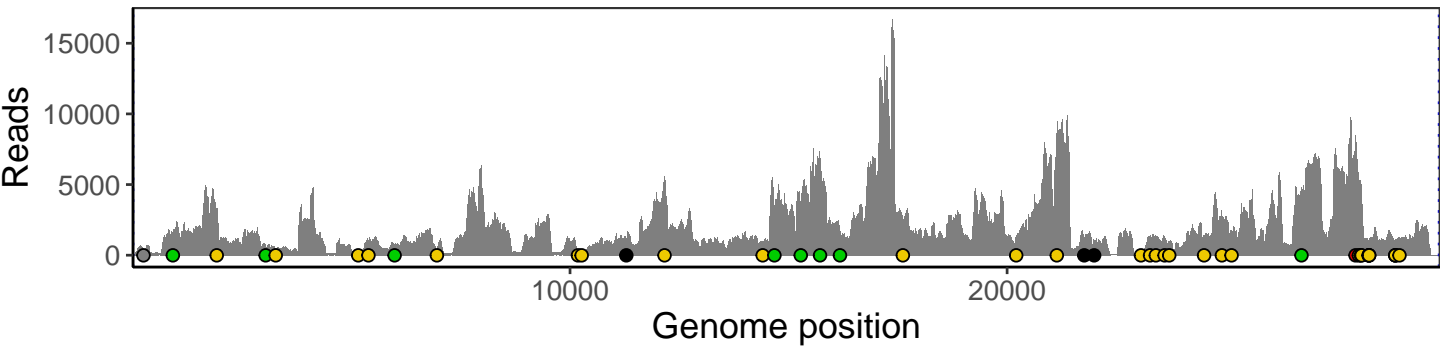
Excerpt from plot above focusing on reads coverage from 0 to 50 NT.
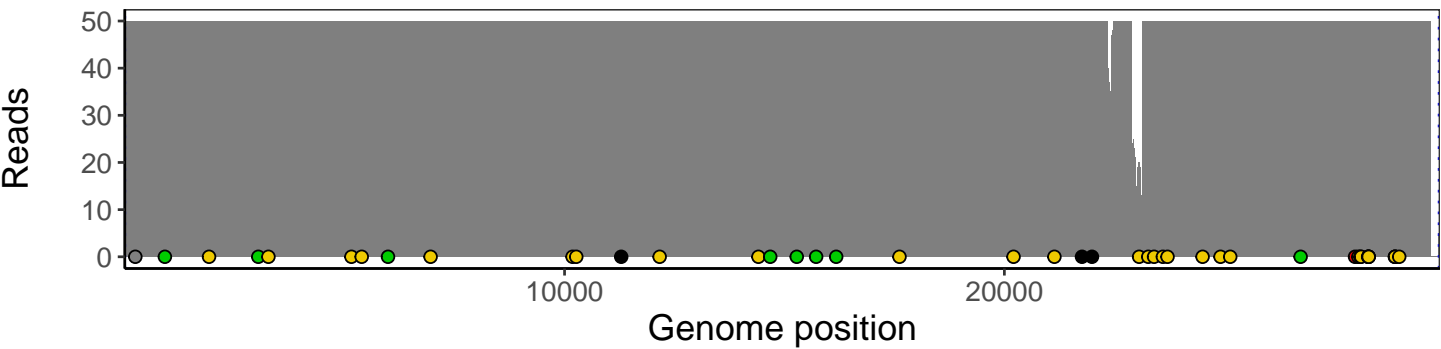
The longest five assembled contigs are shown below colored by their edit distance to the reference genome.

4
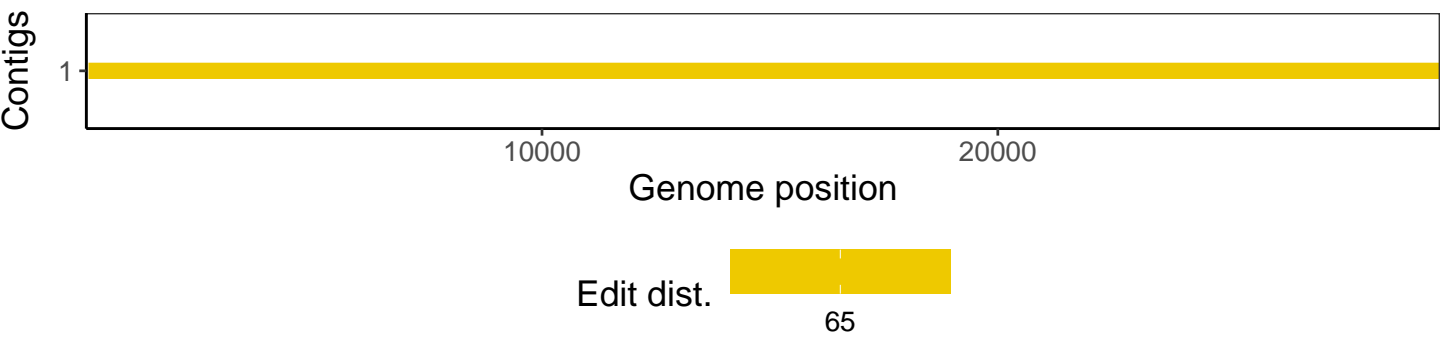
Software environment

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