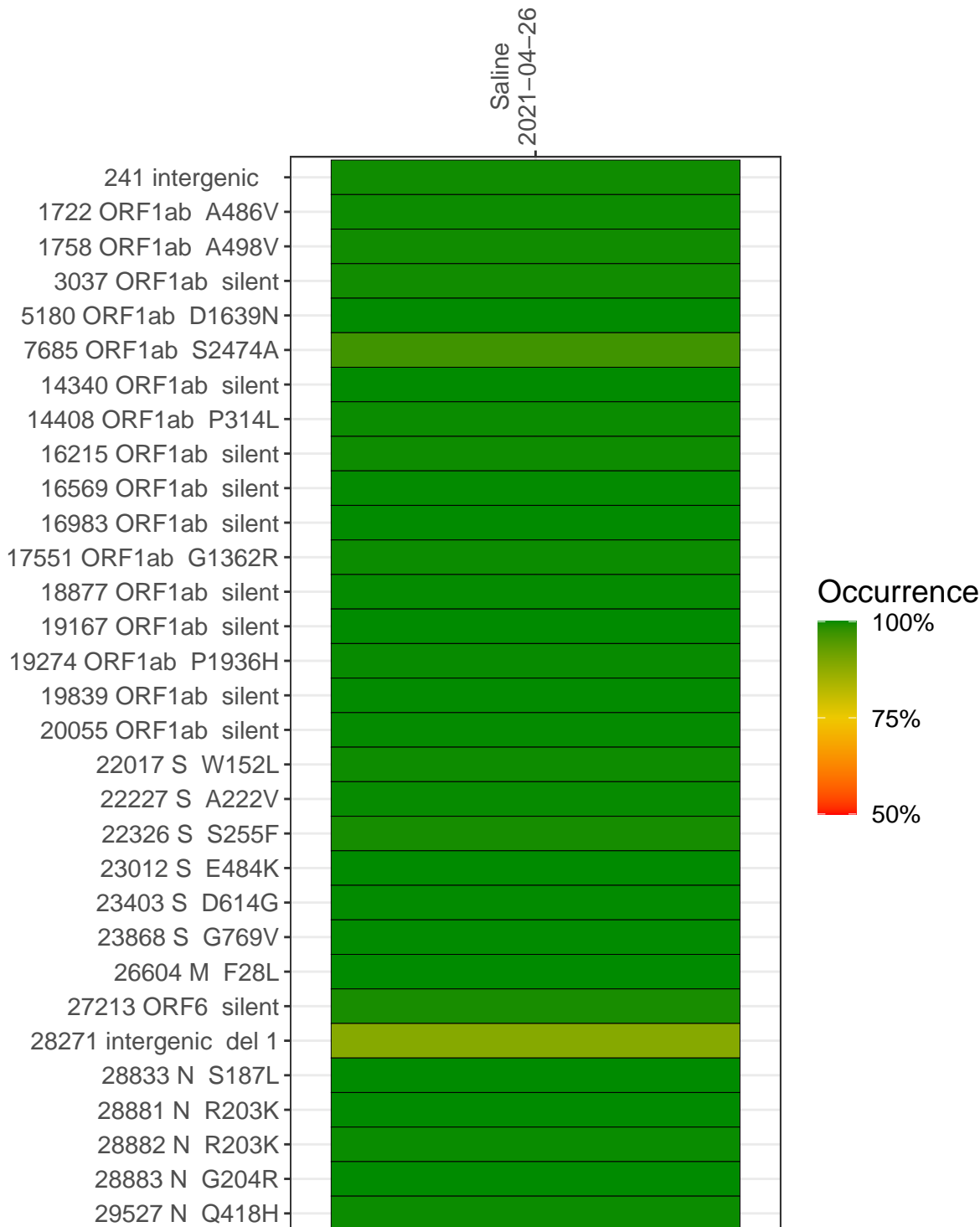# COVID-19 subject UPHS-1348

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.
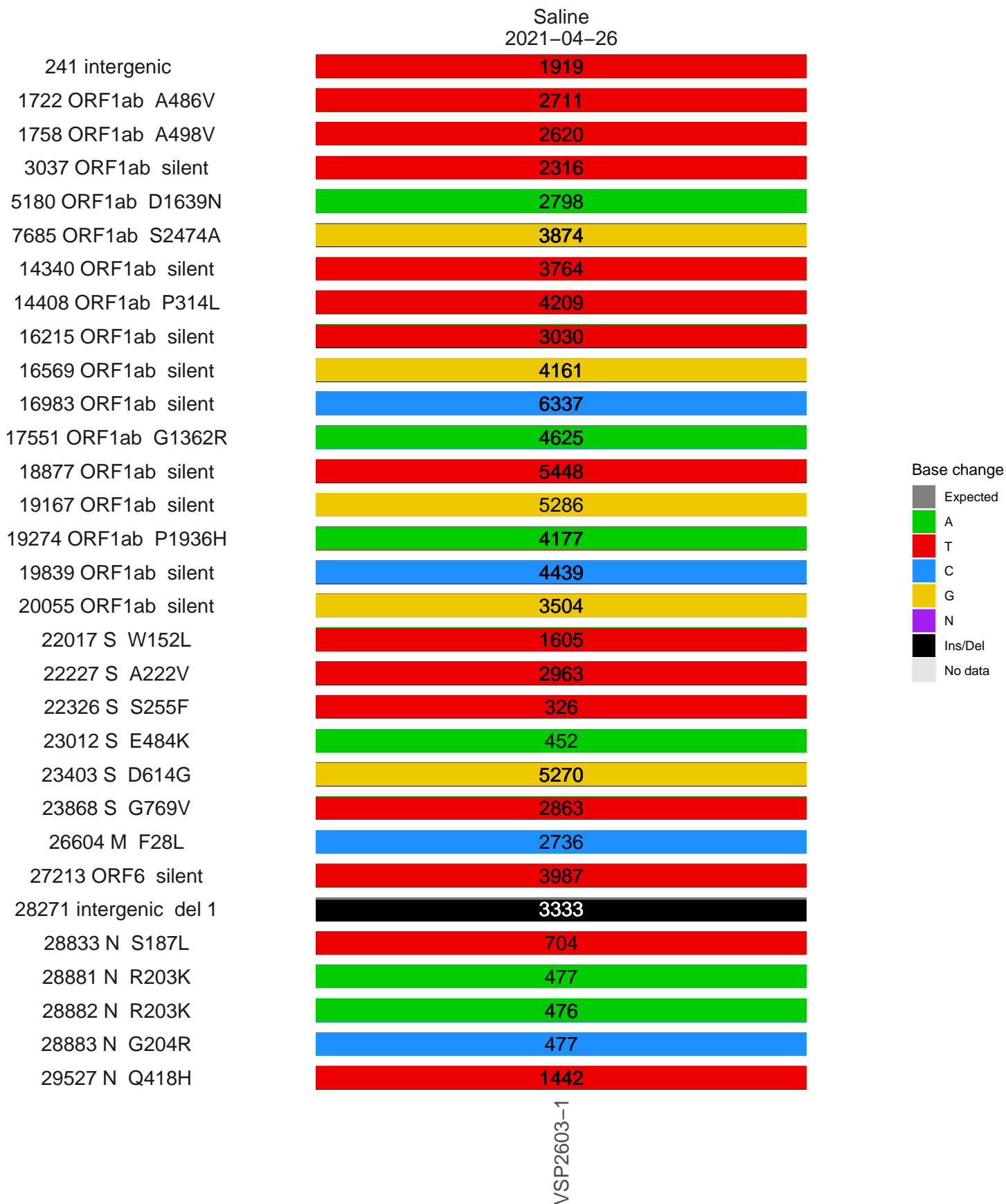
Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP2603-1	single experiment	NA	Saline	2021-04-26	29.82	R.1	99.7%	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.

Saline
2021−04−26

Position	VSP2603−1
241 intergenic	1919
1722 ORF1ab A486V	2711
1758 ORF1ab A498V	2620
3037 ORF1ab silent	2316
5180 ORF1ab D1639N	2798
7685 ORF1ab S2474A	3874
14340 ORF1ab silent	3764
14408 ORF1ab P314L	4209
16215 ORF1ab silent	3030
16569 ORF1ab silent	4161
16983 ORF1ab silent	6337
17551 ORF1ab G1362R	4625
18877 ORF1ab silent	5448
19167 ORF1ab silent	5286
19274 ORF1ab P1936H	4177
19839 ORF1ab silent	4439
20055 ORF1ab silent	3504
22017 S W152L	1605
22227 S A222V	2963
22326 S S255F	326
23012 S E484K	452
23403 S D614G	5270
23868 S G769V	2863
26604 M F28L	2736
27213 ORF6 silent	3987
28271 intergenic del 1	3333
28833 N S187L	704
28881 N R203K	477
28882 N R203K	476
28883 N G204R	477
29527 N Q418H	1442

Base change

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

3

Analyses of individual experiments and composite results

**VSP2603-1 | 2021-04-26 | Saline | UPHS-1348 | 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.
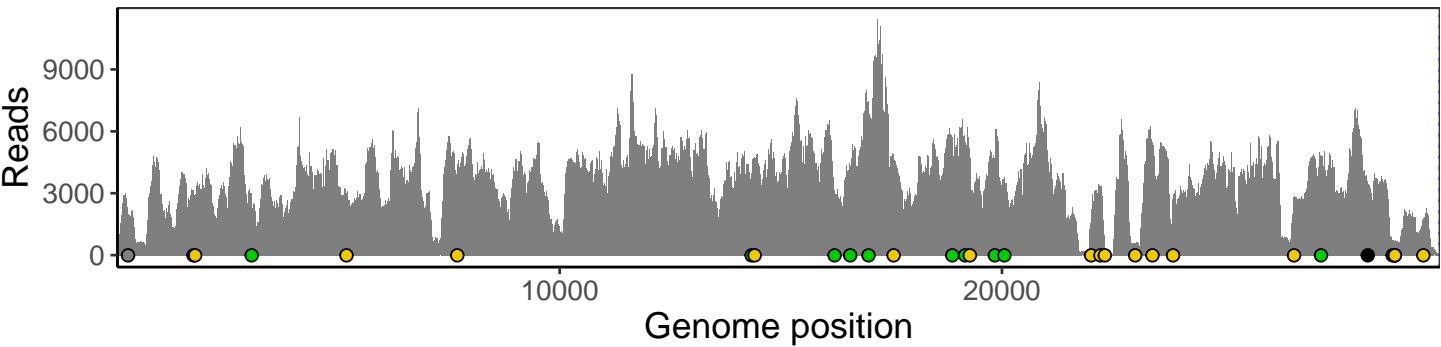
Excerpt from plot above focusing on reads coverage from 0 to 50 NT.
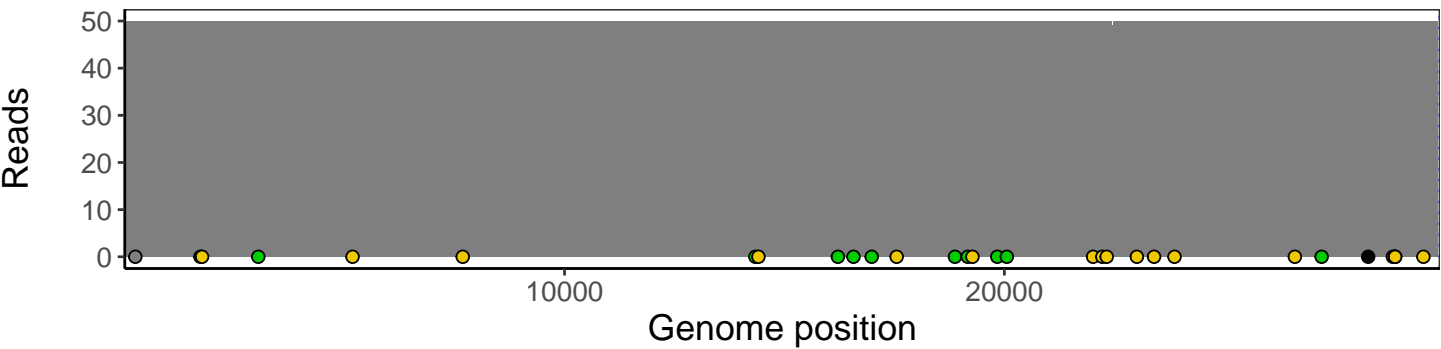
The longest five assembled contigs are shown below colored by their edit distance to the reference genome.

4
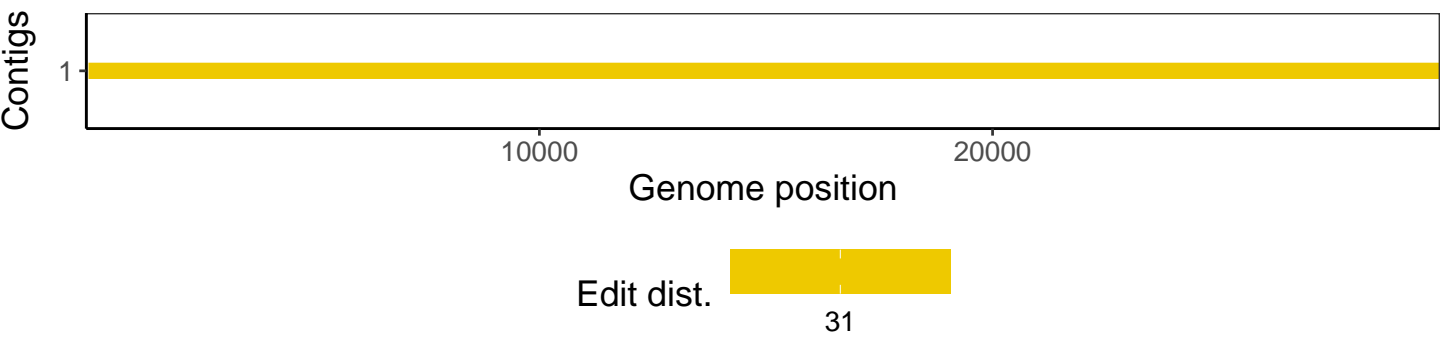
Software environment

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