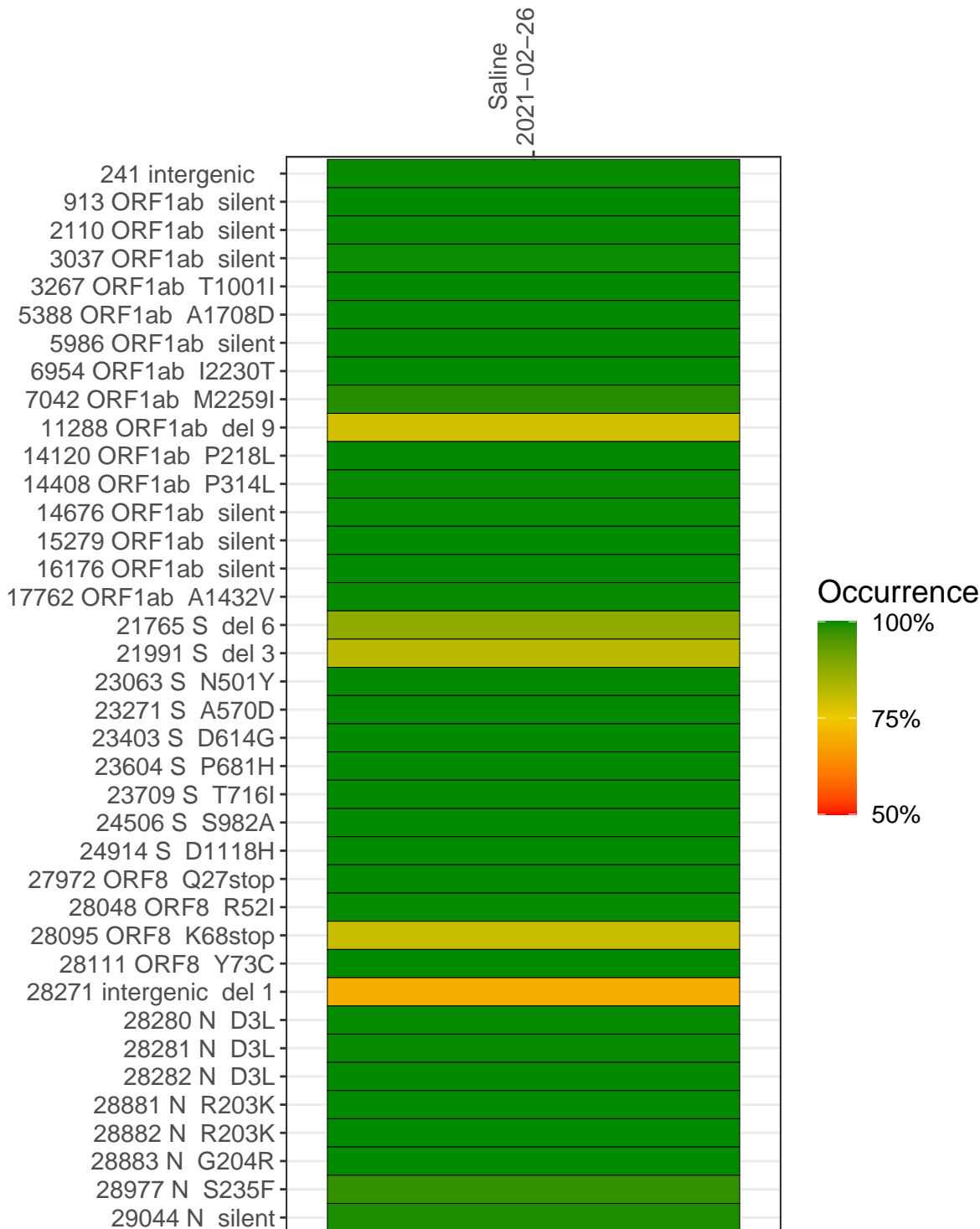# COVID-19 subject HUP Q-0021

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.
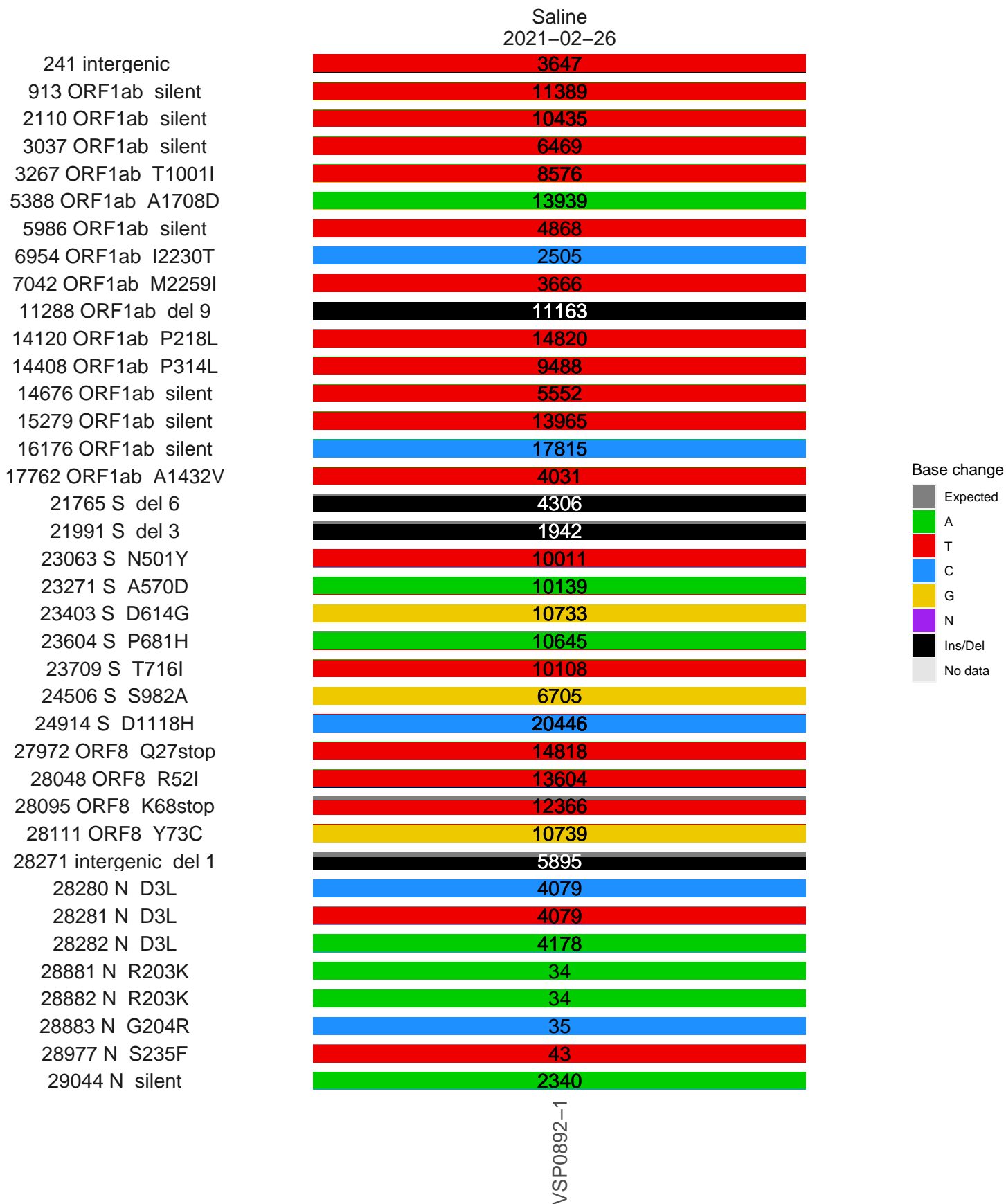
Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP0892-1	single experiment	NA	Saline	2021-02-26	29.87	B.1.1.7	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.

Saline
2021−02−26

Position / Gene / Mutation	Value	Base change
241 intergenic	3647	T
913 ORF1ab silent	11389	T
2110 ORF1ab silent	10435	T
3037 ORF1ab silent	6469	T
3267 ORF1ab T1001I	8576	T
5388 ORF1ab A1708D	13939	A
5986 ORF1ab silent	4868	T
6954 ORF1ab I2230T	2505	C
7042 ORF1ab M2259I	3666	T
11288 ORF1ab del 9	11163	Ins/Del
14120 ORF1ab P218L	14820	T
14408 ORF1ab P314L	9488	T
14676 ORF1ab silent	5552	T
15279 ORF1ab silent	13965	T
16176 ORF1ab silent	17815	C
17762 ORF1ab A1432V	4031	T
21765 S del 6	4306	Ins/Del
21991 S del 3	1942	Ins/Del
23063 S N501Y	10011	T
23271 S A570D	10139	A
23403 S D614G	10733	G
23604 S P681H	10645	A
23709 S T716I	10108	T
24506 S S982A	6705	G
24914 S D1118H	20446	C
27972 ORF8 Q27stop	14818	T
28048 ORF8 R52I	13604	T
28095 ORF8 K68stop	12366	Expected/T
28111 ORF8 Y73C	10739	G
28271 intergenic del 1	5895	Ins/Del
28280 N D3L	4079	C
28281 N D3L	4079	T
28282 N D3L	4178	A
28881 N R203K	34	A
28882 N R203K	34	A
28883 N G204R	35	C
28977 N S235F	43	T
29044 N silent	2340	A

VSP0892−1

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

3

Analyses of individual experiments and composite results

VSP0892-1 | 2021-02-26 | Saline | HUP Q-0021 | 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.
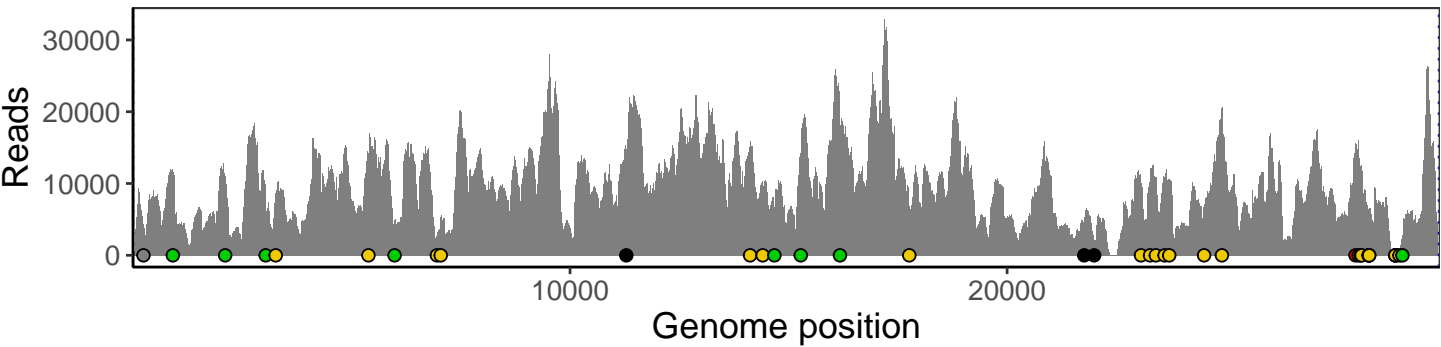
Excerpt from plot above focusing on reads coverage from 0 to 50 NT.
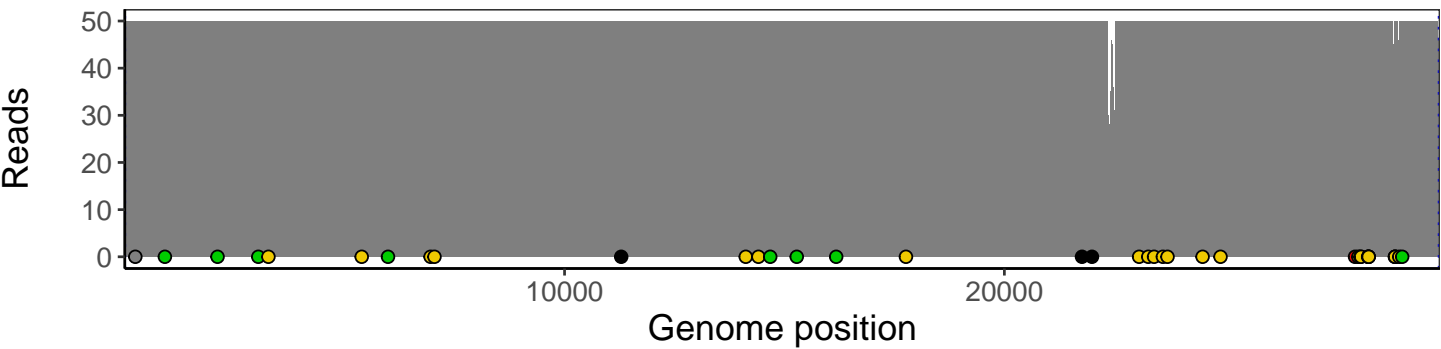
The longest five assembled contigs are shown below colored by their edit distance to the reference genome.

Edit dist.

53

4
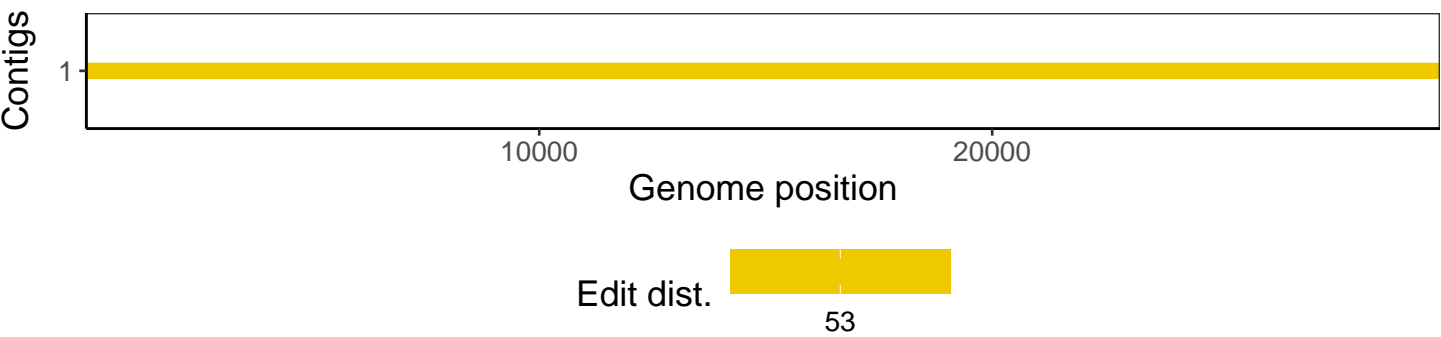
Software environment

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