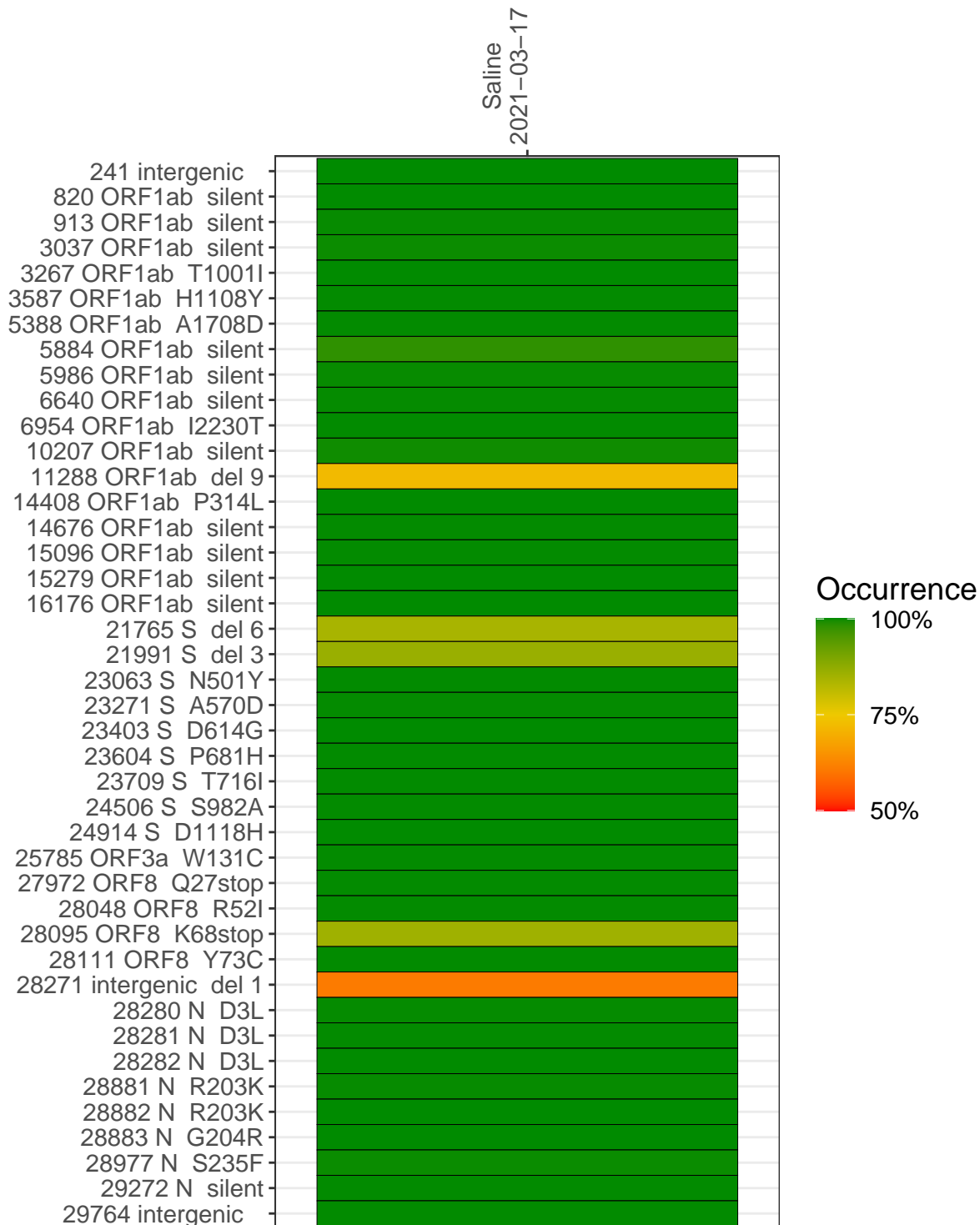# COVID-19 subject HUP Q-0066

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.
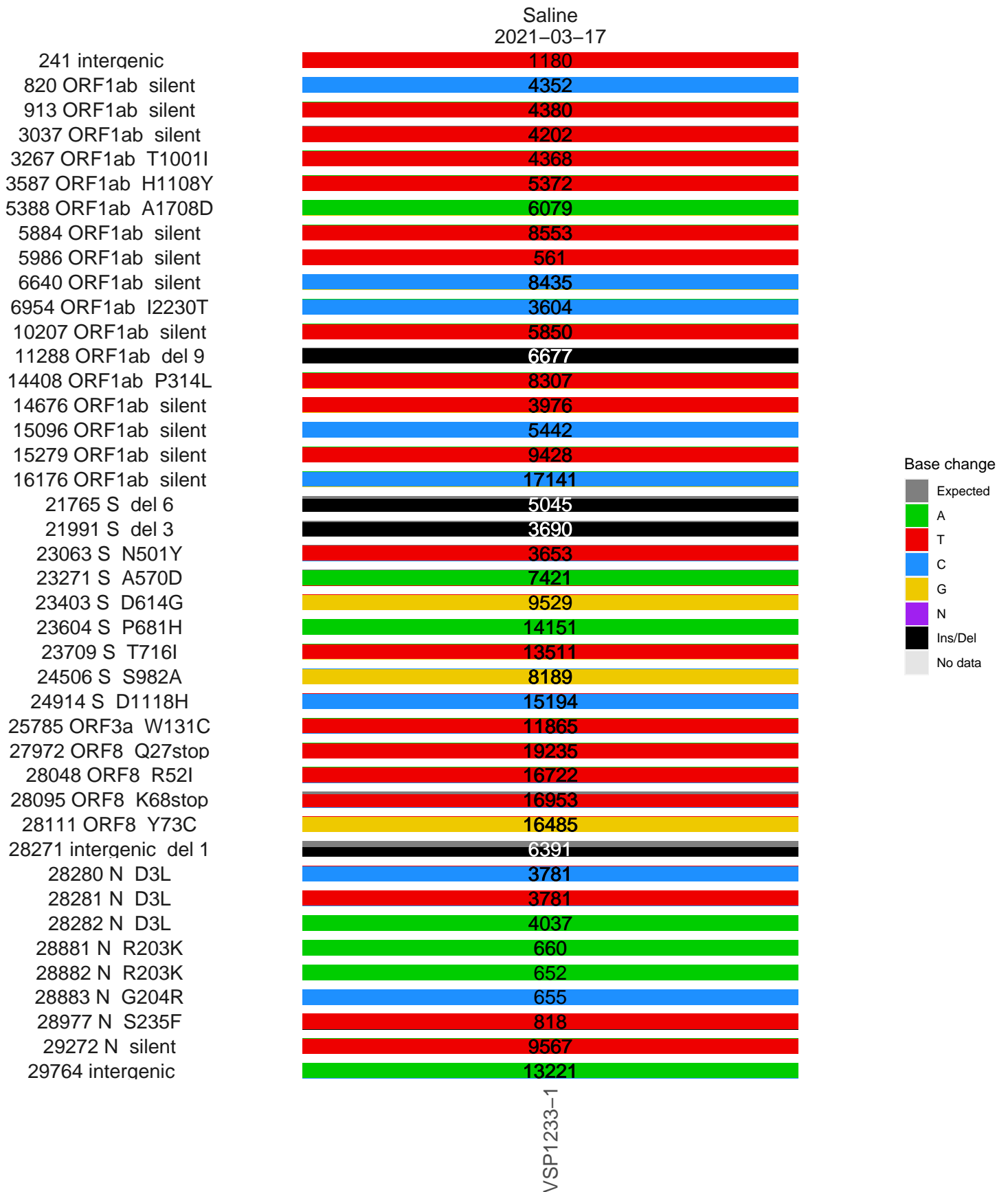
Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP1233-1	single experiment	NA	Saline	2021-03-17	29.84	B.1.1.7	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.

Saline
2021-03-17

Position	Value	Base
241 intergenic	1180	T
820 ORF1ab silent	4352	C
913 ORF1ab silent	4380	T
3037 ORF1ab silent	4202	T
3267 ORF1ab T1001I	4368	T
3587 ORF1ab H1108Y	5372	T
5388 ORF1ab A1708D	6079	A
5884 ORF1ab silent	8553	T
5986 ORF1ab silent	561	T
6640 ORF1ab silent	8435	C
6954 ORF1ab I2230T	3604	C
10207 ORF1ab silent	5850	T
11288 ORF1ab del 9	6677	Ins/Del
14408 ORF1ab P314L	8307	T
14676 ORF1ab silent	3976	T
15096 ORF1ab silent	5442	C
15279 ORF1ab silent	9428	T
16176 ORF1ab silent	17141	C
21765 S del 6	5045	Ins/Del
21991 S del 3	3690	Ins/Del
23063 S N501Y	3653	T
23271 S A570D	7421	A
23403 S D614G	9529	G
23604 S P681H	14151	A
23709 S T716I	13511	T
24506 S S982A	8189	G
24914 S D1118H	15194	C
25785 ORF3a W131C	11865	T
27972 ORF8 Q27stop	19235	T
28048 ORF8 R52I	16722	T
28095 ORF8 K68stop	16953	T
28111 ORF8 Y73C	16485	G
28271 intergenic del 1	6391	Ins/Del
28280 N D3L	3781	C
28281 N D3L	3781	T
28282 N D3L	4037	A
28881 N R203K	660	A
28882 N R203K	652	A
28883 N G204R	655	C
28977 N S235F	818	T
29272 N silent	9567	T
29764 intergenic	13221	A

VSP1233-1

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

VSP1233-1 | 2021-03-17 | Saline | HUP Q-0066 | 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.
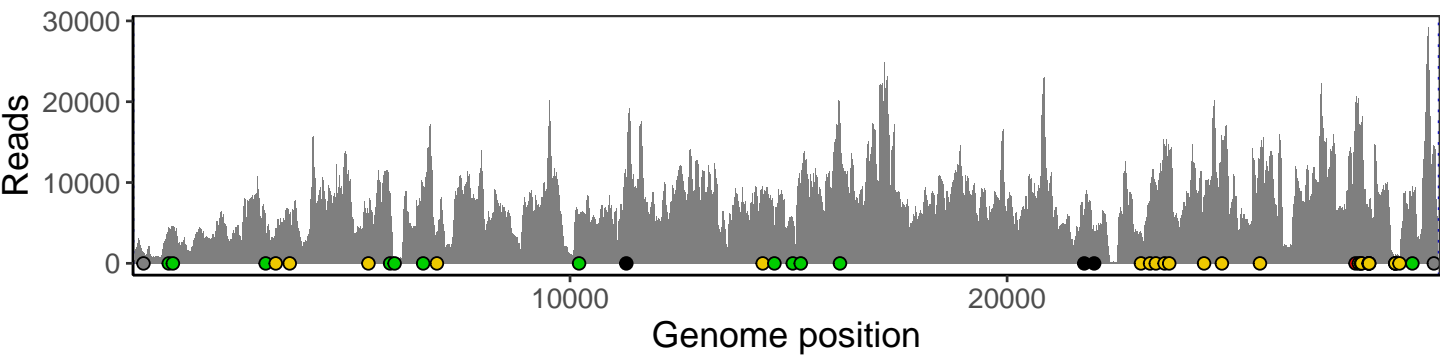
Excerpt from plot above focusing on reads coverage from 0 to 50 NT.
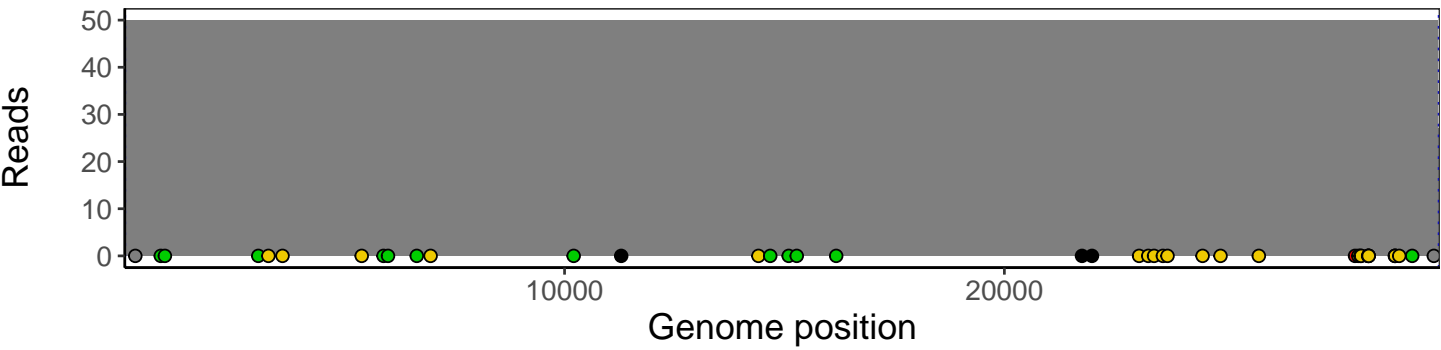
The longest five assembled contigs are shown below colored by their edit distance to the reference genome.
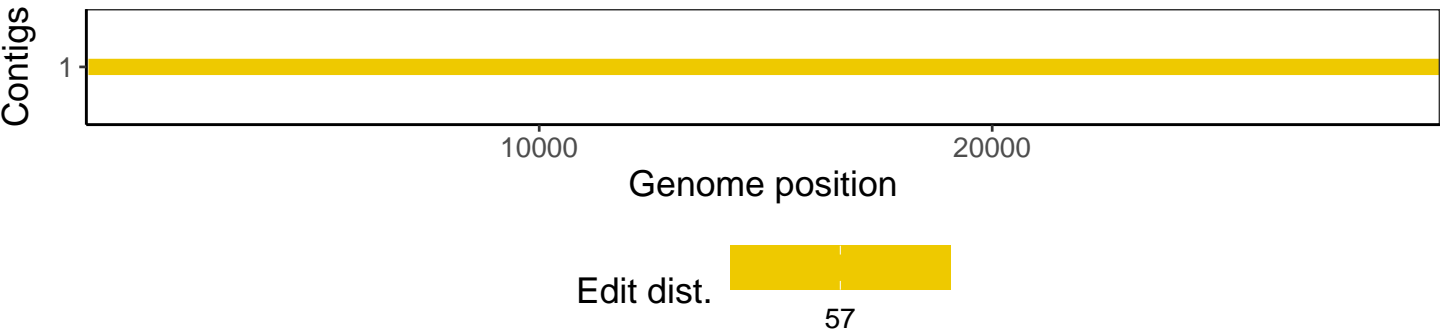
Software environment

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