COVID-19 subject UPHS-0075

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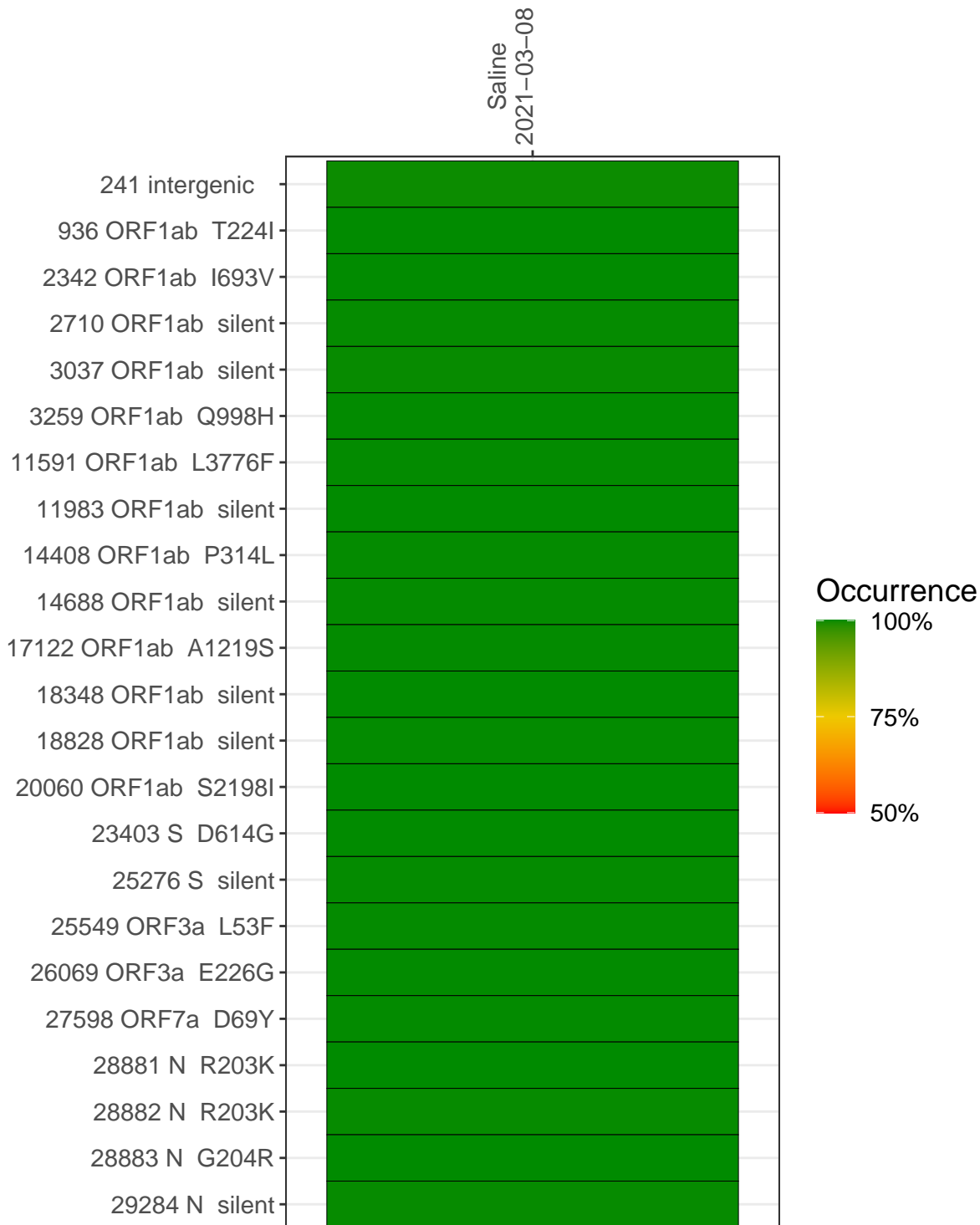 sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software tool (Rambaut et al 2020) for genomes with > 90% sequence coverage.
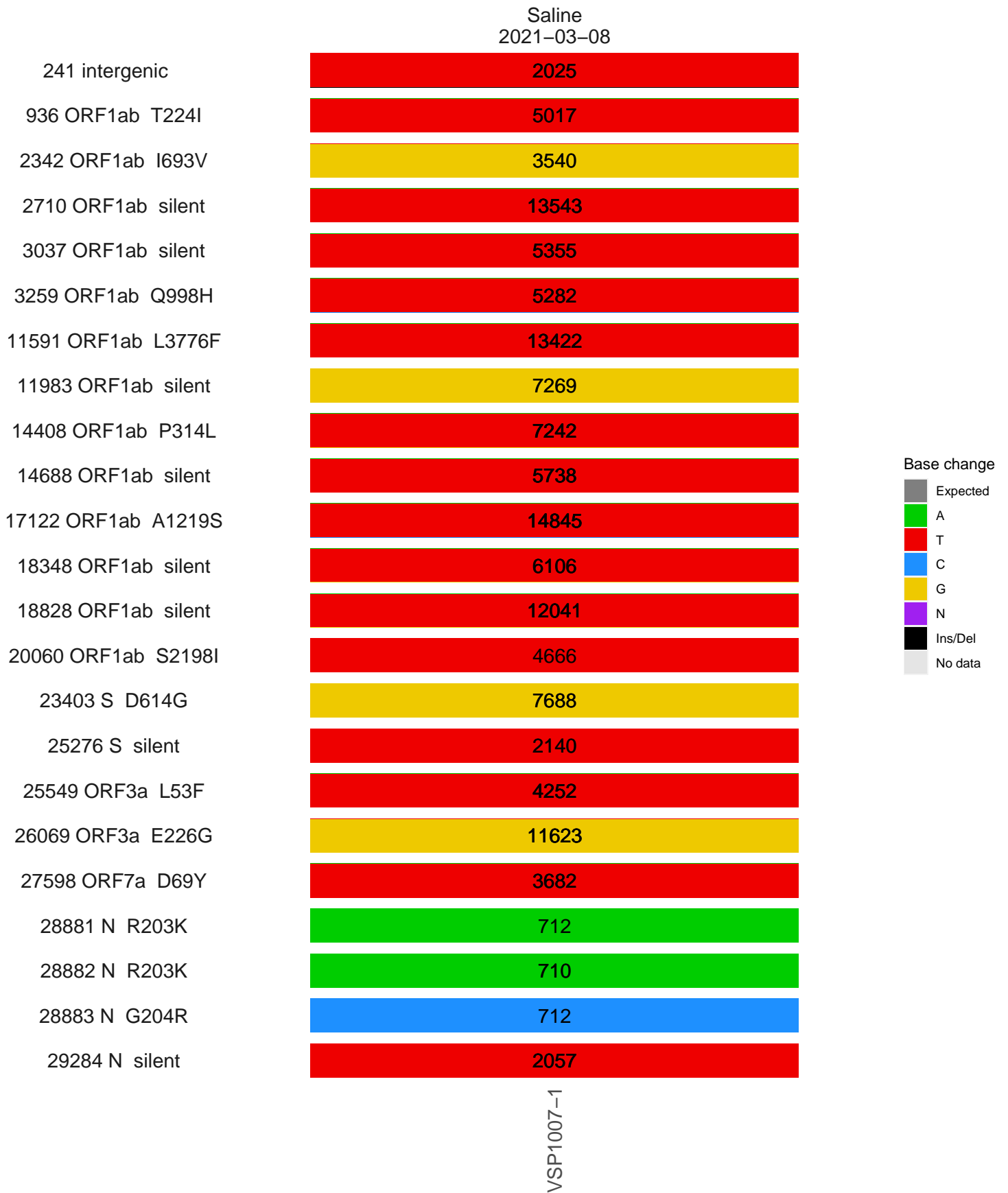
Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP1007-1	single experiment	NA	Saline	2021-03-08	29.83	B.1.1.434	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
2021−03−08

Position	Value	Base change
241 intergenic	2025	T
936 ORF1ab T224I	5017	T
2342 ORF1ab I693V	3540	G
2710 ORF1ab silent	13543	T
3037 ORF1ab silent	5355	T
3259 ORF1ab Q998H	5282	T
11591 ORF1ab L3776F	13422	T
11983 ORF1ab silent	7269	G
14408 ORF1ab P314L	7242	T
14688 ORF1ab silent	5738	T
17122 ORF1ab A1219S	14845	T
18348 ORF1ab silent	6106	T
18828 ORF1ab silent	12041	T
20060 ORF1ab S2198I	4666	T
23403 S D614G	7688	G
25276 S silent	2140	T
25549 ORF3a L53F	4252	T
26069 ORF3a E226G	11623	G
27598 ORF7a D69Y	3682	T
28881 N R203K	712	A
28882 N R203K	710	A
28883 N G204R	712	C
29284 N silent	2057	T

Base change

Expected
A
T
C
G
N
Ins/Del
No data

VSP1007−1

3

Analyses of individual experiments and composite results

**VSP1007-1 | 2021-03-08 | Saline | UPHS-0075 | 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.
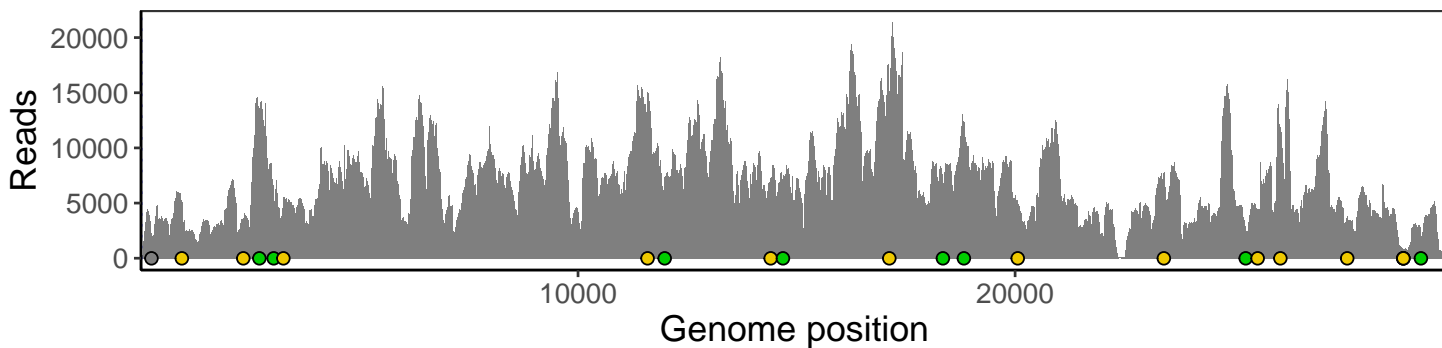
Excerpt from plot above focusing on reads coverage from 0 to 50 NT.
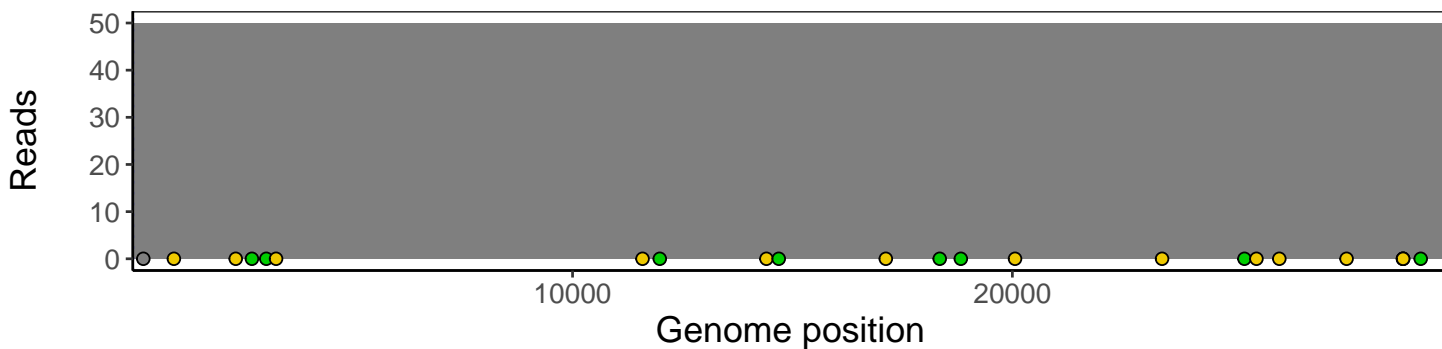
The longest five assembled contigs are shown below colored by their edit distance to the reference genome.
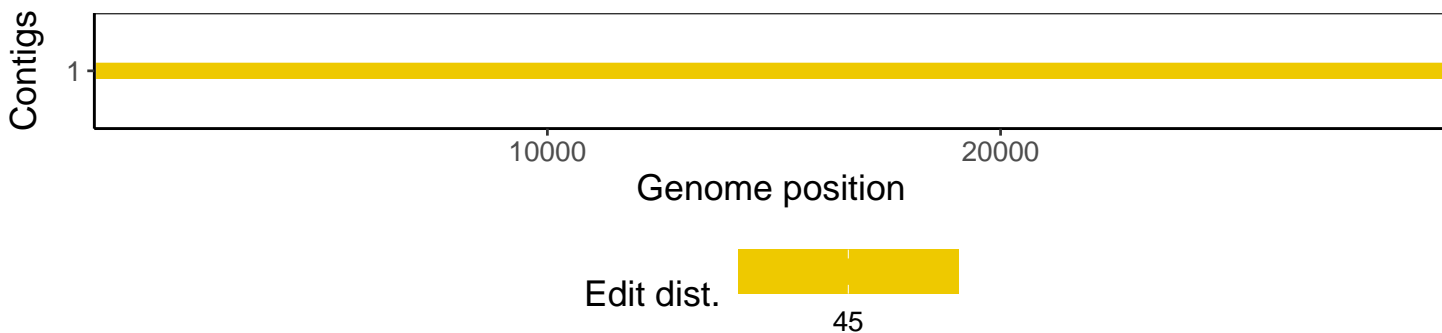
Software environment

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