COVID-19 subject UPHS-0040

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.
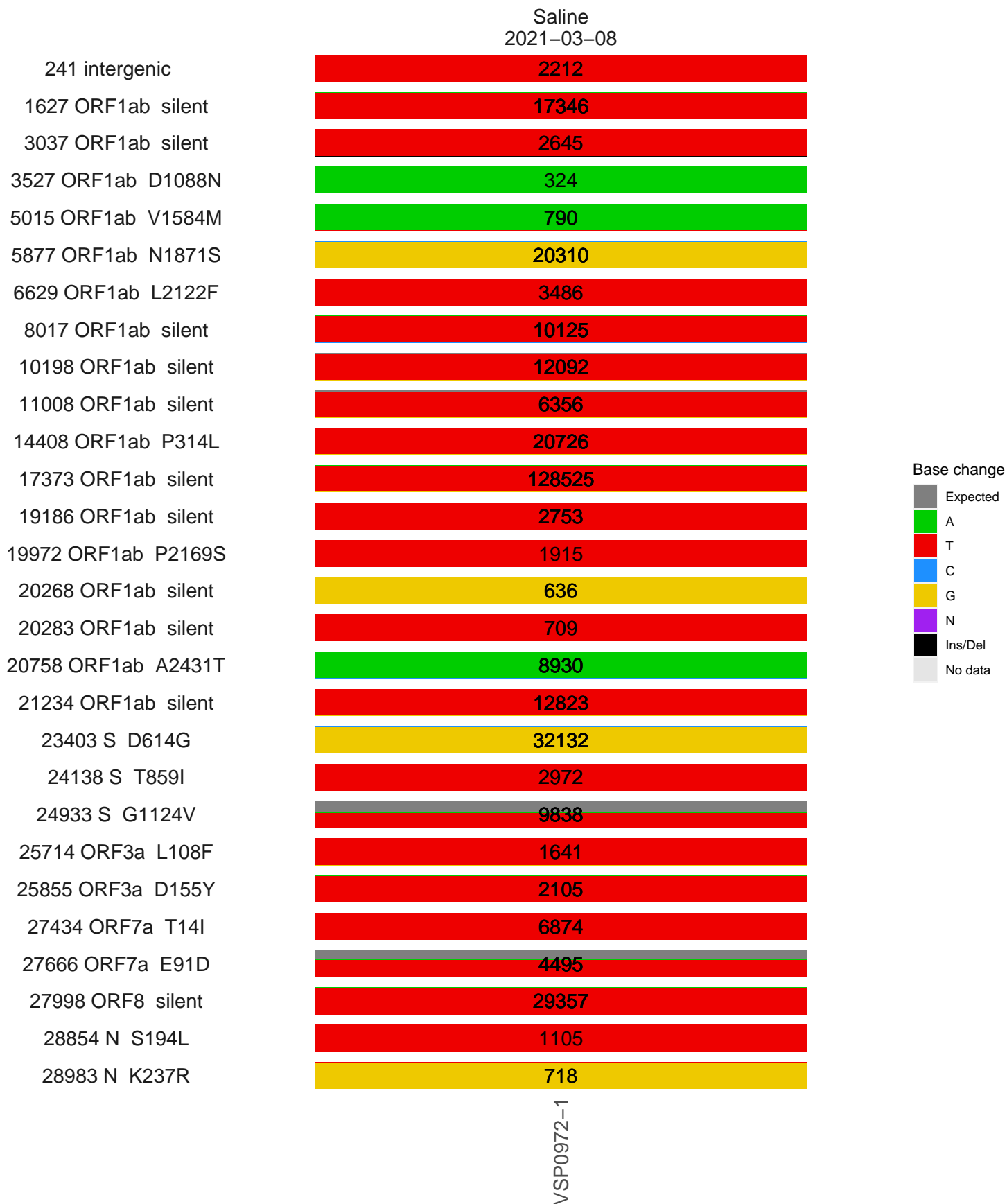
Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP0972-1	single experiment	NA	Saline	2021-03-08	29.86	B.1.409	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	Count
241 intergenic	2212
1627 ORF1ab silent	17346
3037 ORF1ab silent	2645
3527 ORF1ab D1088N	324
5015 ORF1ab V1584M	790
5877 ORF1ab N1871S	20310
6629 ORF1ab L2122F	3486
8017 ORF1ab silent	10125
10198 ORF1ab silent	12092
11008 ORF1ab silent	6356
14408 ORF1ab P314L	20726
17373 ORF1ab silent	128525
19186 ORF1ab silent	2753
19972 ORF1ab P2169S	1915
20268 ORF1ab silent	636
20283 ORF1ab silent	709
20758 ORF1ab A2431T	8930
21234 ORF1ab silent	12823
23403 S D614G	32132
24138 S T859I	2972
24933 S G1124V	9838
25714 ORF3a L108F	1641
25855 ORF3a D155Y	2105
27434 ORF7a T14I	6874
27666 ORF7a E91D	4495
27998 ORF8 silent	29357
28854 N S194L	1105
28983 N K237R	718

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

VSP0972-1

3

Analyses of individual experiments and composite results

**VSP0972-1 | 2021-03-08 | Saline | UPHS-0040 | 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.
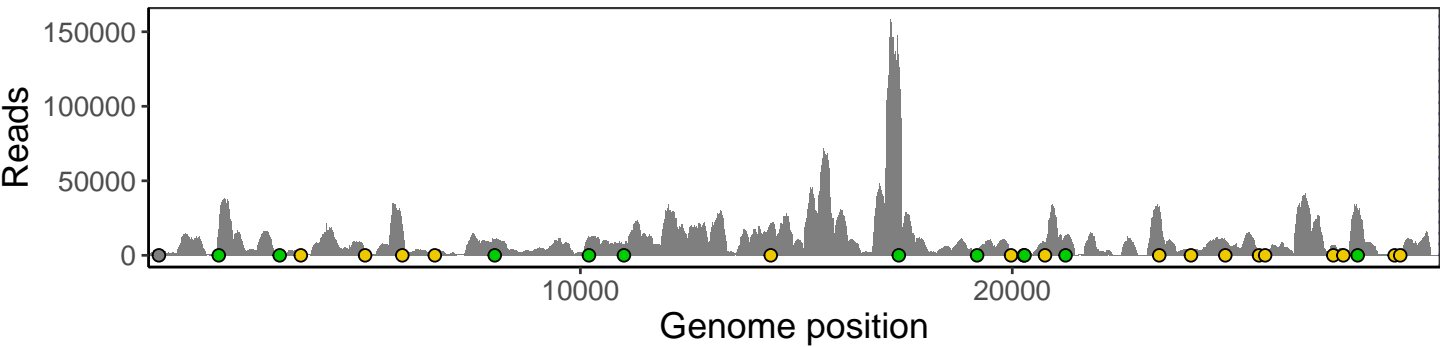
Excerpt from plot above focusing on reads coverage from 0 to 50 NT.
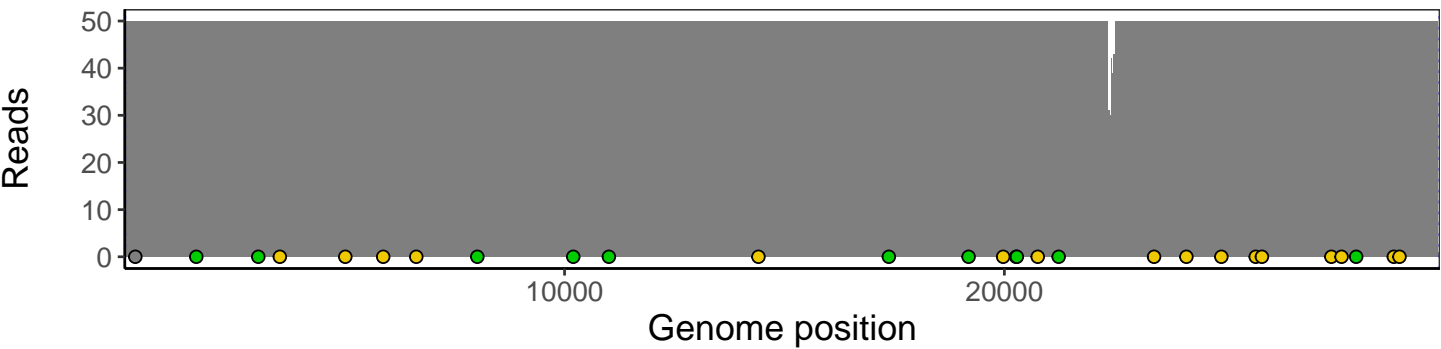
The longest five assembled contigs are shown below colored by their edit distance to the reference genome.
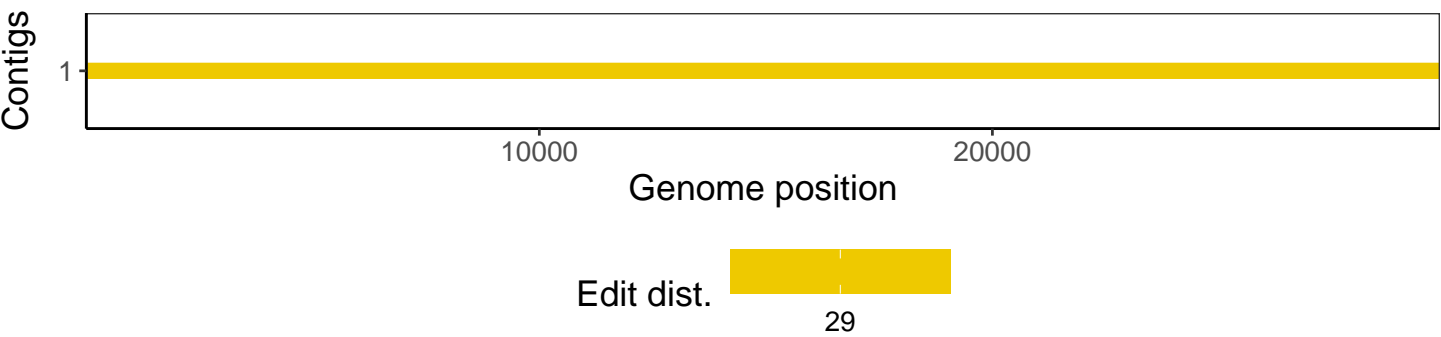
Software environment

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