COVID-19 subject HUP Q-0191

2021-05-05

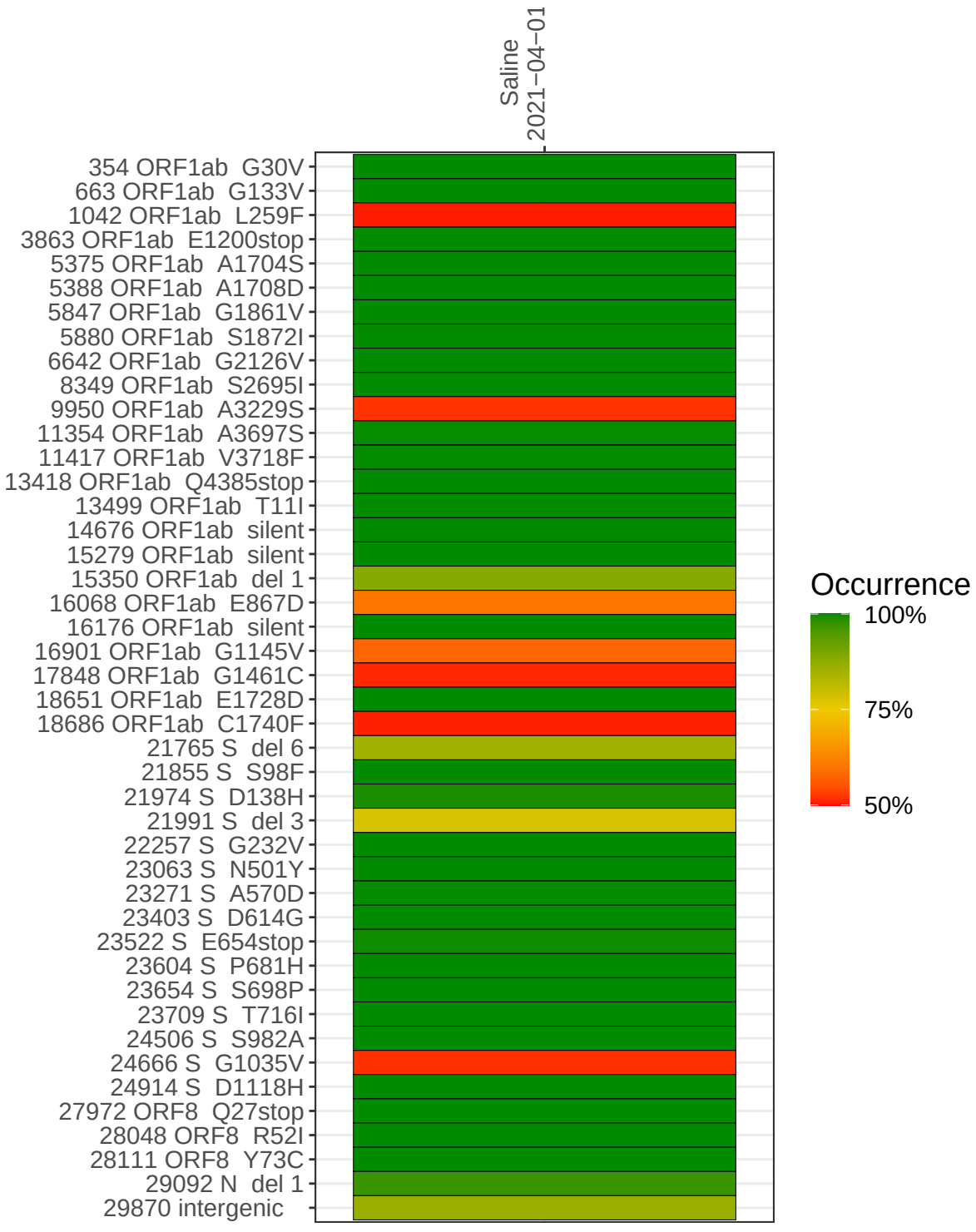
The table below provides a summary of subject samples for which sequencing data is available.
The experiments column shows the number of sequencing experiments performed for each specimen.
Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software tool (Rambaut et al 2020) for genomes with > 90% sequence coverage.

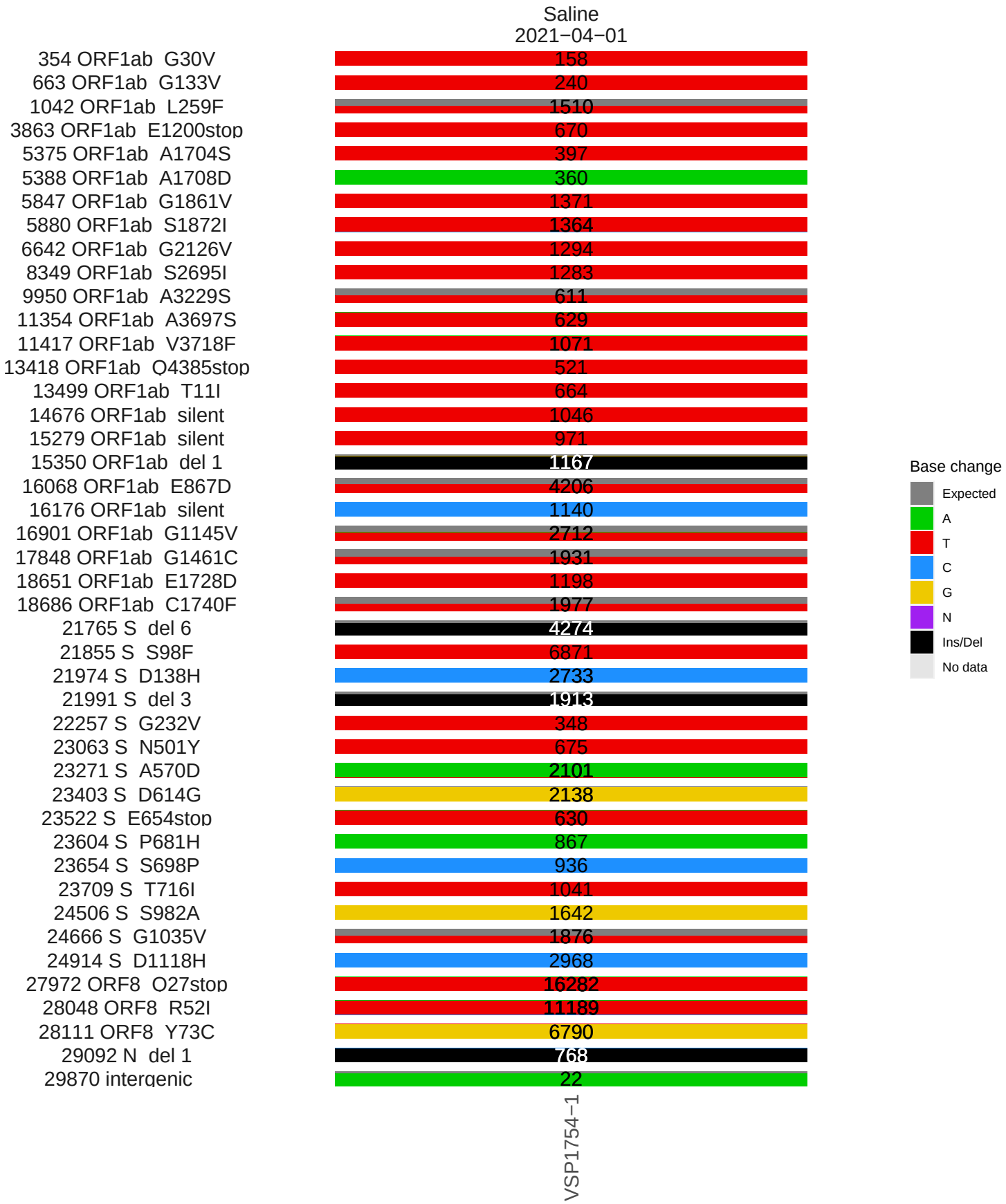
Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP1754-1	single experiment	NA	Saline	2021-04-01	4.15	NA	65.8%	65.2%

Variants shared across samples

The heat map below shows how variants (reference genome /home/everett/projects/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.

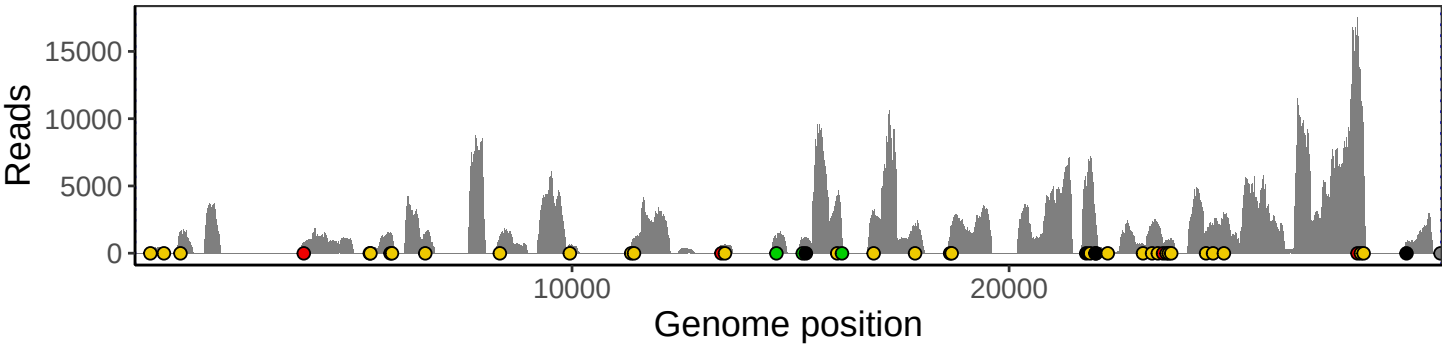



Saline
2021−04−01
354 ORF1ab G30V
663 ORF1ab G133V
1042 ORF1ab L259F
3863 ORF1ab E1200stop
5375 ORF1ab A1704S
5388 ORF1ab A1708D
5847 ORF1ab G1861V
5880 ORF1ab S1872I
6642 ORF1ab G2126V
8349 ORF1ab S2695I
9950 ORF1ab A3229S
11354 ORF1ab A3697S
11417 ORF1ab V3718F
13418 ORF1ab Q4385stop
13499 ORF1ab T11I
14676 ORF1ab silent
15279 ORF1ab silent
15350 ORF1ab del 1
16068 ORF1ab E867D
16176 ORF1ab silent
16901 ORF1ab G1145V
17848 ORF1ab G1461C
18651 ORF1ab E1728D
18686 ORF1ab C1740F
21765 S del 6
21855 S S98F
21974 S D138H
21991 S del 3
22257 S G232V
23063 S N501Y
23271 S A570D
23403 S D614G
23522 S E654stop
23604 S P681H
23654 S S698P
23709 S T716I
24506 S S982A
24666 S G1035V
24914 S D1118H
27972 ORF8 Q27stop
28048 ORF8 R52I
28111 ORF8 Y73C
29092 N del 1
29870 intergenic
158
240
1510
670
397
360
1371
1364
1294
1283
611
629
1071
521
664
1046
971
1167
4206
1140
2712
1931
1198
1977
4274
6871
2733
1913
348
675
2101
2138
630
867
936
1041
1642
1876
2968
16282
11189
6790
768
22
VSP1754−1
Base change
Expected
A
T
C
G
N
Ins/Del
No data

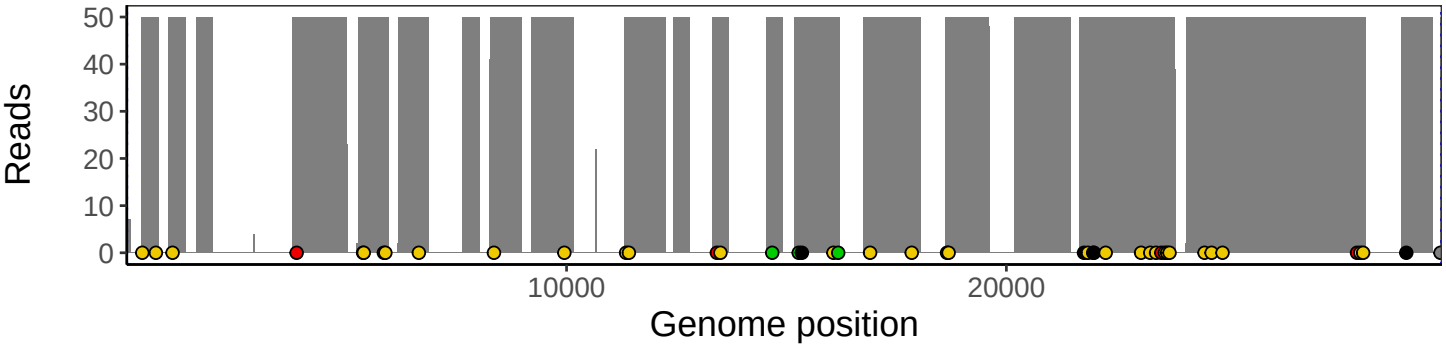
Analyses of individual experiments and composite results

VSP1754-1 | 2021-04-01 | Saline | HUP Q-0191 | genomes | single experiment

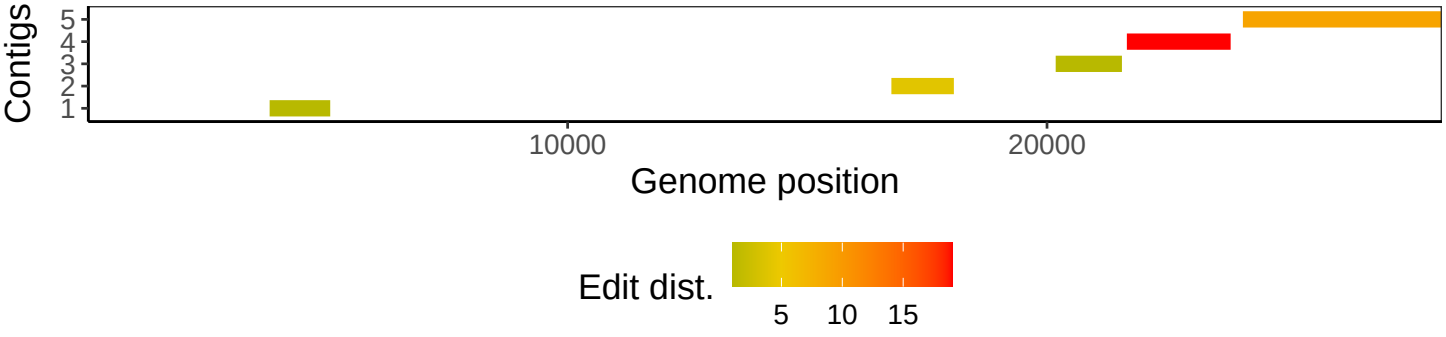
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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	2.3.8
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