COVID-19 subject HUP Q-0194

2021-05-05

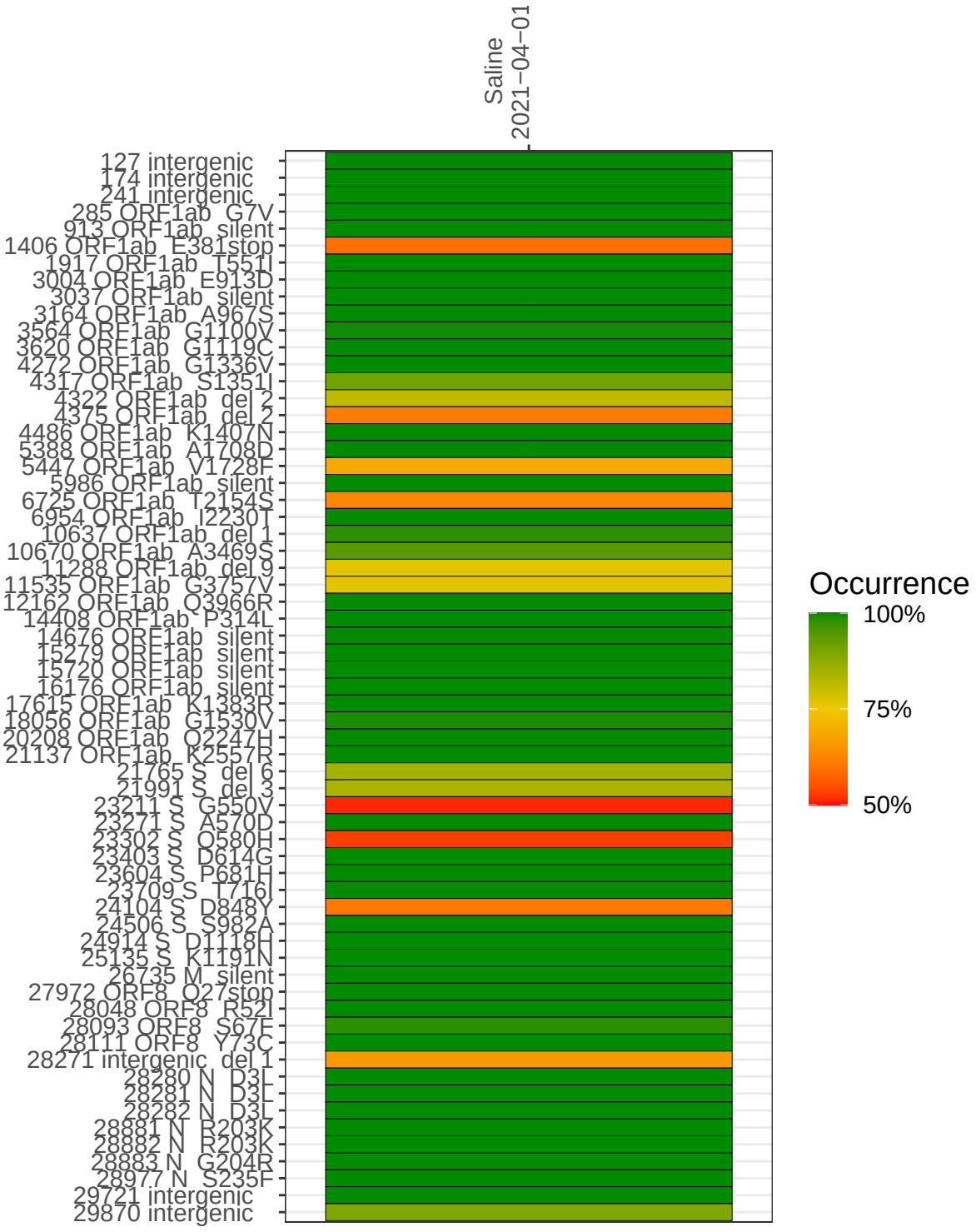
The table below provides a summary of subject samples for which sequencing data is available.
The experiments column shows the number of sequencing experiments performed for each specimen.
Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software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

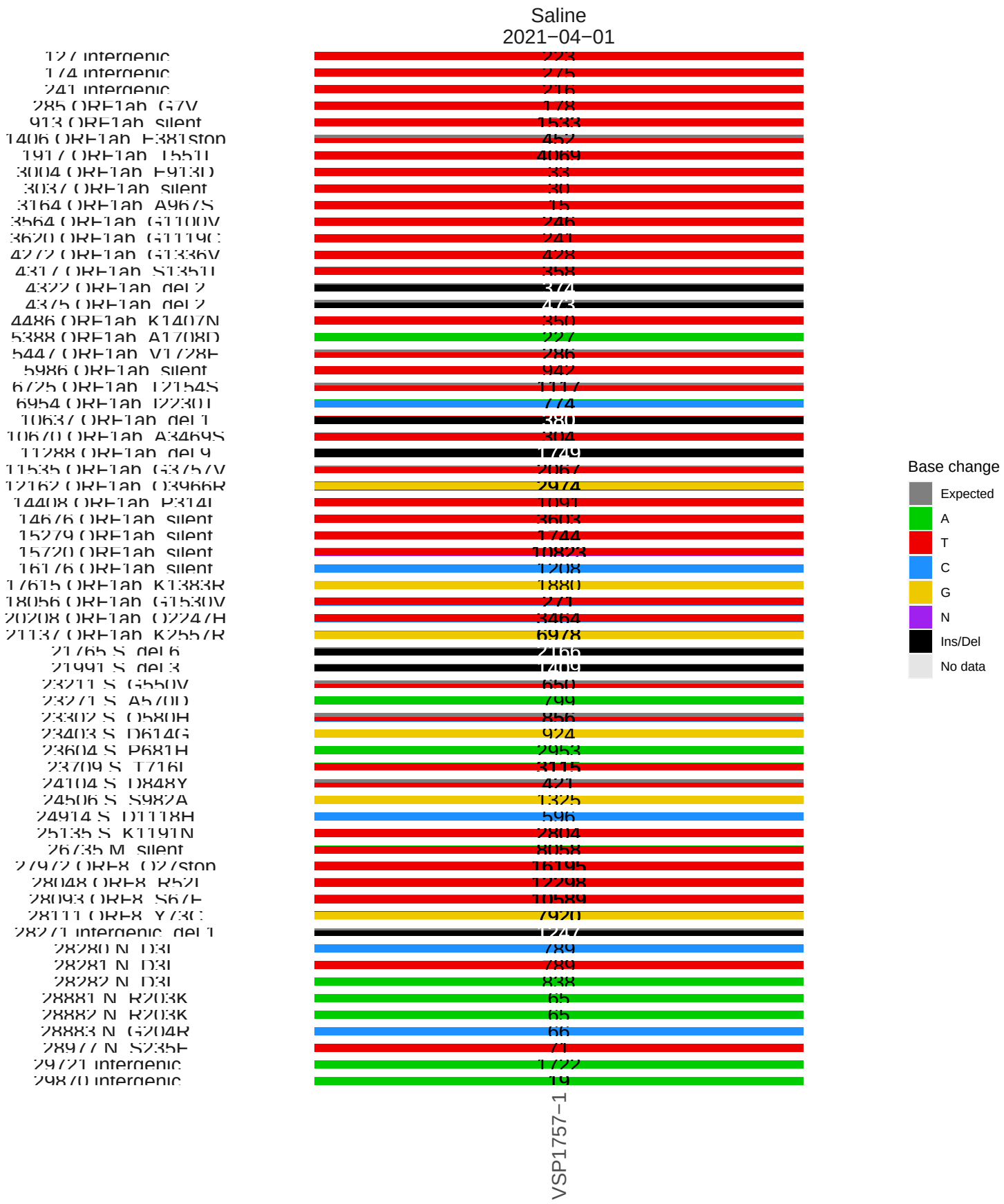
Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP1757-1	single experiment	NA	Saline	2021-04-01	9.14	B.1.1.7	95.1%	95.1%

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.

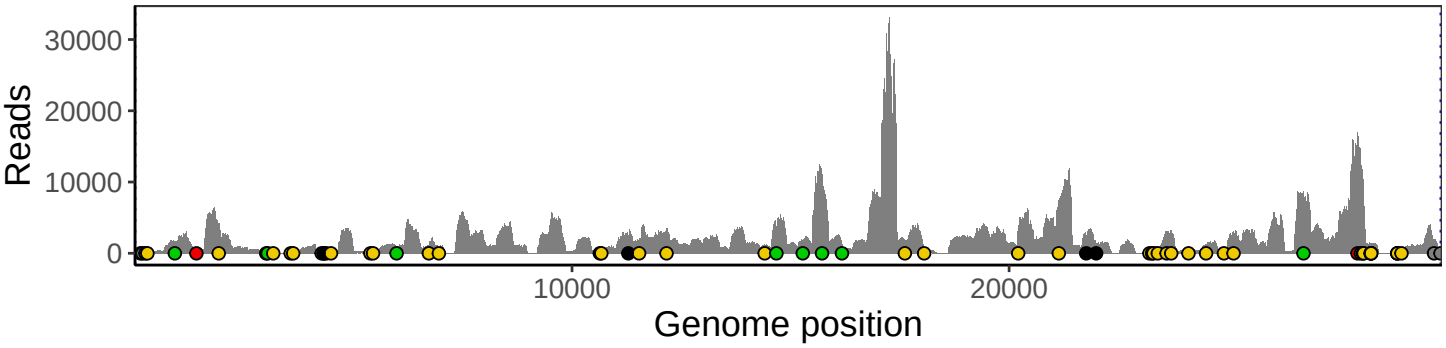



Saline
2021−04−01
127 intergenic
174 intergenic
241 intergenic
285 ORF1ab_G7V
913 ORF1ab_silent
1406 ORF1ab_E381stop
1917 ORF1ab_I551L
3004 ORF1ab_F913D
3037 ORF1ab_silent
3164 ORF1ab_A967S
3564 ORF1ab_G1100V
3620 ORF1ab_G1119C
4272 ORF1ab_G1336V
4317 ORF1ab_S1351L
4322 ORF1ab_del 2
4375 ORF1ab_del 2
4486 ORF1ab_K1407N
5388 ORF1ab_A1708D
5447 ORF1ab_V1728F
5986 ORF1ab_silent
6725 ORF1ab_I2154S
6954 ORF1ab_I2230L
10637 ORF1ab_del 1
10670 ORF1ab_A3469S
11288 ORF1ab_del 9
11535 ORF1ab_G3757V
12162 ORF1ab_G3966R
14408 ORF1ab_P4714L
14676 ORF1ab_silent
15279 ORF1ab_silent
15720 ORF1ab_silent
16176 ORF1ab_silent
17615 ORF1ab_K5383R
18056 ORF1ab_G5530V
20208 ORF1ab_D6247H
21137 ORF1ab_K6557R
21765 S_del 6
21991 S_del 3
23211 S_G550V
23271 S_A570D
23302 S_Q580H
23403 S_D614G
23604 S_P681H
23709 S_T716I
24104 S_D848Y
24506 S_S982A
24914 S_D1118H
25135 S_K1191N
26735 M_silent
27972 ORF8_Q27stop
28048 ORF8_R52I
28093 ORF8_S67F
28111 ORF8_Y73C
28271 intergenic_del 1
28280 N_D3L
28281 N_D3L
28282 N_D3L
28881 N_R203K
28882 N_R203K
28883 N_G204R
28977 N_S235F
29721 intergenic
29870 intergenic
223
275
215
178
1533
452
40869
33
30
15
246
241
428
358
374
473
350
227
286
942
1117
774
380
304
1749
2067
2974
1091
3603
1744
10823
1208
1880
271
3464
6978
2166
1406
650
799
856
924
2953
3115
421
1325
596
2804
8058
16195
12298
10589
7920
1247
789
789
838
65
65
66
71
1722
19
Base change
Expected
A
T
C
G
N
Ins/Del
No data
VSP1757−1

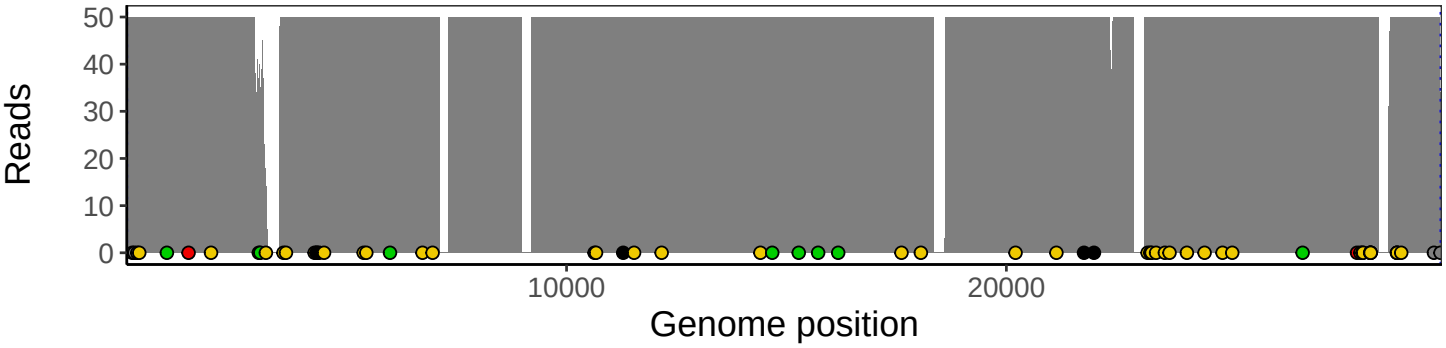
Analyses of individual experiments and composite results

VSP1757-1 | 2021-04-01 | Saline | HUP Q-0194 | genomes | single experiment

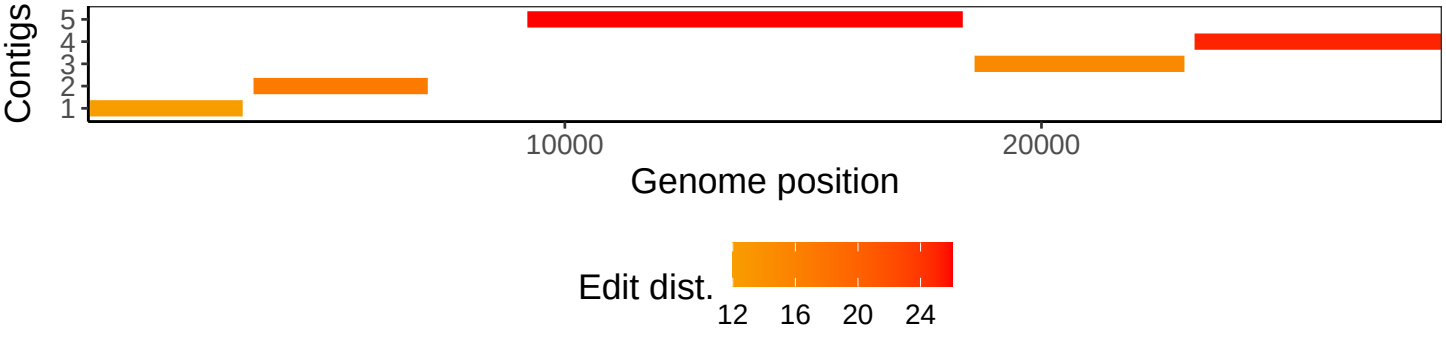
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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