COVID-19 subject HUP Q-0045

2021-05-05

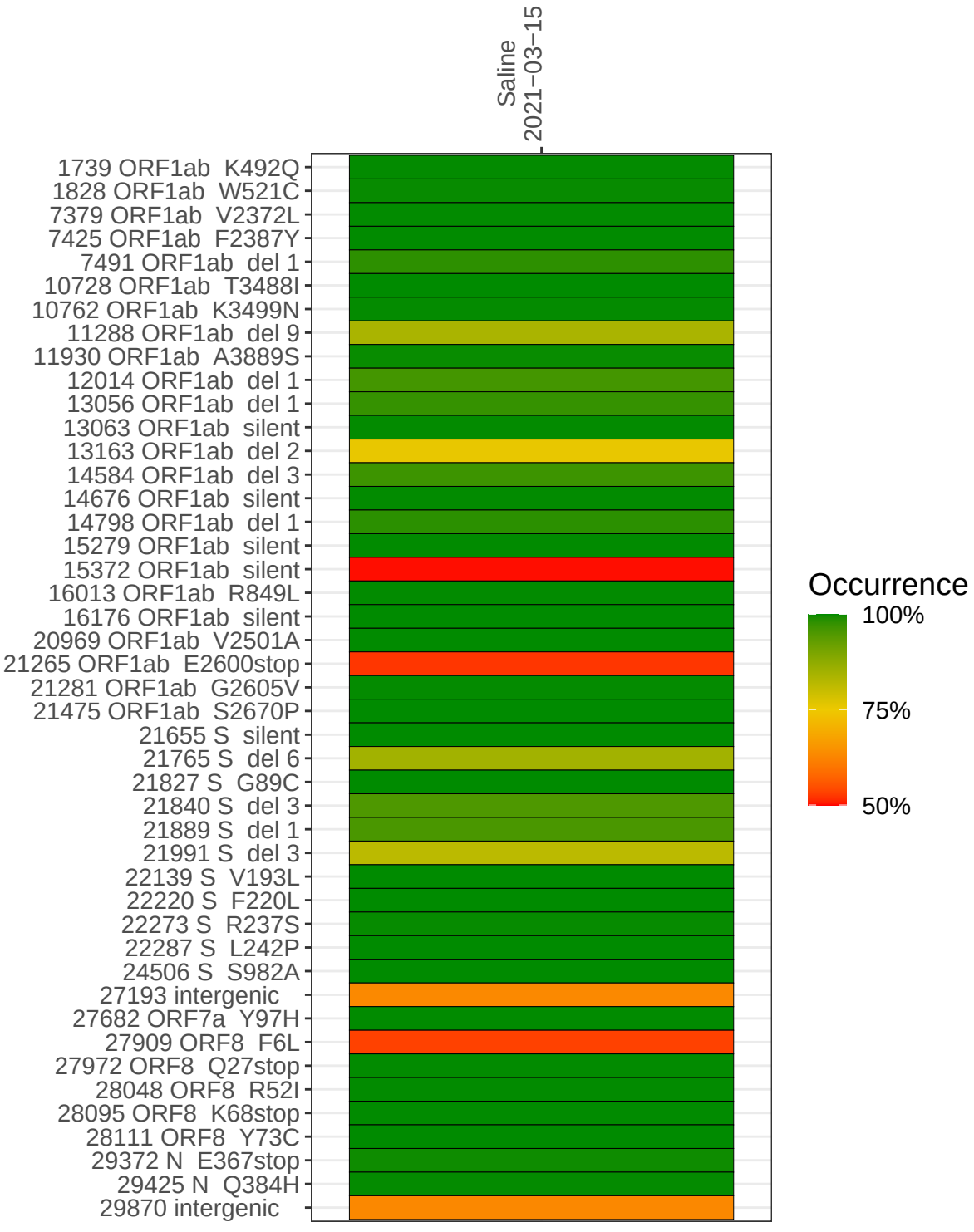
The table below provides a summary of subject samples for which sequencing data is available.
The experiments column shows the number of sequencing experiments performed for each specimen.
Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software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

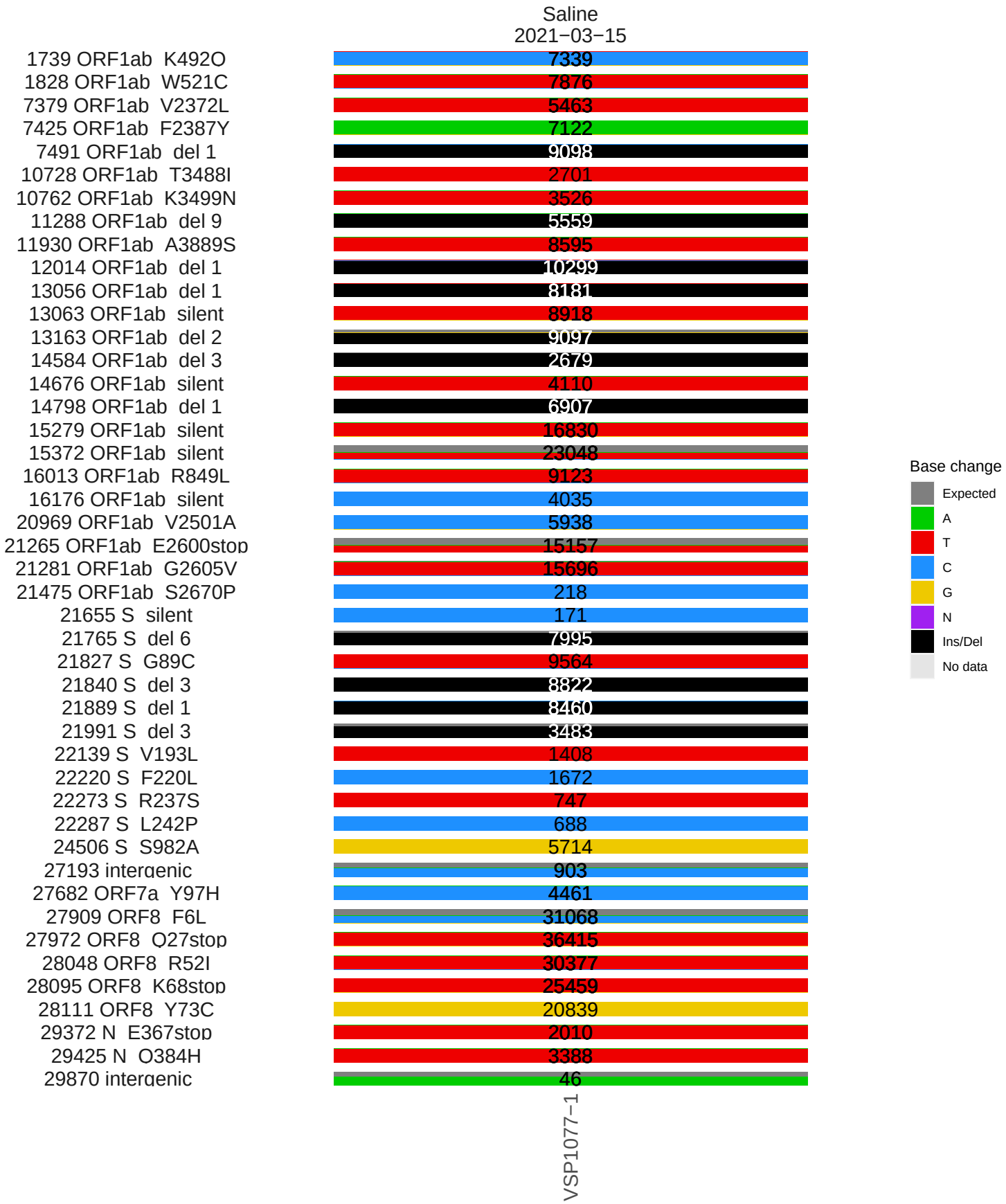
Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP1077-1	single experiment	NA	Saline	2021-03-15	1.53	NA	27.5%	25.6%

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.

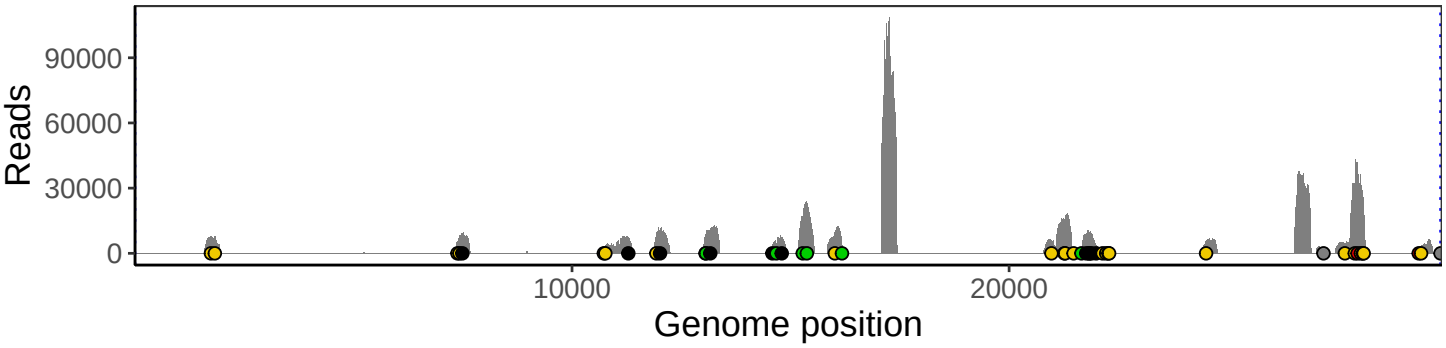



Saline
2021−03−15
1739 ORF1ab K492O
1828 ORF1ab W521C
7379 ORF1ab V2372L
7425 ORF1ab F2387Y
7491 ORF1ab del 1
10728 ORF1ab T3488I
10762 ORF1ab K3499N
11288 ORF1ab del 9
11930 ORF1ab A3889S
12014 ORF1ab del 1
13056 ORF1ab del 1
13063 ORF1ab silent
13163 ORF1ab del 2
14584 ORF1ab del 3
14676 ORF1ab silent
14798 ORF1ab del 1
15279 ORF1ab silent
15372 ORF1ab silent
16013 ORF1ab R849L
16176 ORF1ab silent
20969 ORF1ab V2501A
21265 ORF1ab E2600stop
21281 ORF1ab G2605V
21475 ORF1ab S2670P
21655 S silent
21765 S del 6
21827 S G89C
21840 S del 3
21889 S del 1
21991 S del 3
22139 S V193L
22220 S F220L
22273 S R237S
22287 S L242P
24506 S S982A
27193 intergenic
27682 ORF7a Y97H
27909 ORF8 F6L
27972 ORF8 Q27stop
28048 ORF8 R52I
28095 ORF8 K68stop
28111 ORF8 Y73C
29372 N E367stop
29425 N Q384H
29870 intergenic
7339
7876
5463
7122
9098
2701
3526
5559
8595
10299
8181
8918
9097
2679
4110
6907
16830
23048
9123
4035
5938
15157
15696
218
171
7995
9564
8822
8460
3483
1408
1672
747
688
5714
903
4461
31068
36415
30377
25459
20839
2010
3388
46
VSP1077−1
Base change
Expected
A
T
C
G
N
Ins/Del
No data

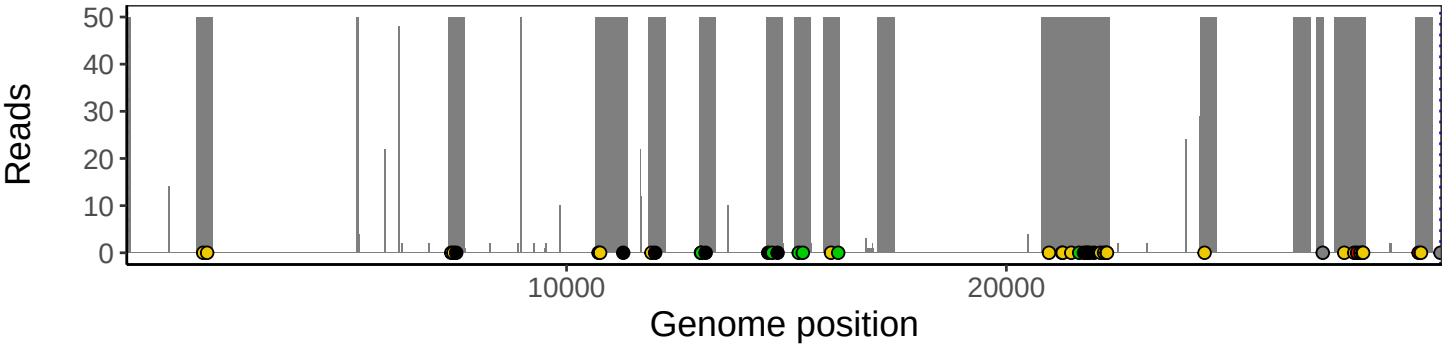
Analyses of individual experiments and composite results

VSP1077-1 | 2021-03-15 | Saline | HUP Q-0045 | genomes | single experiment

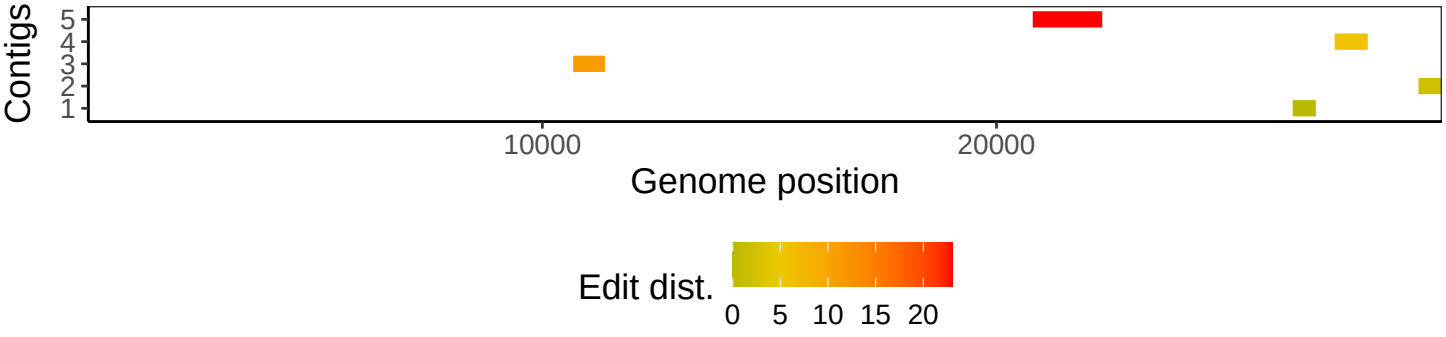
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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	3.0.0
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