COVID-19 subject HUP Q-0047

2021-06-23

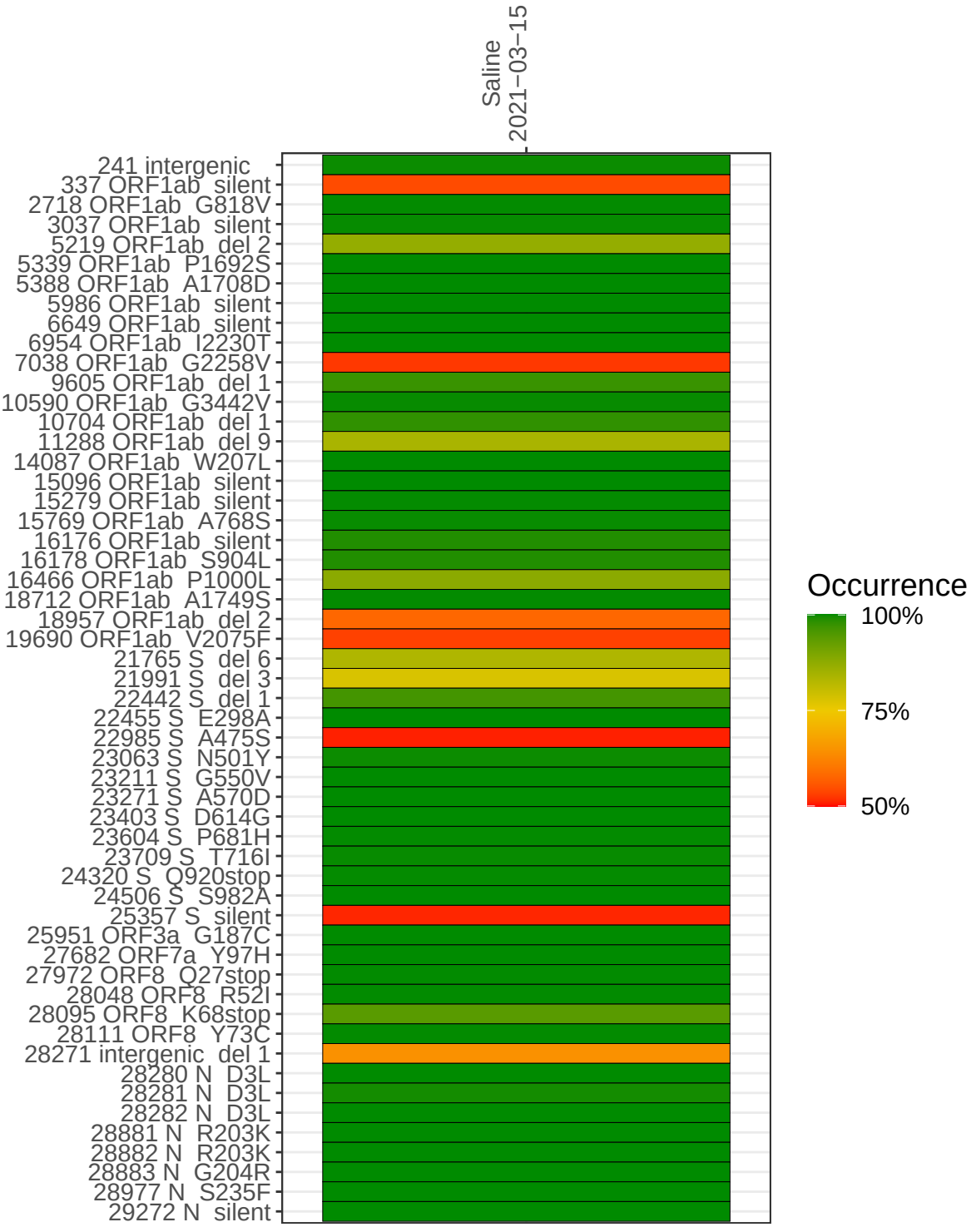
The table below provides a summary of subject samples for which sequencing data is available.
The experiments column shows the number of sequencing experiments performed for each specimen.
Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software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

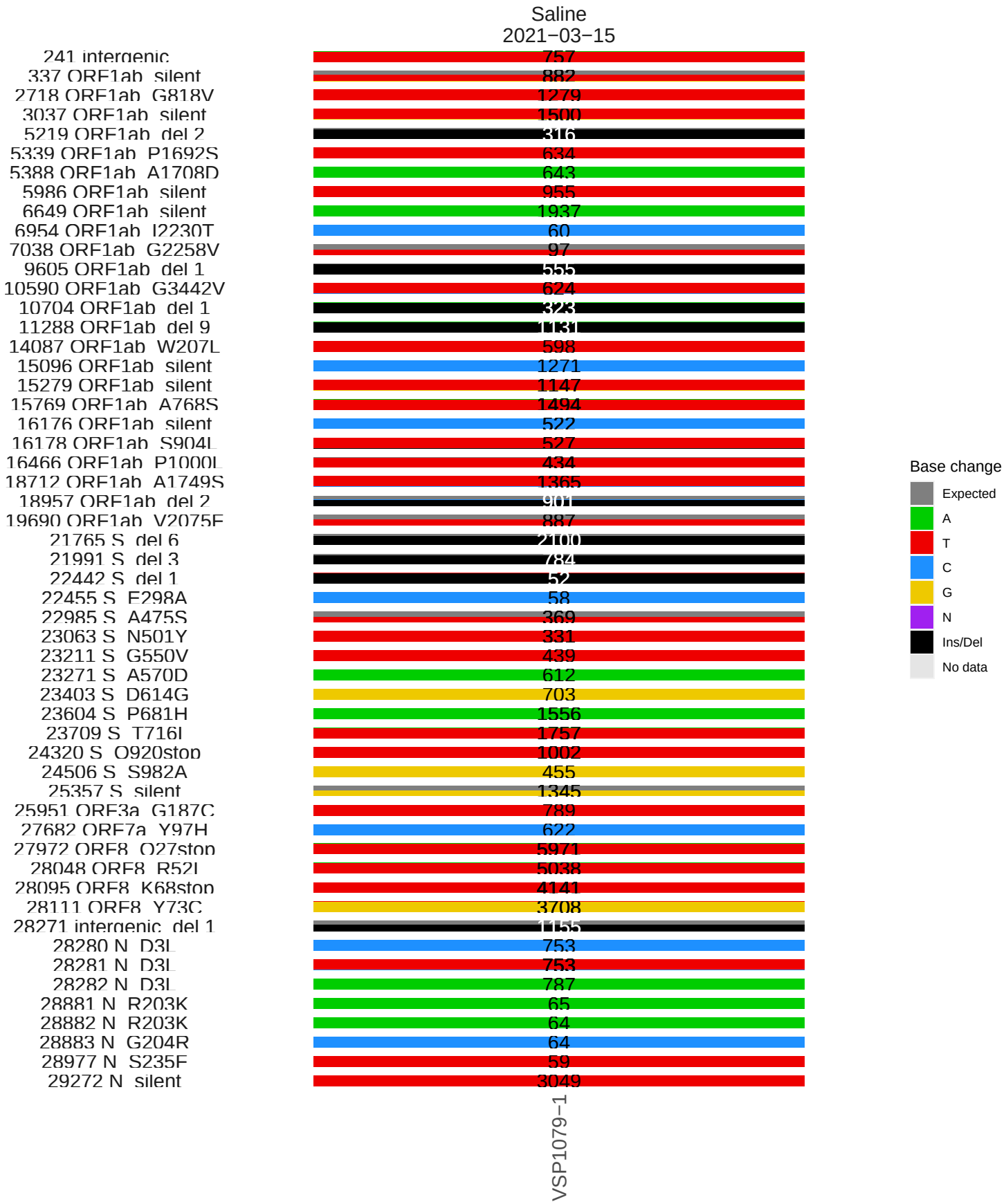
Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP1079-1	single experiment	NA	Saline	2021-03-15	3.11	NA	70.4%	70.0%

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.

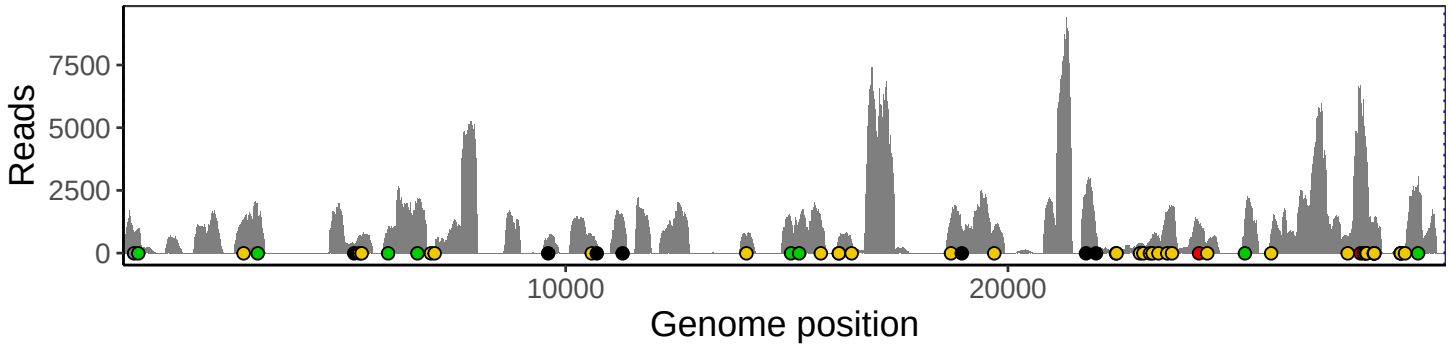



Saline
2021−03−15
241 intergenic
337 ORF1ab silent
2718 ORF1ab G818V
3037 ORF1ab silent
5219 ORF1ab del 2
5339 ORF1ab P1692S
5388 ORF1ab A1708D
5986 ORF1ab silent
6649 ORF1ab silent
6954 ORF1ab I2230T
7038 ORF1ab G2258V
9605 ORF1ab del 1
10590 ORF1ab G3442V
10704 ORF1ab del 1
11288 ORF1ab del 9
14087 ORF1ab W207L
15096 ORF1ab silent
15279 ORF1ab silent
15769 ORF1ab A768S
16176 ORF1ab silent
16178 ORF1ab S904L
16466 ORF1ab P1000L
18712 ORF1ab A1749S
18957 ORF1ab del 2
19690 ORF1ab V2075F
21765 S del 6
21991 S del 3
22442 S del 1
22455 S F298A
22985 S A475S
23063 S N501Y
23211 S G550V
23271 S A570D
23403 S D614G
23604 S P681H
23709 S T716I
24320 S Q920stop
24506 S S982A
25357 S silent
25951 ORF3a G187C
27682 ORF7a Y97H
27972 ORF8 Q27stop
28048 ORF8 R52I
28095 ORF8 K68stop
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
29272 N silent
757
882
1279
1500
316
634
643
955
1937
60
97
555
624
323
1131
598
1271
1147
1494
522
527
434
1365
901
887
2100
784
52
58
369
331
439
612
703
1556
1757
1002
455
1345
789
622
5971
5038
4141
3708
1155
753
753
787
65
64
64
59
3049
VSP1079−1
Base change
Expected
A
T
C
G
N
Ins/Del
No data

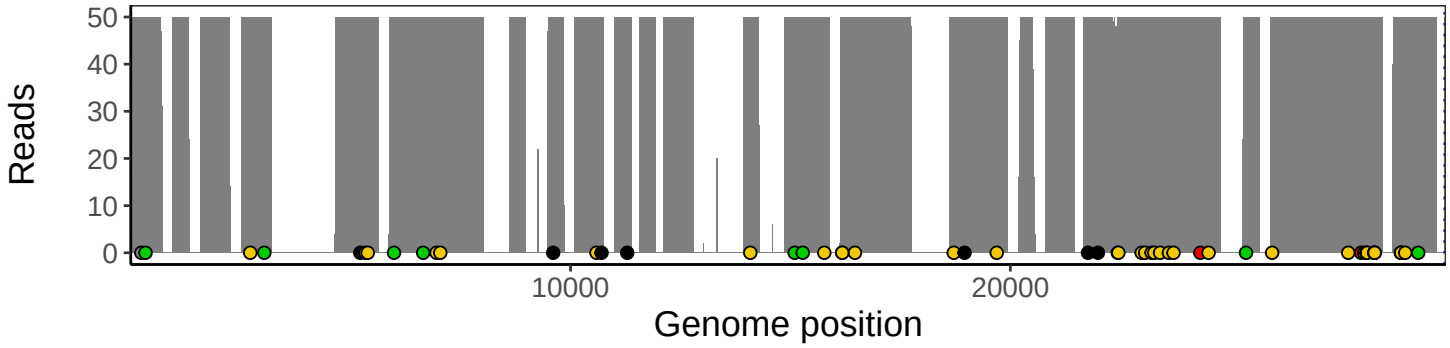
Analyses of individual experiments and composite results

VSP1079-1 | 2021-03-15 | Saline | HUP Q-0047 | genomes | single experiment

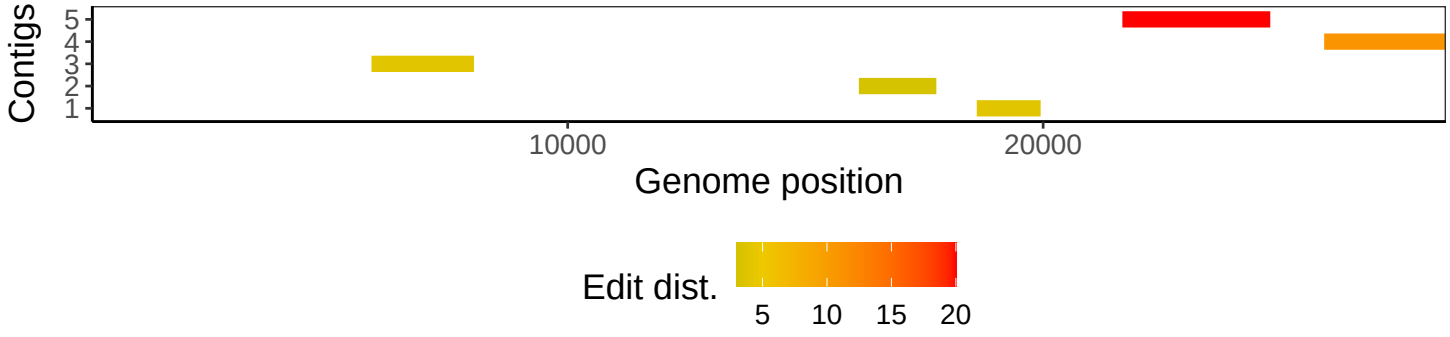
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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