COVID-19 subject UPHS-0153

2021-04-17

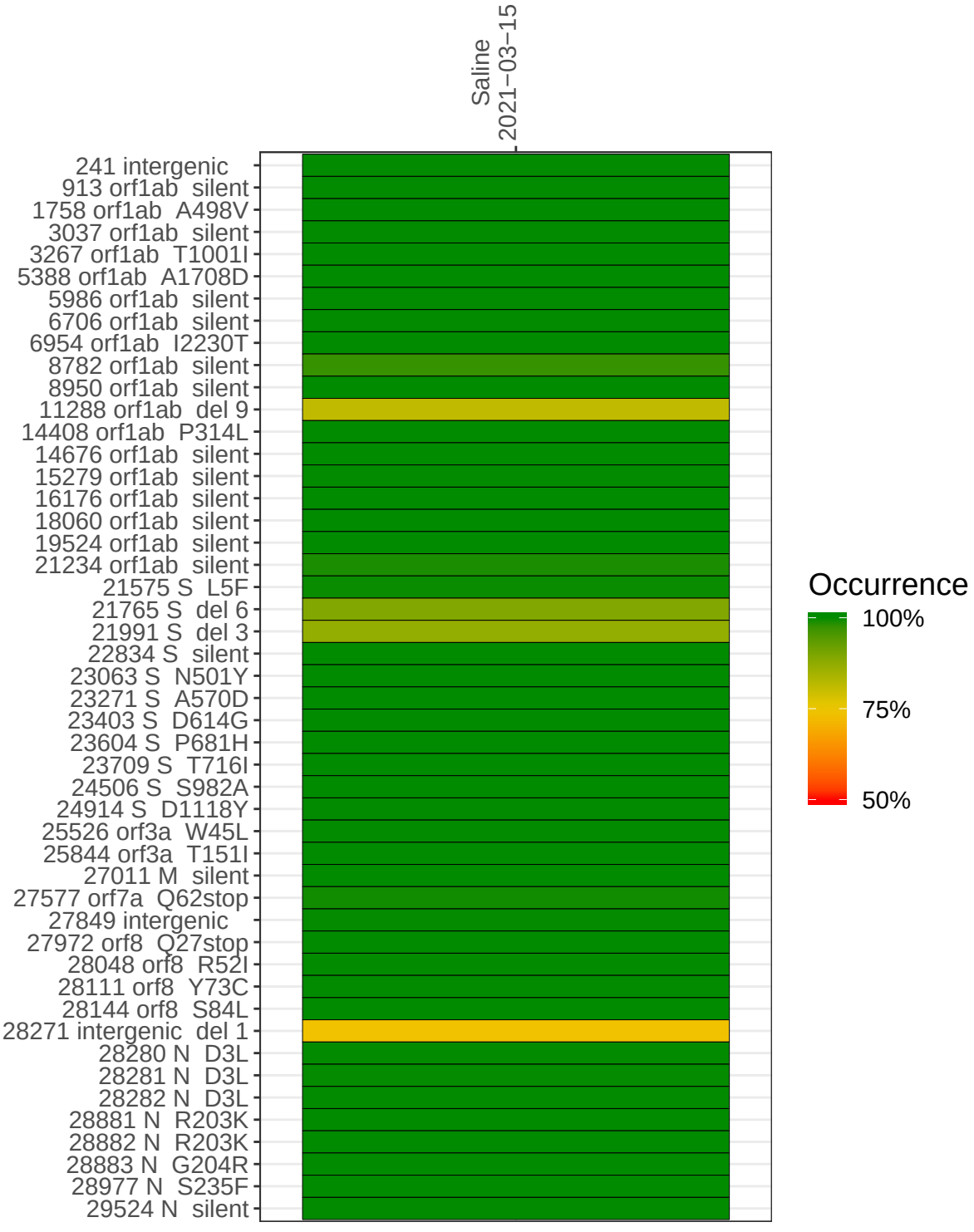
The table below provides a summary of subject samples for which sequencing data is available.
The experiments column shows the number of sequencing experiments performed for each specimen.
Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software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

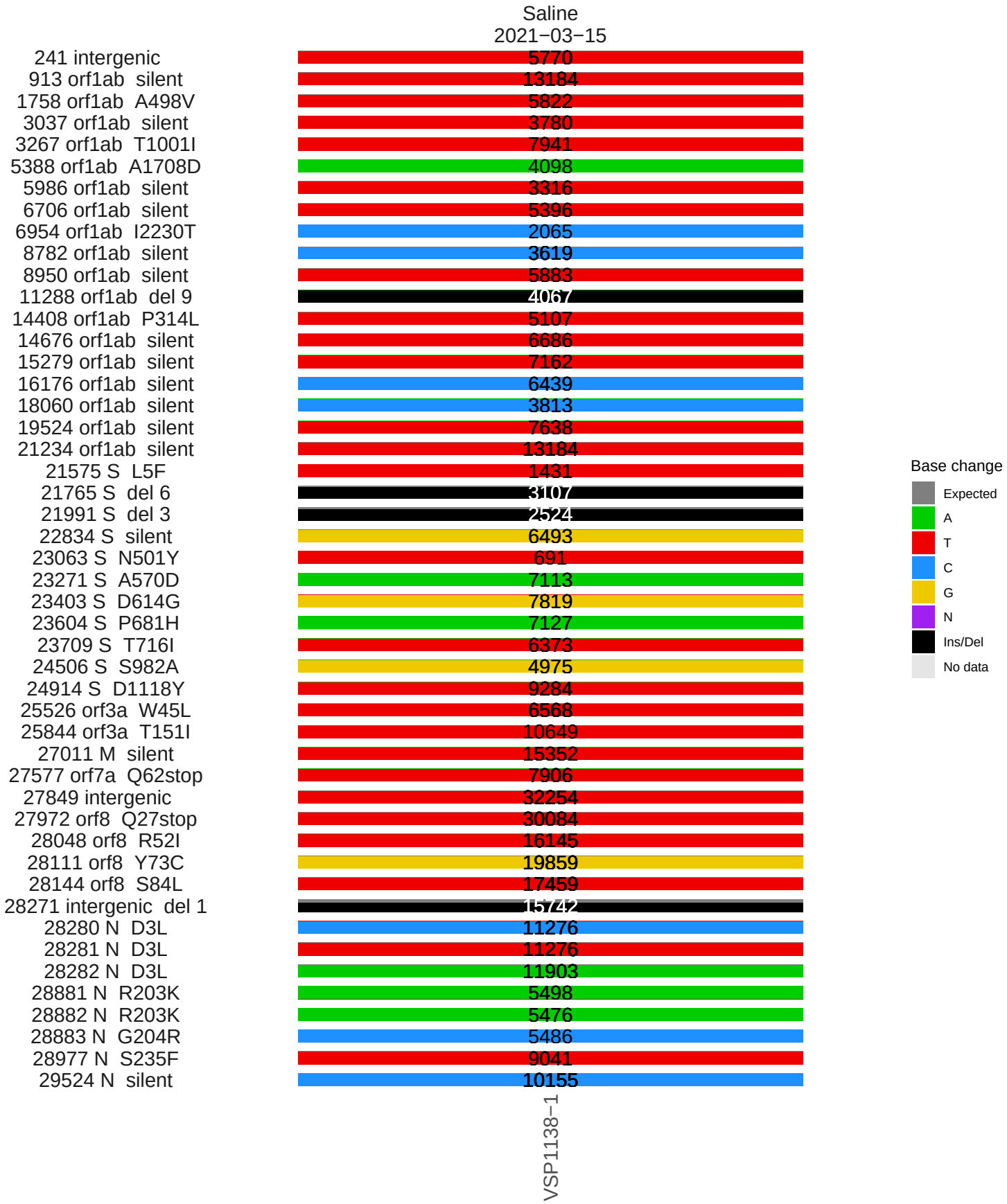
Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP1138-1	single experiment	NA	Saline	2021-03-15	29.82	B.1.1.7	99.8%	99.8%

Variants shared across samples

The heat map below shows how variants (reference genome /home/everett/projects/SARS-CoV-2-Philadelphia/USA-WA1-2020) 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.

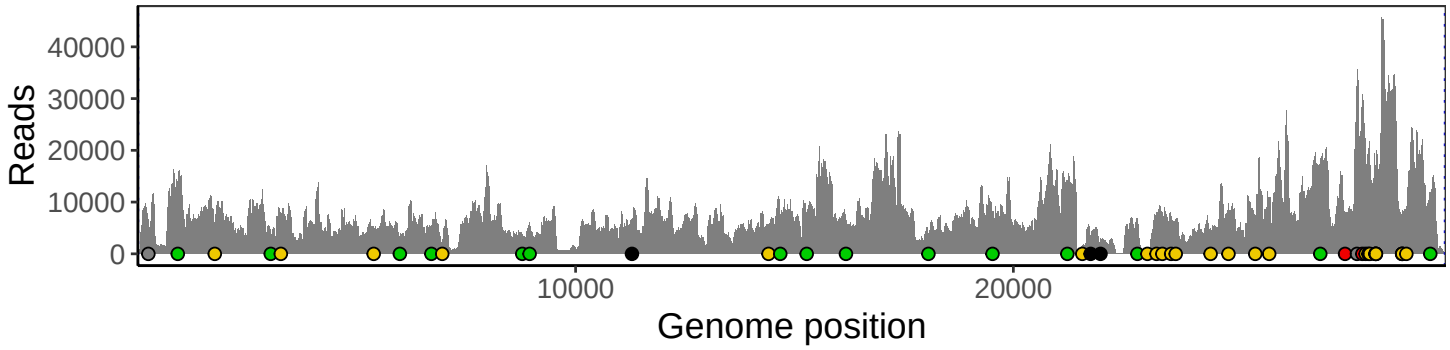



Saline
2021−03−15
241 intergenic 5770
913 orf1ab silent 13184
1758 orf1ab A498V 5822
3037 orf1ab silent 3780
3267 orf1ab T1001I 7941
5388 orf1ab A1708D 4098
5986 orf1ab silent 3316
6706 orf1ab silent 5396
6954 orf1ab I2230T 2065
8782 orf1ab silent 3619
8950 orf1ab silent 5883
11288 orf1ab del 9 4067
14408 orf1ab P314L 5107
14676 orf1ab silent 6686
15279 orf1ab silent 7162
16176 orf1ab silent 6439
18060 orf1ab silent 3813
19524 orf1ab silent 7638
21234 orf1ab silent 13184
21575 S L5F 1431
21765 S del 6 3107
21991 S del 3 2524
22834 S silent 6493
23063 S N501Y 691
23271 S A570D 7113
23403 S D614G 7819
23604 S P681H 7127
23709 S T716I 6373
24506 S S982A 4975
24914 S D1118Y 9284
25526 orf3a W45L 6568
25844 orf3a T151I 10649
27011 M silent 15352
27577 orf7a Q62stop 7906
27849 intergenic 32254
27972 orf8 Q27stop 30084
28048 orf8 R52I 16145
28111 orf8 Y73C 19859
28144 orf8 S84L 17459
28271 intergenic del 1 15742
28280 N D3L 11276
28281 N D3L 11276
28282 N D3L 11903
28881 N R203K 5498
28882 N R203K 5476
28883 N G204R 5486
28977 N S235F 9041
29524 N silent 10155
VSP1138−1
Base change
Expected
A
T
C
G
N
Ins/Del
No data

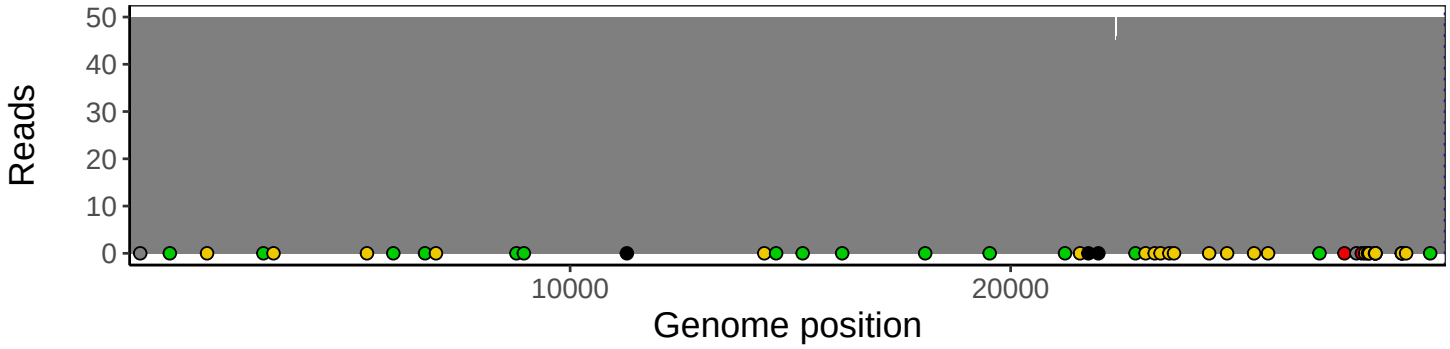
Analyses of individual experiments and composite results

VSP1138-1 | 2021-03-15 | Saline | UPHS-0153 | genomes | single experiment

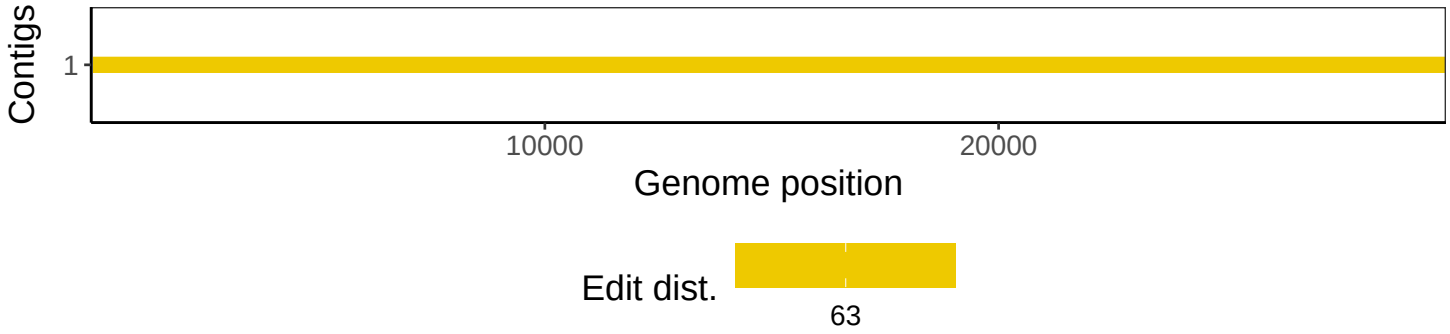
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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