COVID-19 subject UPHS-0834

2021-06-23

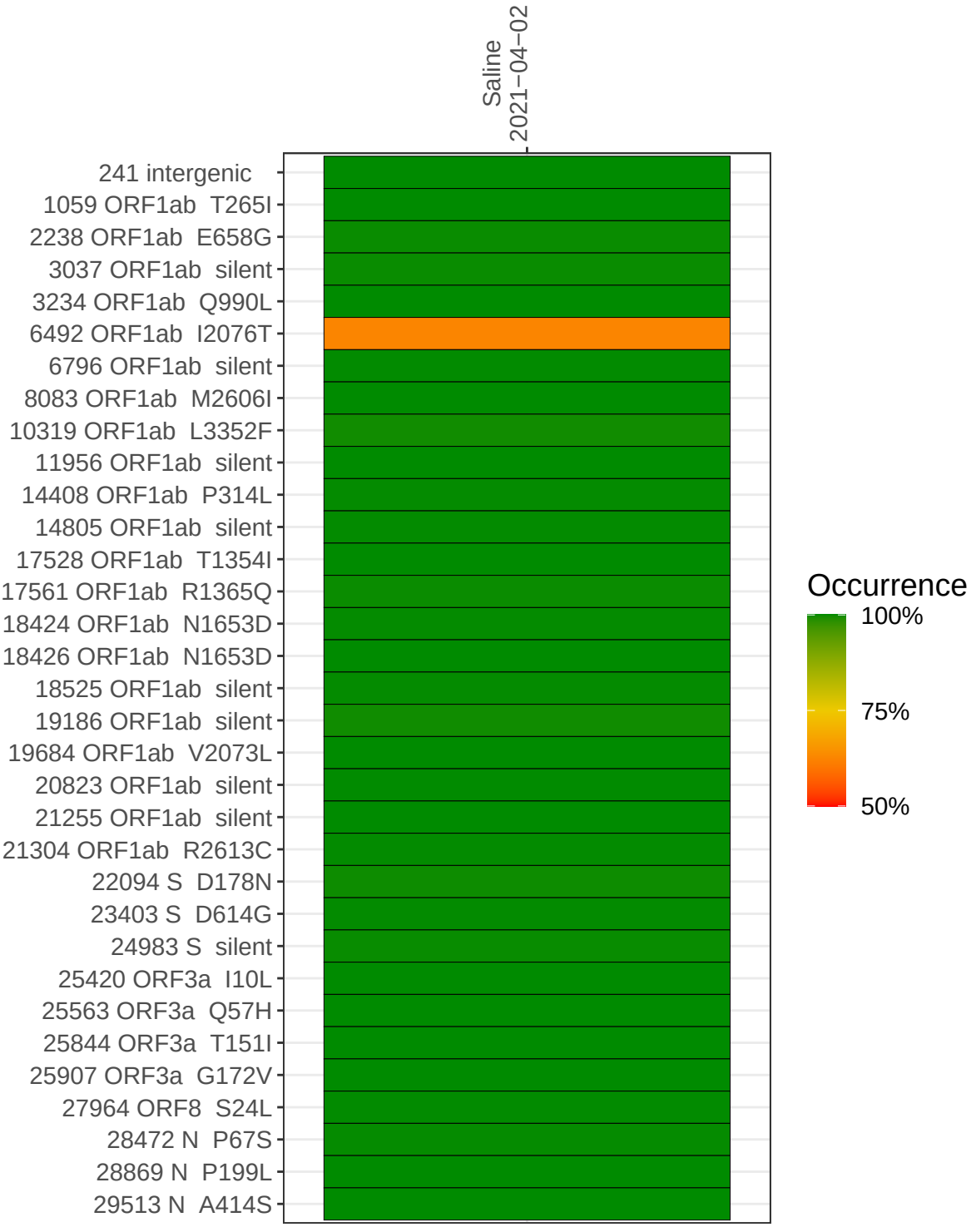
The table below provides a summary of subject samples for which sequencing data is available.
The experiments column shows the number of sequencing experiments performed for each specimen.
Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software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

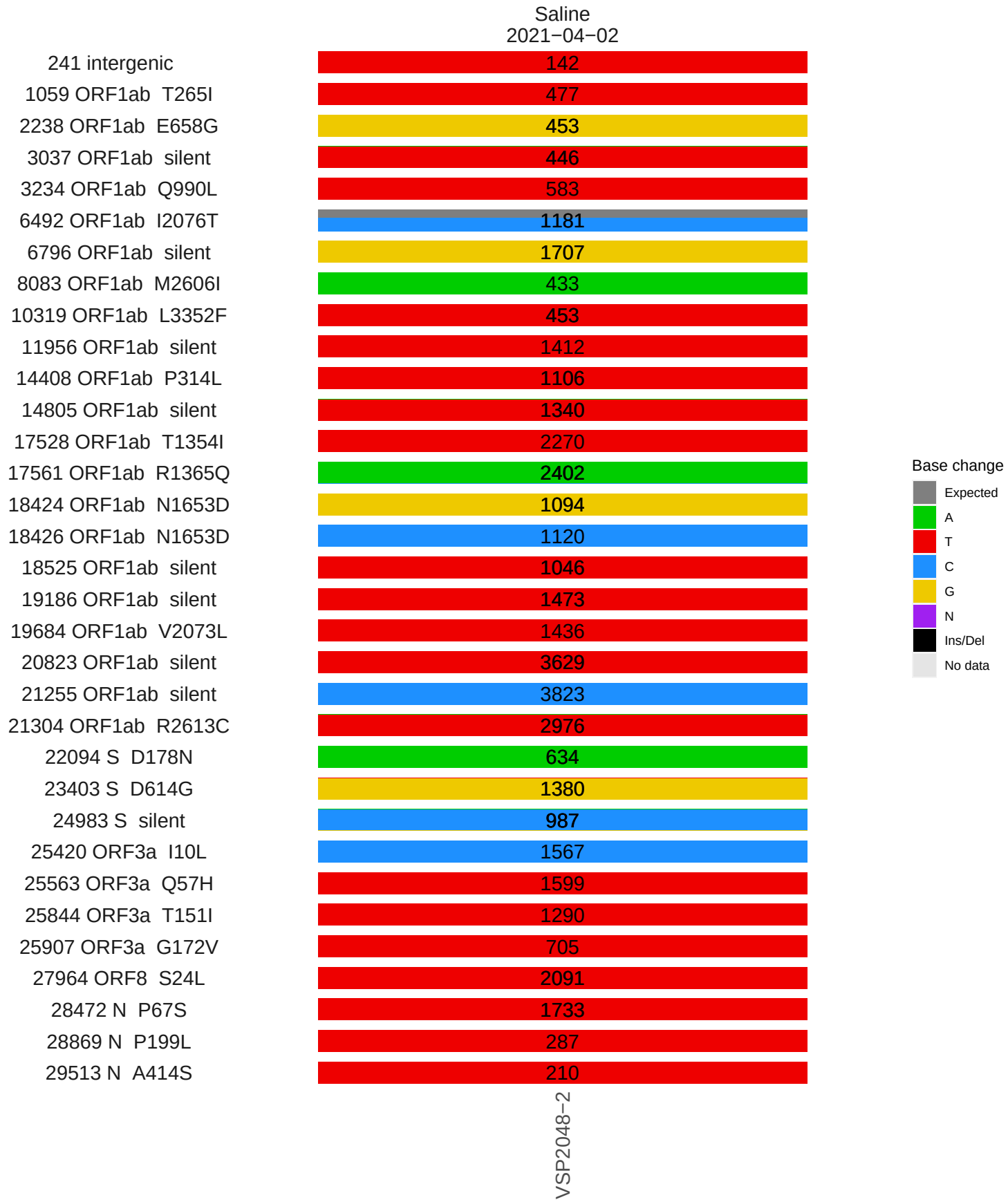
Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP2048-2	single experiment	NA	Saline	2021-04-02	29.82	B.1.2	99.8%	99.8%

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.

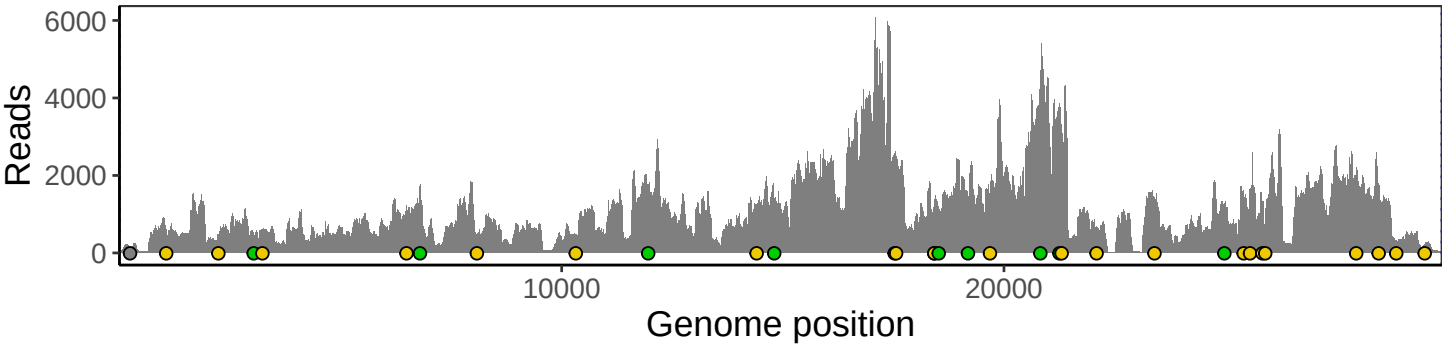



Saline
2021−04−02
241 intergenic
1059 ORF1ab T265I
2238 ORF1ab E658G
3037 ORF1ab silent
3234 ORF1ab Q990L
6492 ORF1ab I2076T
6796 ORF1ab silent
8083 ORF1ab M2606I
10319 ORF1ab L3352F
11956 ORF1ab silent
14408 ORF1ab P314L
14805 ORF1ab silent
17528 ORF1ab T1354I
17561 ORF1ab R1365Q
18424 ORF1ab N1653D
18426 ORF1ab N1653D
18525 ORF1ab silent
19186 ORF1ab silent
19684 ORF1ab V2073L
20823 ORF1ab silent
21255 ORF1ab silent
21304 ORF1ab R2613C
22094 S D178N
23403 S D614G
24983 S silent
25420 ORF3a I10L
25563 ORF3a Q57H
25844 ORF3a T151I
25907 ORF3a G172V
27964 ORF8 S24L
28472 N P67S
28869 N P199L
29513 N A414S
142
477
453
446
583
1181
1707
433
453
1412
1106
1340
2270
2402
1094
1120
1046
1473
1436
3629
3823
2976
634
1380
987
1567
1599
1290
705
2091
1733
287
210
VSP2048−2
Base change
Expected
A
T
C
G
N
Ins/Del
No data

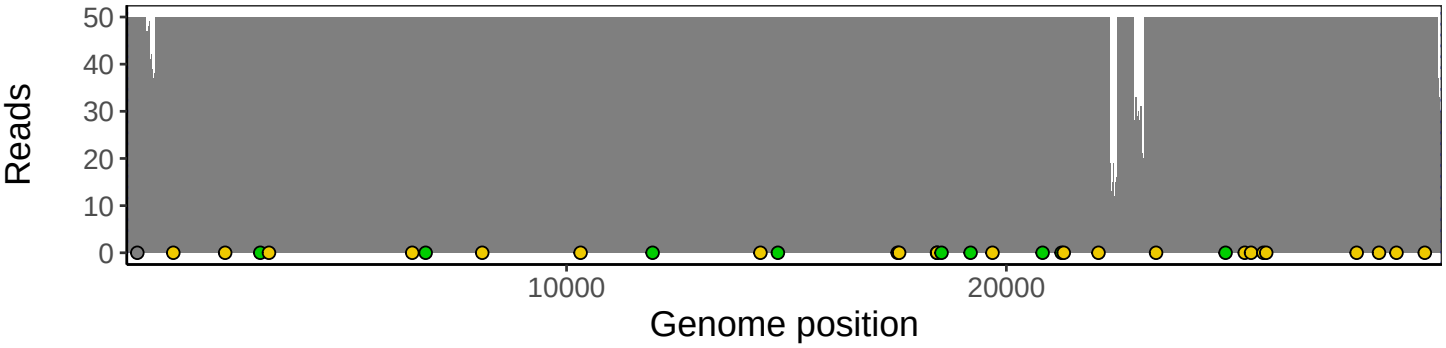
Analyses of individual experiments and composite results

VSP2048-2 | 2021-04-02 | Saline | UPHS-0834 | genomes | single experiment

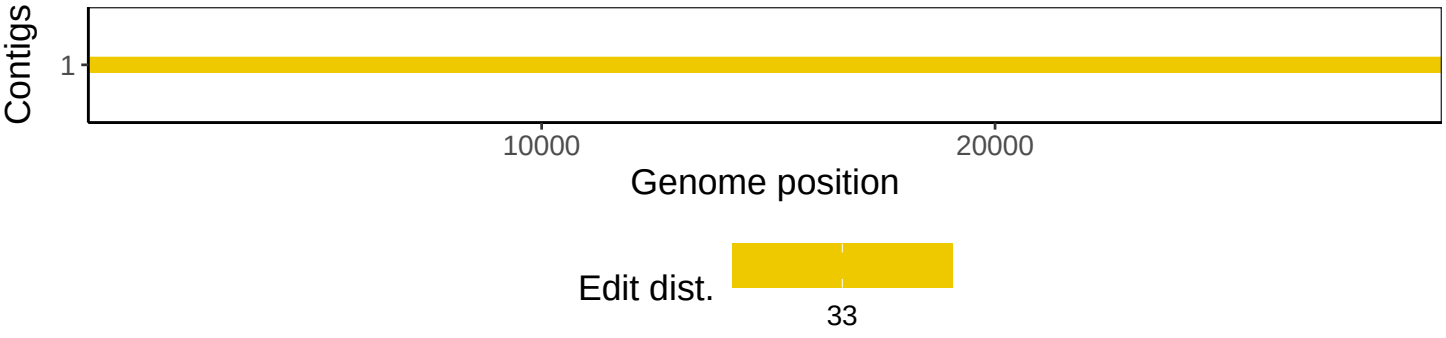
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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