COVID-19 subject UPHS-1034

2021-05-10

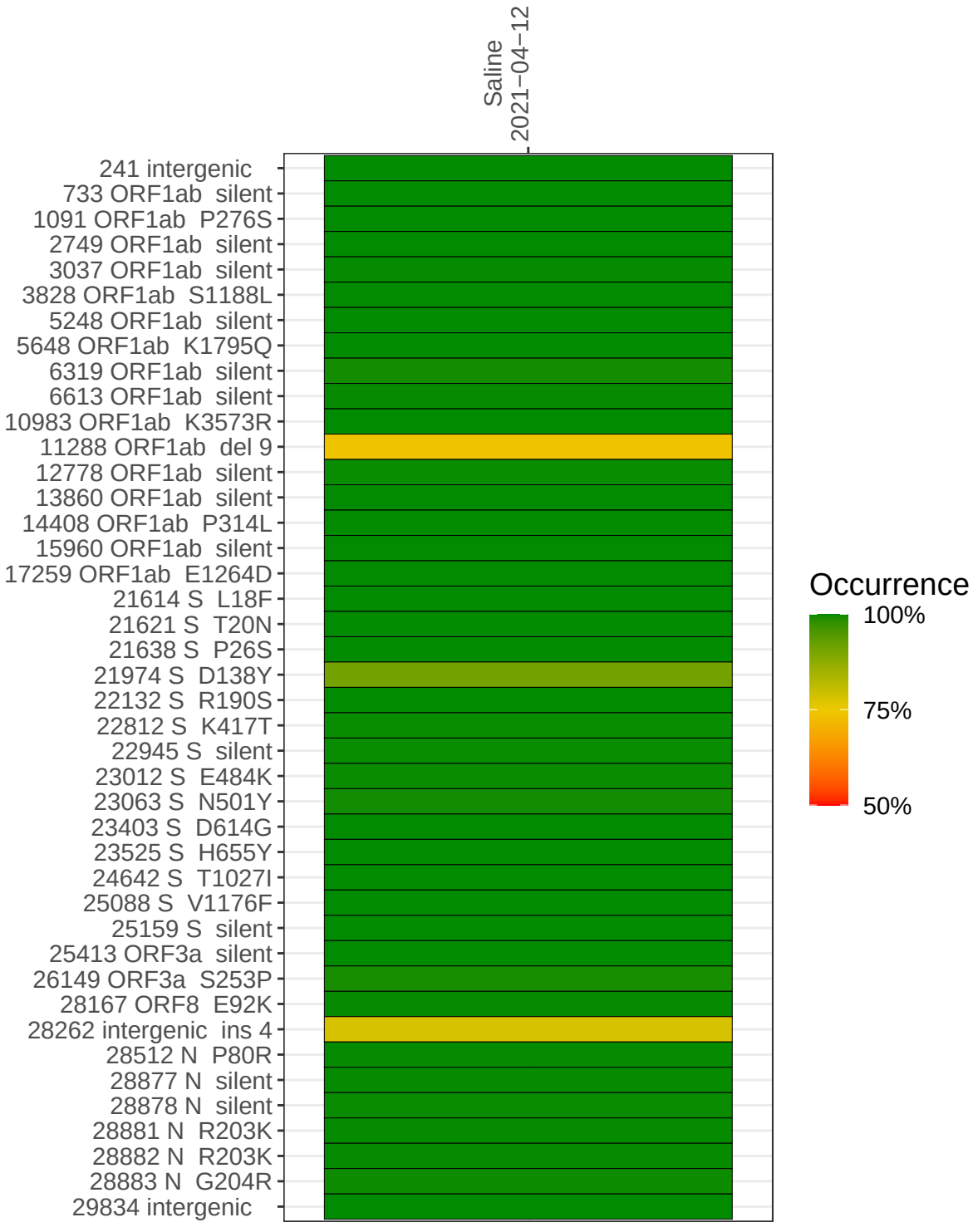
The table below provides a summary of subject samples for which sequencing data is available.
The experiments column shows the number of sequencing experiments performed for each specimen.
Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software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

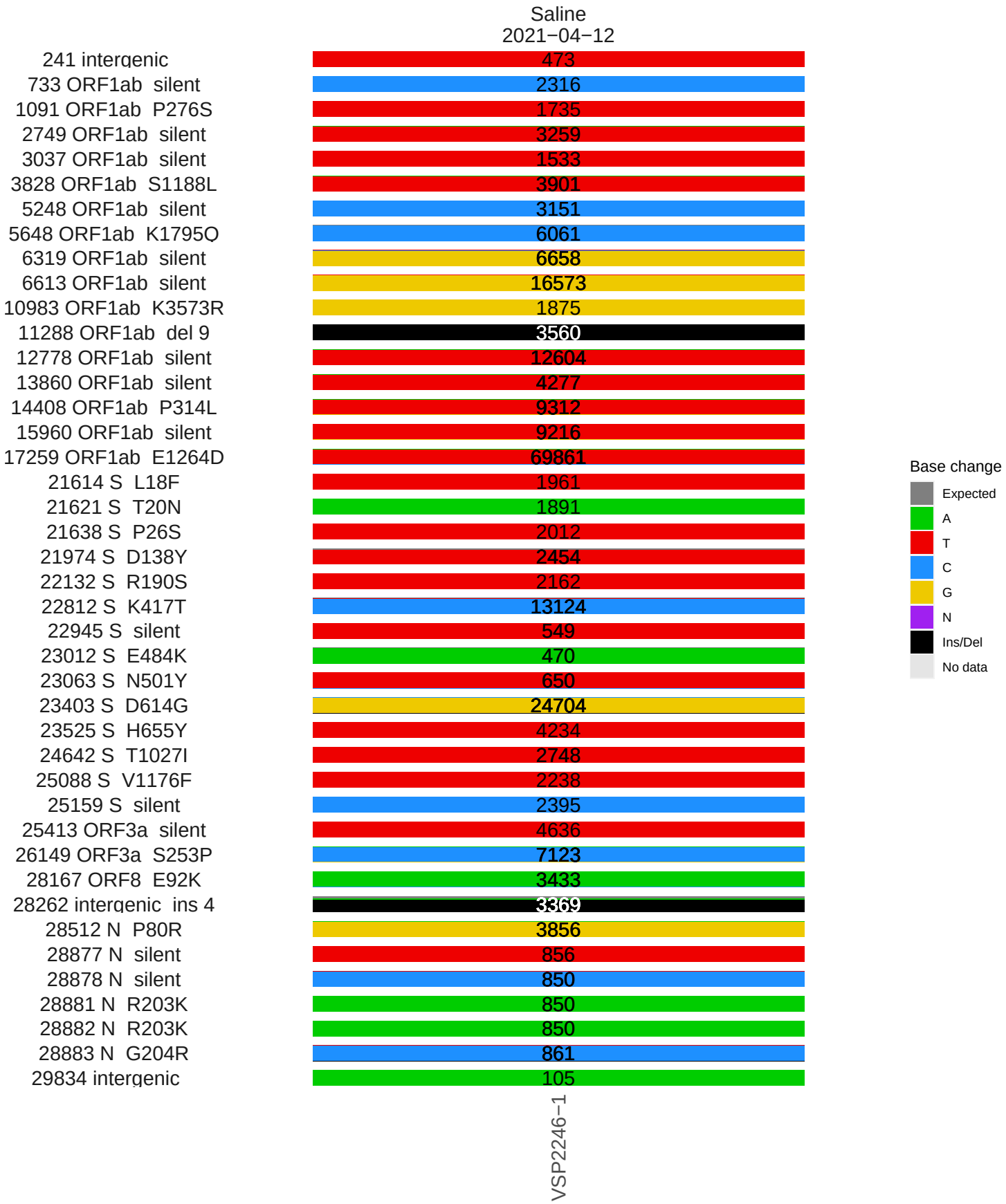
Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP2246-1	single experiment	NA	Saline	2021-04-12	29.81	P.1	99.7%	99.7%

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.

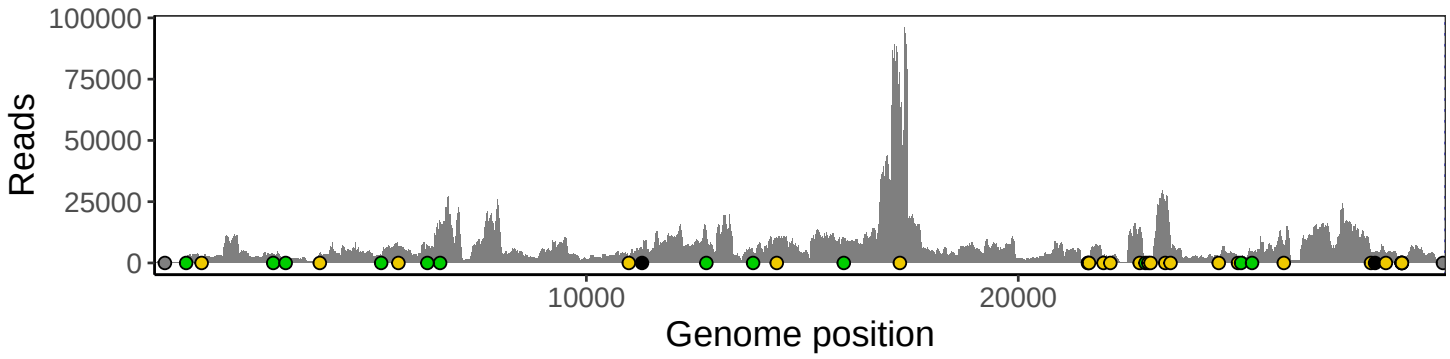



Saline
2021−04−12
241 intergenic
733 ORF1ab silent
1091 ORF1ab P276S
2749 ORF1ab silent
3037 ORF1ab silent
3828 ORF1ab S1188L
5248 ORF1ab silent
5648 ORF1ab K1795Q
6319 ORF1ab silent
6613 ORF1ab silent
10983 ORF1ab K3573R
11288 ORF1ab del 9
12778 ORF1ab silent
13860 ORF1ab silent
14408 ORF1ab P314L
15960 ORF1ab silent
17259 ORF1ab E1264D
21614 S L18F
21621 S T20N
21638 S P26S
21974 S D138Y
22132 S R190S
22812 S K417T
22945 S silent
23012 S E484K
23063 S N501Y
23403 S D614G
23525 S H655Y
24642 S T1027I
25088 S V1176F
25159 S silent
25413 ORF3a silent
26149 ORF3a S253P
28167 ORF8 E92K
28262 intergenic ins 4
28512 N P80R
28877 N silent
28878 N silent
28881 N R203K
28882 N R203K
28883 N G204R
29834 intergenic
473
2316
1735
3259
1533
3901
3151
6061
6658
16573
1875
3560
12604
4277
9312
9216
69861
1961
1891
2012
2454
2162
13124
549
470
650
24704
4234
2748
2238
2395
4636
7123
3433
3369
3856
856
850
850
850
861
105
VSP2246−1
Base change
Expected
A
T
C
G
N
Ins/Del
No data

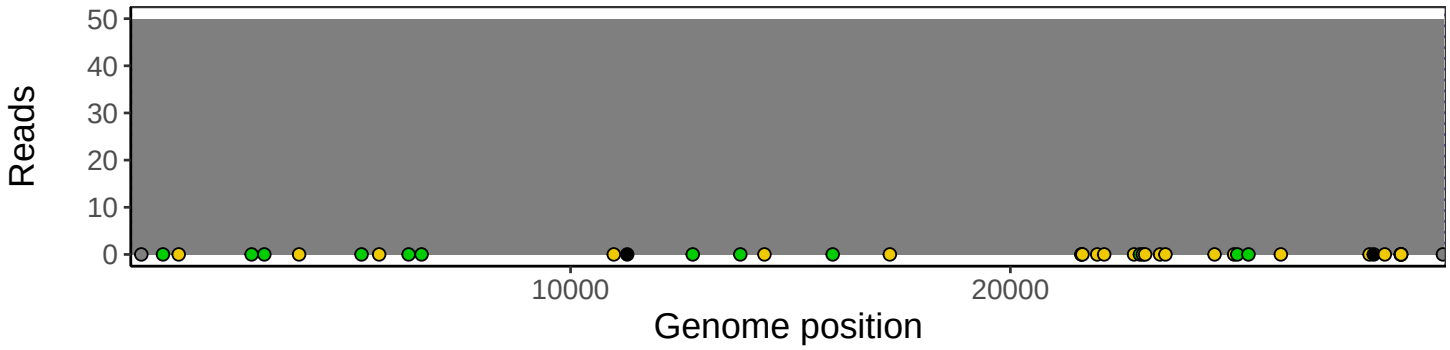
Analyses of individual experiments and composite results

VSP2246-1 | 2021-04-12 | Saline | UPHS-1034 | genomes | single experiment

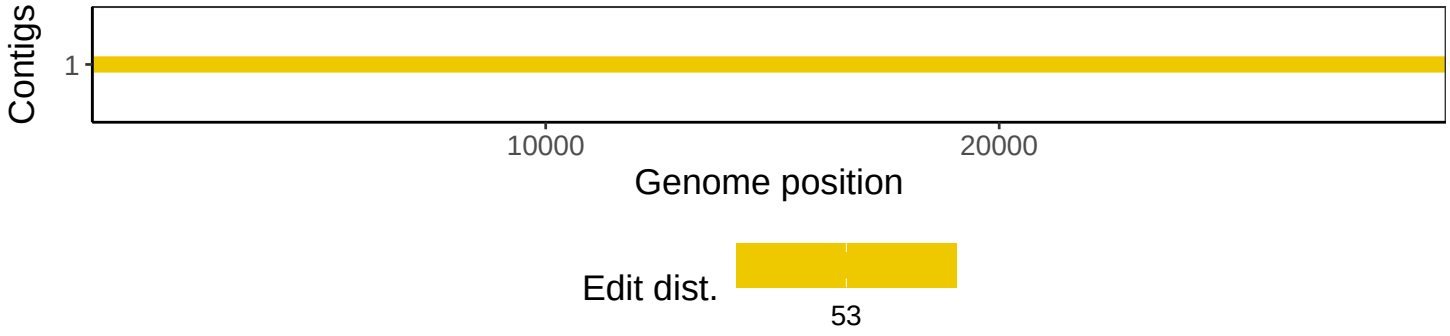
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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