COVID-19 subject UPHS-0103

2021-03-29

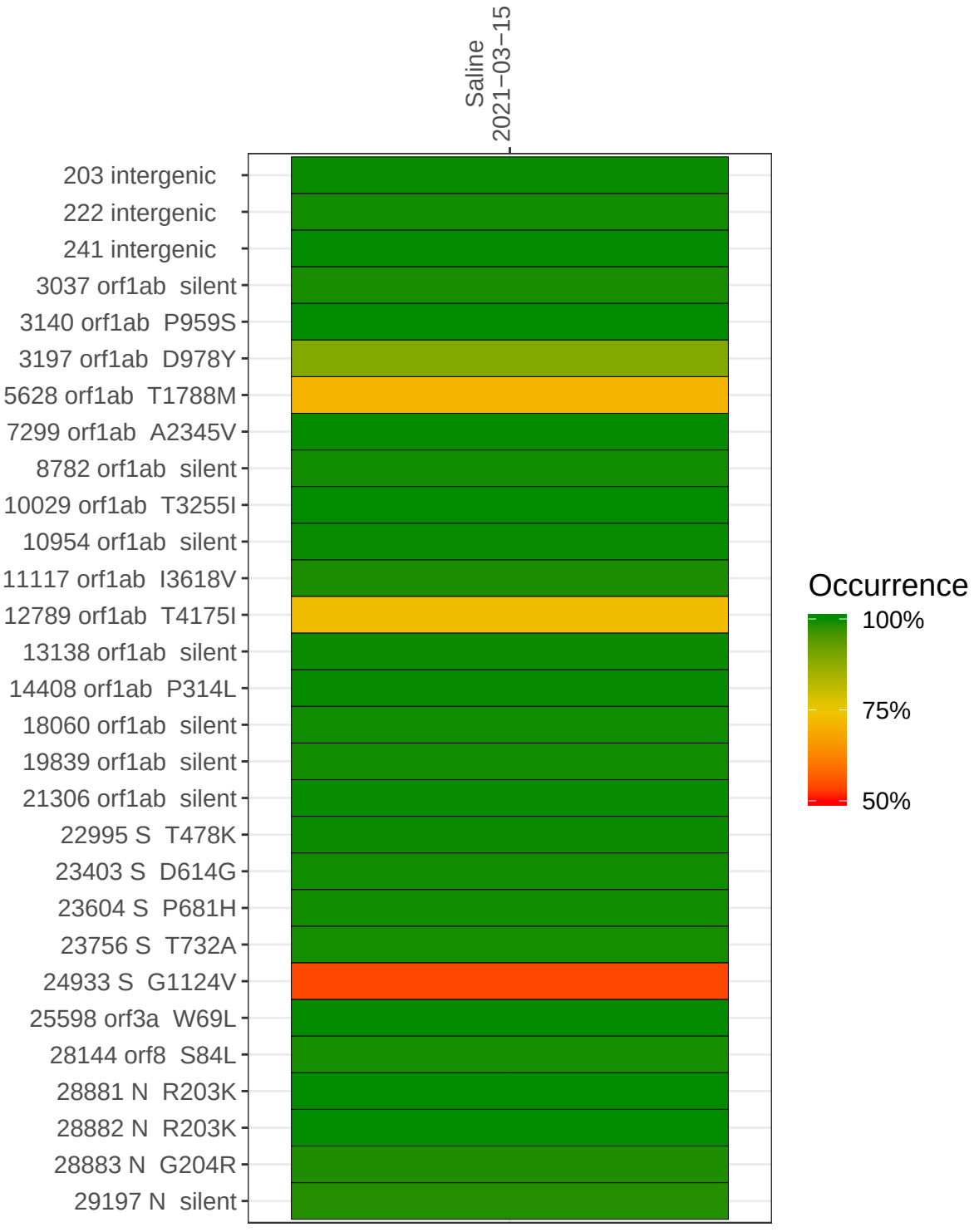
The table below provides a summary of subject samples for which sequencing data is available.
The experiments column shows the number of sequencing experiments performed for each specimen.
Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software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

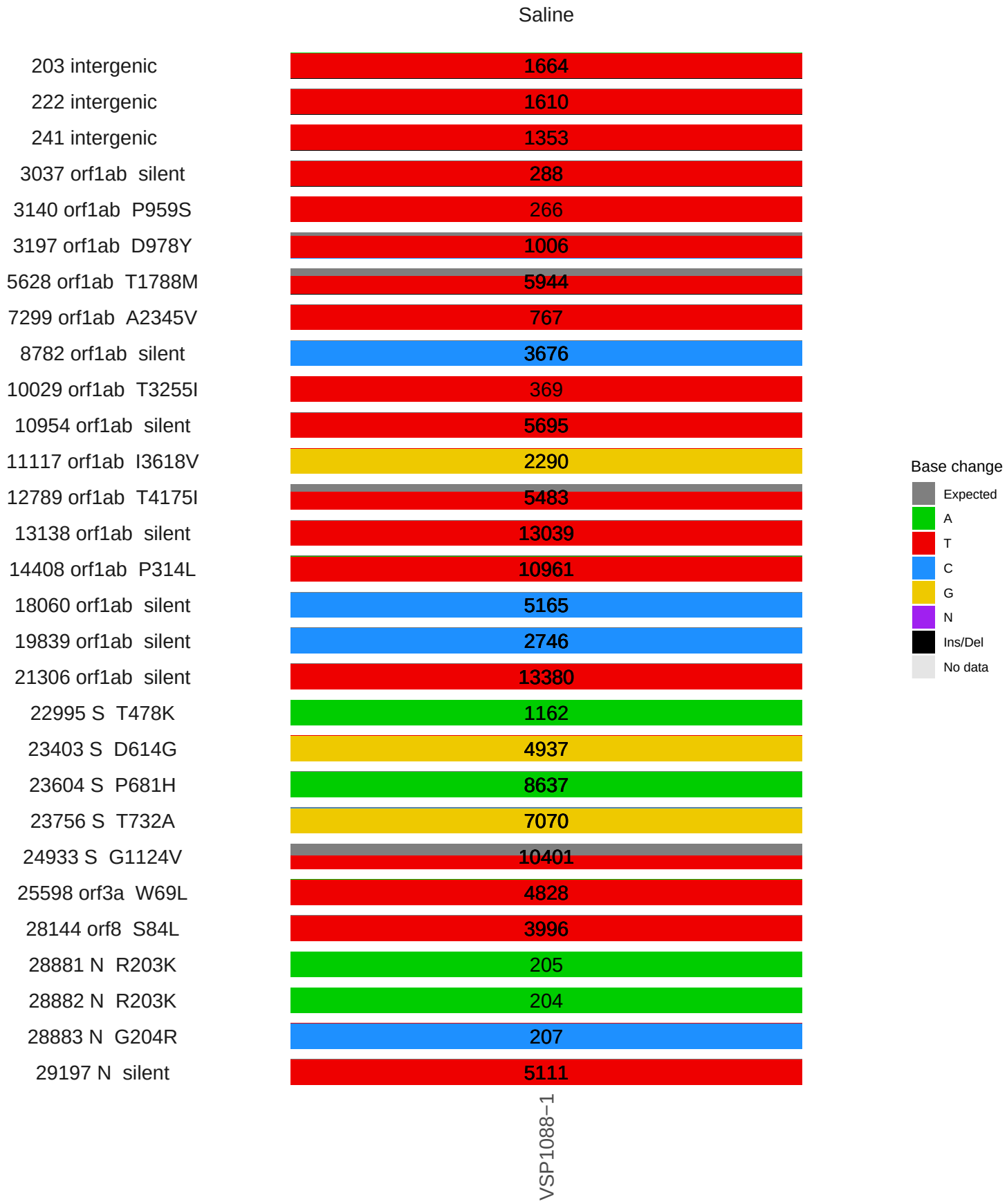
Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP1088-1	single experiment	NA	Saline	2021-03-15	29.88	B.1.1.222	99.8%	99.8%

Variants shared across samples

The heat map below shows how variants (reference genome 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.

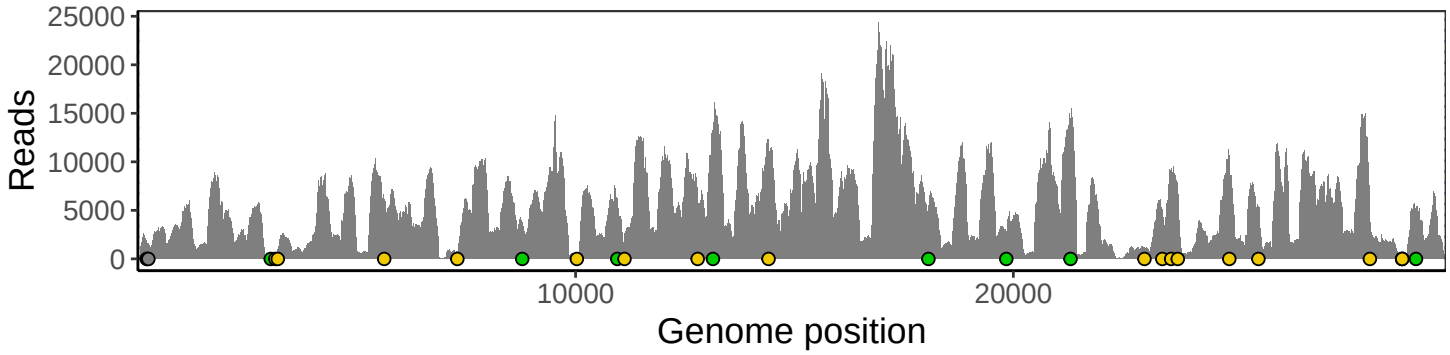



Saline
203 intergenic
222 intergenic
241 intergenic
3037 orf1ab silent
3140 orf1ab P959S
3197 orf1ab D978Y
5628 orf1ab T1788M
7299 orf1ab A2345V
8782 orf1ab silent
10029 orf1ab T3255I
10954 orf1ab silent
11117 orf1ab I3618V
12789 orf1ab T4175I
13138 orf1ab silent
14408 orf1ab P314L
18060 orf1ab silent
19839 orf1ab silent
21306 orf1ab silent
22995 S T478K
23403 S D614G
23604 S P681H
23756 S T732A
24933 S G1124V
25598 orf3a W69L
28144 orf8 S84L
28881 N R203K
28882 N R203K
28883 N G204R
29197 N silent
1664
1610
1353
288
266
1006
5944
767
3676
369
5695
2290
5483
13039
10961
5165
2746
13380
1162
4937
8637
7070
10401
4828
3996
205
204
207
5111
VSP1088−1
Base change
Expected
A
T
C
G
N
Ins/Del
No data

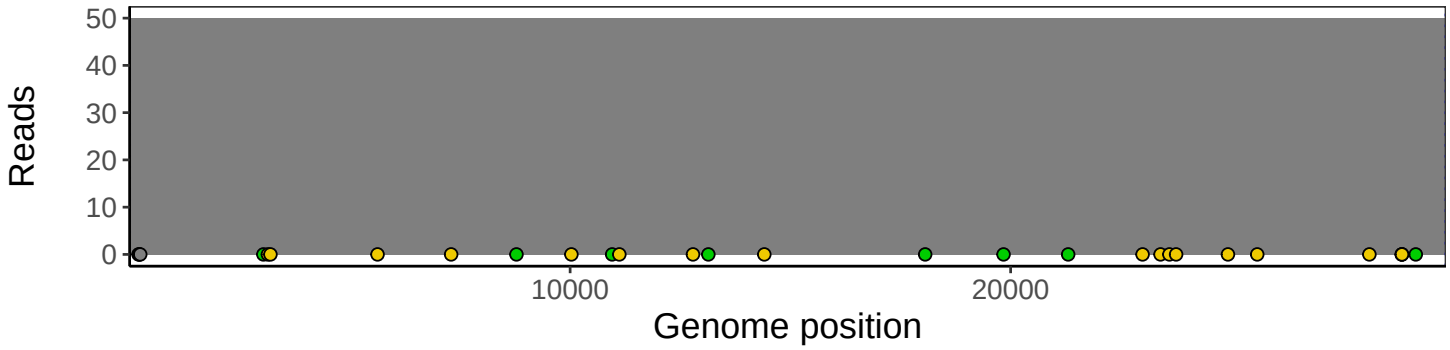
Analyses of individual experiments and composite results

VSP1088-1 | 2021-03-15 | Saline | UPHS-0103 | genomes | single experiment

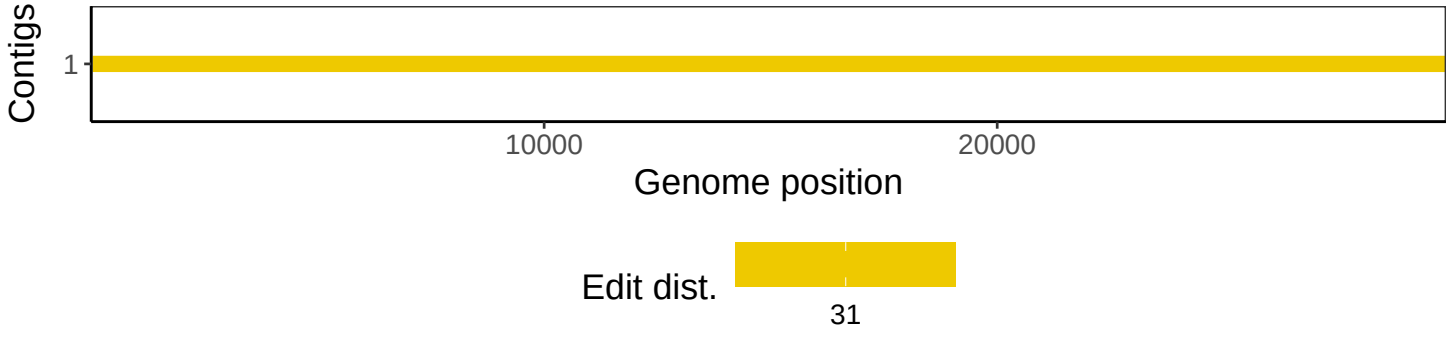
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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	2.3.3
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	3.0.0
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