COVID-19 subject HUP Q-0043

2021-03-29

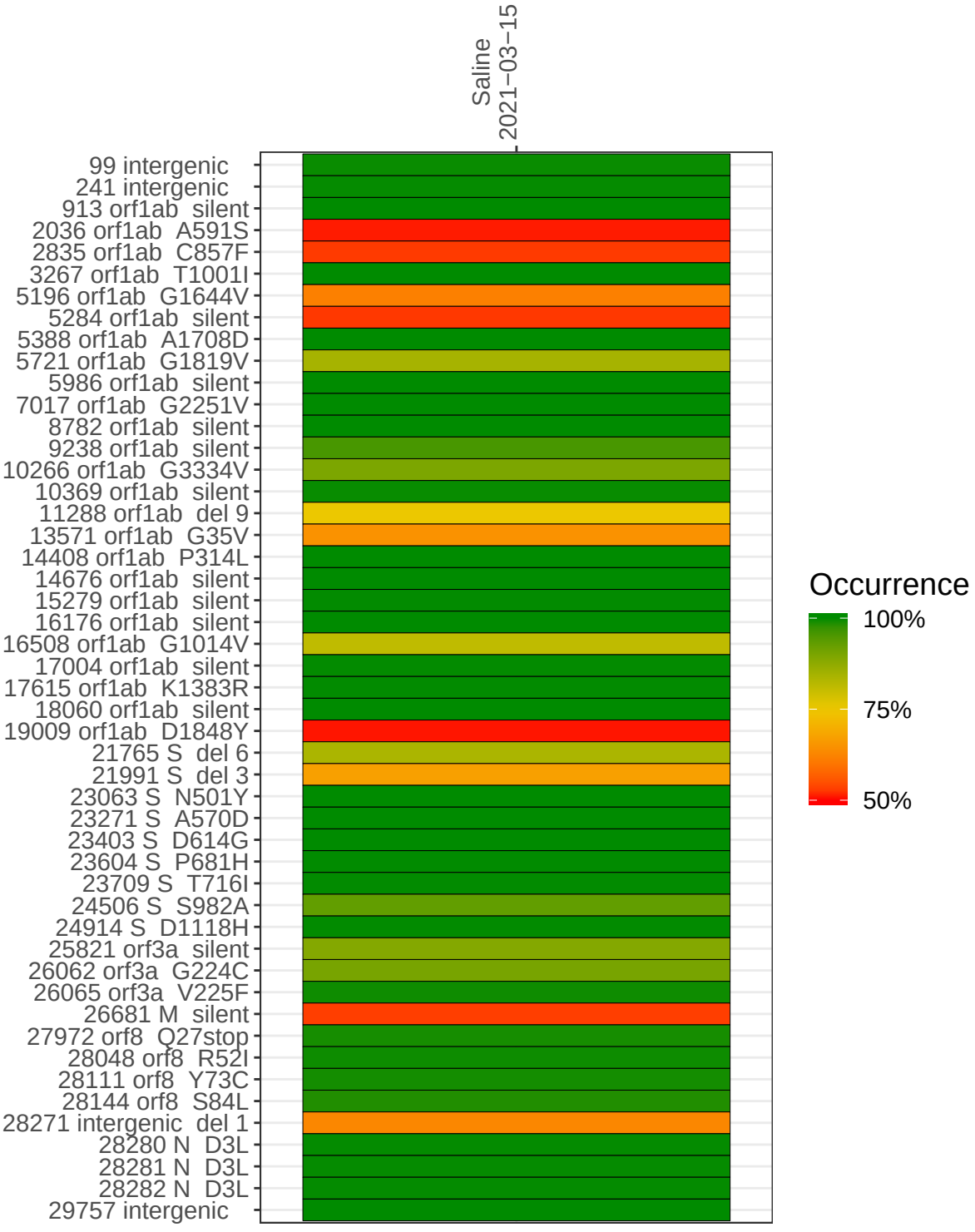
The table below provides a summary of subject samples for which sequencing data is available.
The experiments column shows the number of sequencing experiments performed for each specimen.
Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software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

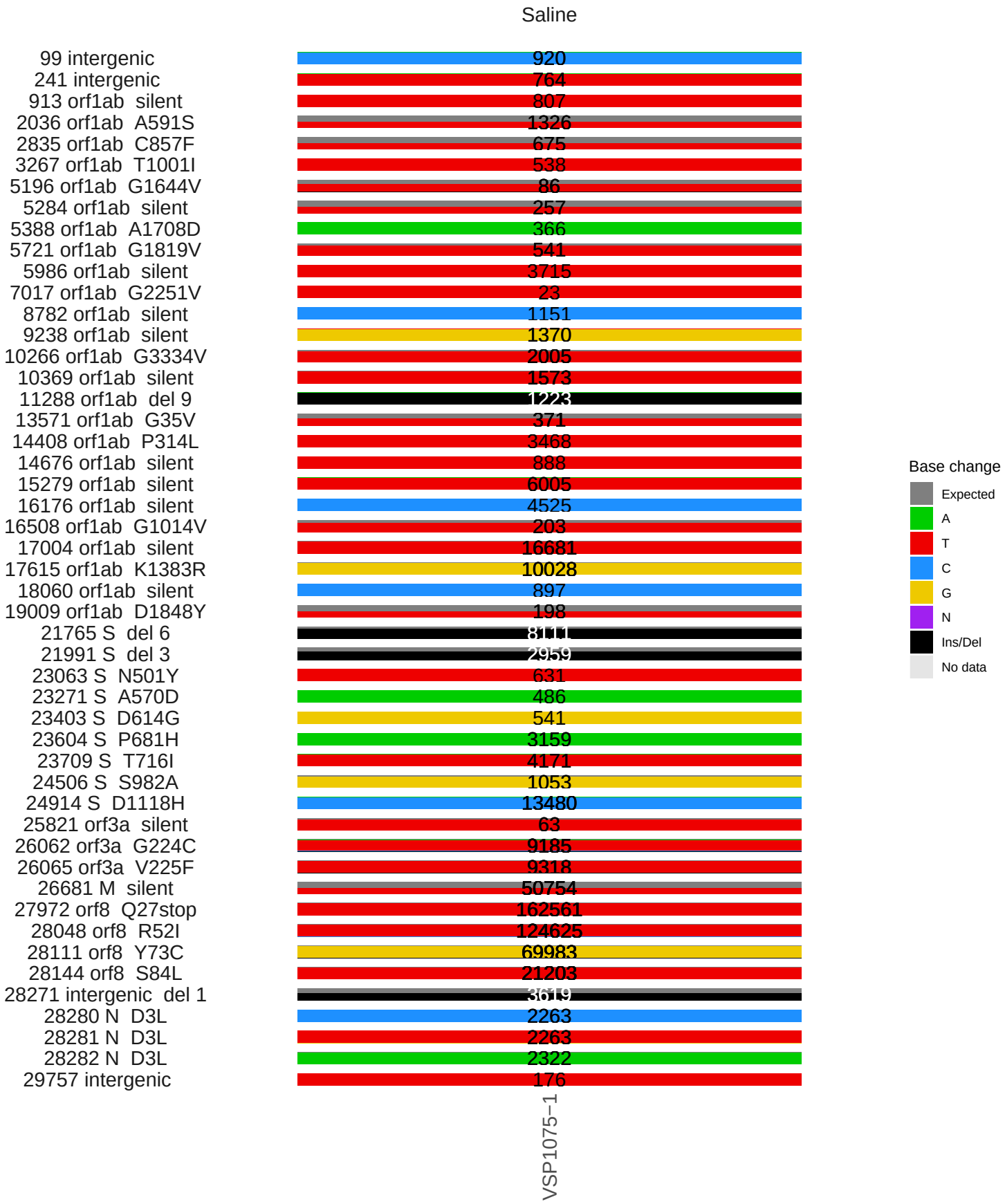
Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP1075-1	single experiment	NA	Saline	2021-03-15	24.61	B.1.1.7	97.3%	97.2%

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.

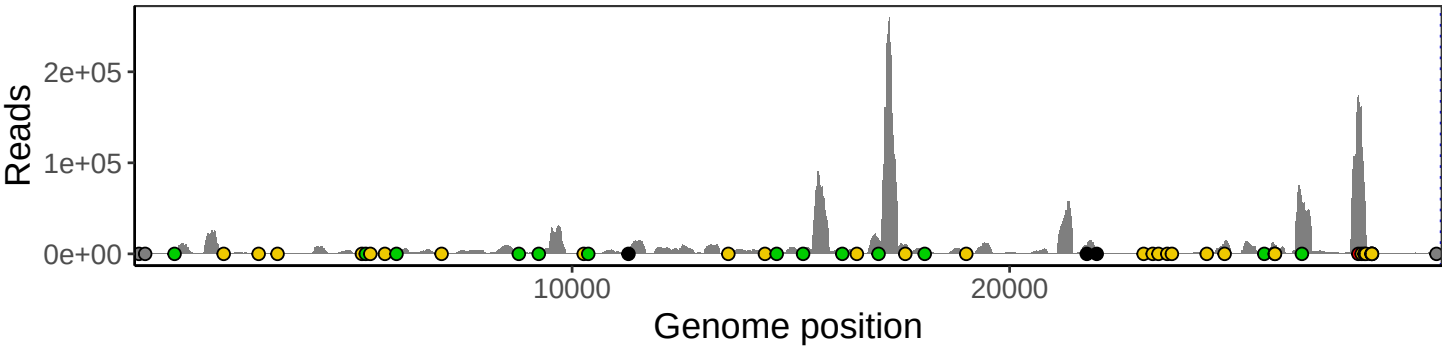



Saline
99 intergenic
241 intergenic
913 orf1ab silent
2036 orf1ab A591S
2835 orf1ab C857F
3267 orf1ab T1001I
5196 orf1ab G1644V
5284 orf1ab silent
5388 orf1ab A1708D
5721 orf1ab G1819V
5986 orf1ab silent
7017 orf1ab G2251V
8782 orf1ab silent
9238 orf1ab silent
10266 orf1ab G3334V
10369 orf1ab silent
11288 orf1ab del 9
13571 orf1ab G35V
14408 orf1ab P314L
14676 orf1ab silent
15279 orf1ab silent
16176 orf1ab silent
16508 orf1ab G1014V
17004 orf1ab silent
17615 orf1ab K1383R
18060 orf1ab silent
19009 orf1ab D1848Y
21765 S del 6
21991 S del 3
23063 S N501Y
23271 S A570D
23403 S D614G
23604 S P681H
23709 S T716I
24506 S S982A
24914 S D1118H
25821 orf3a silent
26062 orf3a G224C
26065 orf3a V225F
26681 M silent
27972 orf8 Q27stop
28048 orf8 R52I
28111 orf8 Y73C
28144 orf8 S84L
28271 intergenic del 1
28280 N D3L
28281 N D3L
28282 N D3L
29757 intergenic
920
764
807
1326
675
538
86
257
366
541
3715
23
1151
1370
2005
1573
1223
371
3468
888
6005
4525
203
16681
10028
897
198
8111
2959
631
486
541
3159
4171
1053
13480
63
9185
9318
50754
162561
124625
69983
21203
3619
2263
2263
2322
176
VSP1075−1
Base change
Expected
A
T
C
G
N
Ins/Del
No data

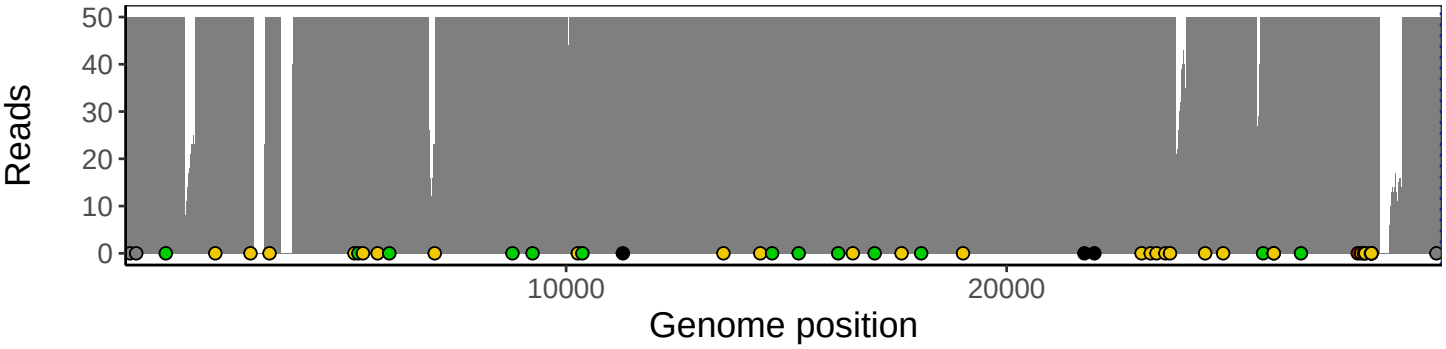
Analyses of individual experiments and composite results

VSP1075-1 | 2021-03-15 | Saline | HUP Q-0043 | genomes | single experiment

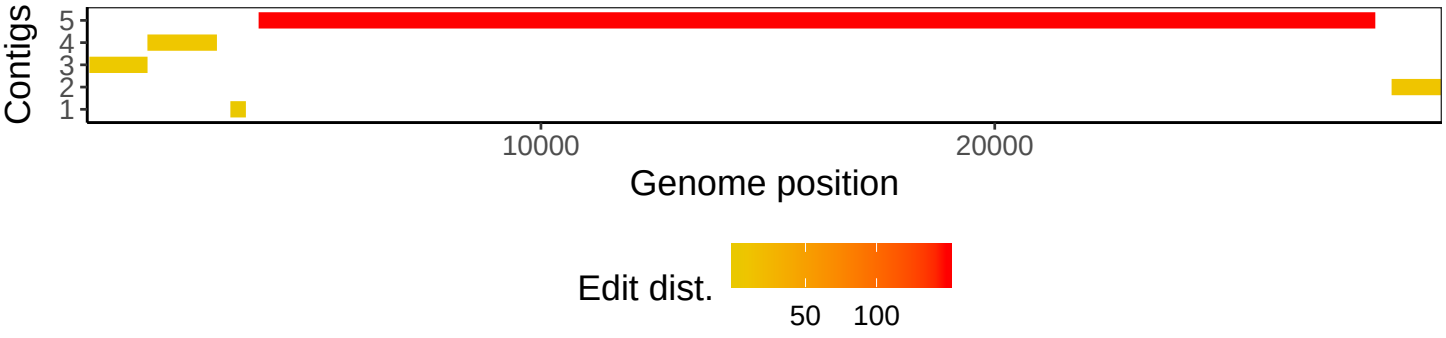
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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