

COVID-19 subject UPHS-0109

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.

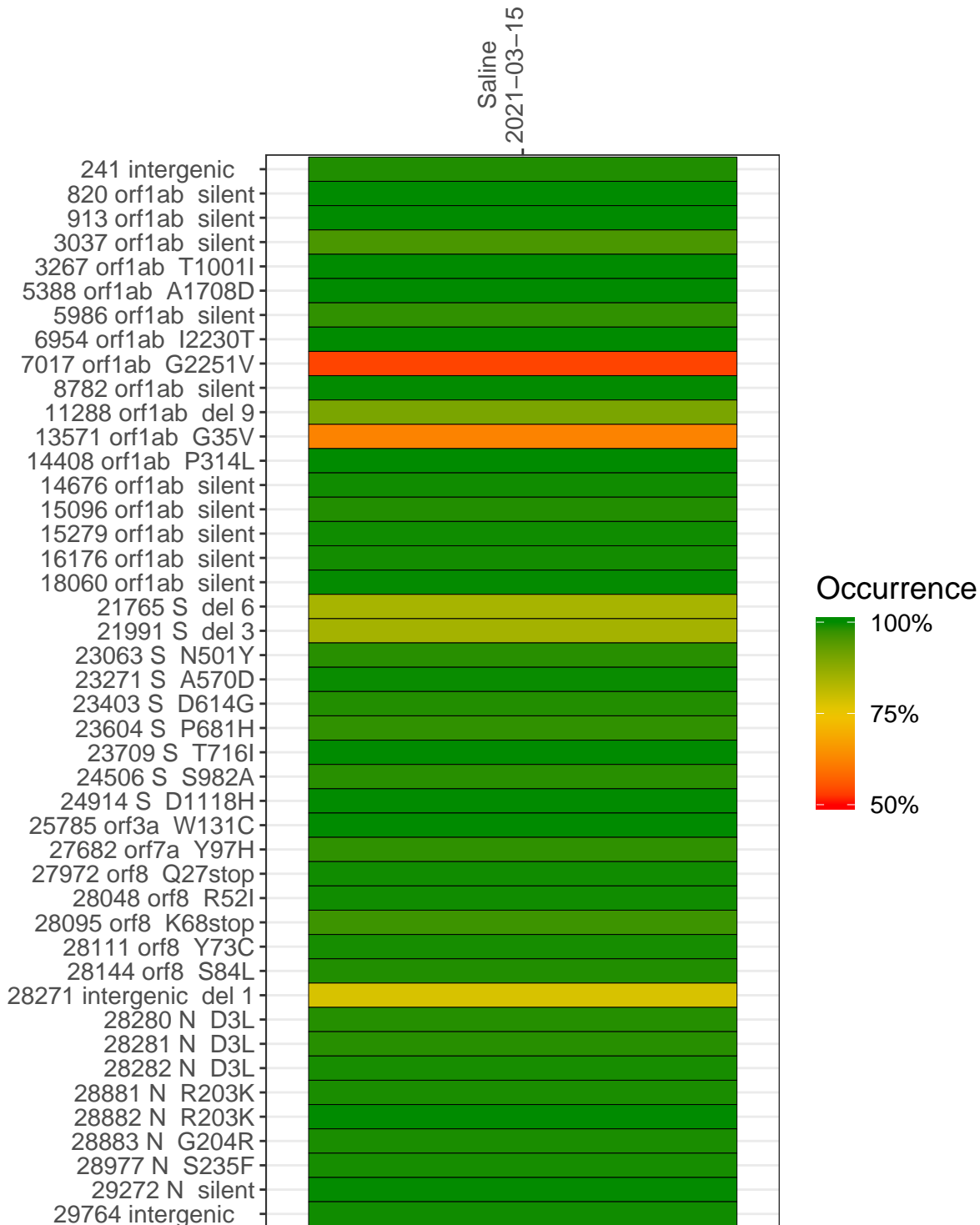
Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software tool (Rambaut et al 2020) for genomes with $> 90\%$ sequence coverage.

Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (≥ 5 reads)
VSP1094-1	single experiment	NA	Saline	2021-03-15	29.79	B.1.1.7	99.9%	99.9%

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

241 intergenic	918
820 orf1ab silent	1390
913 orf1ab silent	1421
3037 orf1ab silent	274
3267 orf1ab T1001I	614
5388 orf1ab A1708D	2314
5986 orf1ab silent	1552
6954 orf1ab I2230T	60
7017 orf1ab G2251V	82
8782 orf1ab silent	1817
11288 orf1ab del 9	3930
13571 orf1ab G35V	370
14408 orf1ab P314L	1288
14676 orf1ab silent	3308
15096 orf1ab silent	3124
15279 orf1ab silent	5519
16176 orf1ab silent	2433
18060 orf1ab silent	961
21765 S del 6	4396
21991 S del 3	2128
23063 S N501Y	126
23271 S A570D	4604
23403 S D614G	5492
23604 S P681H	2367
23709 S T716I	2215
24506 S S982A	2473
24914 S D1118H	4851
25785 orf3a W131C	454
27682 orf7a Y97H	4475
27972 orf8 Q27stop	49926
28048 orf8 R52I	43683
28095 orf8 K68stop	36114
28111 orf8 Y73C	30210
28144 orf8 S84L	11152
28271 intergenic del 1	4845
28280 N D3L	3854
28281 N D3L	3855
28282 N D3L	3981
28881 N R203K	126
28882 N R203K	126
28883 N G204R	127
28977 N S235F	159
29272 N silent	3596
29764 intergenic	749

Base change

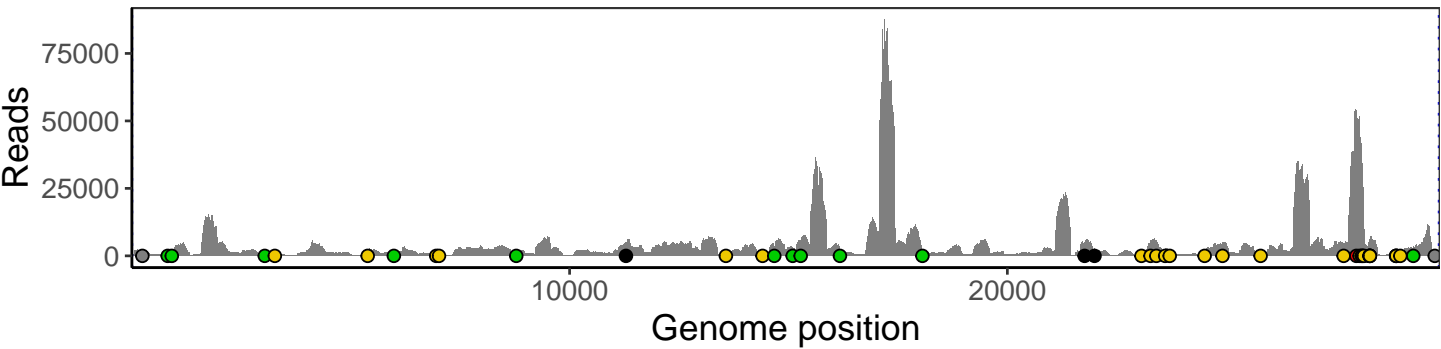
Expected
A
T
C
G
N
Ins/Del
No data

VSP1094-1

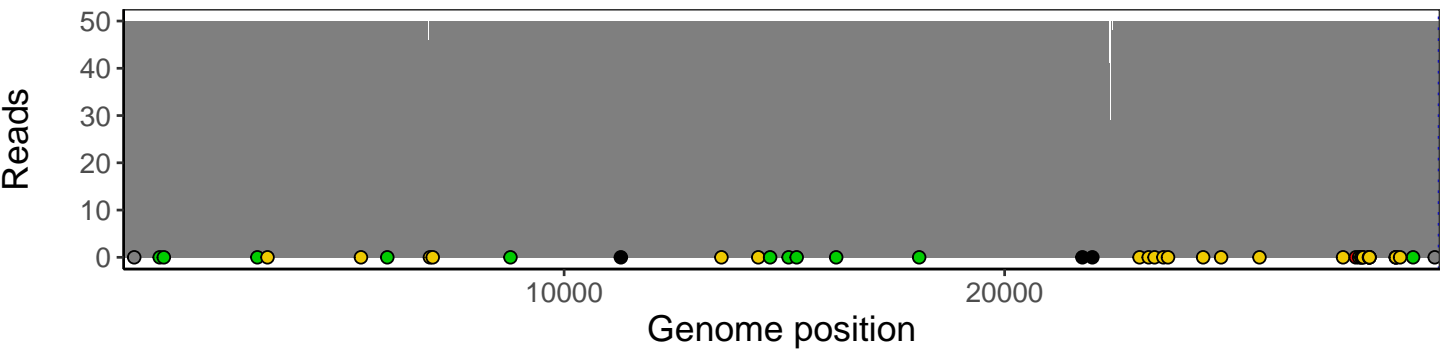
Analyses of individual experiments and composite results

VSP1094-1 | 2021-03-15 | Saline | UPHS-0109 | genomes | single experiment

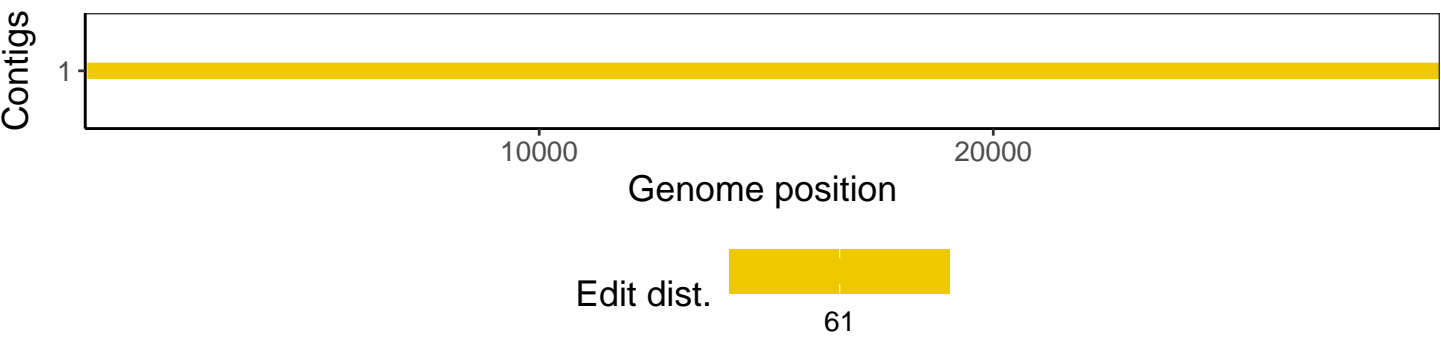
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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 htlib 1.10
bcftools	1.10.2-34-g1a12af0-dirty Using htlib 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