

COVID-19 subject UPHS-0104

2021-05-05

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.

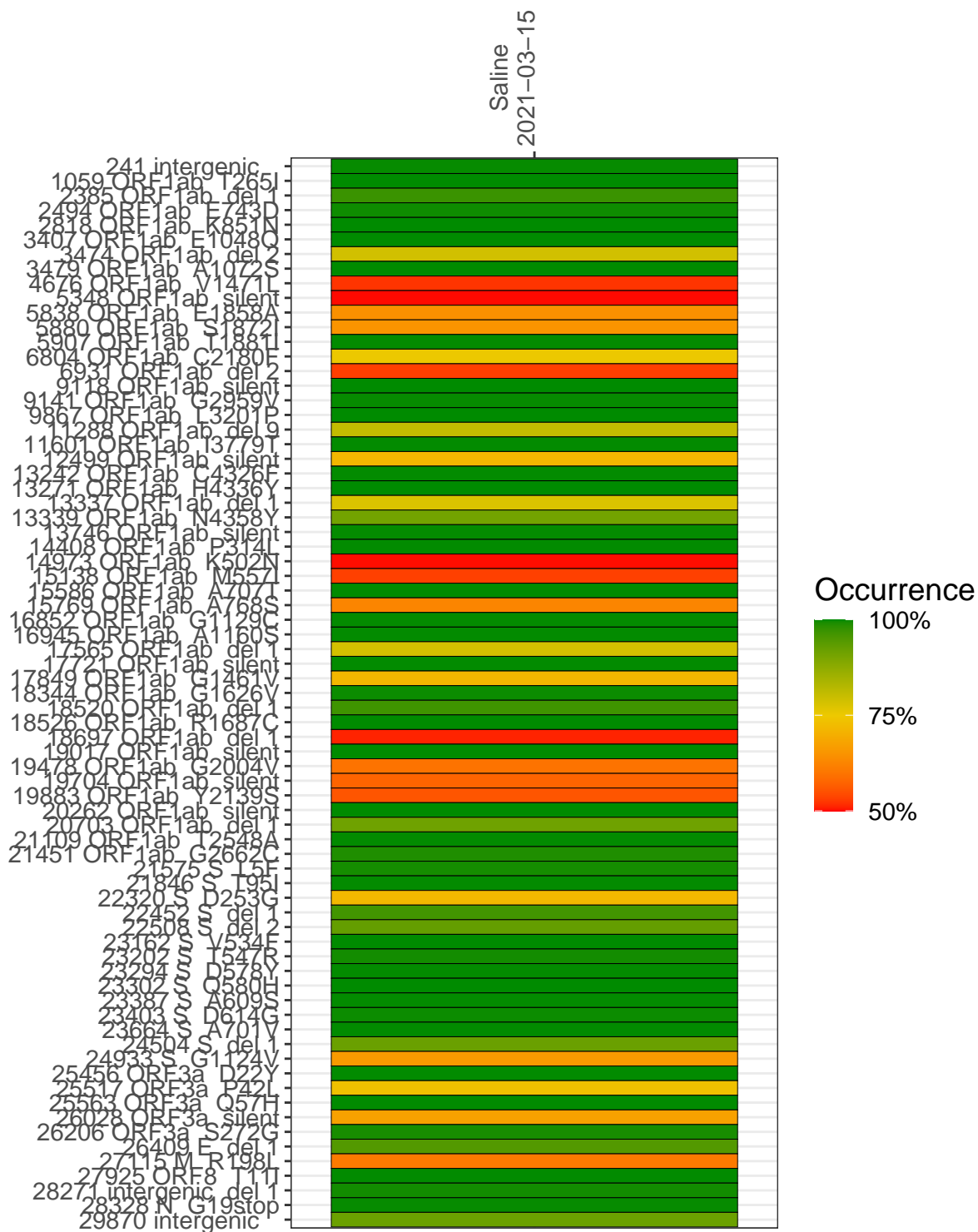
Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software tool (Rambaut et al 2020) for genomes with $> 90\%$ sequence coverage.

Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (≥ 5 reads)
VSP1089-1	single experiment	NA	Saline	2021-03-15	4.63	NA	78.0%	77.0%

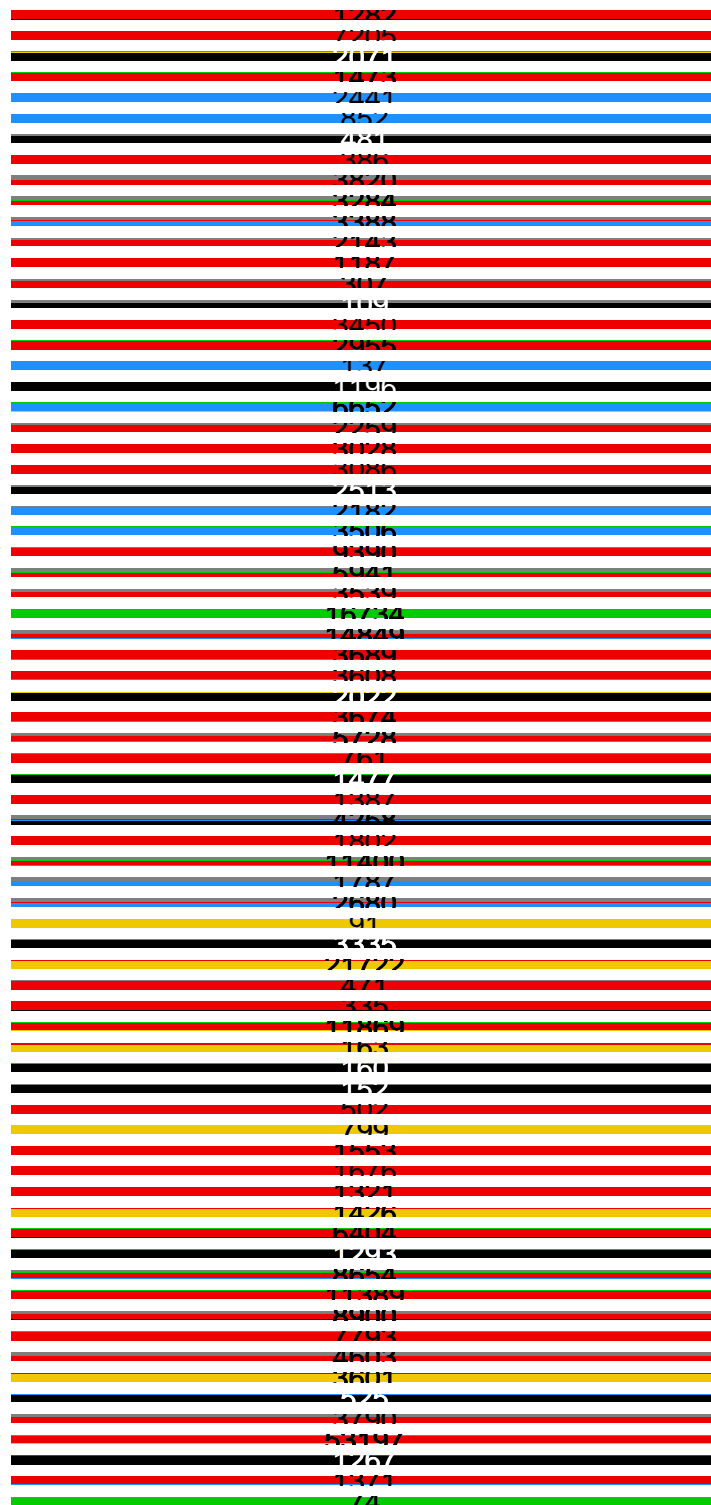
Variants shared across samples

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
2021-03-15

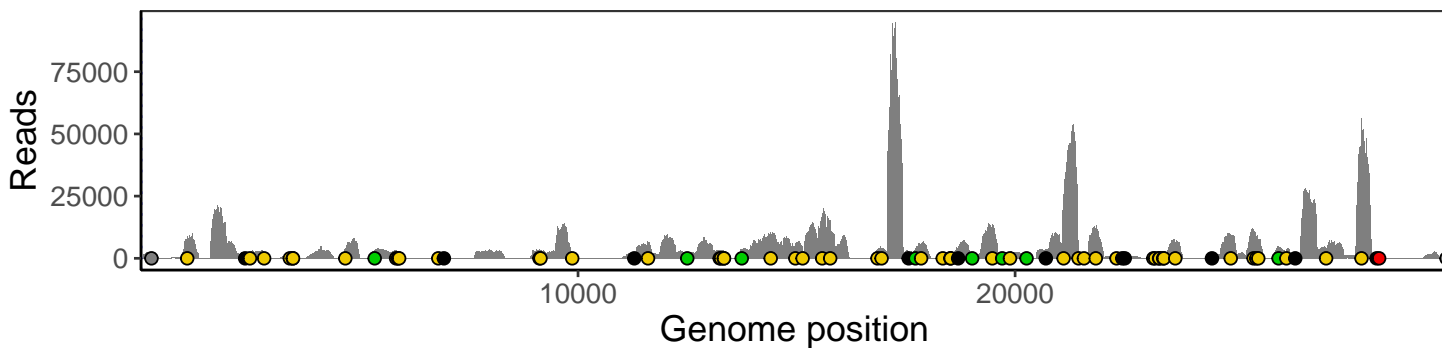
241 intergenic
1054 ORF126
2385 ORF126 del
2494 ORF126 F7431
2818 ORF126 K851N
3407 ORF126 F1048G
3474 ORF126 del
3474 ORF126 A1072S
4676 ORF126 V1471
5348 ORF126 silent
5838 ORF126 F1858A
5880 ORF126 S1872
5907 ORF126 I1881
6804 ORF126 C2080F
6931 ORF126 del
9118 ORF126 silent
9141 ORF126 G2959V
9867 ORF126 I3201P
11288 ORF126 del
11601 ORF126 I3791
12494 ORF126 silent
13242 ORF126 C14326F
13271 ORF126 H4336Y
13337 ORF126 del
13339 ORF126 M4368Y
13746 ORF126 silent
14408 ORF126 P3141
14973 ORF126 K502N
15138 ORF126 M5571
15586 ORF126 A7017
15769 ORF126 A768S
16852 ORF126 G11791
16945 ORF126 A1160S
17565 ORF126 del
17721 ORF126 silent
17849 ORF126 G1461V
18344 ORF126 G1626V
18520 ORF126 del
18526 ORF126 K1687C
18697 ORF126 del
19017 ORF126 silent
19478 ORF126 G2004V
19704 ORF126 silent
19883 ORF126 Y2139S
20262 ORF126 silent
20703 ORF126 del
21109 ORF126 I2588A
21451 ORF126 G2662C
21575 S 15F
21846 S 1951
22320 S 125315
22452 S del
22508 S del
23162 S V534F
23202 S 1547R
23294 S 1578Y
23302 S 1580H
23387 S A609S
23403 S 161404
23664 S A7017V
24504 S del
24933 S G1172V
25456 ORF32 1122Y
25517 ORF32 P421
25563 ORF32 157H
26028 ORF32 silent
26206 ORF32 S27214
26409 F del
27115 M R1981
27925 ORF32 1111
28271 intergenic del
28328 N G1950n
29870 intergenic



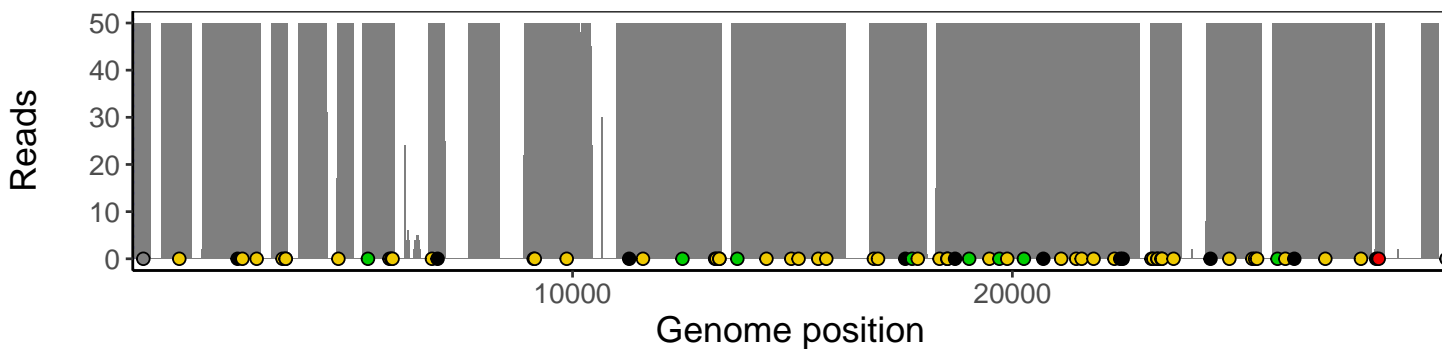
Analyses of individual experiments and composite results

VSP1089-1 | 2021-03-15 | Saline | UPHS-0104 | 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	2.3.8
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	3.0.0
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