

# COVID-19 subject UPHS-0102

*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.

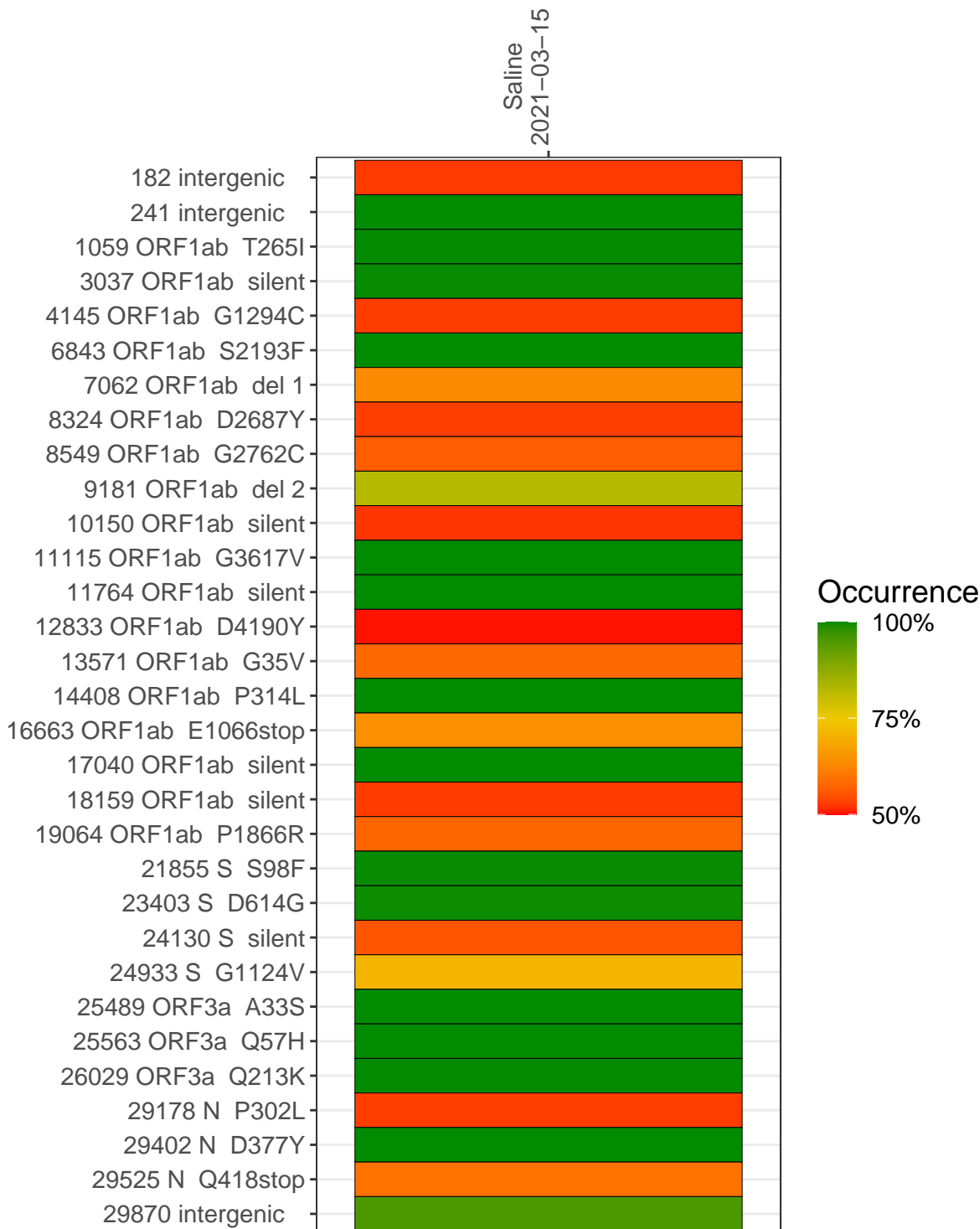
Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software tool (Rambaut et al 2020) for genomes with  $> 90\%$  sequence coverage.

Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage ( $\geq 5$ reads)
VSP1087-1	single experiment	NA	Saline	2021-03-15	11.30	NA	95.4%	94.6%

## 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

182 intergenic	834
241 intergenic	506
1059 ORF1ab T265I	3830
3037 ORF1ab silent	2458
4145 ORF1ab G1294C	2817
6843 ORF1ab S2193F	4481
7062 ORF1ab del 1	1037
8324 ORF1ab D2687Y	1322
8549 ORF1ab G2762C	1385
9181 ORF1ab del 2	1247
10150 ORF1ab silent	2755
11115 ORF1ab G3617V	82
11764 ORF1ab silent	246
12833 ORF1ab D4190Y	1002
13571 ORF1ab G35V	694
14408 ORF1ab P314L	10630
16663 ORF1ab E1066stop	235
17040 ORF1ab silent	12195
18159 ORF1ab silent	2798
19064 ORF1ab P1866R	524
21855 S S98F	4961
23403 S D614G	681
24130 S silent	1231
24933 S G1124V	5915
25489 ORF3a A33S	713
25563 ORF3a Q57H	497
26029 ORF3a Q213K	6218
29178 N P302L	1830
29402 N D377Y	266
29525 N Q418stop	1247
29870 intergenic	43

Base change

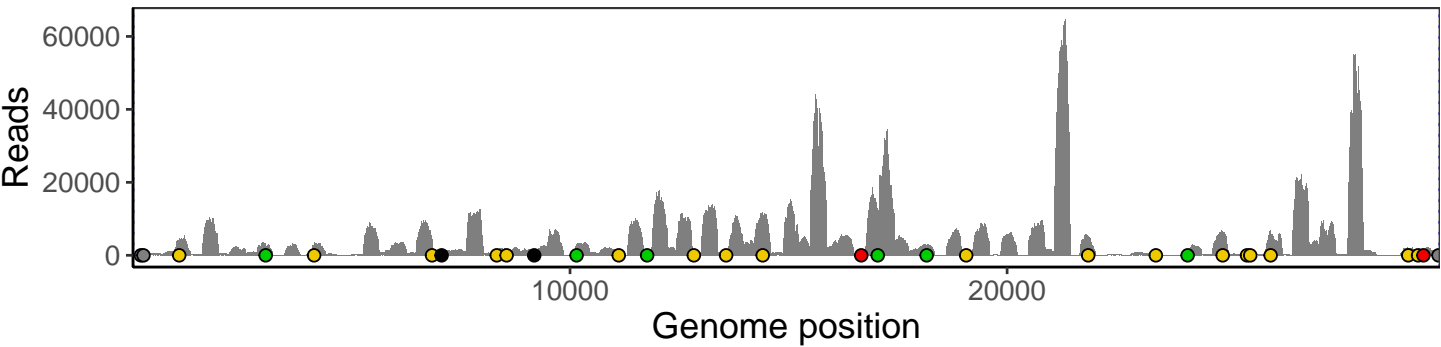


VSP1087-1

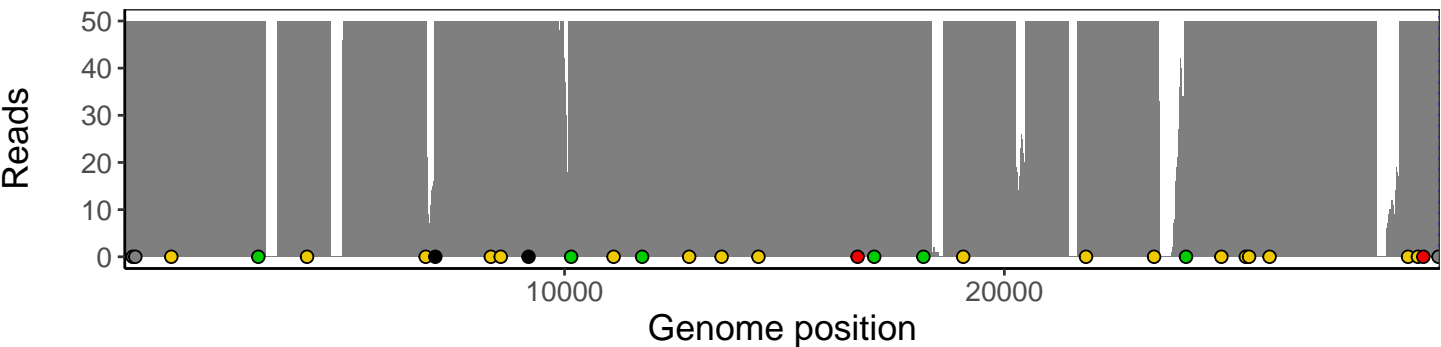
# Analyses of individual experiments and composite results

VSP1087-1 | 2021-03-15 | Saline | UPHS-0102 | genomes | single experiment

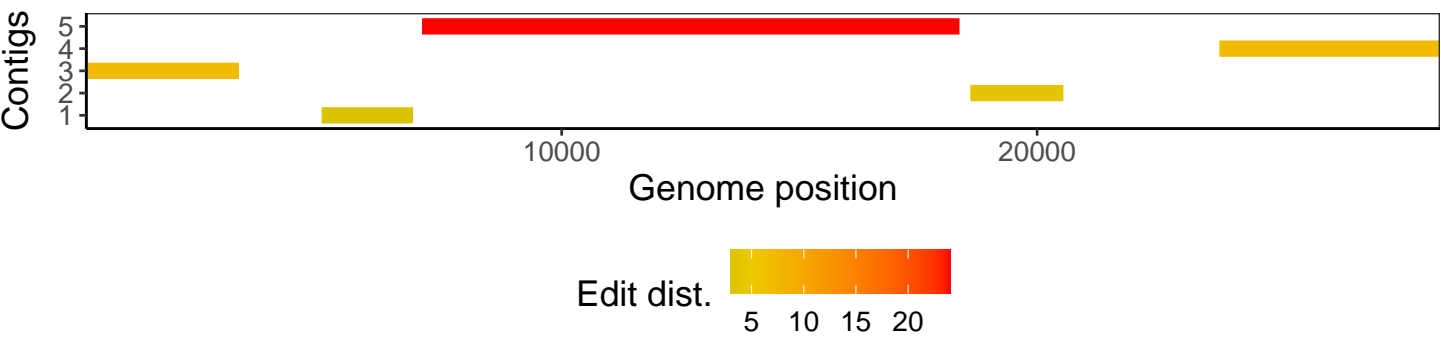
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according to variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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