

# COVID-19 subject UPHS-0819

*2021-06-23*

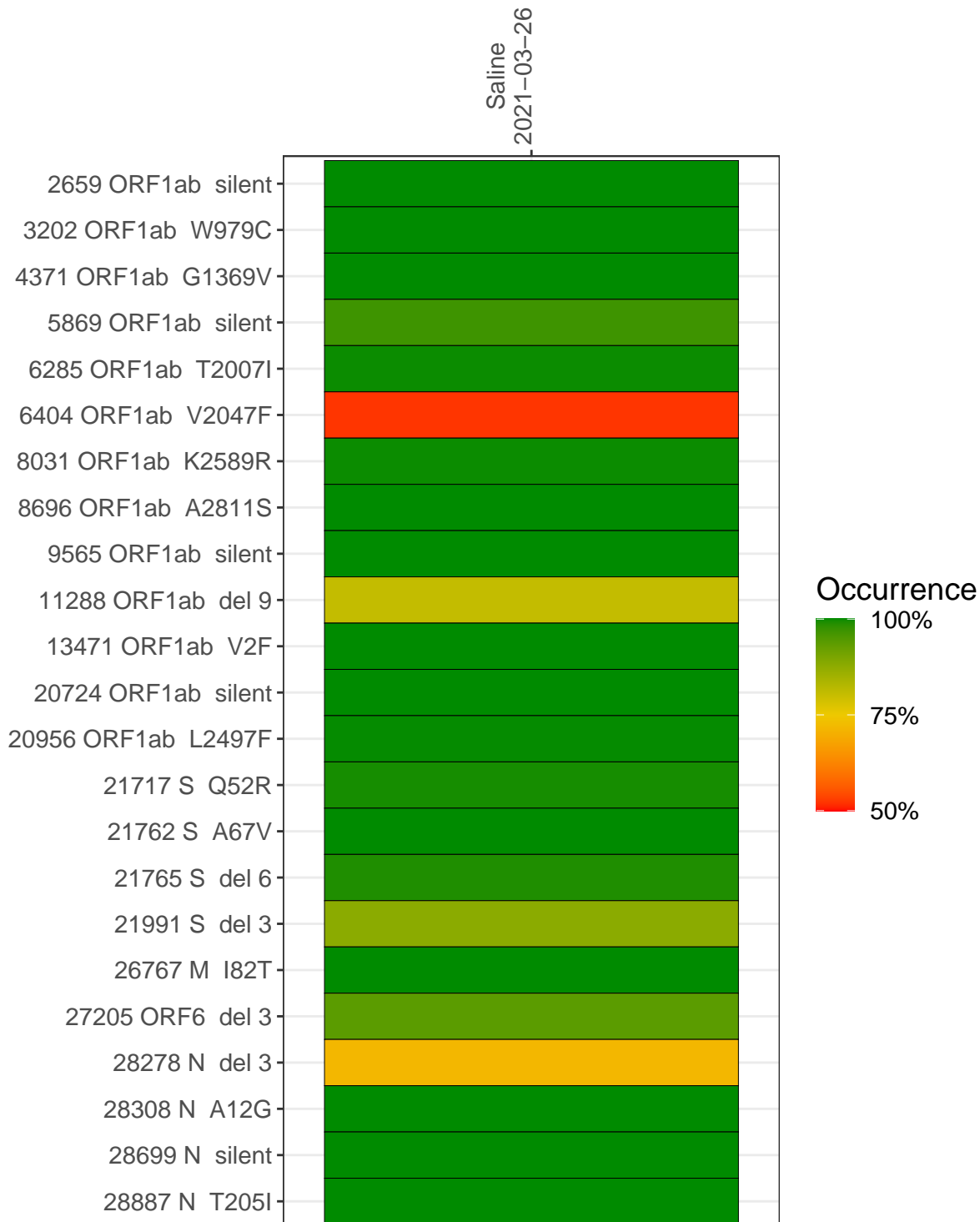
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP2033-2	single experiment	NA	Saline	2021-03-26	3.07	NA	55.9%	54.5%

## Variants shared across samples

The heat map below shows how variants (reference genome /home/common/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  
2021-03-26

2659 ORF1ab silent	277
3202 ORF1ab W979C	128
4371 ORF1ab G1369V	168
5869 ORF1ab silent	61
6285 ORF1ab T2007I	395
6404 ORF1ab V2047F	430
8031 ORF1ab K2589R	368
8696 ORF1ab A2811S	100
9565 ORF1ab silent	12
11288 ORF1ab del 9	154
13471 ORF1ab V2F	148
20724 ORF1ab silent	183
20956 ORF1ab L2497F	2314
21717 S Q52R	158
21762 S A67V	101
21765 S del 6	96
21991 S del 3	50
26767 M I82T	140
27205 ORF6 del 3	370
28278 N del 3	130
28308 N A12G	188
28699 N silent	52
28887 N T205I	122

Base change

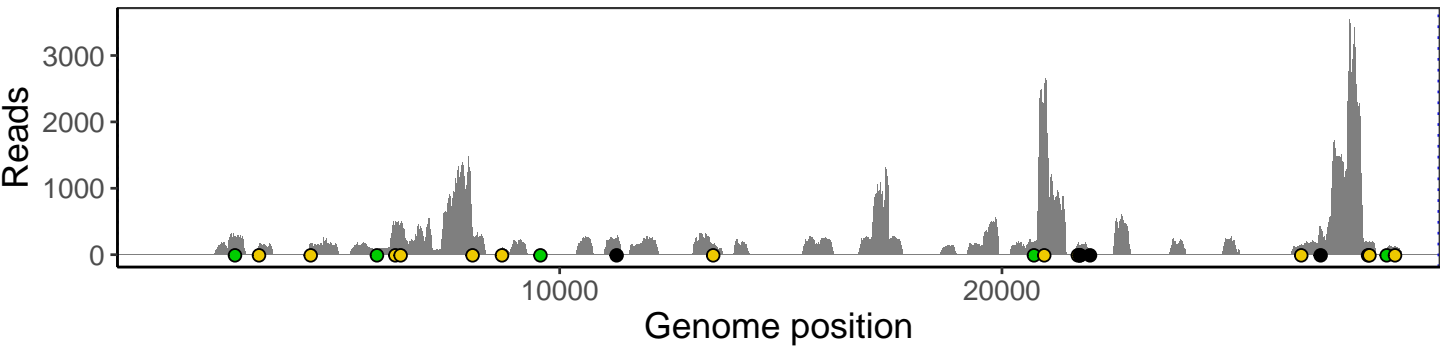


VSP2033-2

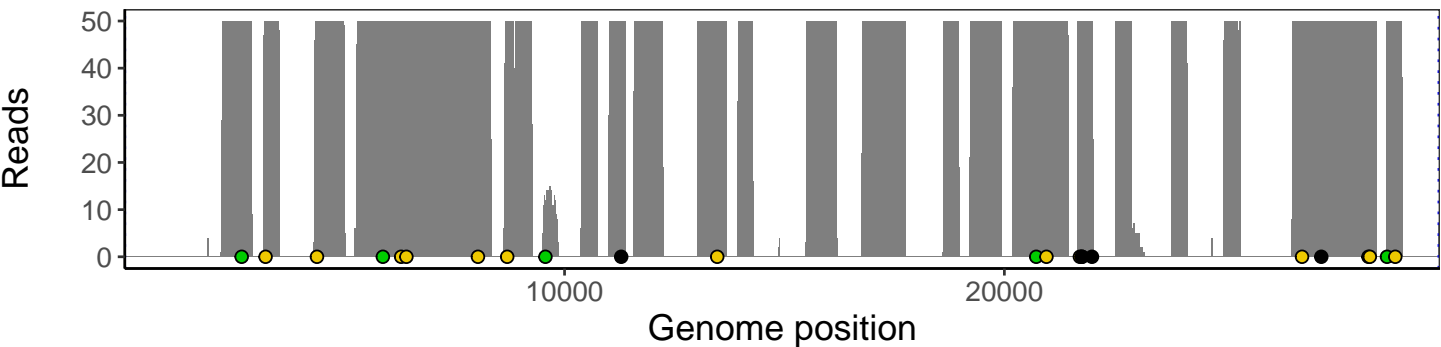
# Analyses of individual experiments and composite results

VSP2033-2 | 2021-03-26 | Saline | UPHS-0819 | genomes | single experiment

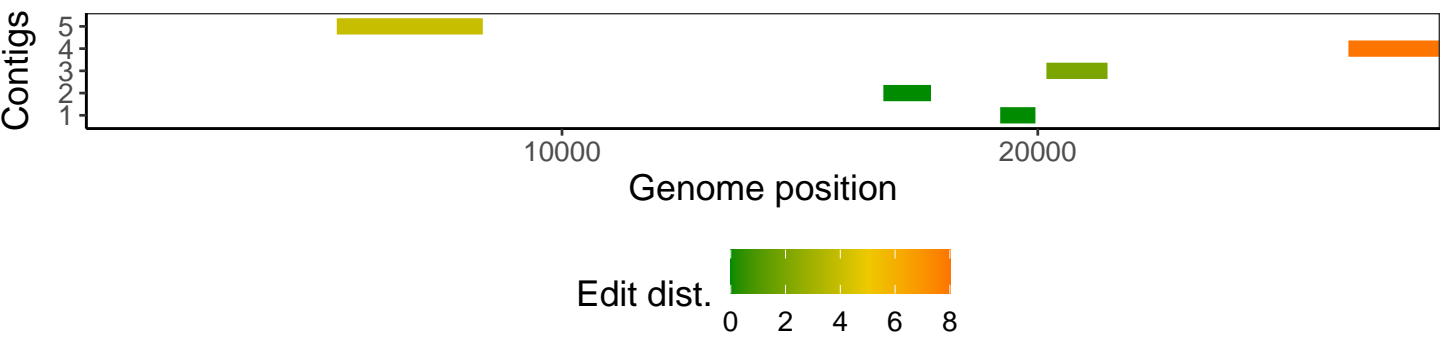
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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	3.1.3
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	3.3.3
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