

# COVID-19 subject UPHS-0449

*2021-06-01*

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.

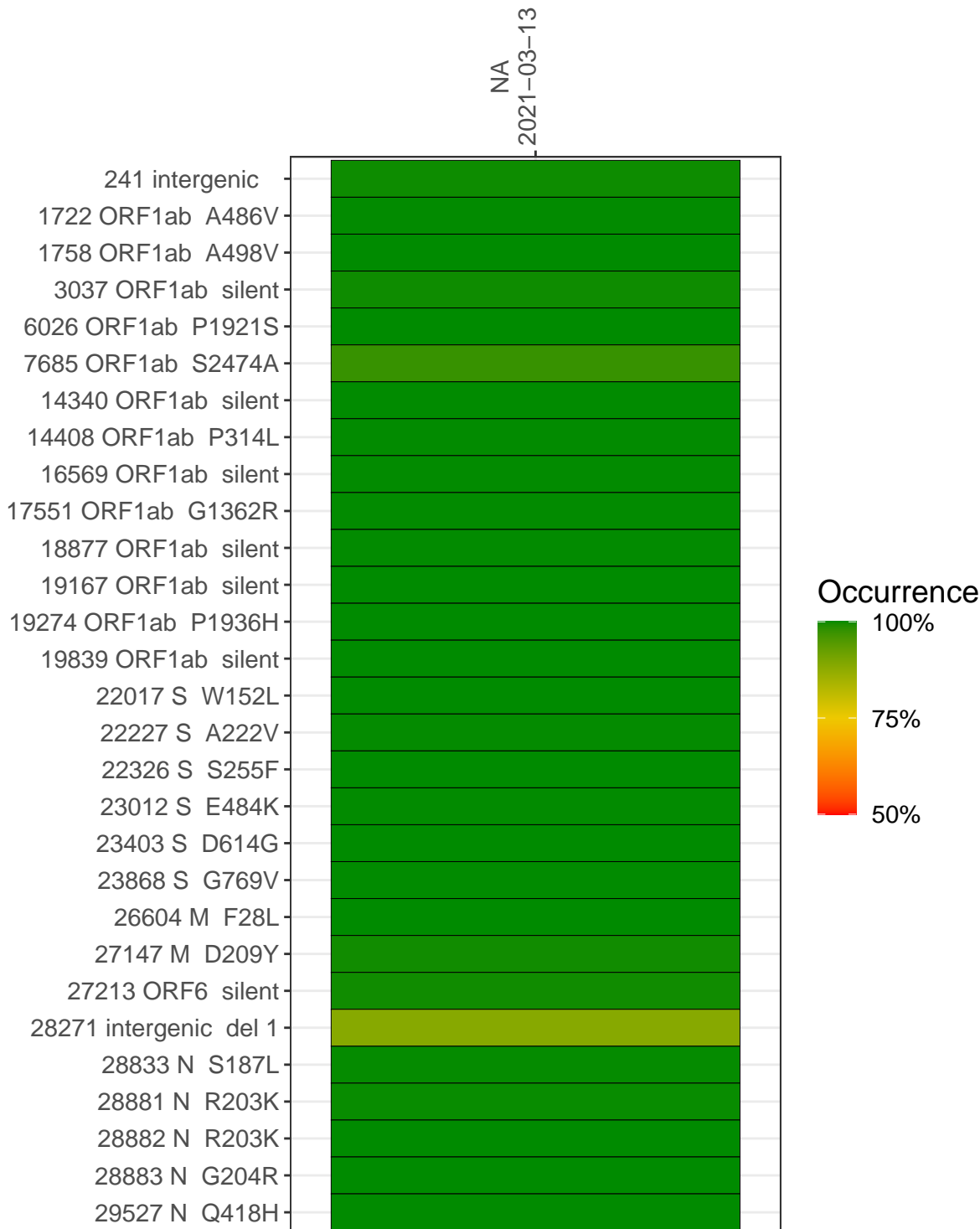
Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software tool (Rambaut et al 2020) for genomes with  $> 90\%$  sequence coverage.

Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage ( $\geq 5$ reads)
VSP1575-1	single experiment	NA	NA	2021-03-13	21.76	R.1	99.6%	99.1%

## 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.



	NA 2021-03-13	
241 intergenic	3445	
1722 ORF1ab A486V	2441	
1758 ORF1ab A498V	2317	
3037 ORF1ab silent	4064	
6026 ORF1ab P1921S	2610	
7685 ORF1ab S2474A	8578	
14340 ORF1ab silent	6294	
14408 ORF1ab P314L	7414	
16569 ORF1ab silent	6700	
17551 ORF1ab G1362R	9317	
18877 ORF1ab silent	13047	
19167 ORF1ab silent	11326	
19274 ORF1ab P1936H	9489	
19839 ORF1ab silent	8449	
22017 S W152L	1439	
22227 S A222V	3852	
22326 S S255F	227	
23012 S E484K	3336	
23403 S D614G	8973	
23868 S G769V	3574	
26604 M F28L	3335	
27147 M D209Y	8588	
27213 ORF6 silent	7319	
28271 intergenic del 1	4839	
28833 N S187L	834	
28881 N R203K	489	
28882 N R203K	486	
28883 N G204R	488	
29527 N Q418H	12286	
	VSP1575-1	

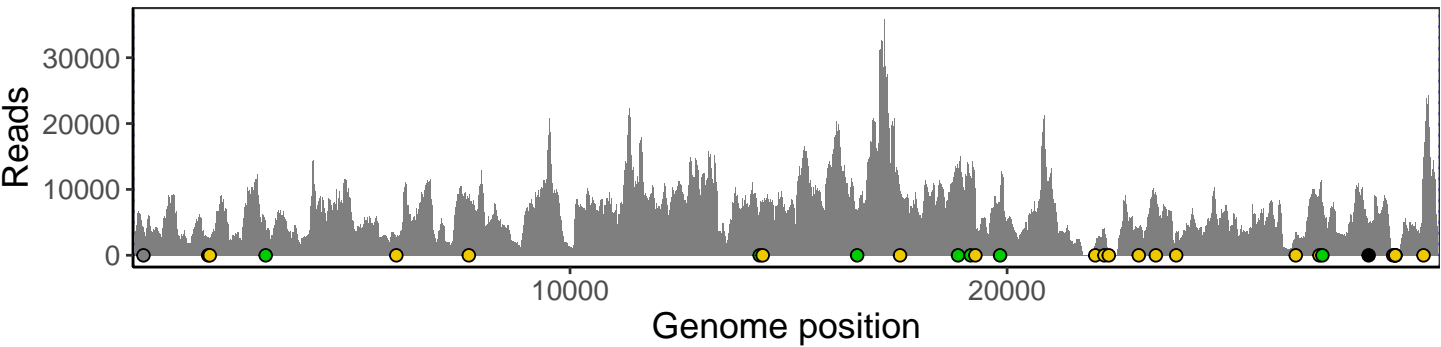
Base change

- Expected
- A
- T
- C
- G
- N
- Ins/Del
- No data

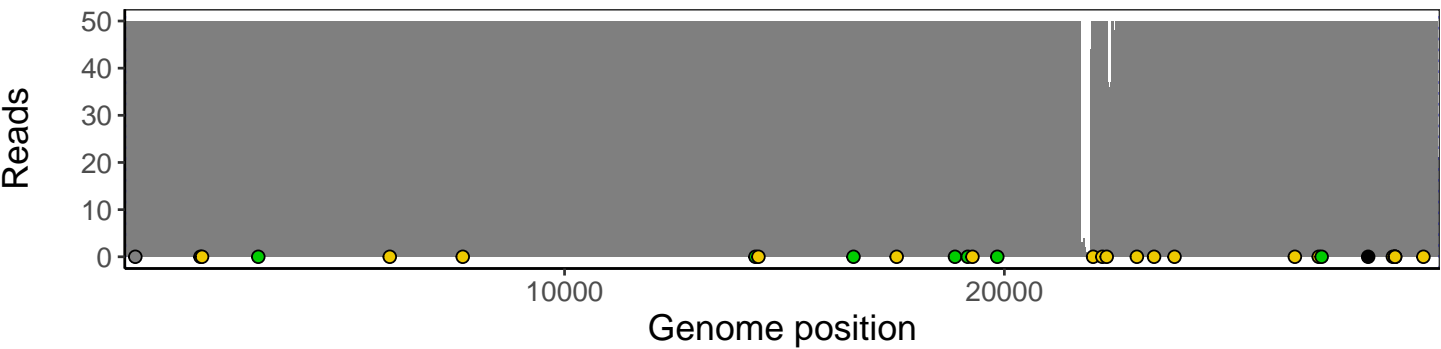
# Analyses of individual experiments and composite results

VSP1575-1 | 2021-03-13 | NA | UPHS-0449 | genomes | single experiment

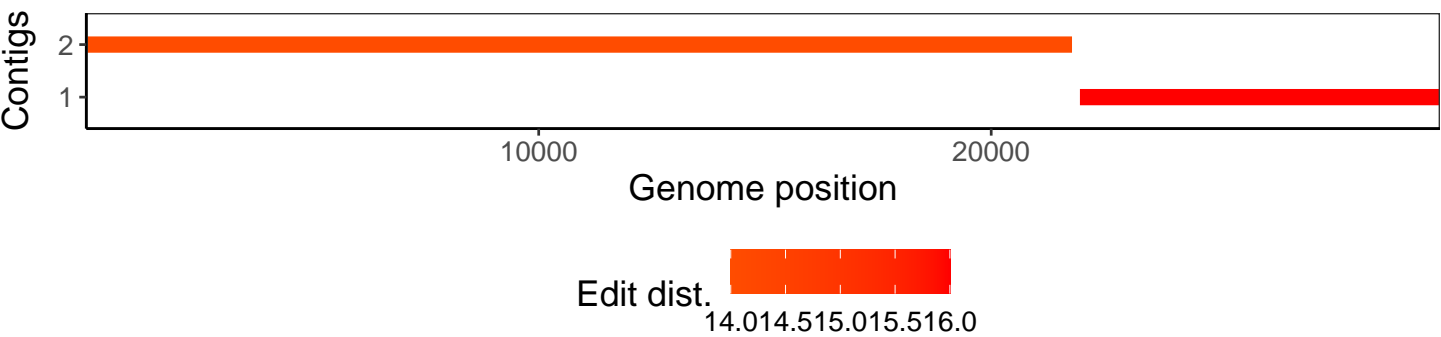
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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	3.3.3
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