# COVID-19 subject UPHS-0393

2021-05-05

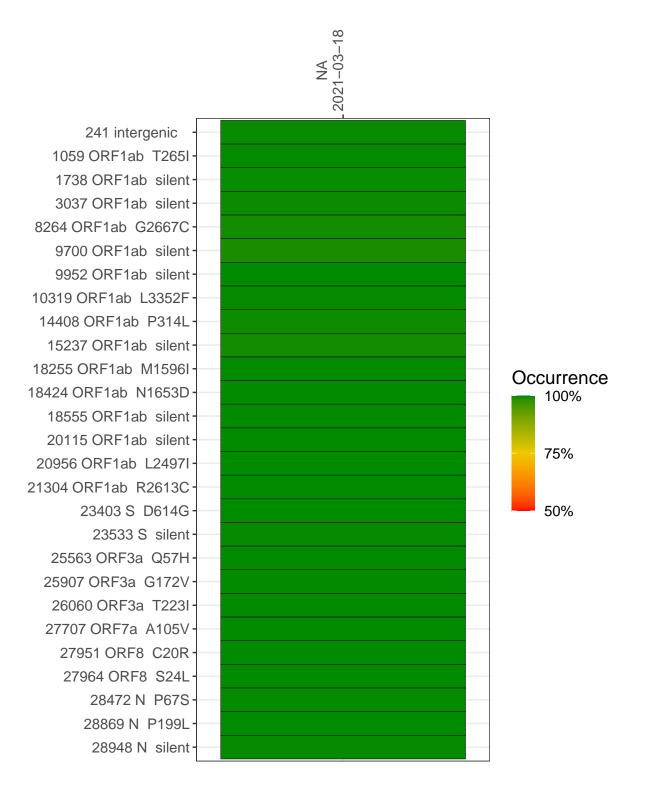
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software tool (Rambaut et al 2020) for genomes with > 90% sequence coverage.

Table 1. Sample summary.

Experiment	Туре	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP1520-1	single experiment	NA	NA	2021-03-18	22.40	B.1.2	99.8%	99.3%

#### 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.



#### NA 2021-03-18

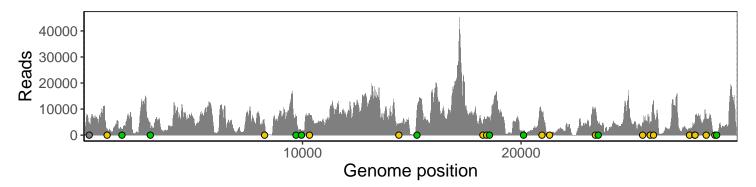
	2021-03-10
241 intergenic	4158
1059 ORF1ab T265I	2314
1738 ORF1ab silent	2069
3037 ORF1ab silent	2736
8264 ORF1ab G2667C	4284
9700 ORF1ab silent	7138
9952 ORF1ab silent	1430
10319 ORF1ab L3352F	8134
14408 ORF1ab P314L	3712
15237 ORF1ab silent	9647
18255 ORF1ab M1596I	4079
18424 ORF1ab N1653D	6067
18555 ORF1ab silent	6541
20115 ORF1ab silent	1678
20956 ORF1ab L2497I	9094
21304 ORF1ab R2613C	489
23403 S D614G	10249
23533 S silent	4087
25563 ORF3a Q57H	4692
25907 ORF3a G172V	4316
26060 ORF3a T223I	12855
27707 ORF7a A105V	2383
27951 ORF8 C20R	4572
27964 ORF8 S24L	6622
28472 N P67S	8007
28869 N P199L	733
28948 N silent	671
	520-1
	22



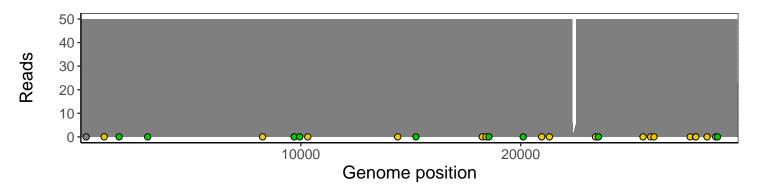
### Analyses of individual experiments and composite results

#### VSP1520-1 | 2021-03-18 | NA | UPHS-0393 | genomes | single experiment

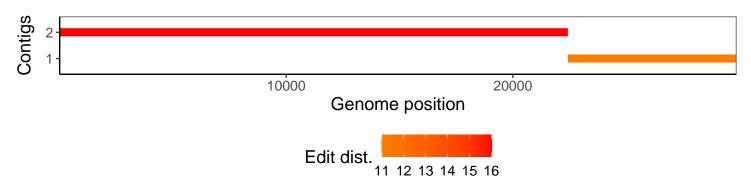
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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 htslib 1.10
bcftools	1.10.2-34-g1a12af0-dirty Using htslib 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
${\it Genomic Alignments}$	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
$\operatorname{GenomeInfoDb}$	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