# COVID-19 subject UPHS-0258

2021-05-05

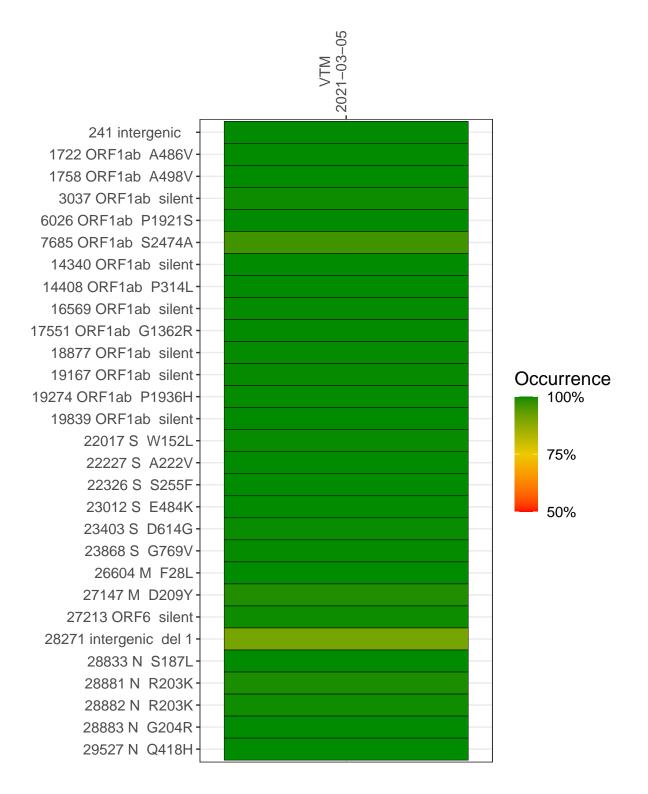
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP1303-1	single experiment	NA	VTM	2021-03-05	29.87	R.1	99.8%	99.7%

#### 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.



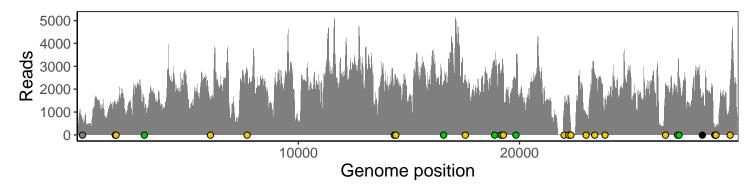
### VTM 2021-03-05

	2021-03-03
241 intergenic	599
1722 ORF1ab A486V	1315
1758 ORF1ab A498V	1183
3037 ORF1ab silent	1195
6026 ORF1ab P1921S	1440
7685 ORF1ab S2474A	2805
14340 ORF1ab silent	2099
14408 ORF1ab P314L	2269
16569 ORF1ab silent	2265
17551 ORF1ab G1362R	2408
18877 ORF1ab silent	3142
19167 ORF1ab silent	2718
19274 ORF1ab P1936H	2554
19839 ORF1ab silent	2372
22017 S W152L	1112
22227 S A222V	1609
22326 S S255F	239
23012 S E484K	679
23403 S D614G	2676
23868 S G769V	1669
26604 M F28L	2148
27147 M D209Y	1694
27213 ORF6 silent	2473
28271 intergenic del 1	1662
28833 N S187L	358
28881 N R203K	240
28882 N R203K	239
28883 N G204R	244
29527 N Q418H	2146
	303–1
	30

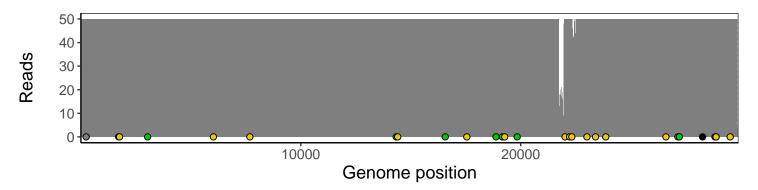
## Analyses of individual experiments and composite results

#### $VSP1303-1 \mid 2021-03-05 \mid VTM \mid UPHS-0258 \mid genomes \mid single experiment$

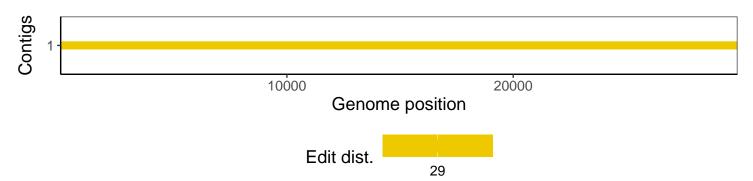
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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	3.0.0
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