COVID-19 subject UPHS-0263

2021-05-05

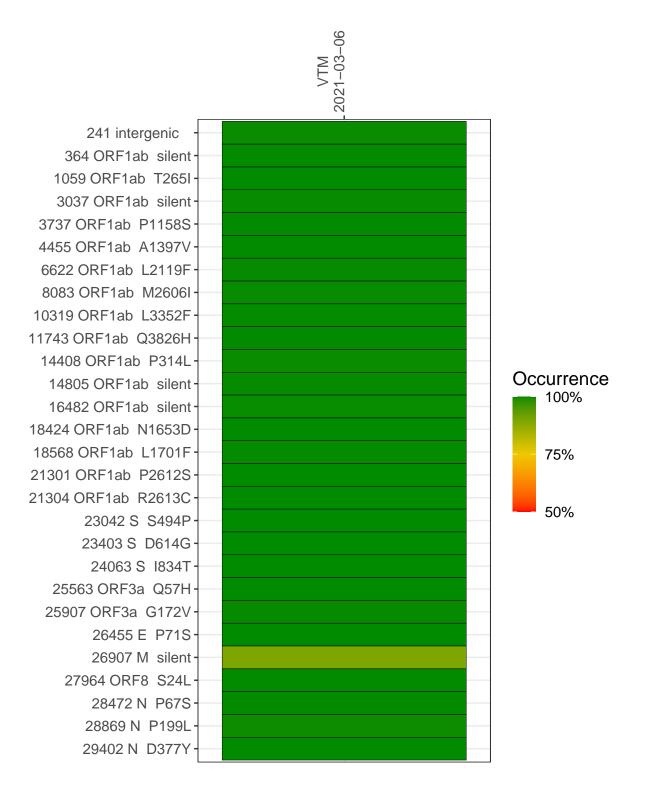
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP1308-1	single experiment	NA	VTM	2021-03-06	29.89	B.1.2	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 2021-03-06

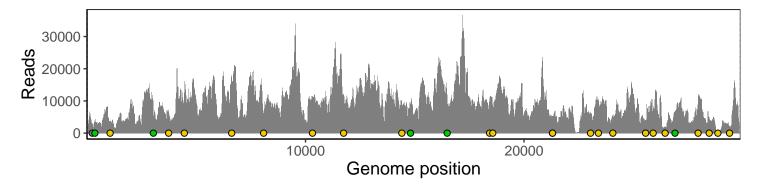
	2021-03-00
241 intergenic	2549
364 ORF1ab silent	3736
1059 ORF1ab T265I	3513
3037 ORF1ab silent	6258
3737 ORF1ab P1158S	6619
4455 ORF1ab A1397V	12004
6622 ORF1ab L2119F	13259
8083 ORF1ab M2606l	6346
10319 ORF1ab L3352F	10958
11743 ORF1ab Q3826H	11183
14408 ORF1ab P314L	8263
14805 ORF1ab silent	8870
16482 ORF1ab silent	10658
18424 ORF1ab N1653D	10090
18568 ORF1ab L1701F	8024
21301 ORF1ab P2612S	4275
21304 ORF1ab R2613C	4241
23042 S S494P	6393
23403 S D614G	9994
24063 S 1834T	4757
25563 ORF3a Q57H	4024
25907 ORF3a G172V	6522
26455 E P71S	1760
26907 M silent	6256
27964 ORF8 S24L	10488
28472 N P67S	4761
28869 N P199L	651
29402 N D377Y	2210
	08–1



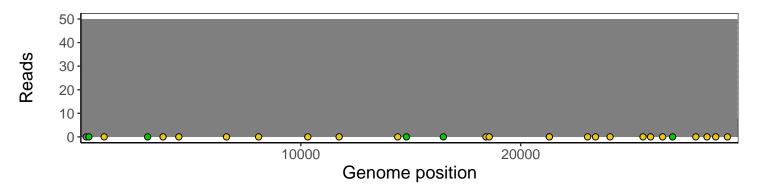
Analyses of individual experiments and composite results

$VSP1308-1 \mid 2021-03-06 \mid VTM \mid UPHS-0263 \mid genomes \mid single \; experiment$

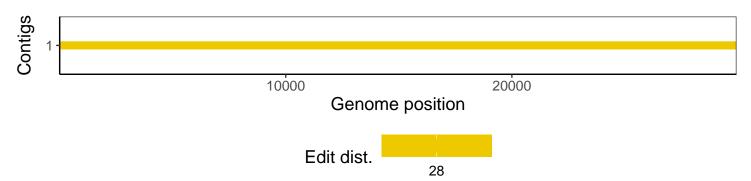
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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