COVID-19 subject UPHS-0816

2021-05-21

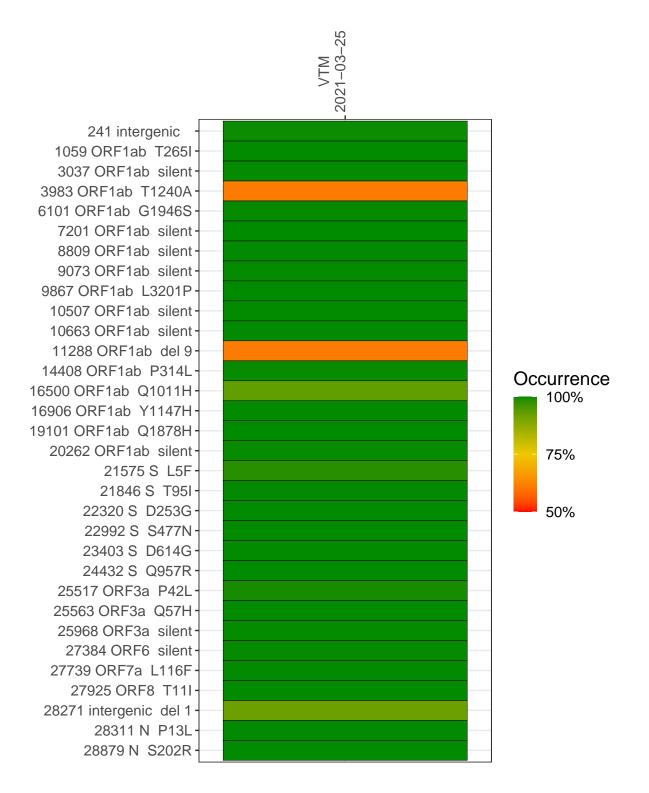
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP2030-2	single experiment	NA	VTM	2021-03-25	29.80	B.1.526.2	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 2021-03-25

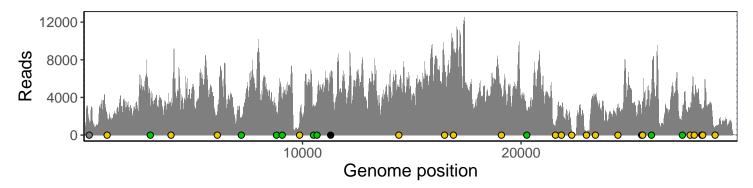
	2021-03-23
241 intergenic	1355
1059 ORF1ab T265I	1901
3037 ORF1ab silent	3285
3983 ORF1ab T1240A	3536
6101 ORF1ab G1946S	2384
7201 ORF1ab silent	1655
8809 ORF1ab silent	2435
9073 ORF1ab silent	3460
9867 ORF1ab L3201P	1274
10507 ORF1ab silent	2828
10663 ORF1ab silent	2681
11288 ORF1ab del 9	3476
14408 ORF1ab P314L	4303
16500 ORF1ab Q1011H	7265
16906 ORF1ab Y1147H	6847
19101 ORF1ab Q1878H	4846
20262 ORF1ab silent	2375
21575 S L5F	978
21846 S T95I	2641
22320 S D253G	437
22992 S S477N	305
23403 S D614G	4160
24432 S Q957R	2807
25517 ORF3a P42L	2062
25563 ORF3a Q57H	3518
25968 ORF3a silent	4172
27384 ORF6 silent	2233
27739 ORF7a L116F	2477
27925 ORF8 T11I	3774
28271 intergenic del 1	2773
28311 N P13L	2620
28879 N S202R	855
	1-2
	2030–2



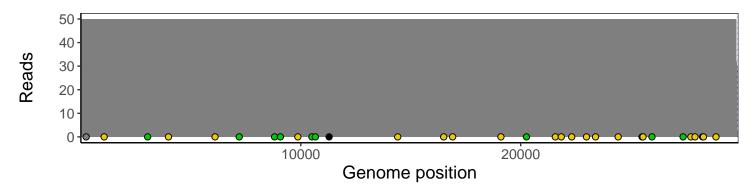
Analyses of individual experiments and composite results

$VSP2030\text{-}2 \mid 2021\text{-}03\text{-}25 \mid VTM \mid UPHS\text{-}0816 \mid genomes \mid single \ experiment$

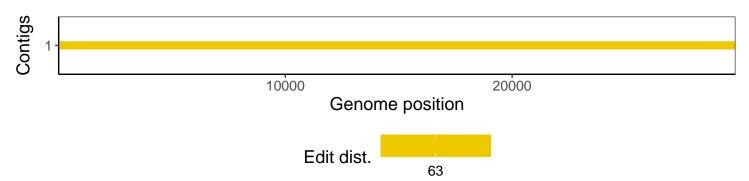
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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	3.3.3
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