COVID-19 subject UPHS-0824

2021-05-21

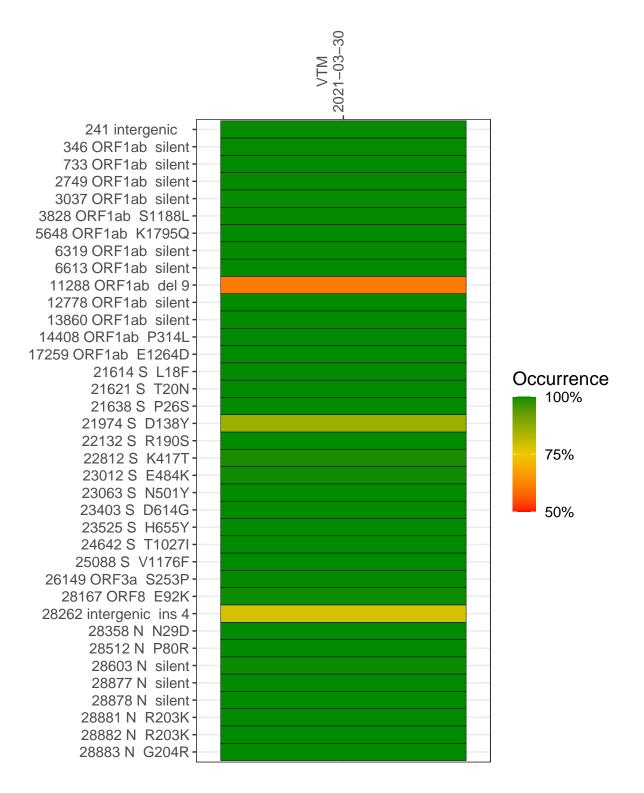
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP2038-2	single experiment	NA	VTM	2021-03-30	29.84	P.1	99.7%	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.



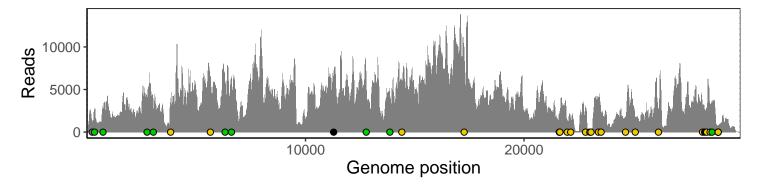
VTM 2021-03-30

	2021-03-30
241 intergenic	1151
346 ORF1ab silent	2618
733 ORF1ab silent	2825
2749 ORF1ab silent	4022
3037 ORF1ab silent	2926
3828 ORF1ab S1188L	3737
5648 ORF1ab K1795Q	6005
6319 ORF1ab silent	5701
6613 ORF1ab silent	6316
11288 ORF1ab del 9	3682
12778 ORF1ab silent	6534
13860 ORF1ab silent	4343
14408 ORF1ab P314L	5081
17259 ORF1ab E1264D	9012
21614 S L18F	2076
21621 S T20N	2119
21638 S P26S	2407
21974 S D138Y	1083
22132 S R190S	1634
22812 S K417T	2811
23012 S E484K	259
23063 S N501Y	411
23403 S D614G	4009
23525 S H655Y	2033
24642 S T1027I	1837
25088 S V1176F	1674
26149 ORF3a S253P	2644
28167 ORF8 E92K	2782
28262 intergenic ins 4	2204
28358 N N29D	2764
28512 N P80R	3232
28603 N silent	3456
28877 N silent	353
28878 N silent	351
28881 N R203K	351
28882 N R203K	351
28883 N G204R	360
	7
	938
	VSP2038-2
	<u>⊗</u>
	r

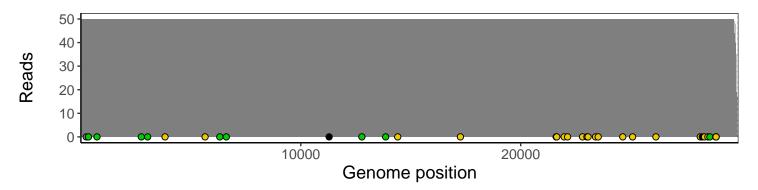
Analyses of individual experiments and composite results

$VSP2038-2 \mid 2021\text{-}03\text{-}30 \mid VTM \mid UPHS\text{-}0824 \mid genomes \mid single \ experiment$

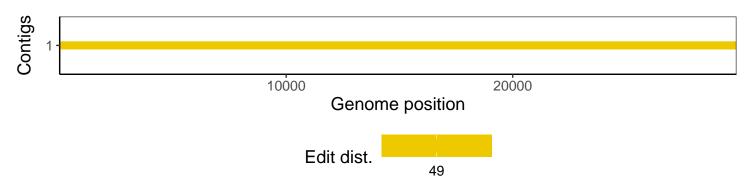
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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