COVID-19 subject UPHS-1233

2021-06-23

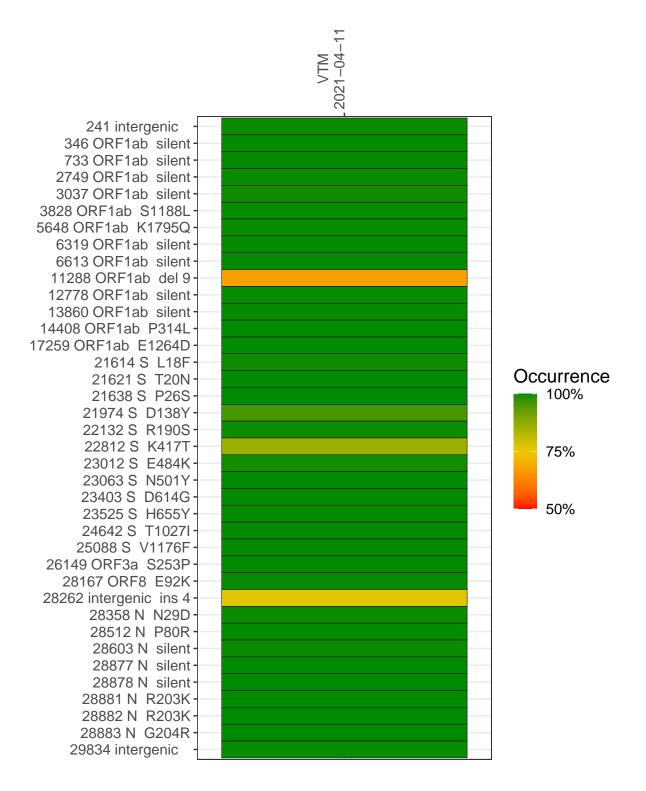
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP2487-1	single experiment	NA	VTM	2021-04-11	29.84	P.1	99.8%	99.7%

Variants shared across samples

The heat map below shows how variants (reference genome /home/common/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.



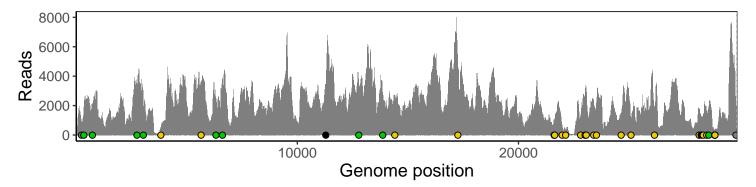
VTM 2021-04-11

	2021-04-11
241 intergenic	870
346 ORF1ab silent	1810
733 ORF1ab silent	1892
2749 ORF1ab silent	3292
3037 ORF1ab silent	2038
3828 ORF1ab S1188L	516
5648 ORF1ab K1795Q	3168
6319 ORF1ab silent	2406
6613 ORF1ab silent	3721
11288 ORF1ab del 9	1932
12778 ORF1ab silent	3614
13860 ORF1ab silent	2660
14408 ORF1ab P314L	2319
17259 ORF1ab E1264D	4795
21614 S L18F	961
21621 S T20N	932
21638 S P26S	1042
21974 S D138Y	656
22132 S R190S	338
22812 S K417T	1205
23012 S E484K	2378
23063 S N501Y	3179
23403 S D614G	2309
23525 S H655Y	2138
24642 S T1027I	1472
25088 S V1176F	1133
26149 ORF3a S253P	1853
28167 ORF8 E92K	1594
28262 intergenic ins 4	987
28358 N N29D	1375
28512 N P80R	1548
28603 N silent	1827
28877 N silent	181
28878 N silent	181
28881 N R203K	181
28882 N R203K	181
28883 N G204R	183
29834 intergenic	2048
	187
	VSP2487-1
	$\overline{\otimes}$

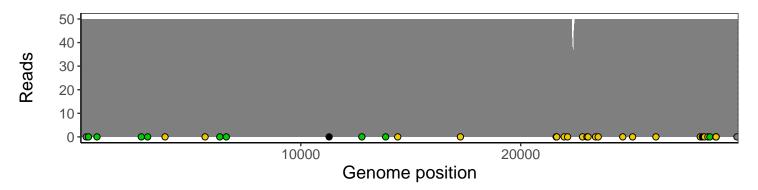
Analyses of individual experiments and composite results

$VSP2487\text{-}1 \mid 2021\text{-}04\text{-}11 \mid VTM \mid UPHS\text{-}1233 \mid genomes \mid single \ experiment$

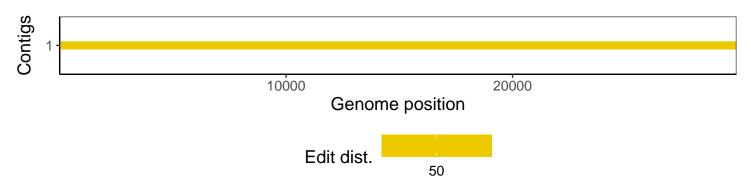
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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	3.1.3				
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				
GenomicAlignments	1.12.2				
${\bf Summarized Experiment}$	1.6.5				
DelayedArray	0.2.7				
matrixStats	0.54.0				
Biobase	2.36.2				
Rsamtools	1.28.0				
GenomicRanges	1.28.6				
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				