COVID-19 subject UPHS-0825

2021-06-23

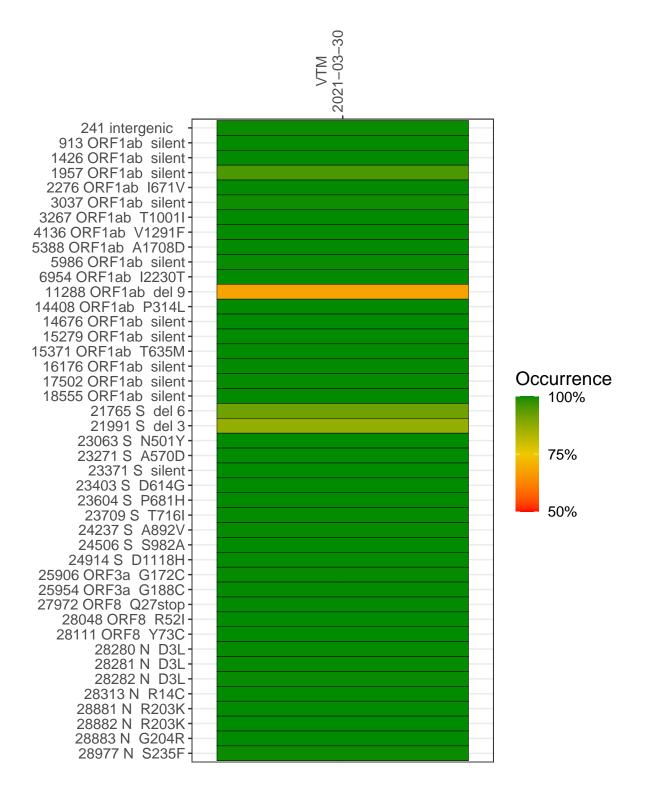
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP2039-2	single experiment	NA	VTM	2021-03-30	29.80	B.1.1.7	99.7%	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 2021-03-30

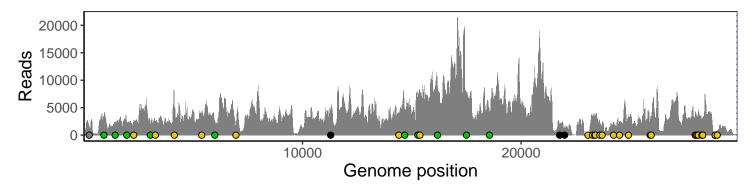
	2021-03-30
241 intergenic	1400
913 ORF1ab silent	3304
1426 ORF1ab silent	2219
1957 ORF1ab silent	1975
2276 ORF1ab I671V	2581
3037 ORF1ab silent	2050
3267 ORF1ab T1001I	2648
4136 ORF1ab V1291F	5406
5388 ORF1ab A1708D	3416
5986 ORF1ab silent	2488
6954 ORF1ab I2230T	1360
11288 ORF1ab del 9	2803
14408 ORF1ab P314L	3609
14676 ORF1ab silent	3768
15279 ORF1ab silent	6150
15371 ORF1ab T635M	7350
16176 ORF1ab silent	8788
17502 ORF1ab silent	6479
18555 ORF1ab silent	4192
21765 S del 6	1646
21991 S del 3	1223
23063 S N501Y	69
23271 S A570D	3218
23371 S silent	3318
23403 S D614G	3589
23604 S P681H	2426
23709 S T716I	1935
24237 S A892V	2284
24506 S S982A	2428
24914 S D1118H	3579
25906 ORF3a G172C	2309
25954 ORF3a G188C	3407
27972 ORF8 Q27stop	4198
28048 ORF8 R52I	2815
28111 ORF8 Y73C	4194
28280 N D3L	1696
28281 N D3L	1696
28282 N D3L	1840
28313 N R14C	3076
28881 N R203K	395
28882 N R203K	394
28883 N G204R	395
28977 N S235F	752
	-5
	939
	20
	VSP2039-2
	>



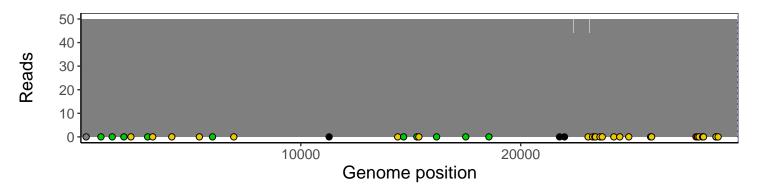
Analyses of individual experiments and composite results

$VSP2039-2 \mid 2021-03-30 \mid VTM \mid UPHS-0825 \mid genomes \mid single \; experiment$

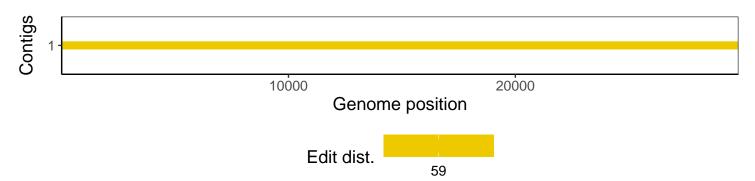
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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				