COVID-19 subject UPHS-0141

2021-04-17

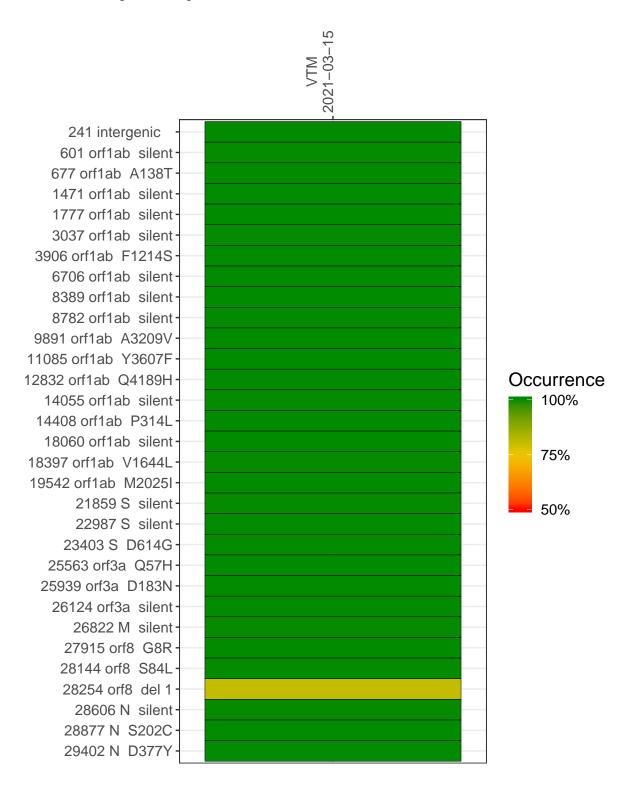
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software tool (Rambaut et al 2020) for genomes with > 90% sequence coverage.

Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP1126-1	single experiment	NA	VTM	2021-03-15	29.85	B.1.110.3	99.9%	99.8%

Variants shared across samples

The heat map below shows how variants (reference genome /home/everett/projects/SARS-CoV-2-Philadelphia/USA-WA1-2020) 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 2021-03-15

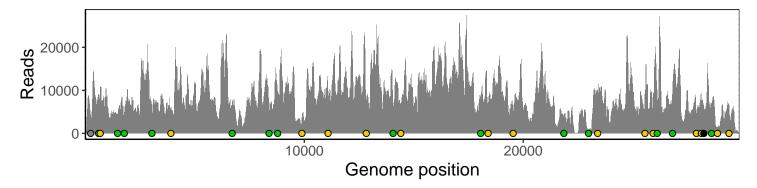
	2021-00-10
241 intergenic	3791
601 orf1ab silent	8143
677 orf1ab A138T	9777
1471 orf1ab silent	4431
1777 orf1ab silent	4521
3037 orf1ab silent	4983
3906 orf1ab F1214S	5881
6706 orf1ab silent	6775
8389 orf1ab silent	6928
8782 orf1ab silent	9019
9891 orf1ab A3209V	3695
11085 orf1ab Y3607F	8421
12832 orf1ab Q4189H	7491
14055 orf1ab silent	7085
14408 orf1ab P314L	5906
18060 orf1ab silent	8287
18397 orf1ab V1644L	6706
19542 orf1ab M2025I	5767
21859 S silent	3938
22987 S silent	761
23403 S D614G	9997
25563 orf3a Q57H	9096
25939 orf3a D183N	8326
26124 orf3a silent	9075
26822 M silent	9384
27915 orf8 G8R	4391
28144 orf8 S84L	7665
28254 orf8 del 1	5678
28606 N silent	6682
28877 N S202C	1481
29402 N D377Y	6755
	$\overline{\mathbf{x}}$



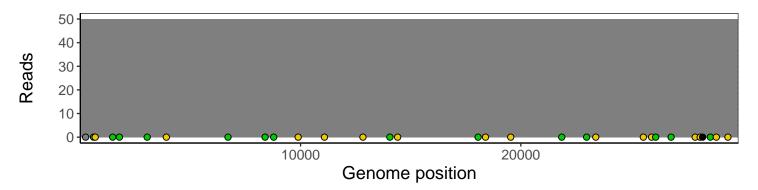
Analyses of individual experiments and composite results

VSP1126-1 | 2021-03-15 | VTM | UPHS-0141 | genomes | single experiment

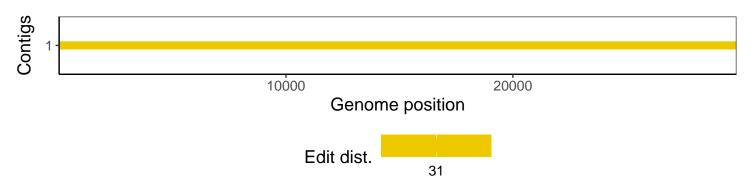
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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