COVID-19 subject UPHS-1616

2021-06-03

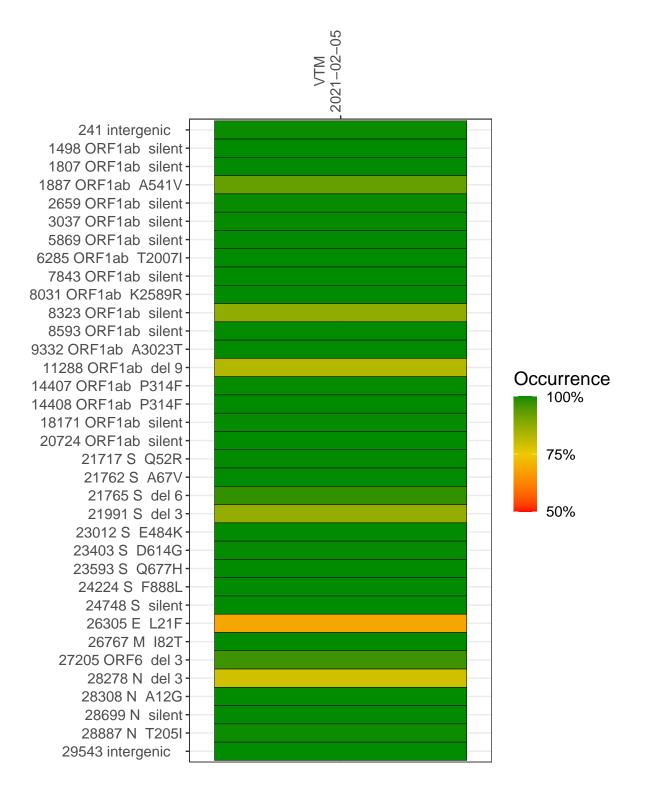
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP2917-1	single experiment	NA	VTM	2021-02-05	29.75	B.1.525	99.7%	99.5%

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 2021-02-05

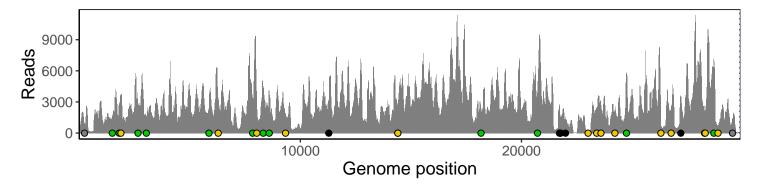
	2021-02-03
241 intergenic	837
1498 ORF1ab silent	1356
1807 ORF1ab silent	2690
1887 ORF1ab A541V	3527
2659 ORF1ab silent	2837
3037 ORF1ab silent	1401
5869 ORF1ab silent	2369
6285 ORF1ab T2007I	2910
7843 ORF1ab silent	3273
8031 ORF1ab K2589R	3557
8323 ORF1ab silent	4172
8593 ORF1ab silent	3806
9332 ORF1ab A3023T	1797
11288 ORF1ab del 9	2140
14407 ORF1ab P314F	2186
14408 ORF1ab P314F	2232
18171 ORF1ab silent	2304
20724 ORF1ab silent	3776
21717 S Q52R	2564
21762 S A67V	994
21765 S del 6	972
21991 S del 3	1297
23012 S E484K	72
23403 S D614G	2344
23593 S Q677H	2647
24224 S F888L	2150
24748 S silent	5480
26305 E L21F	1072
26767 M 182T	2137
27205 ORF6 del 3	3359
28278 N del 3	1669
28308 N A12G	2092
28699 N silent	6698
28887 N T205I	686
29543 intergenic	1406
	7-1
	1~



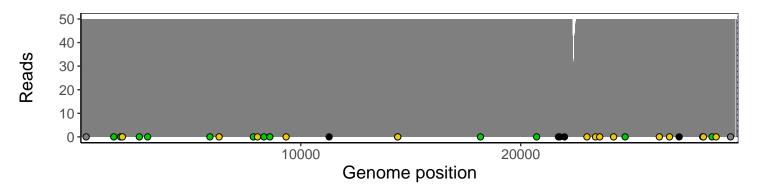
Analyses of individual experiments and composite results

$VSP2917\text{-}1 \mid 2021\text{-}02\text{-}05 \mid VTM \mid UPHS\text{-}1616 \mid genomes \mid single \ experiment$

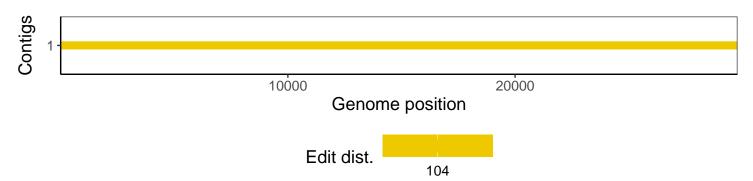
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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