COVID-19 subject UPHS-1665

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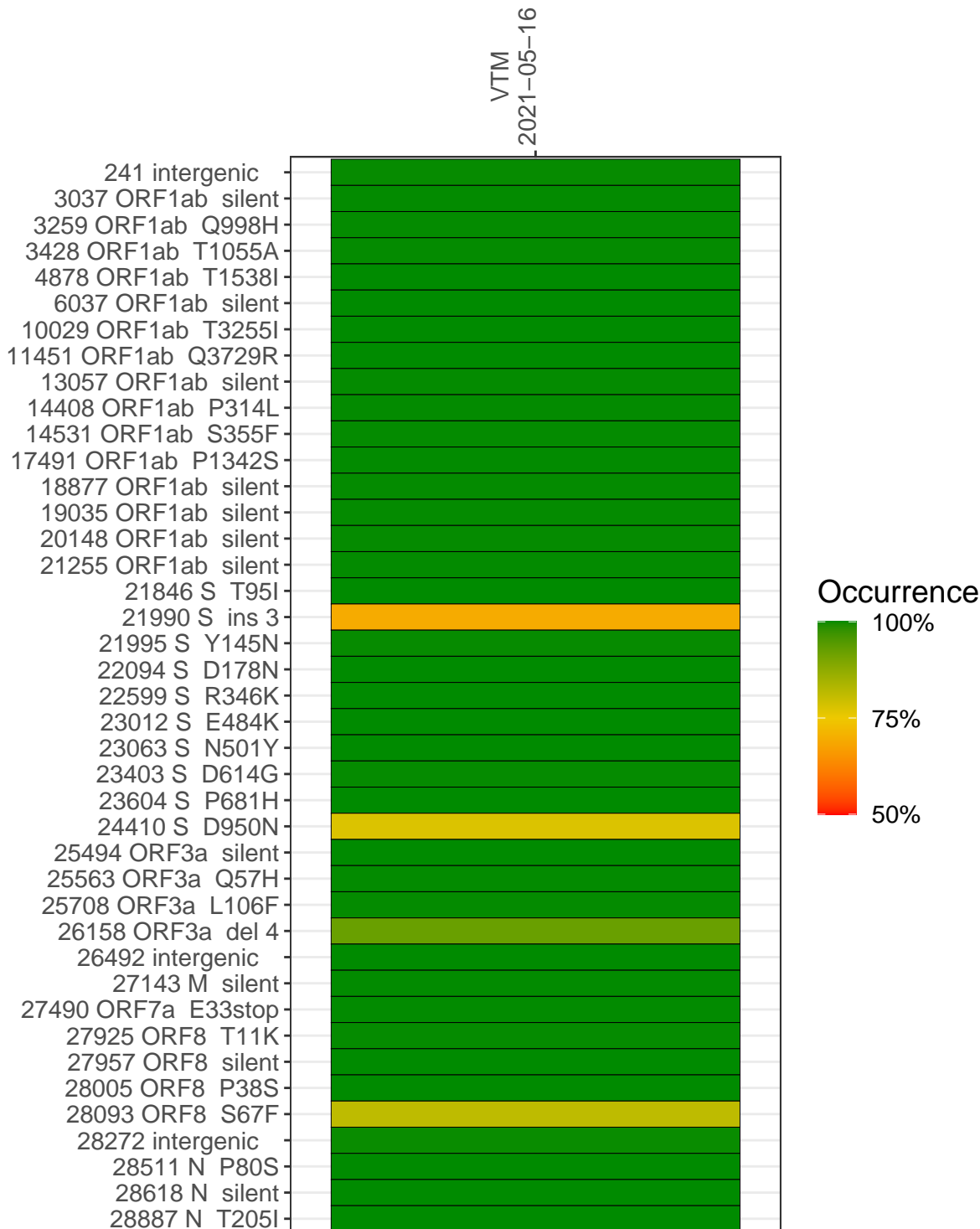 sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software tool (Rambaut et al 2020) for genomes with > 90% sequence coverage.
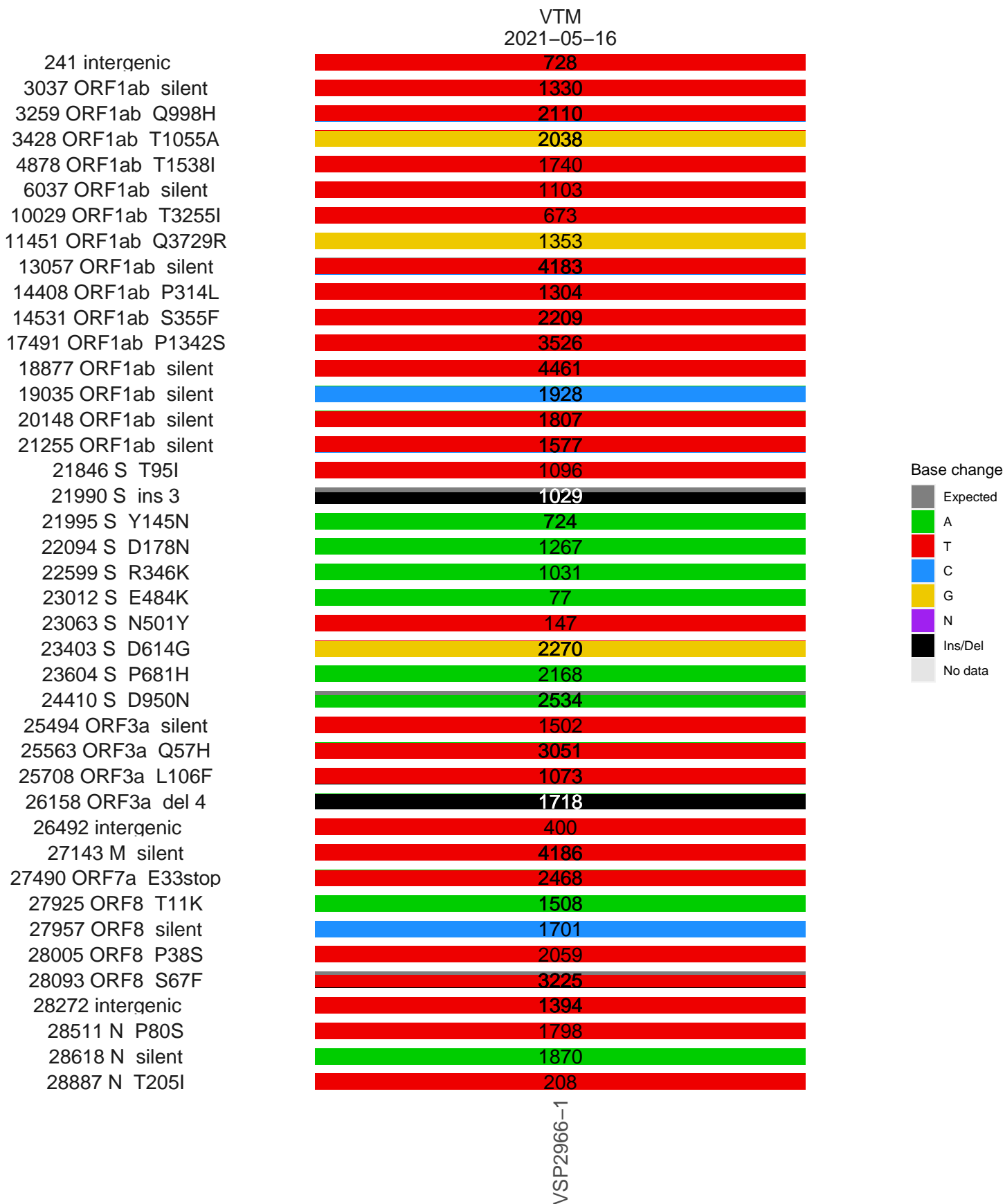
Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP2966-1	single experiment	NA	VTM	2021-05-16	29.86	B.1.621	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
2021−05−16

Variant	VSP2966−1
241 intergenic	728
3037 ORF1ab silent	1330
3259 ORF1ab Q998H	2110
3428 ORF1ab T1055A	2038
4878 ORF1ab T1538I	1740
6037 ORF1ab silent	1103
10029 ORF1ab T3255I	673
11451 ORF1ab Q3729R	1353
13057 ORF1ab silent	4183
14408 ORF1ab P314L	1304
14531 ORF1ab S355F	2209
17491 ORF1ab P1342S	3526
18877 ORF1ab silent	4461
19035 ORF1ab silent	1928
20148 ORF1ab silent	1807
21255 ORF1ab silent	1577
21846 S T95I	1096
21990 S ins 3	1029
21995 S Y145N	724
22094 S D178N	1267
22599 S R346K	1031
23012 S E484K	77
23063 S N501Y	147
23403 S D614G	2270
23604 S P681H	2168
24410 S D950N	2534
25494 ORF3a silent	1502
25563 ORF3a Q57H	3051
25708 ORF3a L106F	1073
26158 ORF3a del 4	1718
26492 intergenic	400
27143 M silent	4186
27490 ORF7a E33stop	2468
27925 ORF8 T11K	1508
27957 ORF8 silent	1701
28005 ORF8 P38S	2059
28093 ORF8 S67F	3225
28272 intergenic	1394
28511 N P80S	1798
28618 N silent	1870
28887 N T205I	208

Base change

Color	Label
Expected	
A	
T	
C	
G	
N	
Ins/Del	
No data	

Analyses of individual experiments and composite results

## VSP2966-1 | 2021-05-16 | VTM | UPHS-1665 | 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.
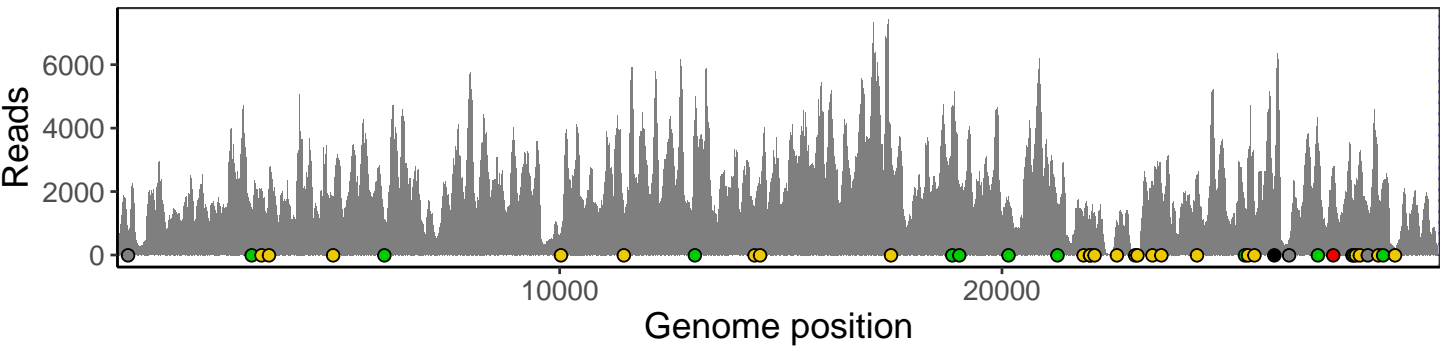
Excerpt from plot above focusing on reads coverage from 0 to 50 NT.
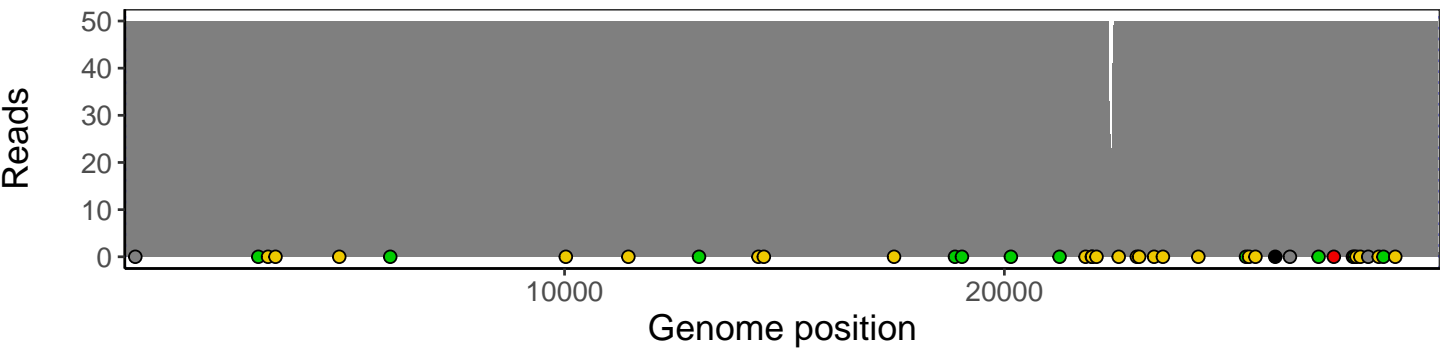
The longest five assembled contigs are shown below colored by their edit distance to the reference genome.

4
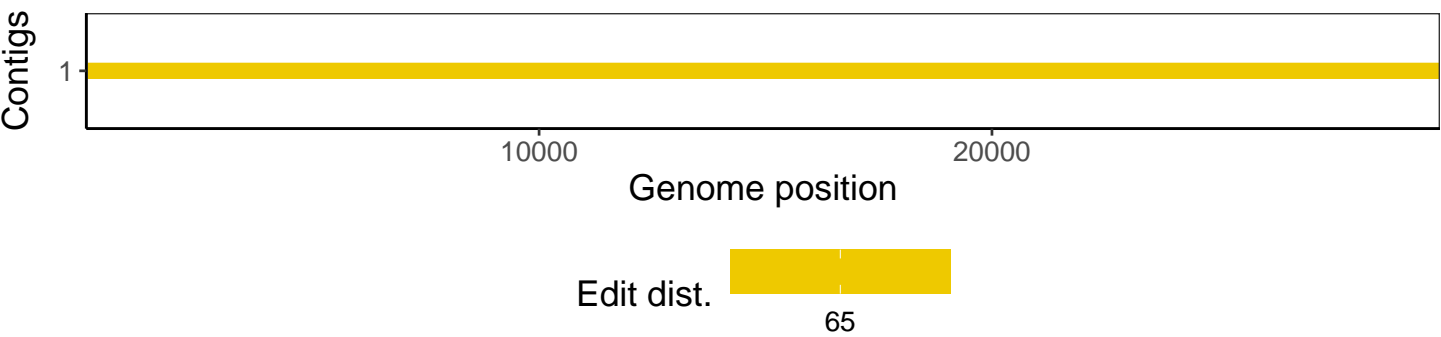
Software environment

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
SummarizedExperiment	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