COVID-19 subject UPHS-0154

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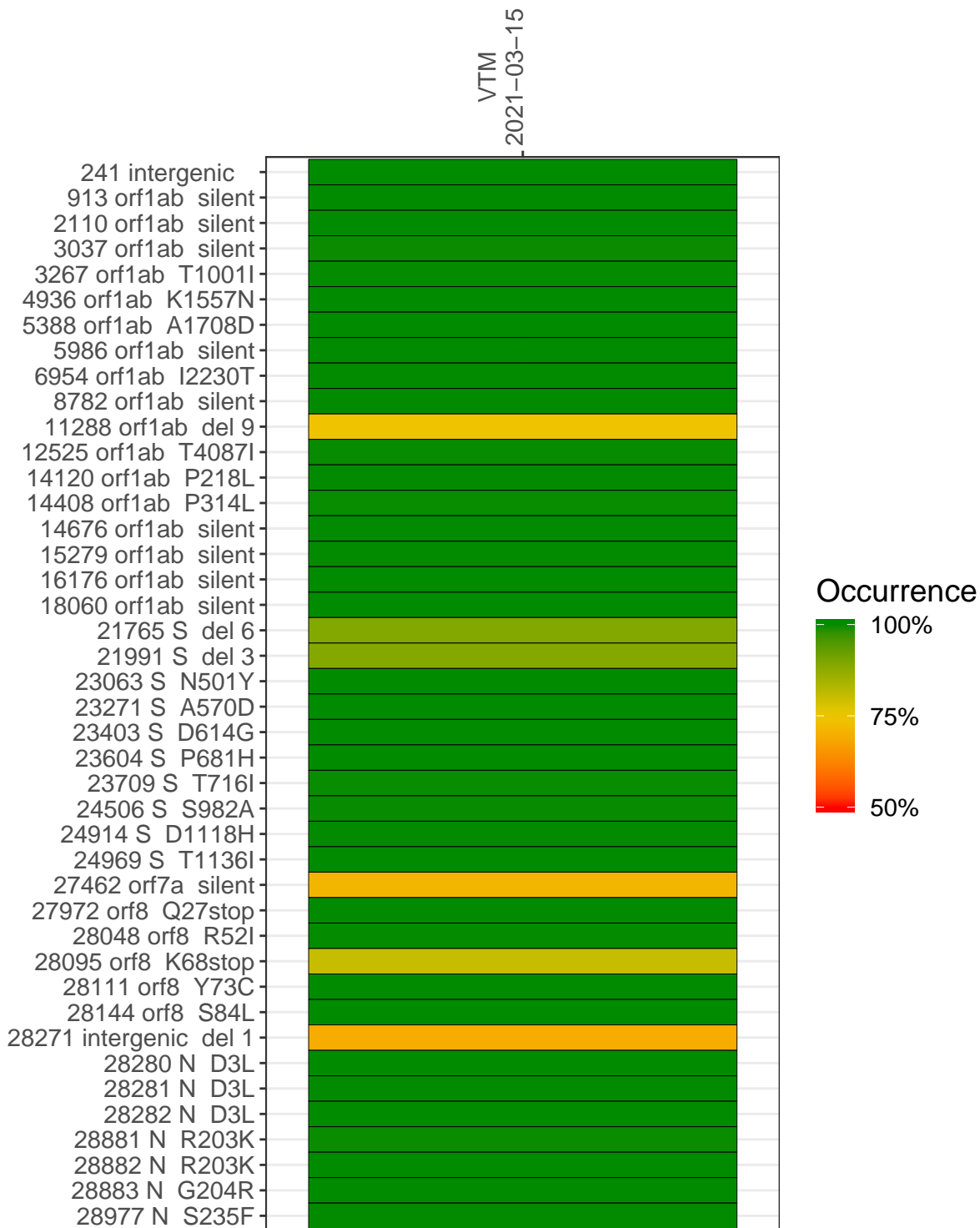 sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software tool (Rambaut et al 2020) for genomes with > 90% sequence coverage.
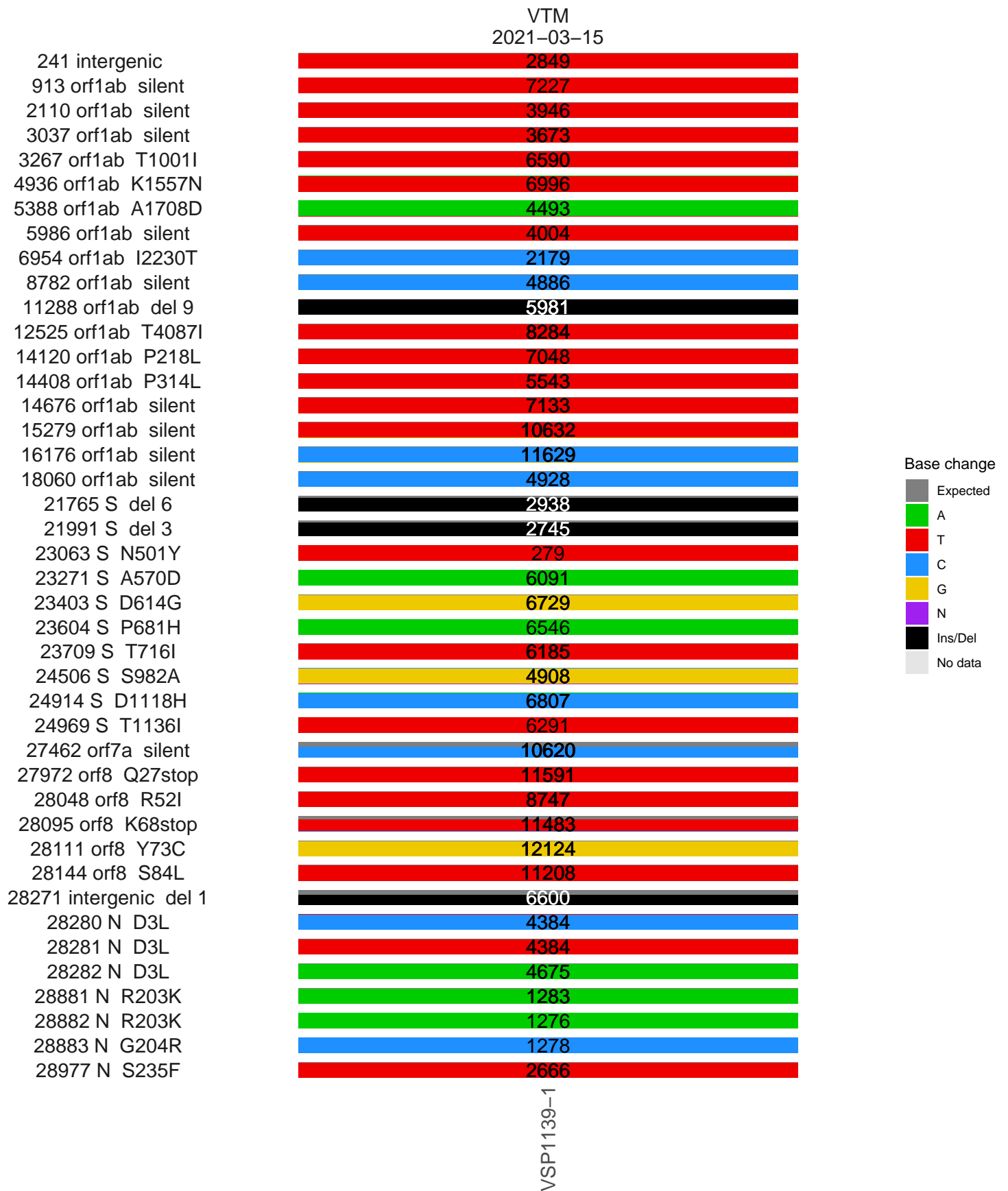
Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP1139-1	single experiment	NA	VTM	2021-03-15	29.91	B.1.1.7	99.9%	99.9%

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

Position	Value
241 intergenic	2849
913 orf1ab silent	7227
2110 orf1ab silent	3946
3037 orf1ab silent	3673
3267 orf1ab T1001I	6590
4936 orf1ab K1557N	6996
5388 orf1ab A1708D	4493
5986 orf1ab silent	4004
6954 orf1ab I2230T	2179
8782 orf1ab silent	4886
11288 orf1ab del 9	5981
12525 orf1ab T4087I	8284
14120 orf1ab P218L	7048
14408 orf1ab P314L	5543
14676 orf1ab silent	7133
15279 orf1ab silent	10632
16176 orf1ab silent	11629
18060 orf1ab silent	4928
21765 S del 6	2938
21991 S del 3	2745
23063 S N501Y	279
23271 S A570D	6091
23403 S D614G	6729
23604 S P681H	6546
23709 S T716I	6185
24506 S S982A	4908
24914 S D1118H	6807
24969 S T1136I	6291
27462 orf7a silent	10620
27972 orf8 Q27stop	11591
28048 orf8 R52I	8747
28095 orf8 K68stop	11483
28111 orf8 Y73C	12124
28144 orf8 S84L	11208
28271 intergenic del 1	6600
28280 N D3L	4384
28281 N D3L	4384
28282 N D3L	4675
28881 N R203K	1283
28882 N R203K	1276
28883 N G204R	1278
28977 N S235F	2666

VSP1139−1

Base change
- Expected
- A
- T
- C
- G
- N
- Ins/Del
- No data

3

Analyses of individual experiments and composite results

VSP1139-1 | 2021-03-15 | VTM | UPHS-0154 | 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.
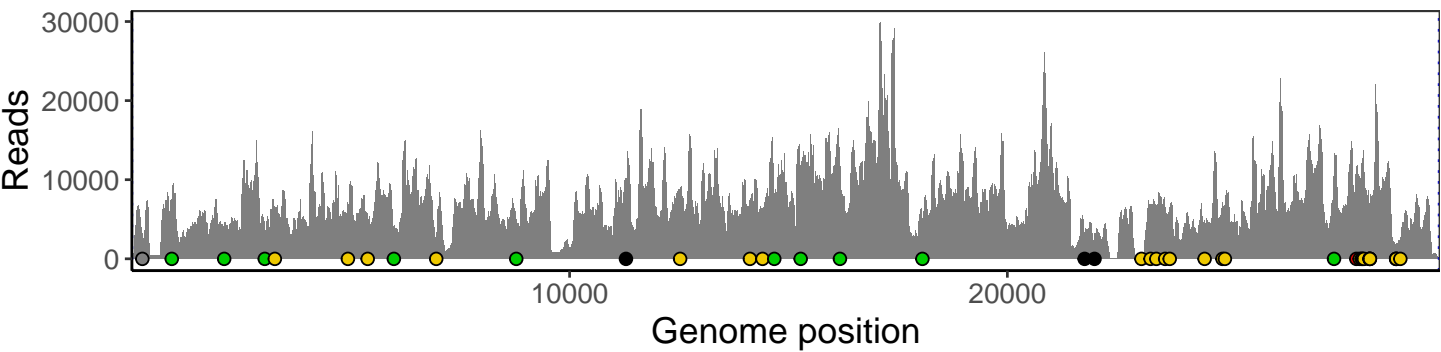
Excerpt from plot above focusing on reads coverage from 0 to 50 NT.
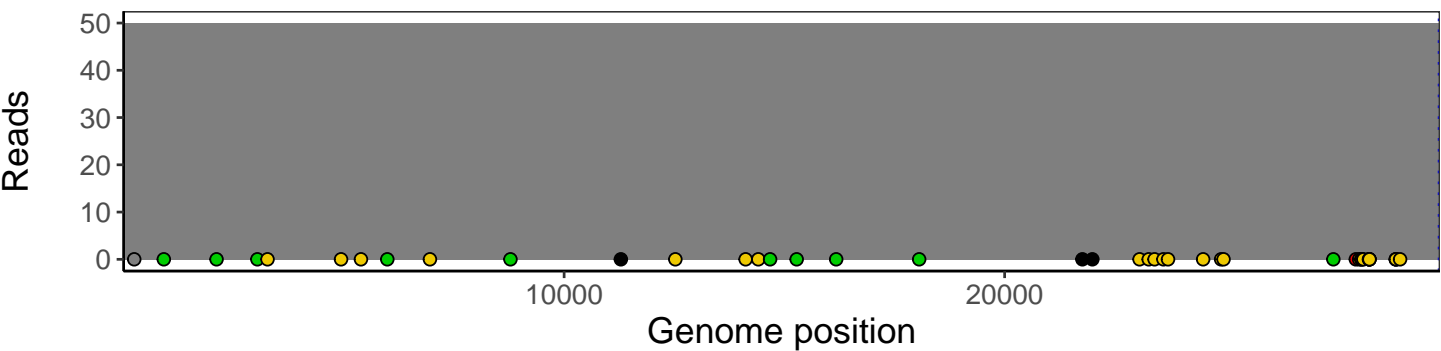
The longest five assembled contigs are shown below colored by their edit distance to the reference genome.

Edit dist.   57
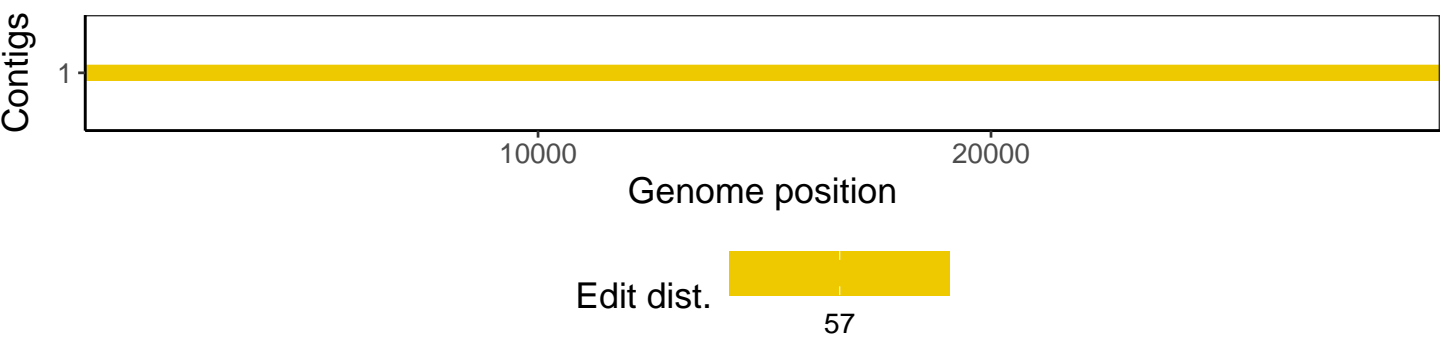
Software environment

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
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