COVID-19 subject UPHS-0252

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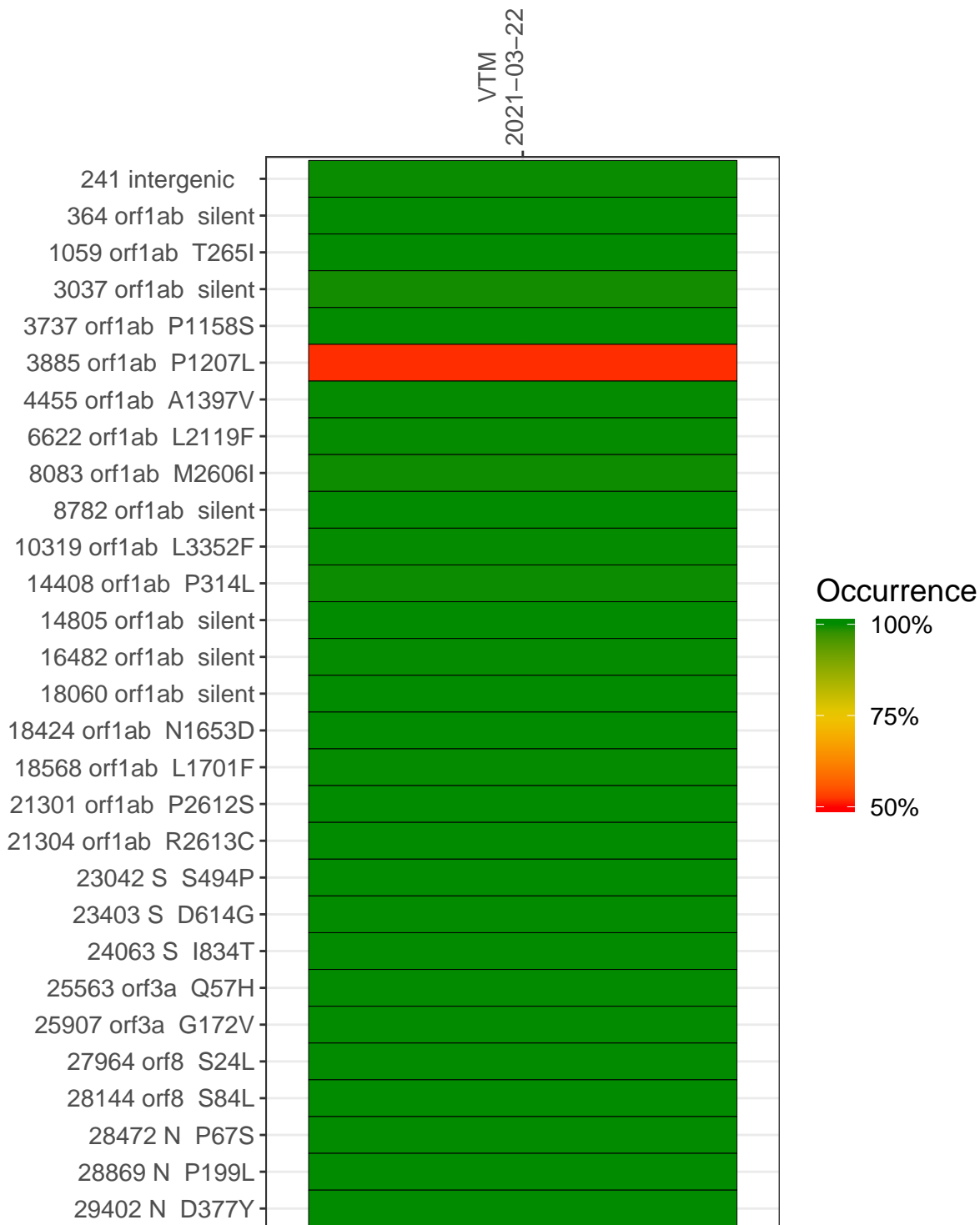 sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software tool (Rambaut et al 2020) for genomes with > 90% sequence coverage.
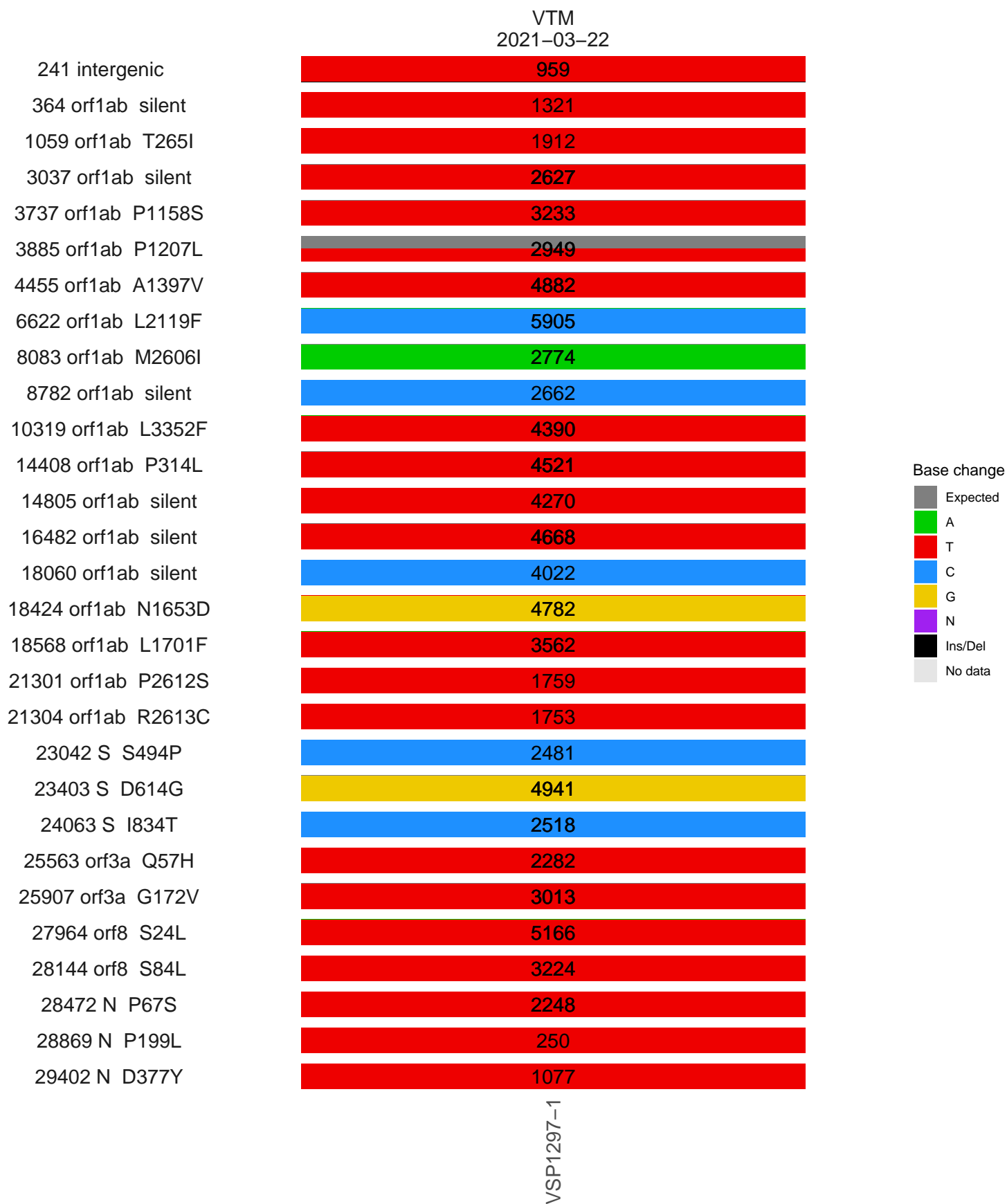
Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP1297-1	single experiment	NA	VTM	2021-03-22	29.89	B.1.2	99.8%	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
2021−03−22

Position	VSP1297−1
241 intergenic	959
364 orf1ab silent	1321
1059 orf1ab T265I	1912
3037 orf1ab silent	2627
3737 orf1ab P1158S	3233
3885 orf1ab P1207L	2949
4455 orf1ab A1397V	4882
6622 orf1ab L2119F	5905
8083 orf1ab M2606I	2774
8782 orf1ab silent	2662
10319 orf1ab L3352F	4390
14408 orf1ab P314L	4521
14805 orf1ab silent	4270
16482 orf1ab silent	4668
18060 orf1ab silent	4022
18424 orf1ab N1653D	4782
18568 orf1ab L1701F	3562
21301 orf1ab P2612S	1759
21304 orf1ab R2613C	1753
23042 S S494P	2481
23403 S D614G	4941
24063 S I834T	2518
25563 orf3a Q57H	2282
25907 orf3a G172V	3013
27964 orf8 S24L	5166
28144 orf8 S84L	3224
28472 N P67S	2248
28869 N P199L	250
29402 N D377Y	1077

Base change

Expected	(grey)
A	(green)
T	(red)
C	(blue)
G	(yellow)
N	(purple)
Ins/Del	(black)
No data	(light grey)

3

Analyses of individual experiments and composite results

**VSP1297-1 | 2021-03-22 | VTM | UPHS-0252 | 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.
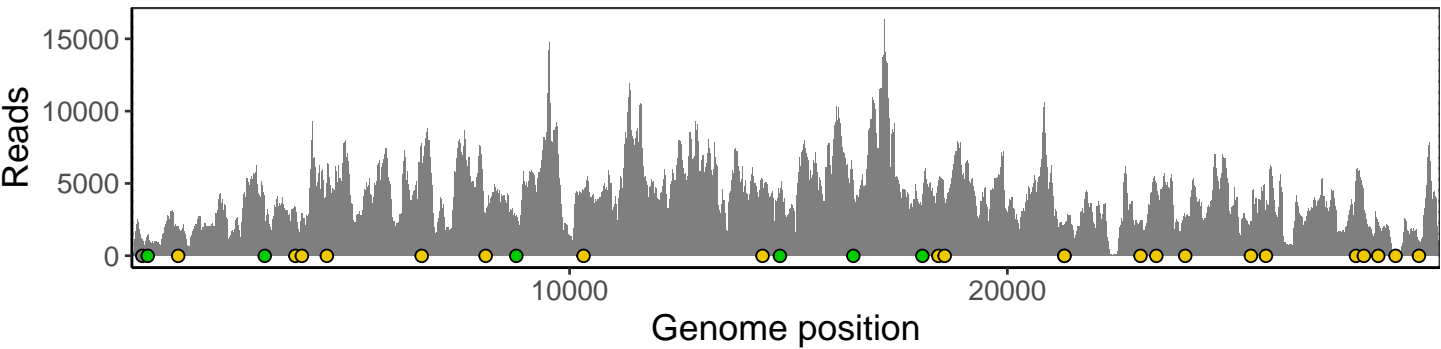
Excerpt from plot above focusing on reads coverage from 0 to 50 NT.
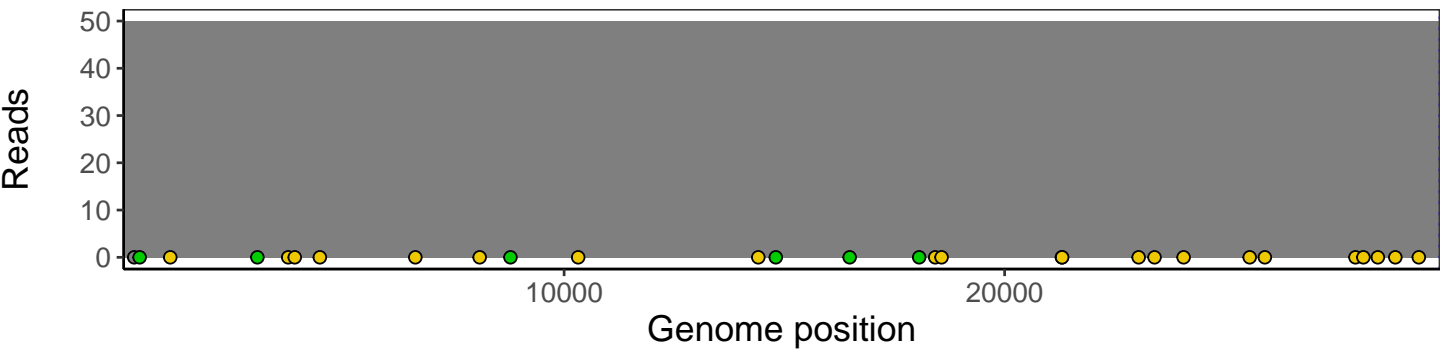
The longest five assembled contigs are shown below colored by their edit distance to the reference genome.
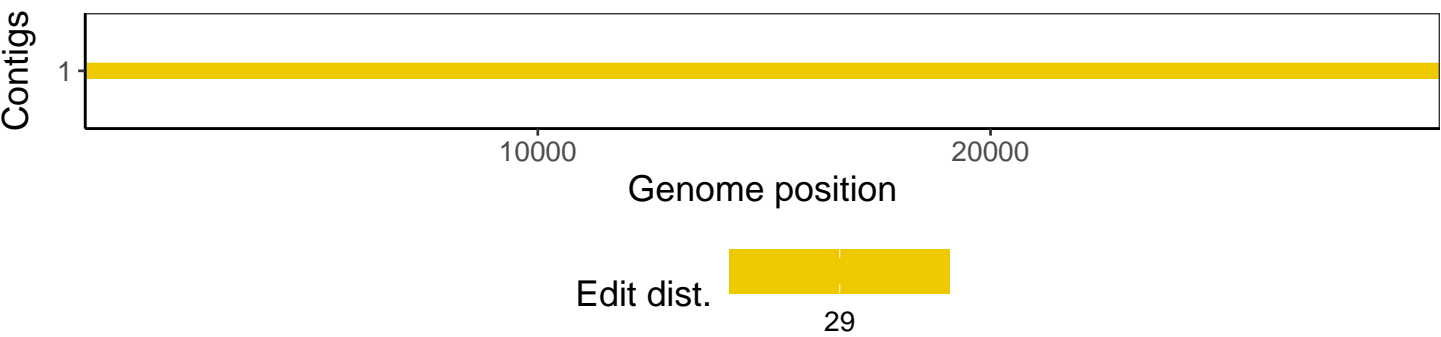
Software environment

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
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