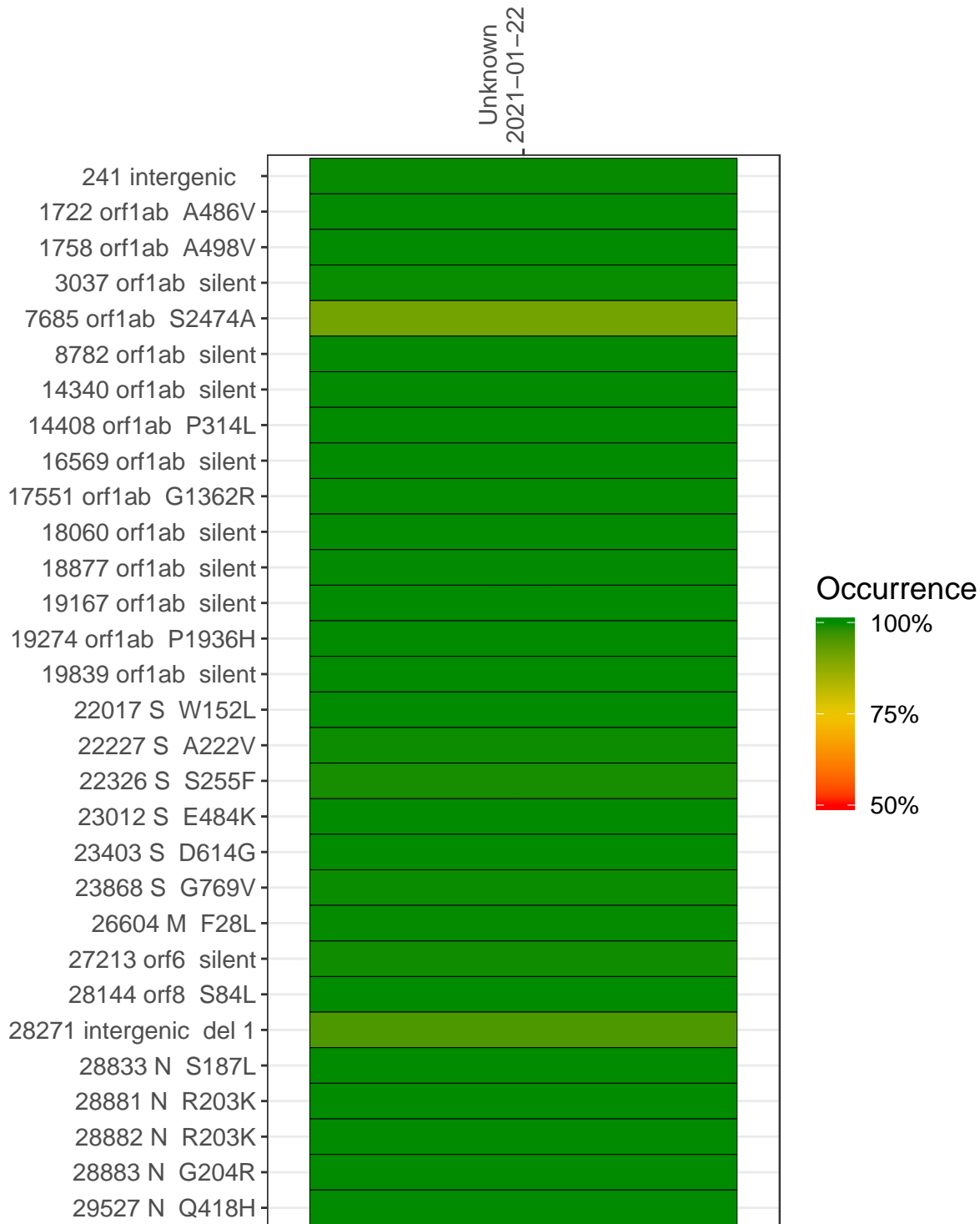# COVID-19 subject molpath-seq1

2021-01-31

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.

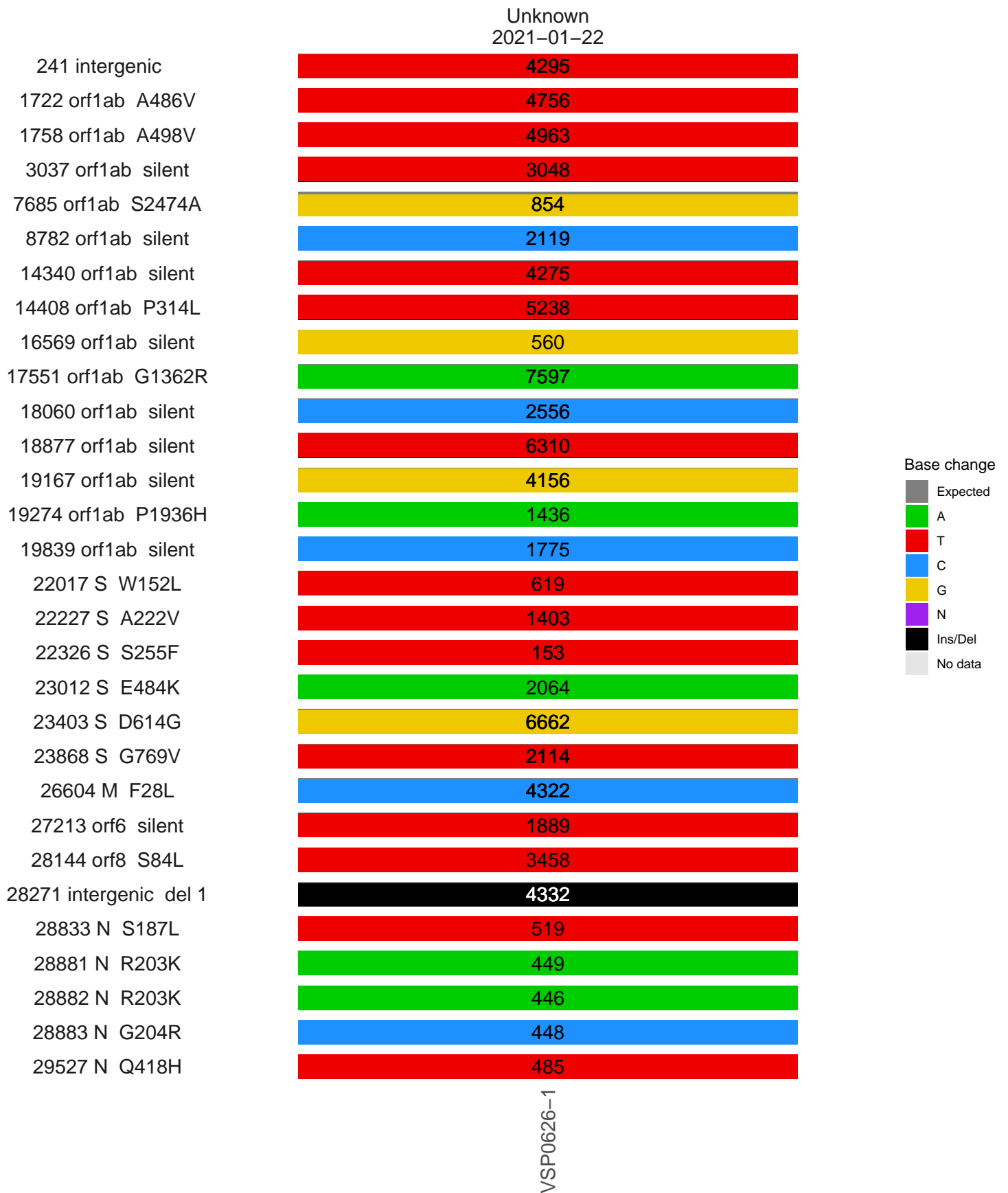Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP0626-1	single experiment	NA	Unknown	2021-01-22	21.48	B.1.1	99.0%	99.0%

Variants shared across samples

The heat map below shows how variants (reference genome 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.

Unknown
2021−01−22

Position	Value
241 intergenic	4295
1722 orf1ab A486V	4756
1758 orf1ab A498V	4963
3037 orf1ab silent	3048
7685 orf1ab S2474A	854
8782 orf1ab silent	2119
14340 orf1ab silent	4275
14408 orf1ab P314L	5238
16569 orf1ab silent	560
17551 orf1ab G1362R	7597
18060 orf1ab silent	2556
18877 orf1ab silent	6310
19167 orf1ab silent	4156
19274 orf1ab P1936H	1436
19839 orf1ab silent	1775
22017 S W152L	619
22227 S A222V	1403
22326 S S255F	153
23012 S E484K	2064
23403 S D614G	6662
23868 S G769V	2114
26604 M F28L	4322
27213 orf6 silent	1889
28144 orf8 S84L	3458
28271 intergenic del 1	4332
28833 N S187L	519
28881 N R203K	449
28882 N R203K	446
28883 N G204R	448
29527 N Q418H	485

VSP0626−1

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

3

Analyses of individual experiments and composite results

VSP0626-1 | 2021-01-22 | Unknown | molpath-seq1 | 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.
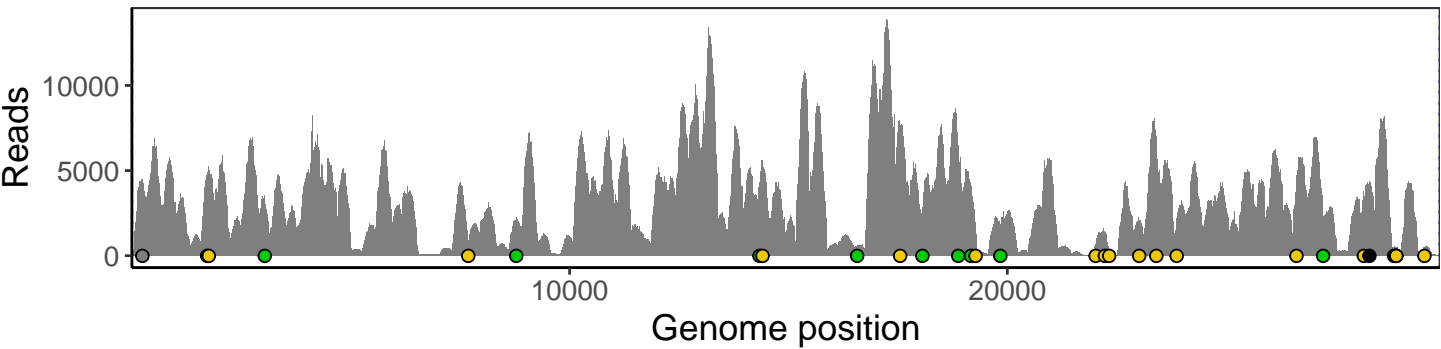
Excerpt from plot above focusing on reads coverage from 0 to 50 NT.
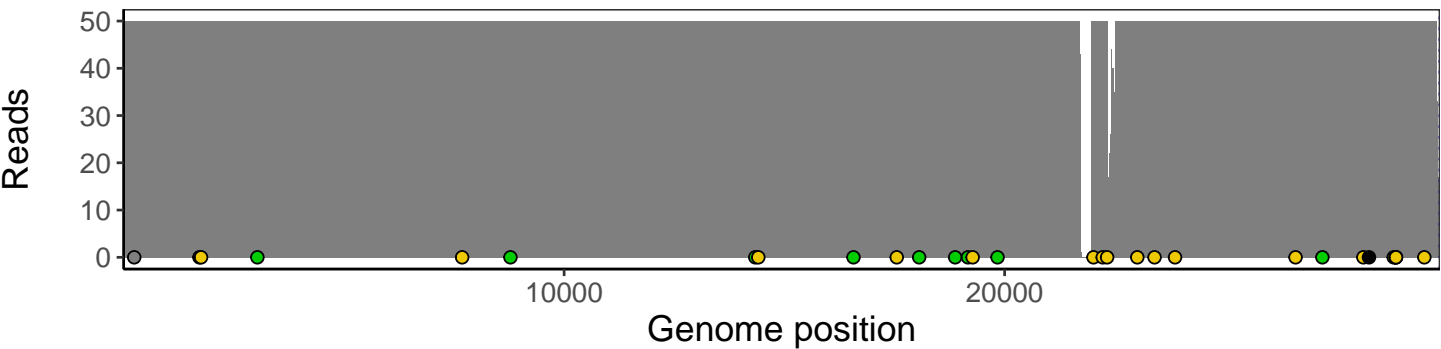
The longest five assembled contigs are shown below colored by their edit distance to the reference genome.
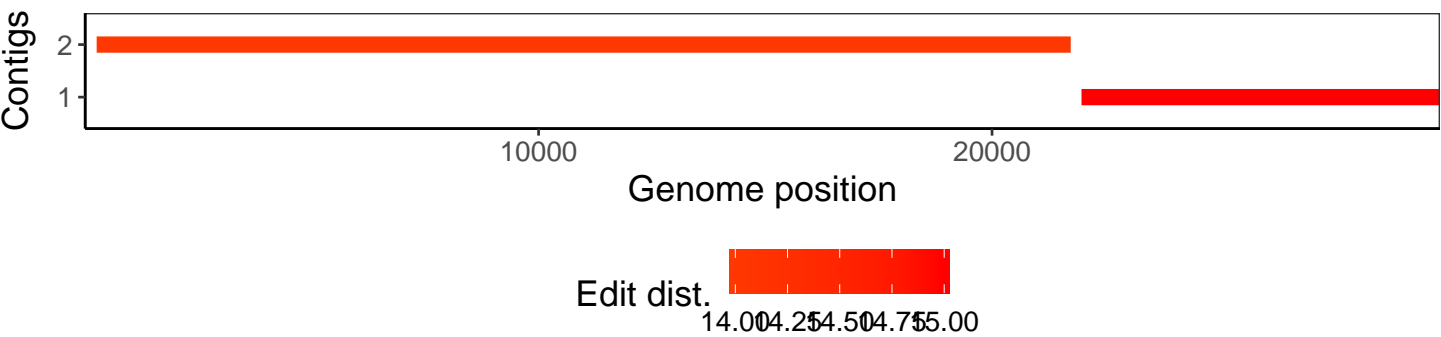
Software environment

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	2.1.7
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