COVID-19 subject UPHS-0394

2021-05-05

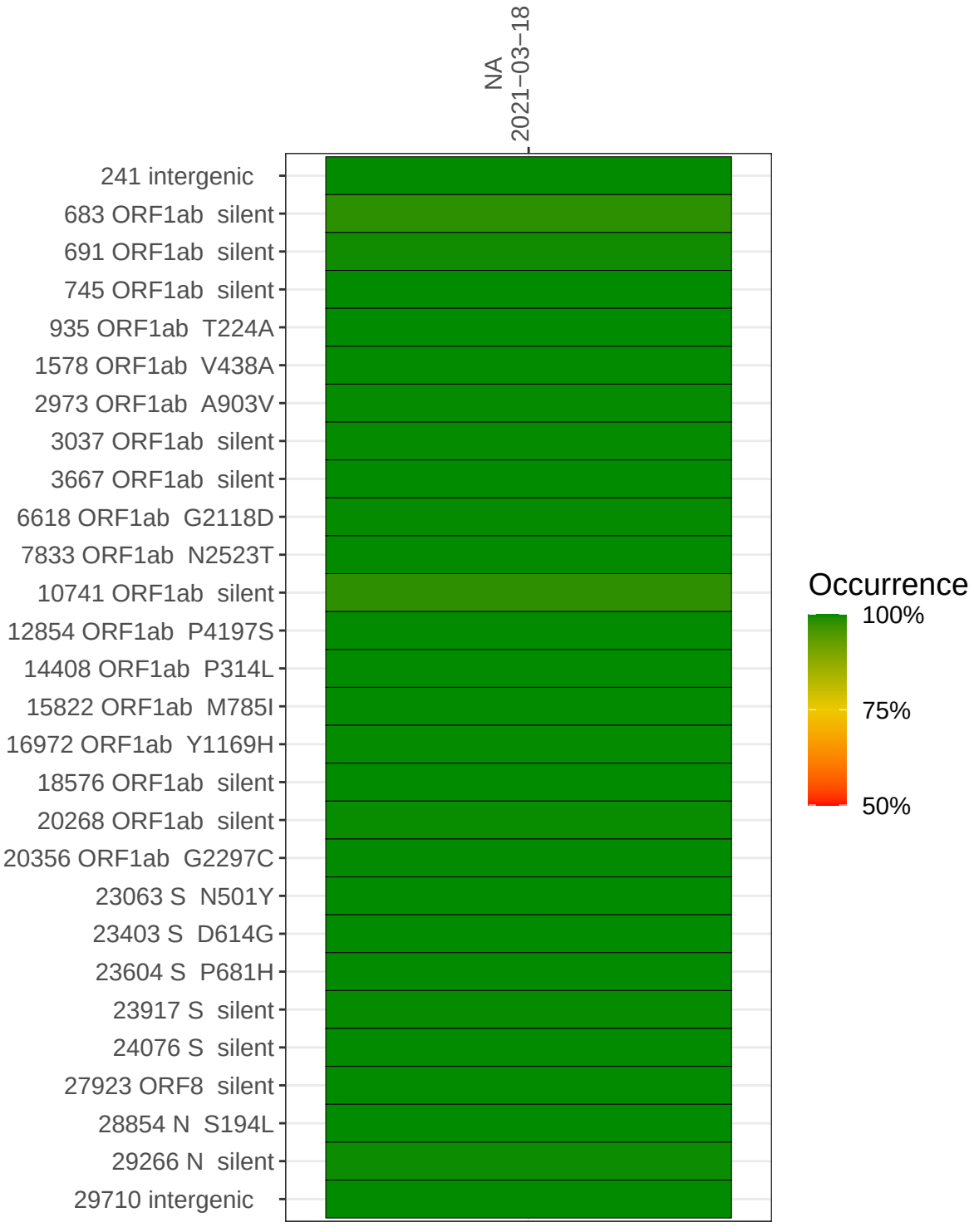
The table below provides a summary of subject samples for which sequencing data is available.
The experiments column shows the number of sequencing experiments performed for each specimen.
Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software tool (Rambaut et al 2020) for genomes with > 90% sequence coverage.

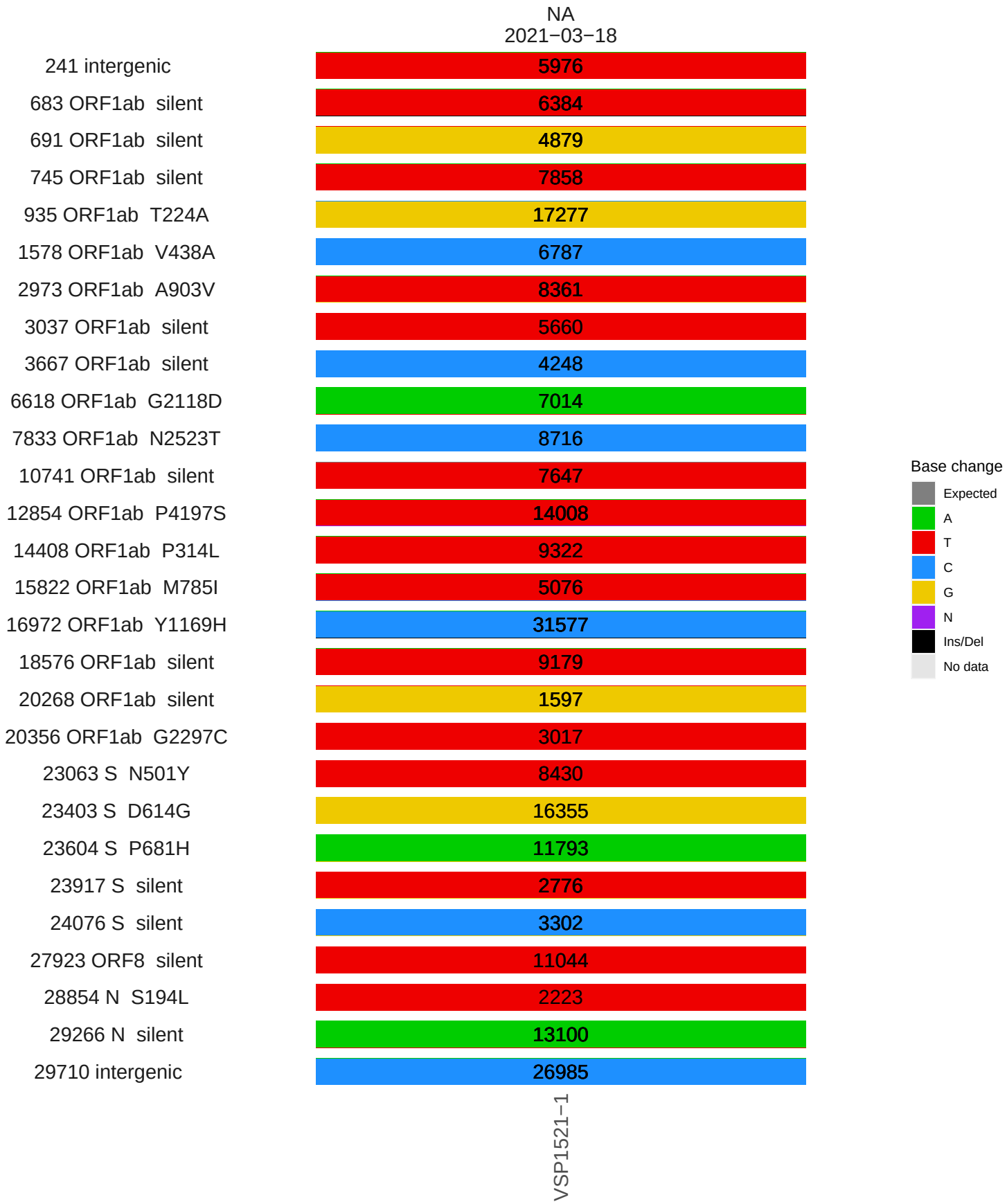
Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP1521-1	single experiment	NA	NA	2021-03-18	29.87	B.1.243	99.8%	99.7%

Variants shared across samples

The heat map below shows how variants (reference genome /home/everett/projects/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.

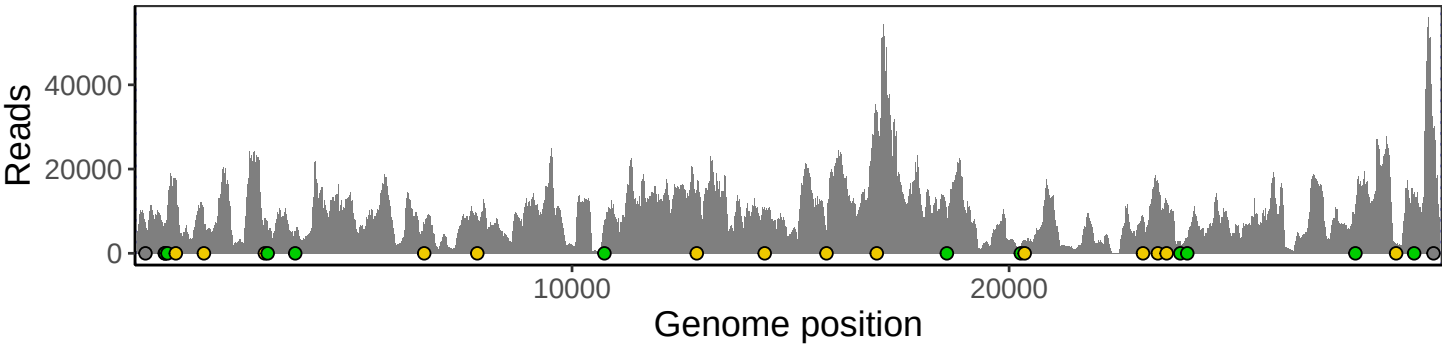



NA
2021−03−18
241 intergenic
5976
683 ORF1ab silent
6384
691 ORF1ab silent
4879
745 ORF1ab silent
7858
935 ORF1ab T224A
17277
1578 ORF1ab V438A
6787
2973 ORF1ab A903V
8361
3037 ORF1ab silent
5660
3667 ORF1ab silent
4248
6618 ORF1ab G2118D
7014
7833 ORF1ab N2523T
8716
10741 ORF1ab silent
7647
12854 ORF1ab P4197S
14008
14408 ORF1ab P314L
9322
15822 ORF1ab M785I
5076
16972 ORF1ab Y1169H
31577
18576 ORF1ab silent
9179
20268 ORF1ab silent
1597
20356 ORF1ab G2297C
3017
23063 S N501Y
8430
23403 S D614G
16355
23604 S P681H
11793
23917 S silent
2776
24076 S silent
3302
27923 ORF8 silent
11044
28854 N S194L
2223
29266 N silent
13100
29710 intergenic
26985
VSP1521−1
Base change
Expected
A
T
C
G
N
Ins/Del
No data

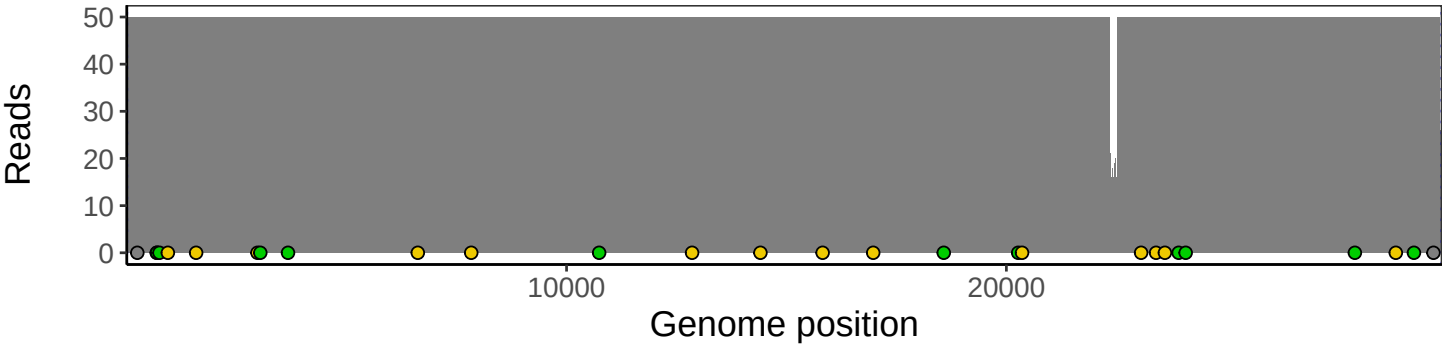
Analyses of individual experiments and composite results

VSP1521-1 | 2021-03-18 | NA | UPHS-0394 | genomes | single experiment

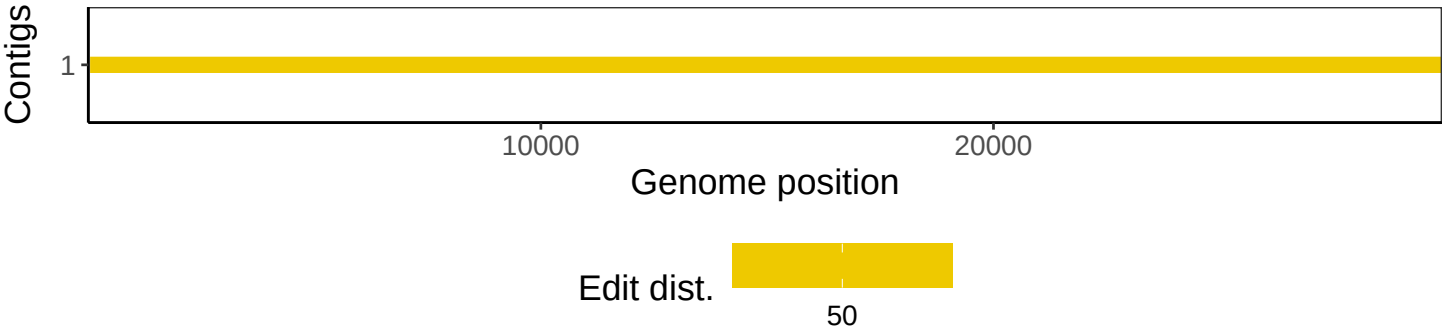
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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	3.3.3
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