COVID-19 subject SARS_CoV_205

2021-06-29

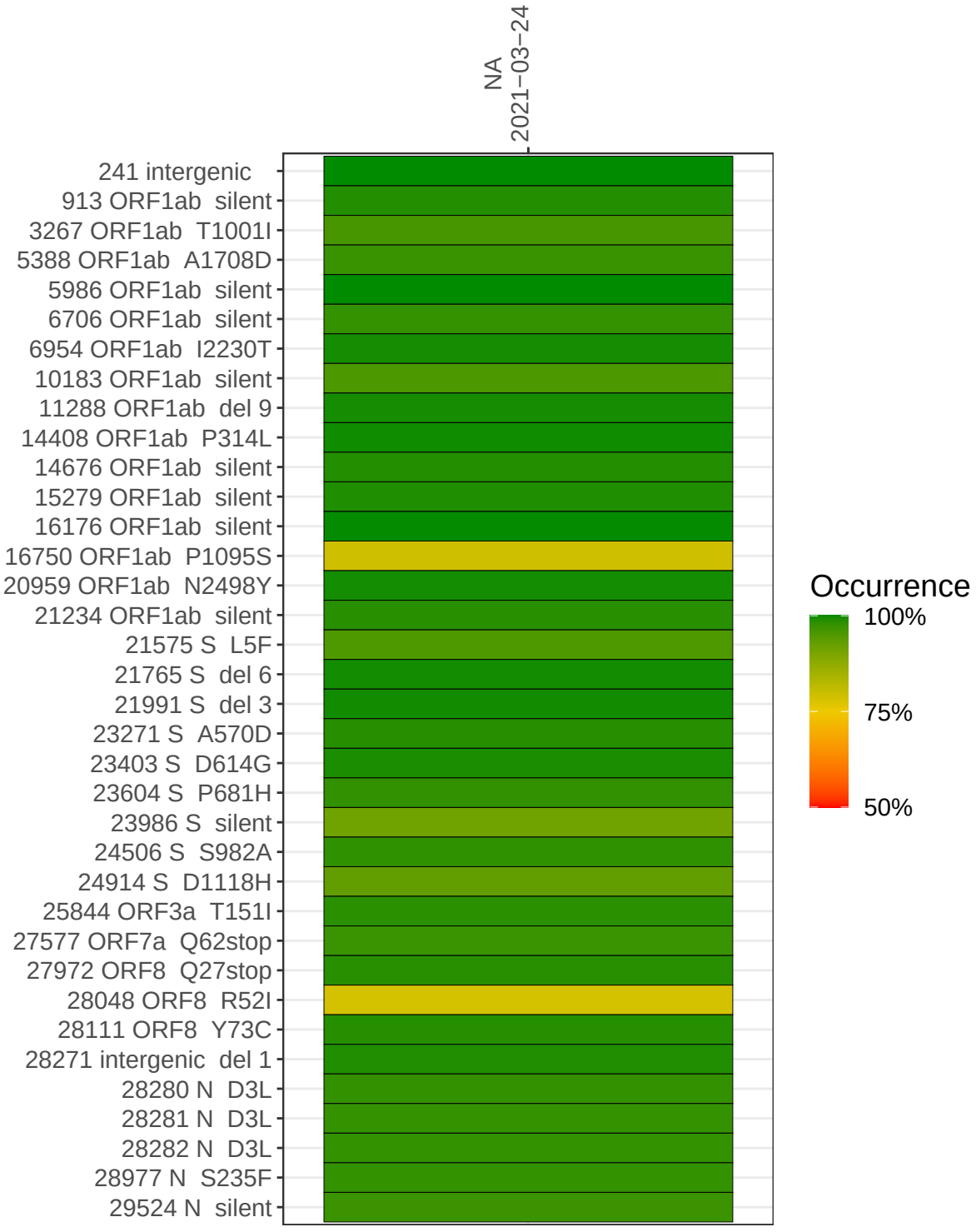
The table below provides a summary of subject samples for which sequencing data is available.
The experiments column shows the number of sequencing experiments performed for each specimen.
Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software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

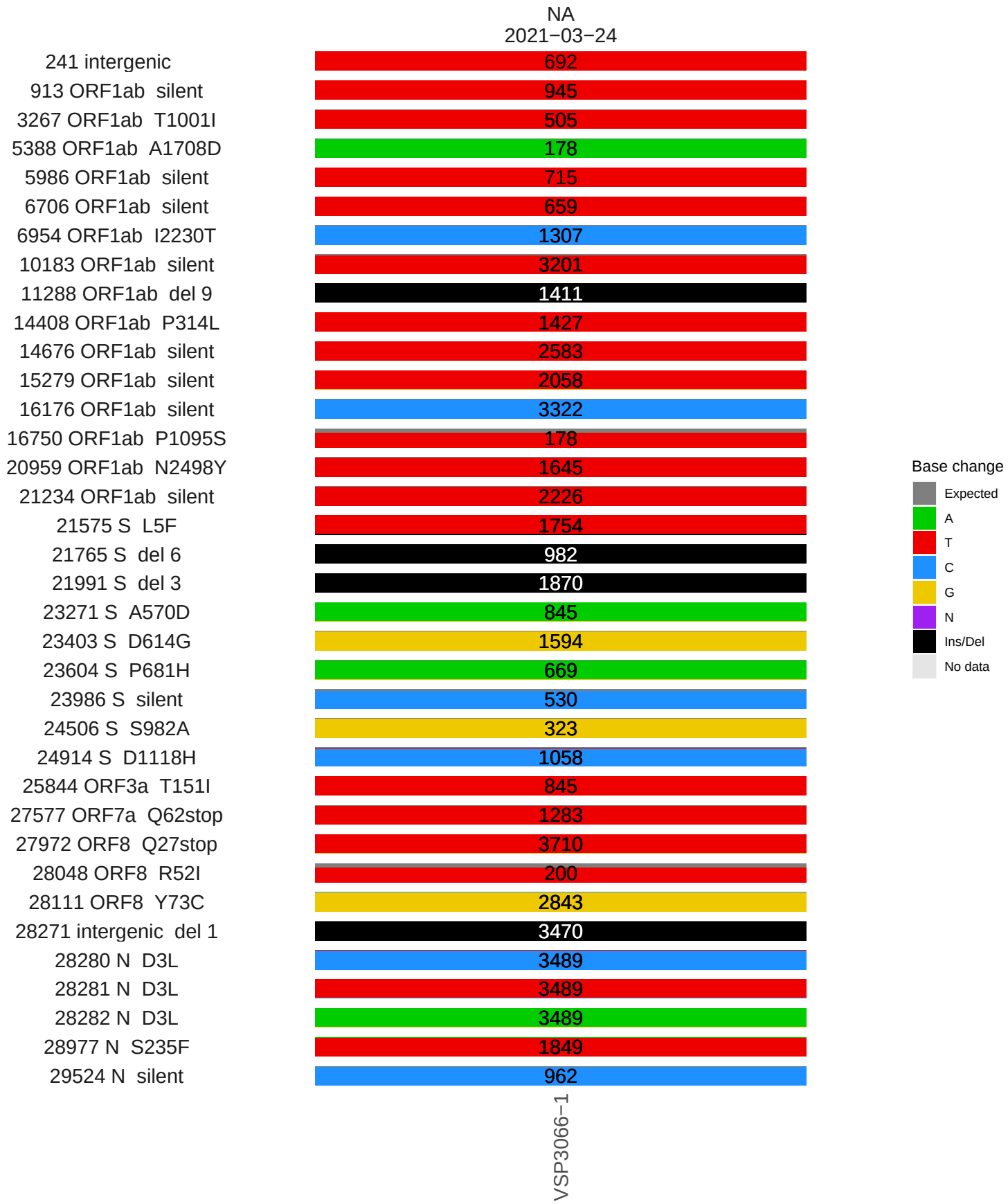
Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP3066-1	single experiment	NA	NA	2021-03-24	29.88	B.1.1.7	99.9%	99.9%

Variants shared across samples

The heat map below shows how variants (reference genome /home/common/SARS-CoV-2-Philadelphia/NC_04551 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.

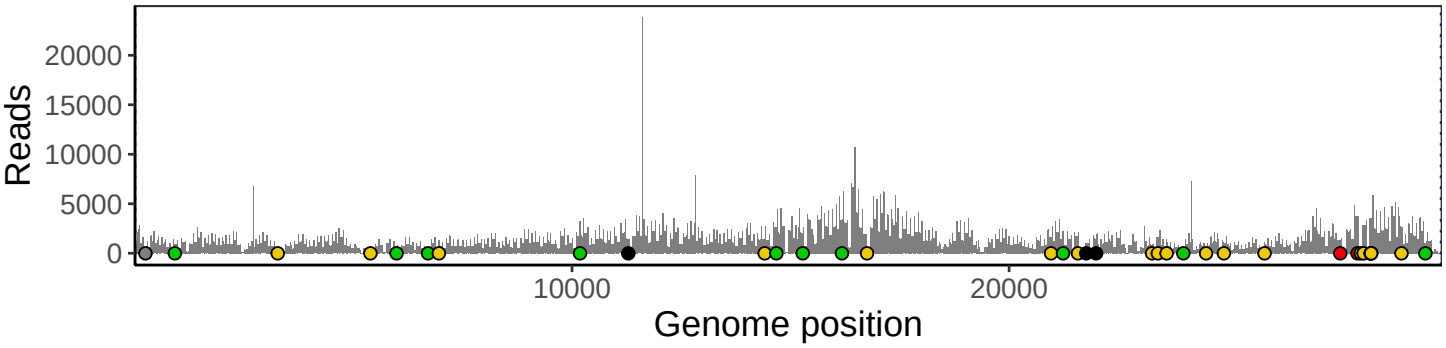



NA
2021−03−24
241 intergenic
913 ORF1ab silent
3267 ORF1ab T1001I
5388 ORF1ab A1708D
5986 ORF1ab silent
6706 ORF1ab silent
6954 ORF1ab I2230T
10183 ORF1ab silent
11288 ORF1ab del 9
14408 ORF1ab P314L
14676 ORF1ab silent
15279 ORF1ab silent
16176 ORF1ab silent
16750 ORF1ab P1095S
20959 ORF1ab N2498Y
21234 ORF1ab silent
21575 S L5F
21765 S del 6
21991 S del 3
23271 S A570D
23403 S D614G
23604 S P681H
23986 S silent
24506 S S982A
24914 S D1118H
25844 ORF3a T151I
27577 ORF7a Q62stop
27972 ORF8 Q27stop
28048 ORF8 R52I
28111 ORF8 Y73C
28271 intergenic del 1
28280 N D3L
28281 N D3L
28282 N D3L
28977 N S235F
29524 N silent
692
945
505
178
715
659
1307
3201
1411
1427
2583
2058
3322
178
1645
2226
1754
982
1870
845
1594
669
530
323
1058
845
1283
3710
200
2843
3470
3489
3489
3489
1849
962
VSP3066−1
Base change
Expected
A
T
C
G
N
Ins/Del
No data

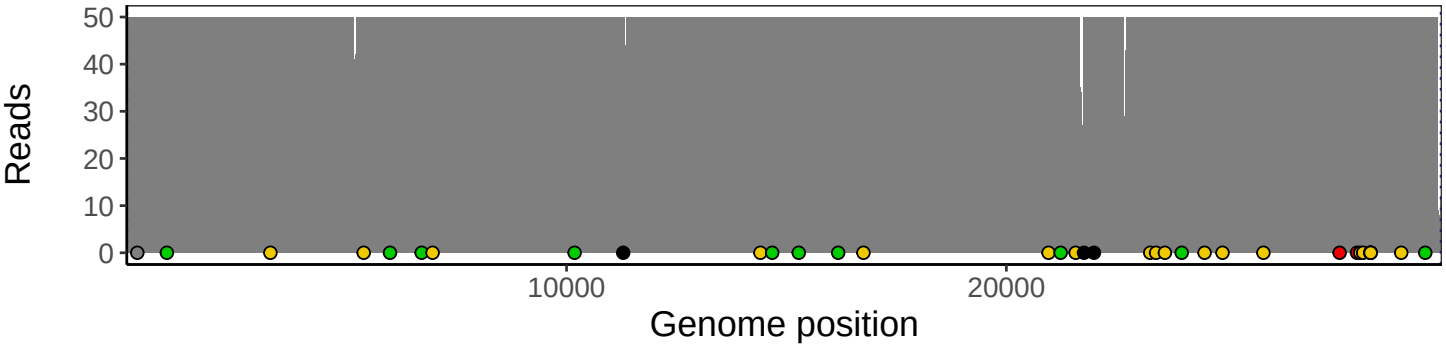
Analyses of individual experiments and composite results

VSP3066-1 | 2021-03-24 | NA | SARS_CoV_205 | genomes | single experiment

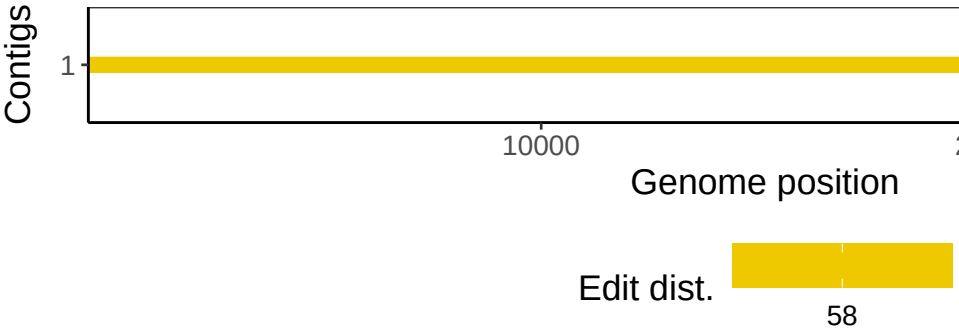
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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	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