COVID-19 subject UPHS-1075

2021-05-10

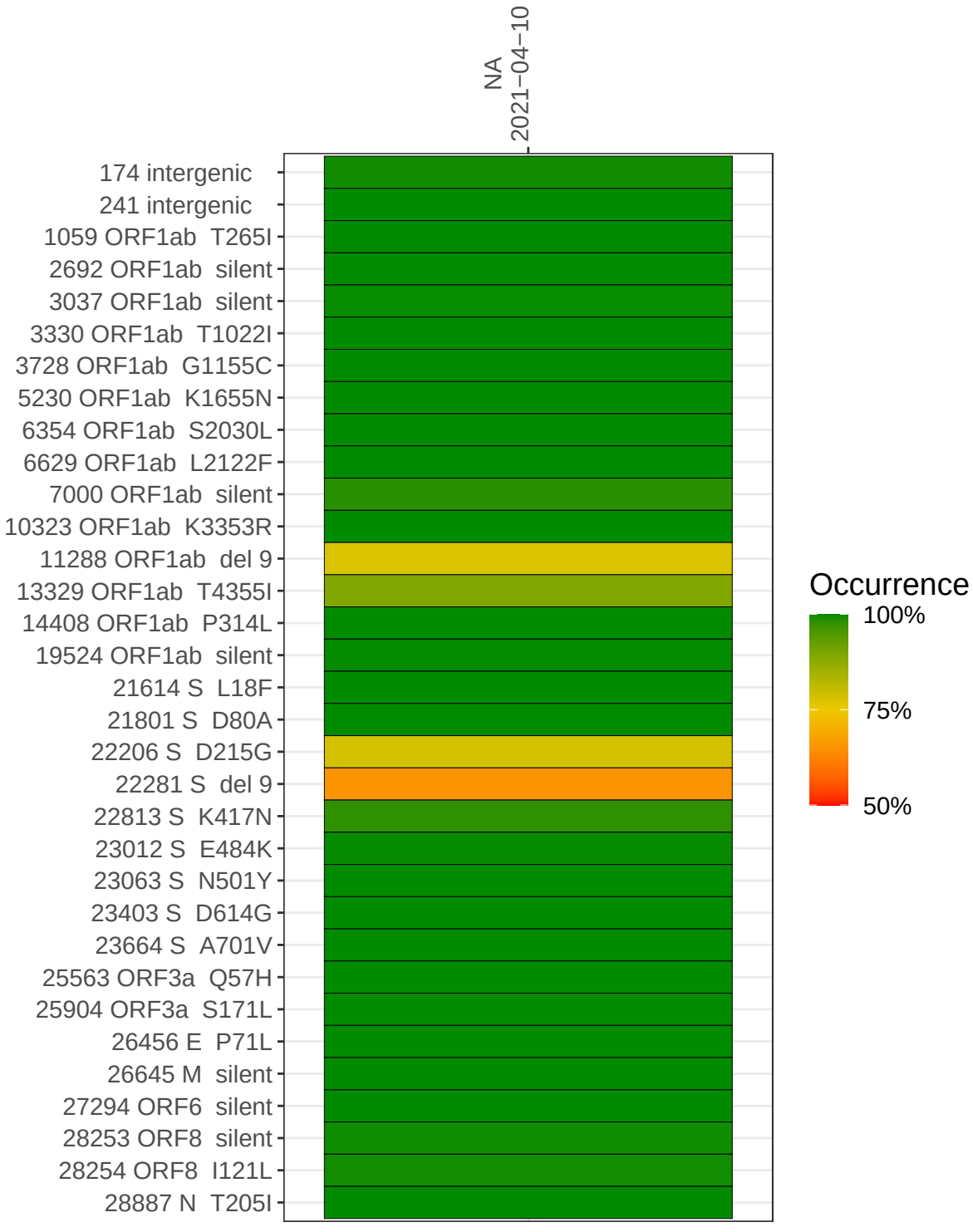
The table below provides a summary of subject samples for which sequencing data is available.
The experiments column shows the number of sequencing experiments performed for each specimen.
Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software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

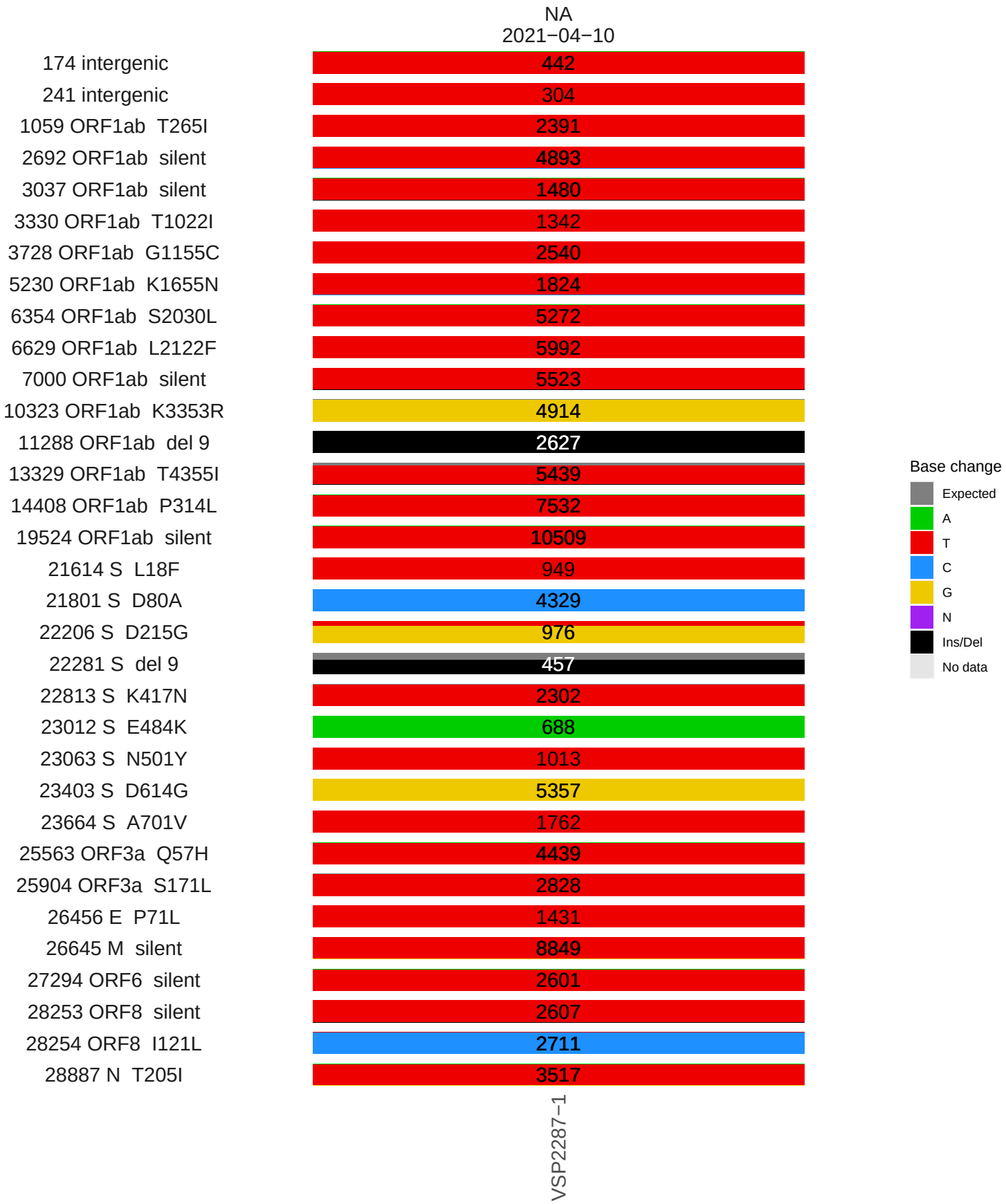
Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP2287-1	single experiment	NA	NA	2021-04-10	22.28	B.1.351	99.6%	99.1%

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.

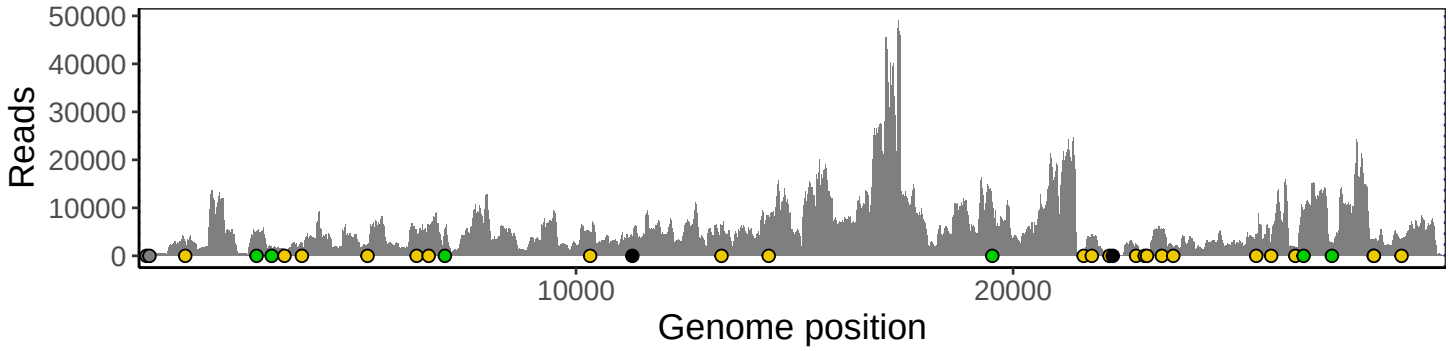



NA
2021−04−10
174 intergenic
241 intergenic
1059 ORF1ab T265I
2692 ORF1ab silent
3037 ORF1ab silent
3330 ORF1ab T1022I
3728 ORF1ab G1155C
5230 ORF1ab K1655N
6354 ORF1ab S2030L
6629 ORF1ab L2122F
7000 ORF1ab silent
10323 ORF1ab K3353R
11288 ORF1ab del 9
13329 ORF1ab T4355I
14408 ORF1ab P314L
19524 ORF1ab silent
21614 S L18F
21801 S D80A
22206 S D215G
22281 S del 9
22813 S K417N
23012 S E484K
23063 S N501Y
23403 S D614G
23664 S A701V
25563 ORF3a Q57H
25904 ORF3a S171L
26456 E P71L
26645 M silent
27294 ORF6 silent
28253 ORF8 silent
28254 ORF8 I121L
28887 N T205I
442
304
2391
4893
1480
1342
2540
1824
5272
5992
5523
4914
2627
5439
7532
10509
949
4329
976
457
2302
688
1013
5357
1762
4439
2828
1431
8849
2601
2607
2711
3517
VSP2287−1
Base change
Expected
A
T
C
G
N
Ins/Del
No data

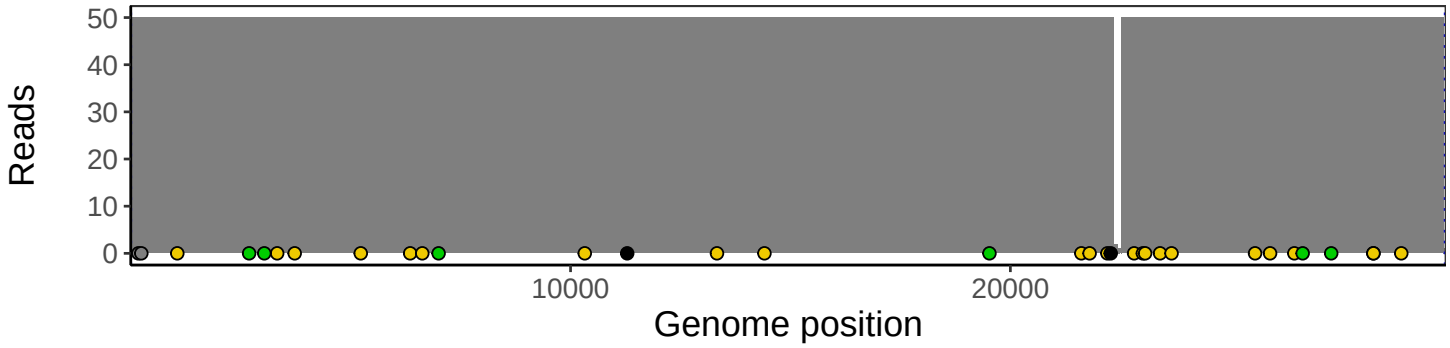
Analyses of individual experiments and composite results

VSP2287-1 | 2021-04-10 | NA | UPHS-1075 | genomes | single experiment

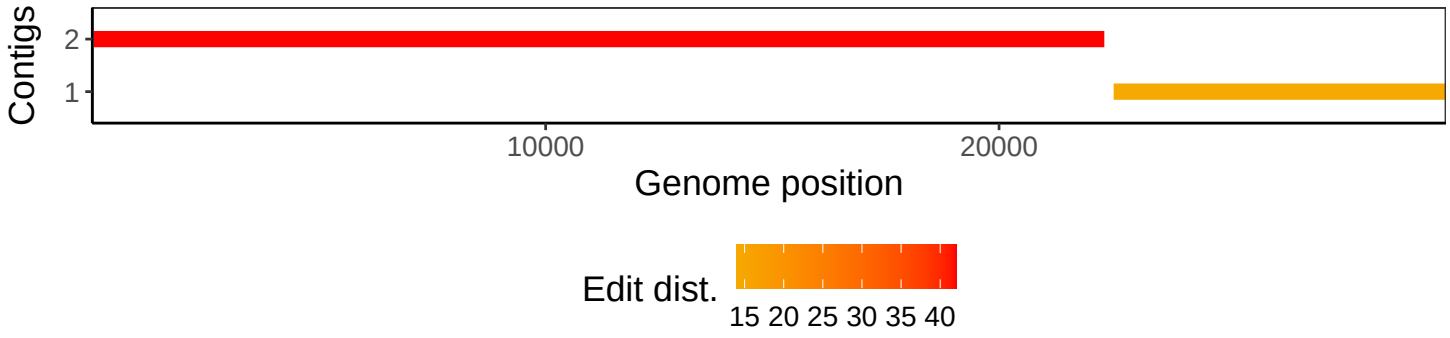
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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