COVID-19 subject UPHS-0481

2021-06-23

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.

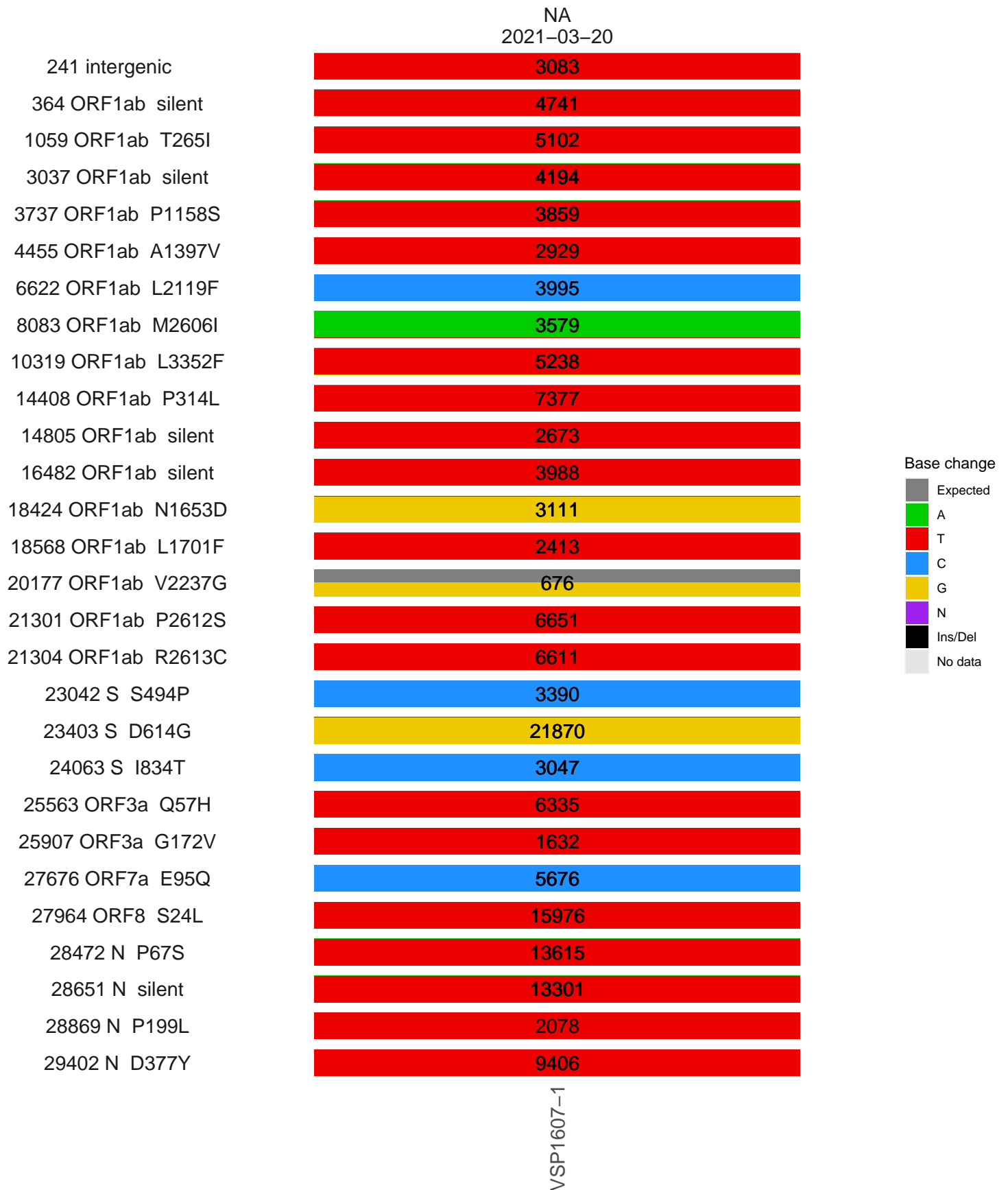Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP1607-1	single experiment	NA	NA	2021-03-20	29.92	B.1.2	99.9%	99.9%

Variants shared across samples

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

Position / Annotation	Value	Base change
241 intergenic	3083	T
364 ORF1ab silent	4741	T
1059 ORF1ab T265I	5102	T
3037 ORF1ab silent	4194	T
3737 ORF1ab P1158S	3859	T
4455 ORF1ab A1397V	2929	T
6622 ORF1ab L2119F	3995	C
8083 ORF1ab M2606I	3579	A
10319 ORF1ab L3352F	5238	T
14408 ORF1ab P314L	7377	T
14805 ORF1ab silent	2673	T
16482 ORF1ab silent	3988	T
18424 ORF1ab N1653D	3111	G
18568 ORF1ab L1701F	2413	T
20177 ORF1ab V2237G	676	Expected
21301 ORF1ab P2612S	6651	T
21304 ORF1ab R2613C	6611	T
23042 S S494P	3390	C
23403 S D614G	21870	G
24063 S I834T	3047	C
25563 ORF3a Q57H	6335	T
25907 ORF3a G172V	1632	T
27676 ORF7a E95Q	5676	C
27964 ORF8 S24L	15976	T
28472 N P67S	13615	T
28651 N silent	13301	T
28869 N P199L	2078	T
29402 N D377Y	9406	T

Base change

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

VSP1607−1

3

Analyses of individual experiments and composite results

**VSP1607-1 | 2021-03-20 | NA | UPHS-0481 | 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.
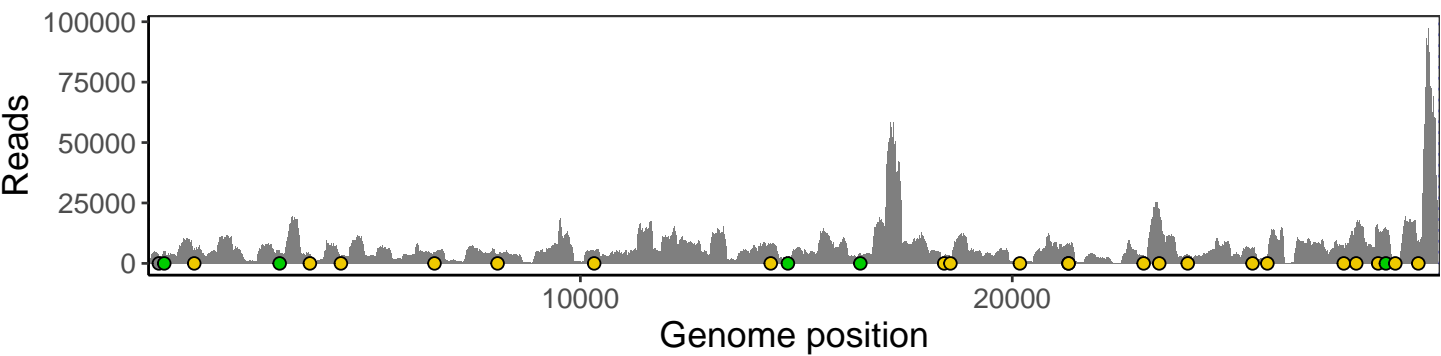
Excerpt from plot above focusing on reads coverage from 0 to 50 NT.
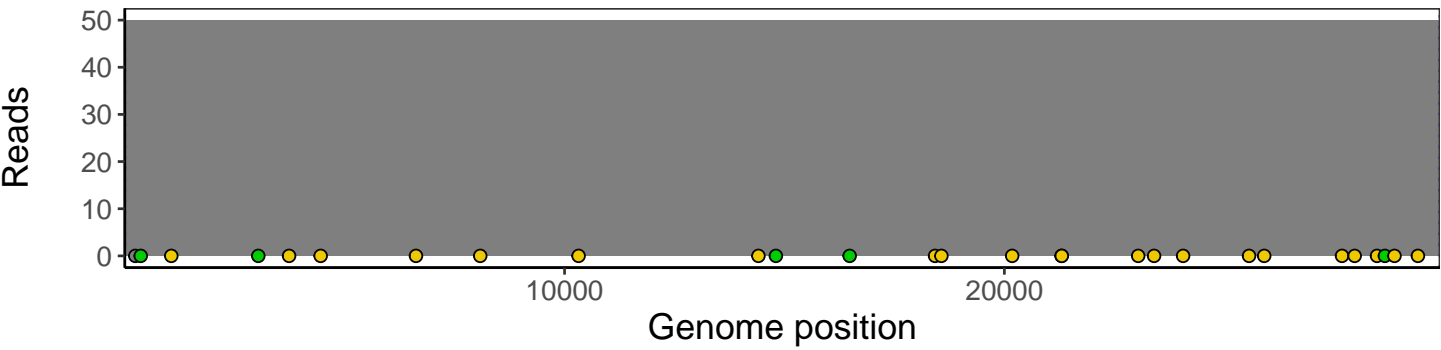
The longest five assembled contigs are shown below colored by their edit distance to the reference genome.
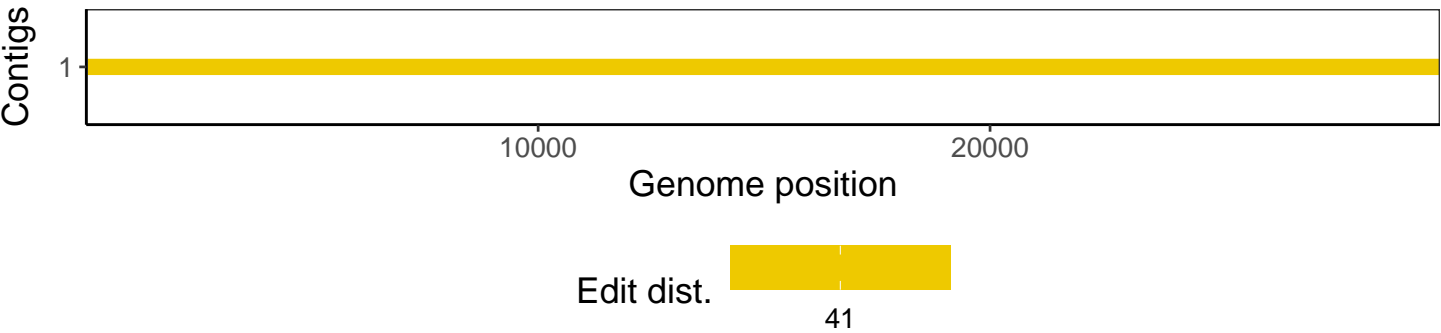
Software environment

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