COVID-19 subject UPHS-0378

2021-04-17

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.
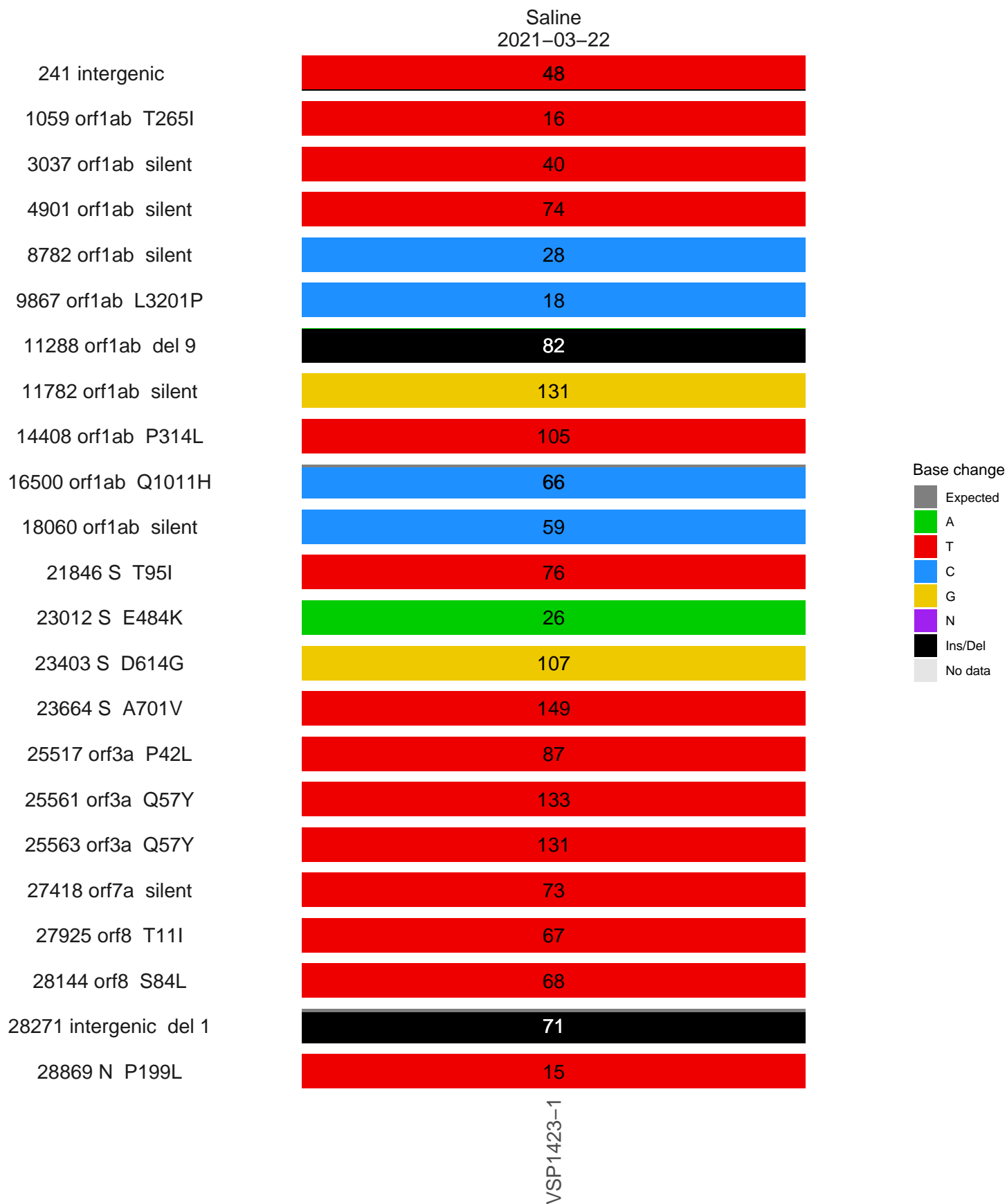
Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP1423-1	single experiment	NA	Saline	2021-03-22	22.29	B.1.526	99.6%	99.1%

Variants shared across samples

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

Saline
2021−03−22

Position	VSP1423−1
241 intergenic	48
1059 orf1ab T265I	16
3037 orf1ab silent	40
4901 orf1ab silent	74
8782 orf1ab silent	28
9867 orf1ab L3201P	18
11288 orf1ab del 9	82
11782 orf1ab silent	131
14408 orf1ab P314L	105
16500 orf1ab Q1011H	66
18060 orf1ab silent	59
21846 S T95I	76
23012 S E484K	26
23403 S D614G	107
23664 S A701V	149
25517 orf3a P42L	87
25561 orf3a Q57Y	133
25563 orf3a Q57Y	131
27418 orf7a silent	73
27925 orf8 T11I	67
28144 orf8 S84L	68
28271 intergenic del 1	71
28869 N P199L	15

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

3

Analyses of individual experiments and composite results

## VSP1423-1 | 2021-03-22 | Saline | UPHS-0378 | 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.
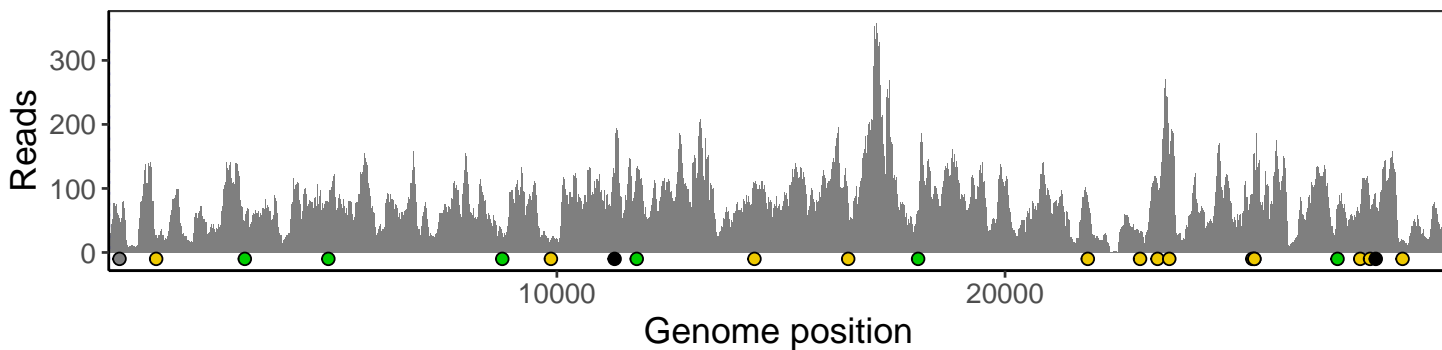
Excerpt from plot above focusing on reads coverage from 0 to 50 NT.
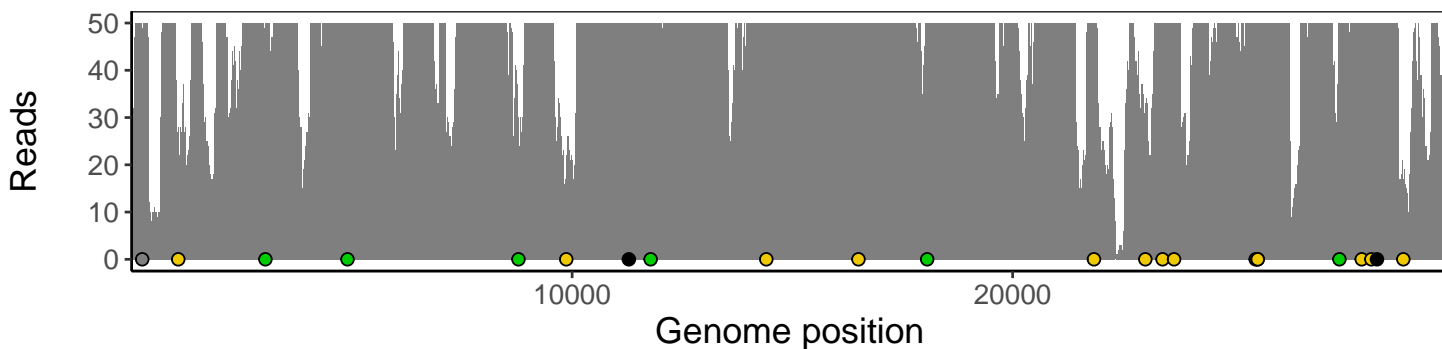
The longest five assembled contigs are shown below colored by their edit distance to the reference genome.
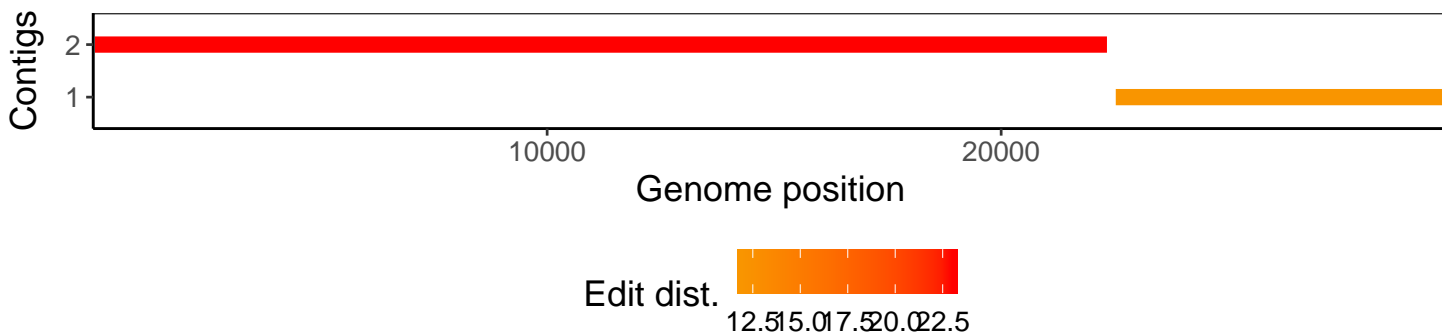
Software environment

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	3.0.0
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