COVID-19 subject HUP-PH-0002

2021-03-05

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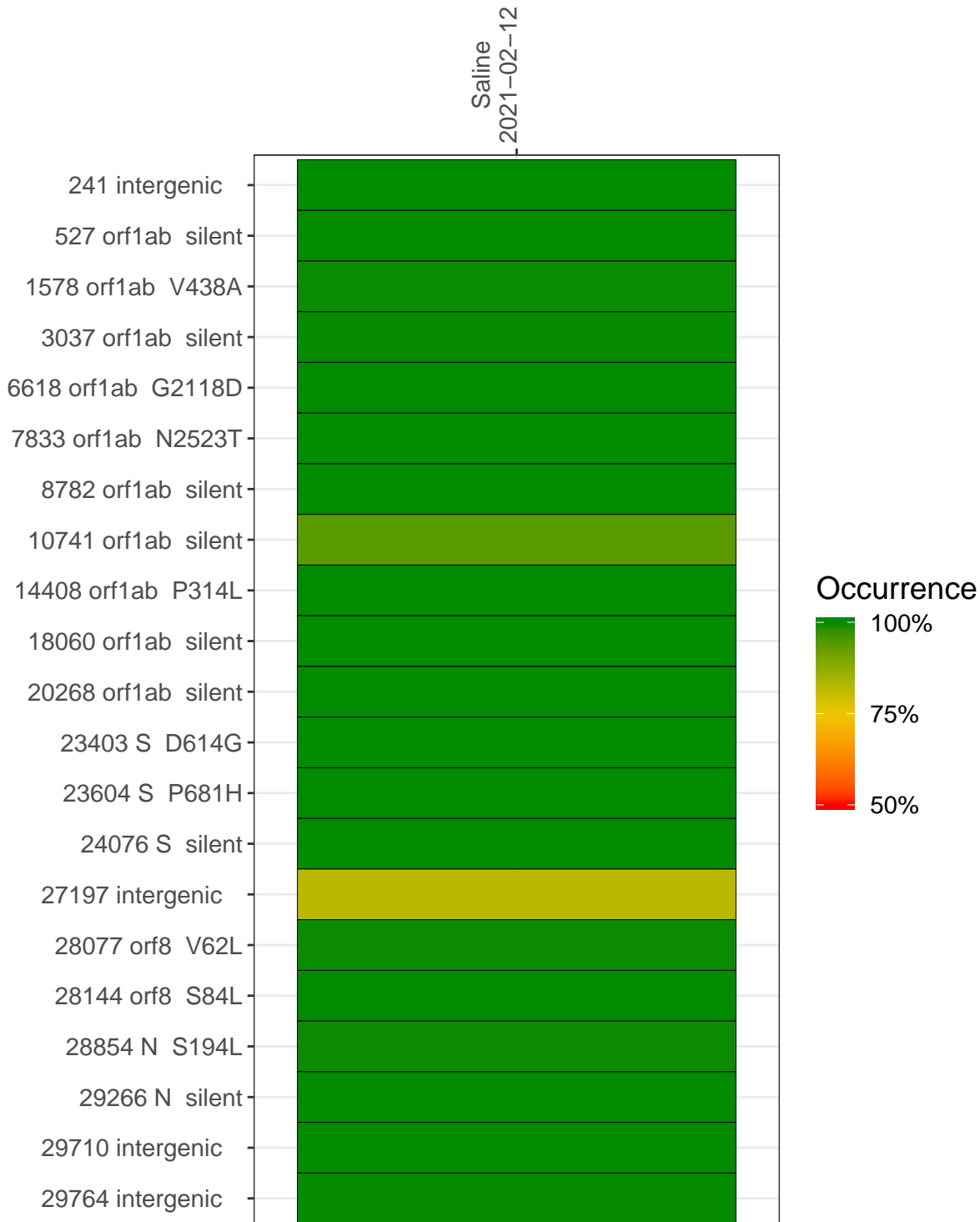of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software tool (Rambaut et al 2020) for genomes with > 90% sequence coverage.

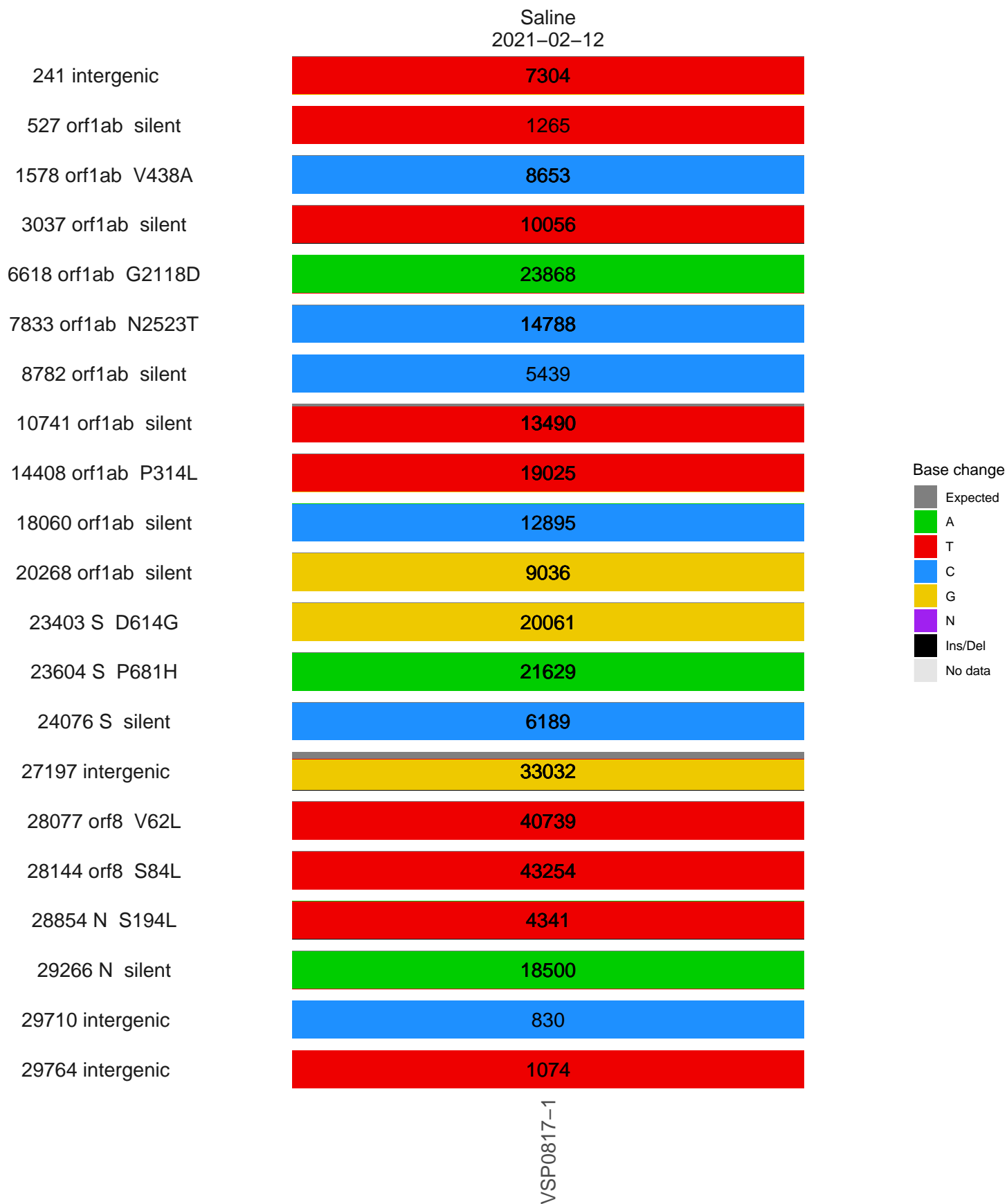Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP0817-1	single experiment	NA	Saline	2021-02-12	29.78	B.1.243	99.8%	99.8%

Variants shared across samples

The heat map below shows how variants (reference genome 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
2021−02−12

Position	Value	Base
241 intergenic	7304	T
527 orf1ab silent	1265	T
1578 orf1ab V438A	8653	C
3037 orf1ab silent	10056	T
6618 orf1ab G2118D	23868	A
7833 orf1ab N2523T	14788	C
8782 orf1ab silent	5439	C
10741 orf1ab silent	13490	T
14408 orf1ab P314L	19025	T
18060 orf1ab silent	12895	C
20268 orf1ab silent	9036	G
23403 S D614G	20061	G
23604 S P681H	21629	A
24076 S silent	6189	C
27197 intergenic	33032	G
28077 orf8 V62L	40739	T
28144 orf8 S84L	43254	T
28854 N S194L	4341	T
29266 N silent	18500	A
29710 intergenic	830	C
29764 intergenic	1074	T

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

VSP0817−1

3

Analyses of individual experiments and composite results

VSP0817-1 | 2021-02-12 | Saline | HUP-PH-0002 | 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.
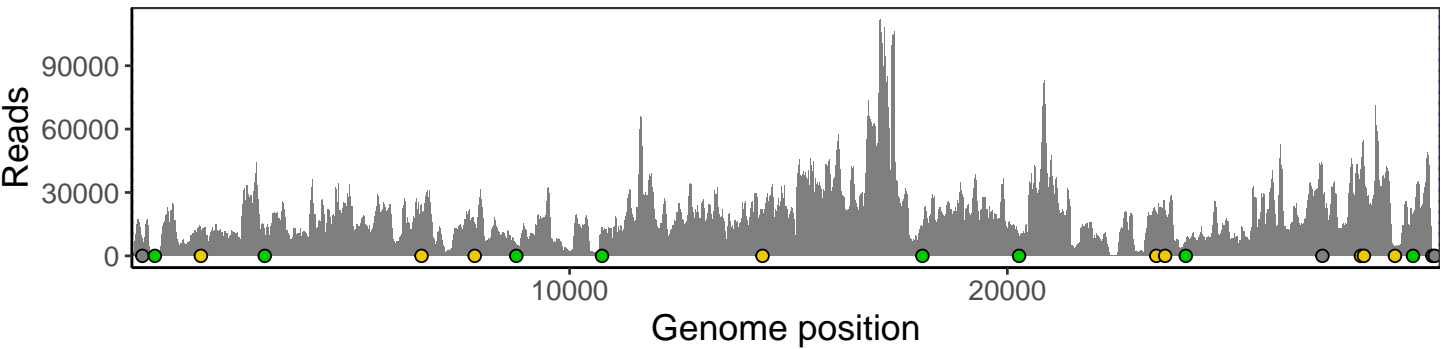
Excerpt from plot above focusing on reads coverage from 0 to 50 NT.
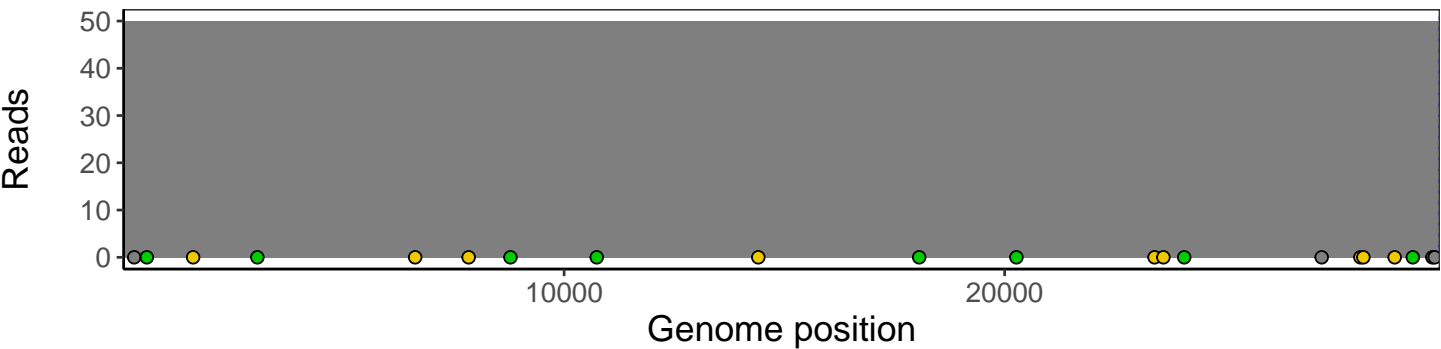
The longest five assembled contigs are shown below colored by their edit distance to the reference genome.
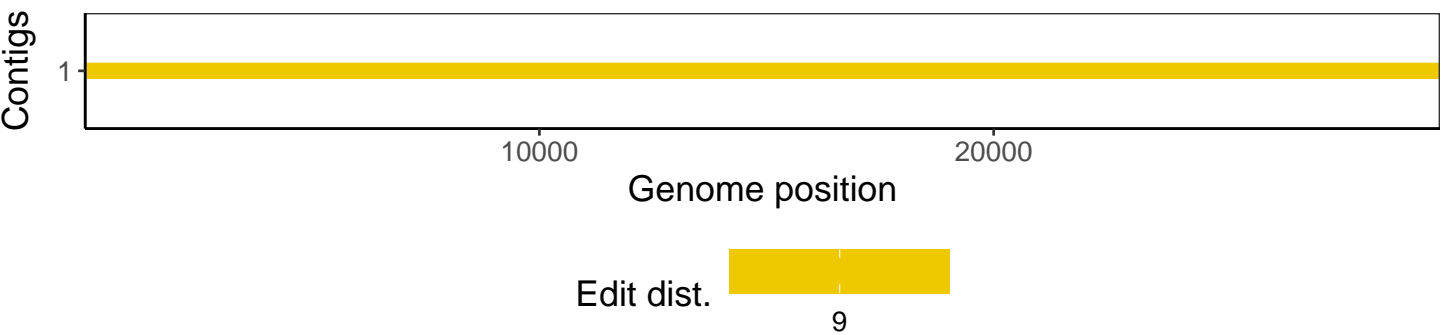
Software environment

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	2.3.3
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	3.0.0
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