COVID-19 subject UPHS-0302

2021-04-17

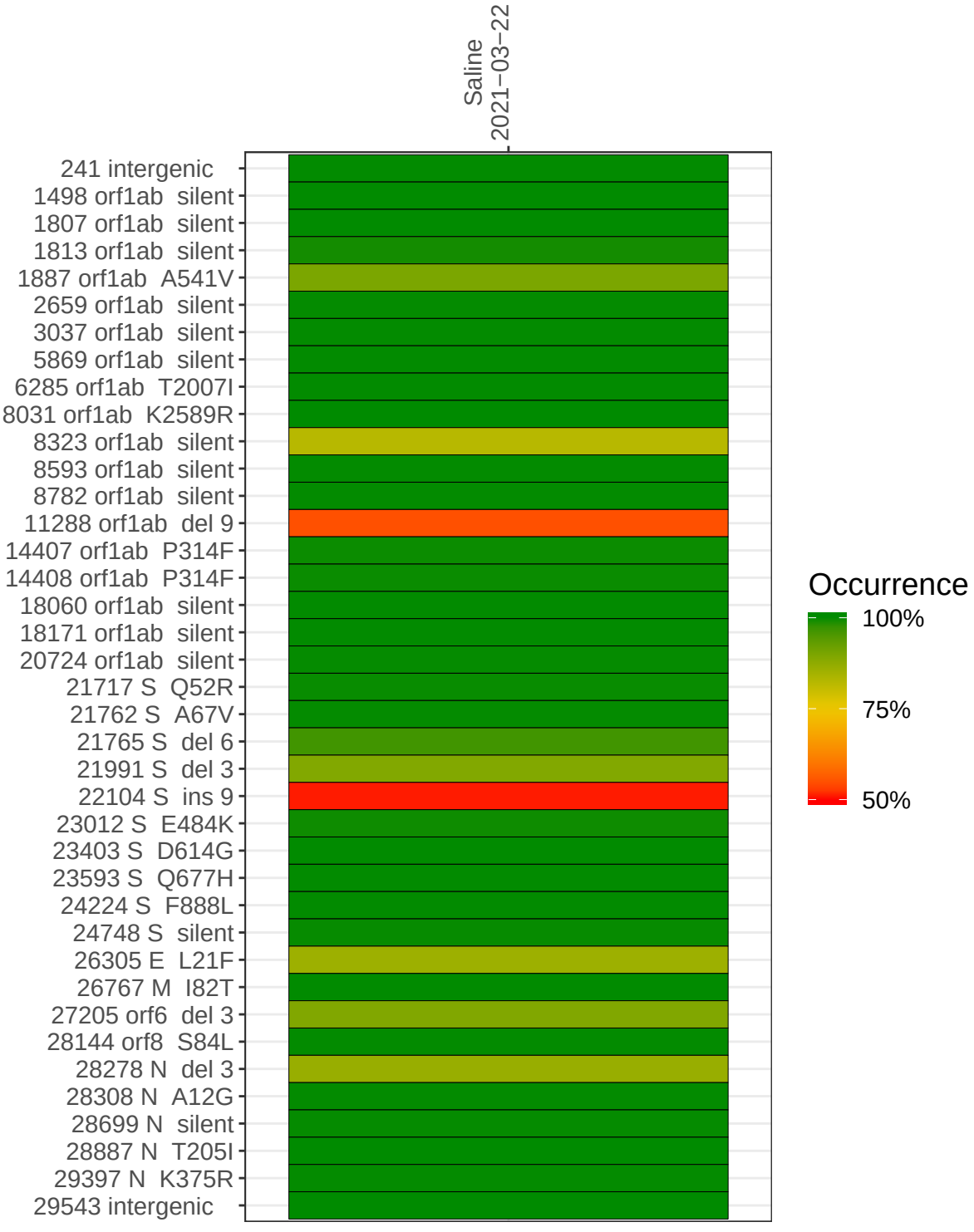
The table below provides a summary of subject samples for which sequencing data is available.
The experiments column shows the number of sequencing experiments performed for each specimen.
Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software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

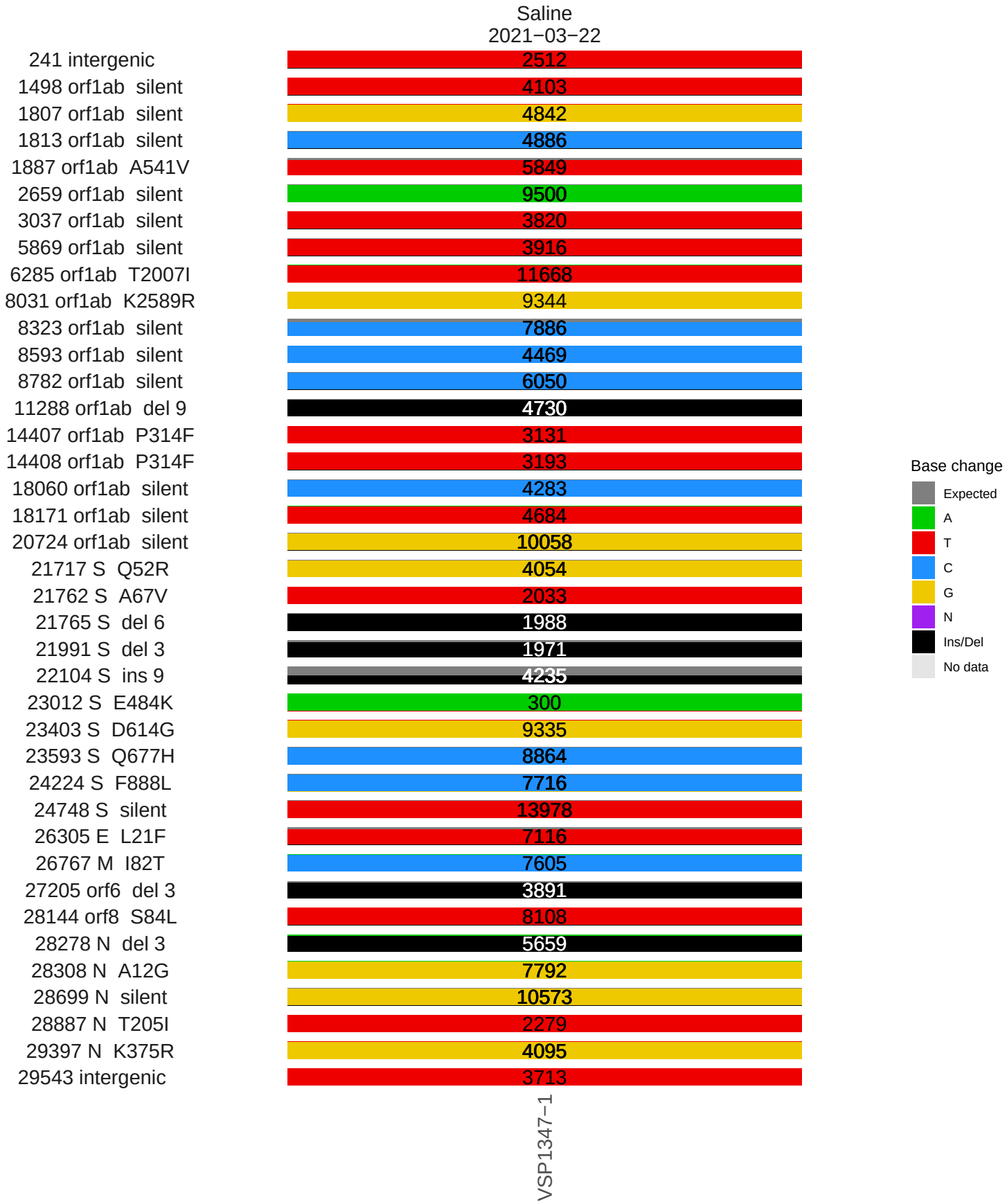
Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP1347-1	single experiment	NA	Saline	2021-03-22	29.79	B.1.525	99.7%	99.6%

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.

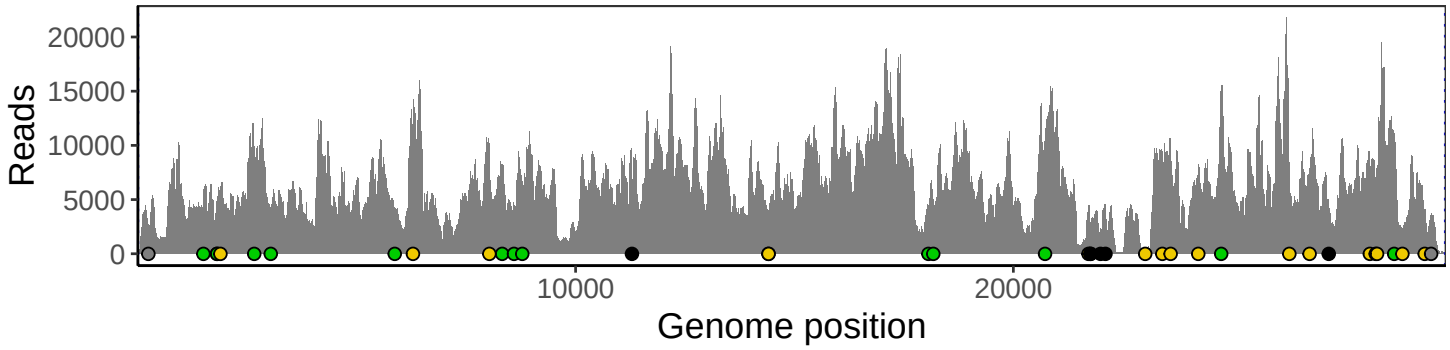



Saline
2021−03−22
241 intergenic
1498 orf1ab silent
1807 orf1ab silent
1813 orf1ab silent
1887 orf1ab A541V
2659 orf1ab silent
3037 orf1ab silent
5869 orf1ab silent
6285 orf1ab T2007I
8031 orf1ab K2589R
8323 orf1ab silent
8593 orf1ab silent
8782 orf1ab silent
11288 orf1ab del 9
14407 orf1ab P314F
14408 orf1ab P314F
18060 orf1ab silent
18171 orf1ab silent
20724 orf1ab silent
21717 S Q52R
21762 S A67V
21765 S del 6
21991 S del 3
22104 S ins 9
23012 S E484K
23403 S D614G
23593 S Q677H
24224 S F888L
24748 S silent
26305 E L21F
26767 M I82T
27205 orf6 del 3
28144 orf8 S84L
28278 N del 3
28308 N A12G
28699 N silent
28887 N T205I
29397 N K375R
29543 intergenic
2512
4103
4842
4886
5849
9500
3820
3916
11668
9344
7886
4469
6050
4730
3131
3193
4283
4684
10058
4054
2033
1988
1971
4235
300
9335
8864
7716
13978
7116
7605
3891
8108
5659
7792
10573
2279
4095
3713
VSP1347−1
Base change
Expected
A
T
C
G
N
Ins/Del
No data

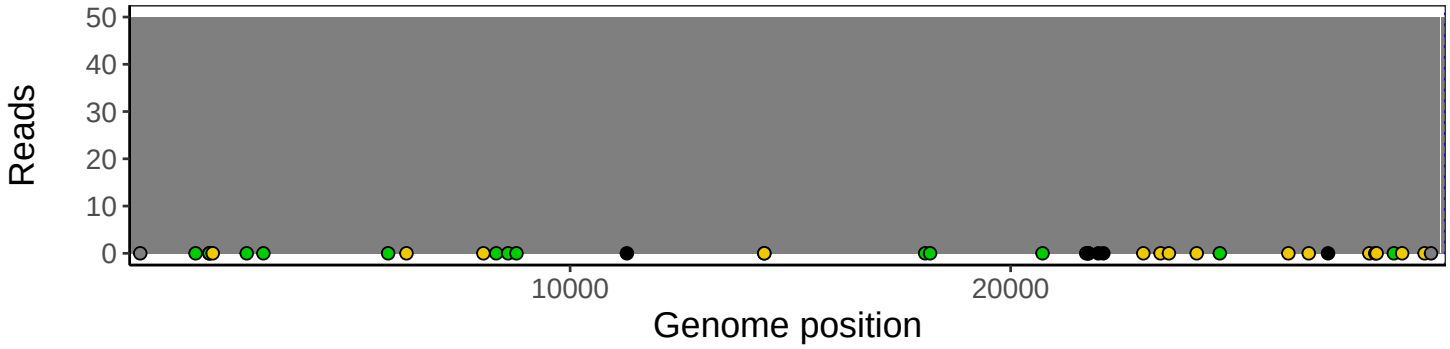
Analyses of individual experiments and composite results

VSP1347-1 | 2021-03-22 | Saline | UPHS-0302 | genomes | single experiment

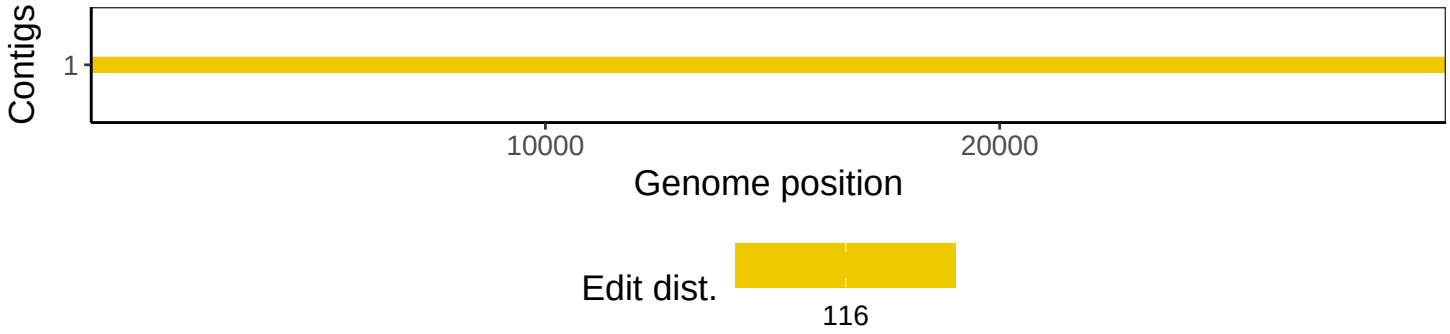
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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