COVID-19 subject UPHS-0342

2021-04-17

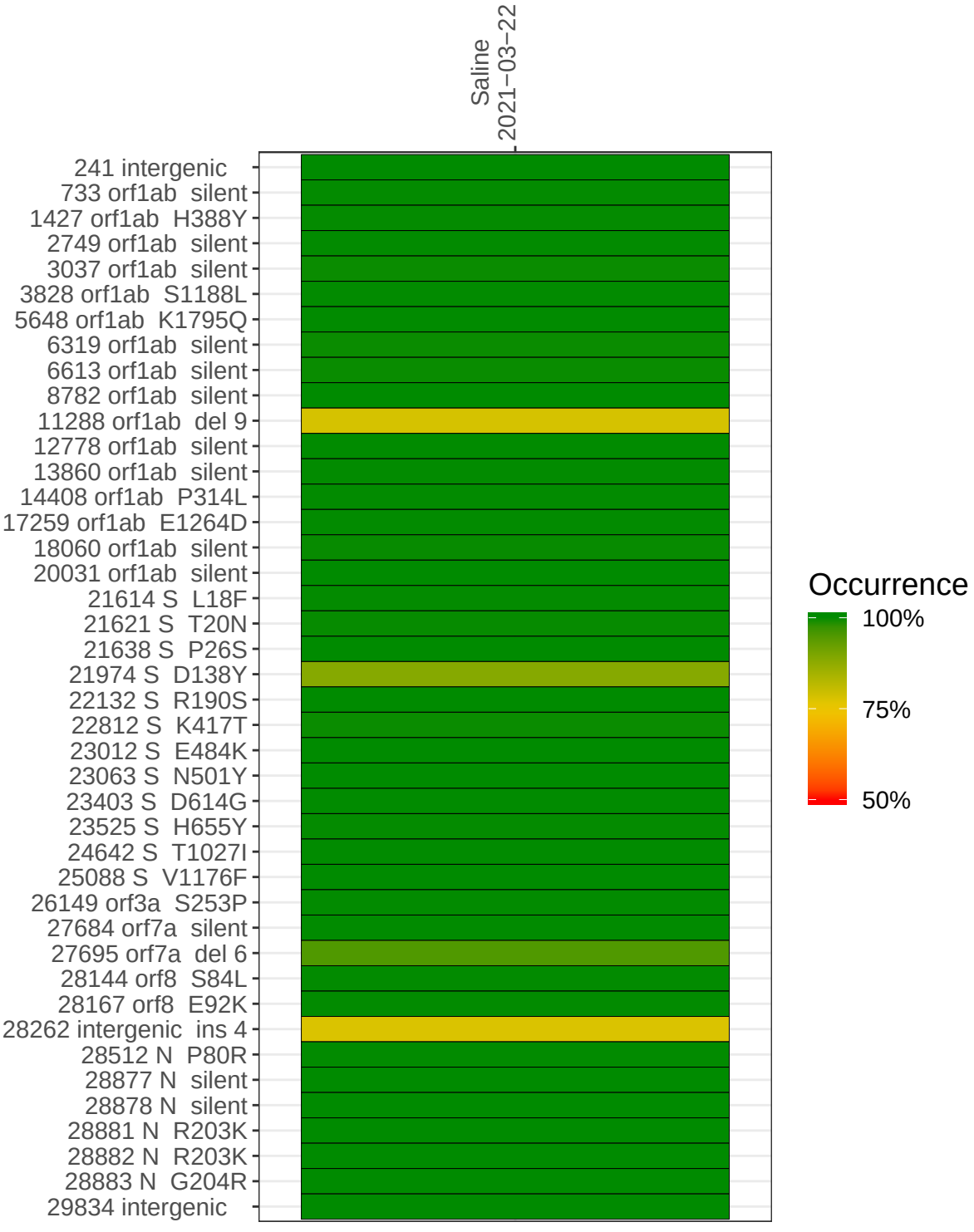
The table below provides a summary of subject samples for which sequencing data is available.
The experiments column shows the number of sequencing experiments performed for each specimen.
Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software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

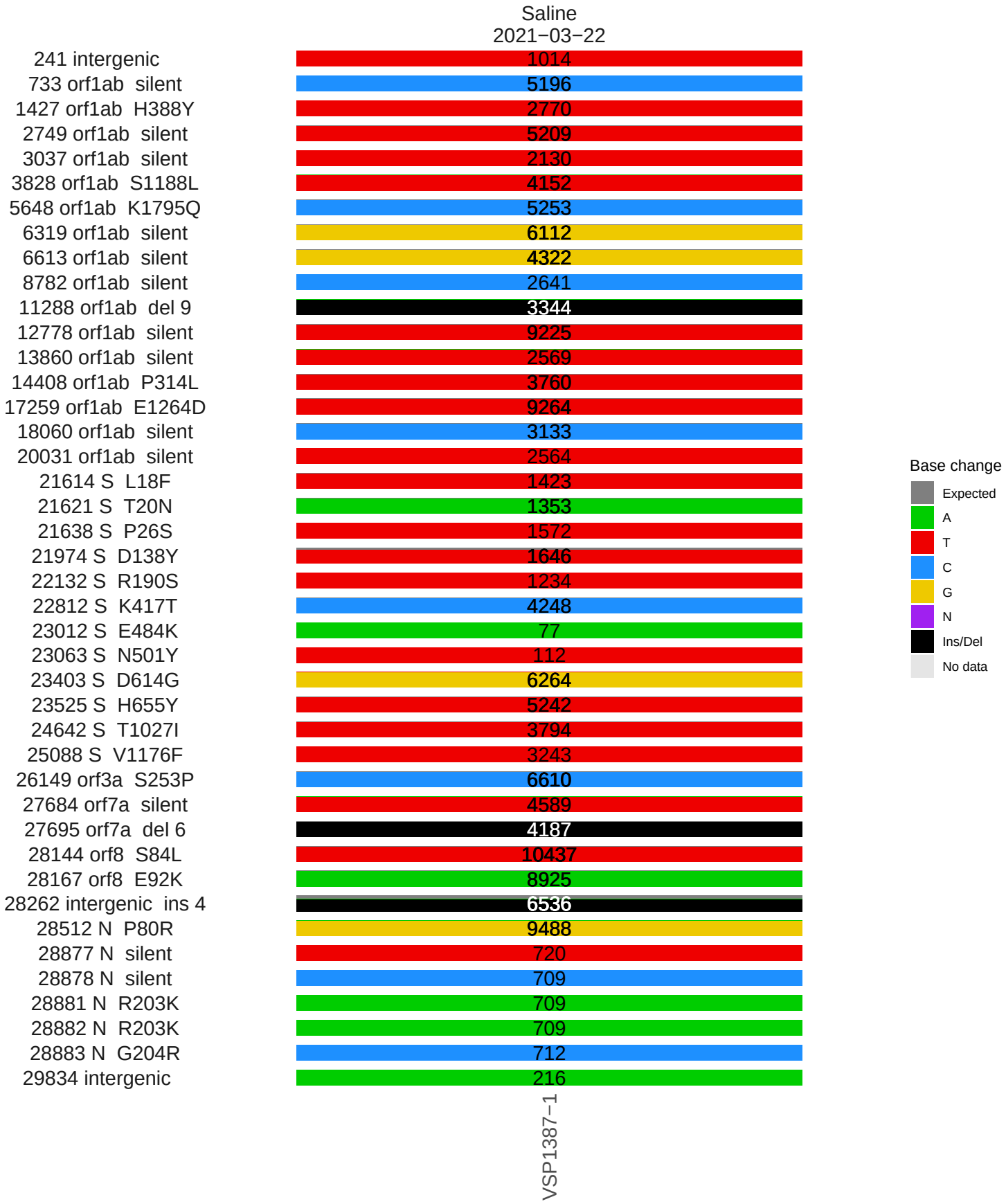
Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP1387-1	single experiment	NA	Saline	2021-03-22	29.84	P.1	99.8%	99.8%

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.

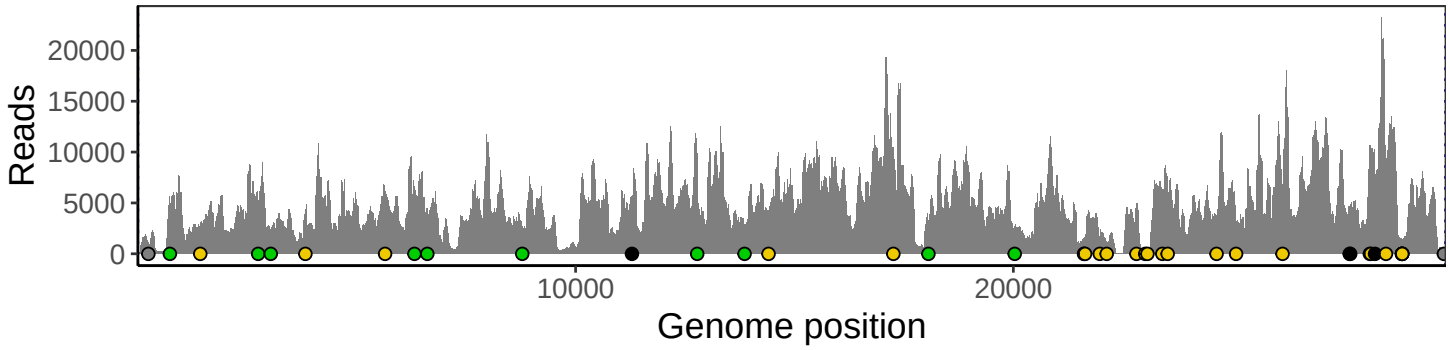



Saline
2021−03−22
241 intergenic 1014
733 orf1ab silent 5196
1427 orf1ab H388Y 2770
2749 orf1ab silent 5209
3037 orf1ab silent 2130
3828 orf1ab S1188L 4152
5648 orf1ab K1795Q 5253
6319 orf1ab silent 6112
6613 orf1ab silent 4322
8782 orf1ab silent 2641
11288 orf1ab del 9 3344
12778 orf1ab silent 9225
13860 orf1ab silent 2569
14408 orf1ab P314L 3760
17259 orf1ab E1264D 9264
18060 orf1ab silent 3133
20031 orf1ab silent 2564
21614 S L18F 1423
21621 S T20N 1353
21638 S P26S 1572
21974 S D138Y 1646
22132 S R190S 1234
22812 S K417T 4248
23012 S E484K 77
23063 S N501Y 112
23403 S D614G 6264
23525 S H655Y 5242
24642 S T1027I 3794
25088 S V1176F 3243
26149 orf3a S253P 6610
27684 orf7a silent 4589
27695 orf7a del 6 4187
28144 orf8 S84L 10437
28167 orf8 E92K 8925
28262 intergenic ins 4 6536
28512 N P80R 9488
28877 N silent 720
28878 N silent 709
28881 N R203K 709
28882 N R203K 709
28883 N G204R 712
29834 intergenic 216
VSP1387−1
Base change
Expected
A
T
C
G
N
Ins/Del
No data

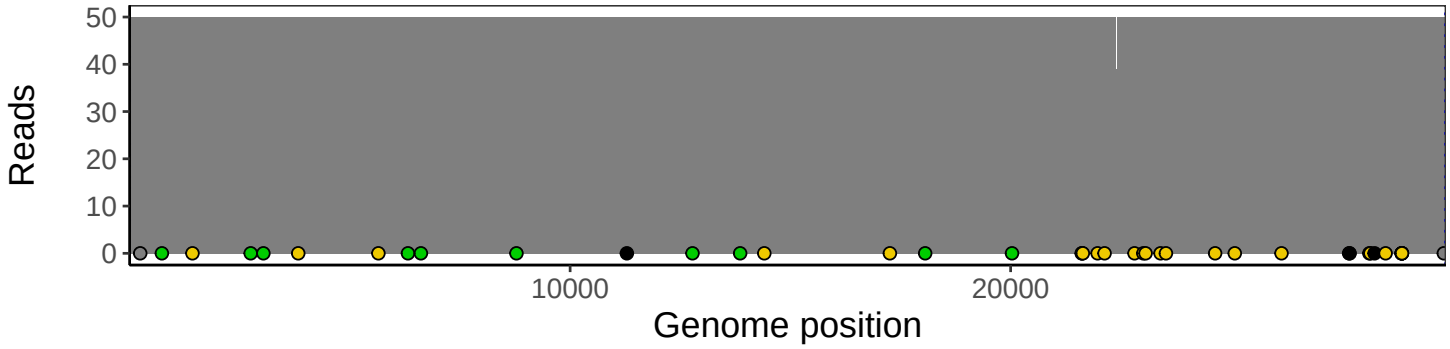
Analyses of individual experiments and composite results

VSP1387-1 | 2021-03-22 | Saline | UPHS-0342 | genomes | single experiment

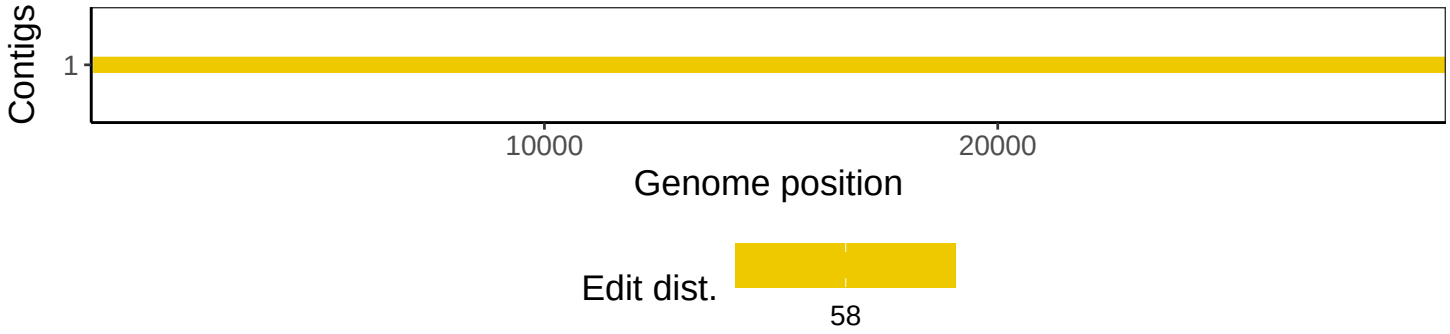
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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