COVID-19 subject UPHS-0159

2021-05-05

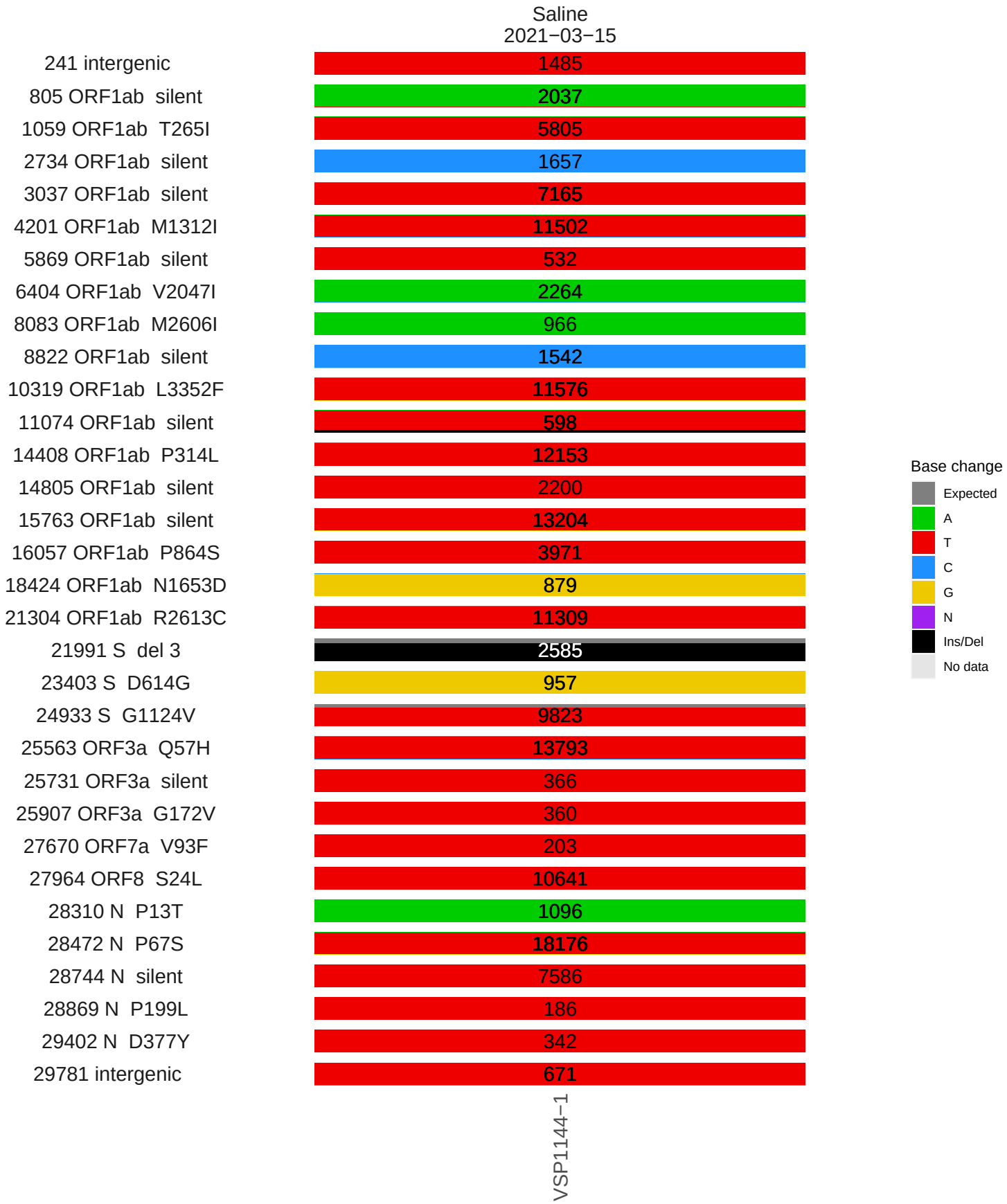
The table below provides a summary of subject samples for which sequencing data is available.
The experiments column shows the number of sequencing experiments performed for each specimen.
Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP1144-1	single experiment	NA	Saline	2021-03-15	29.84	B.1.2	99.8%	99.8%

Variants shared across samples

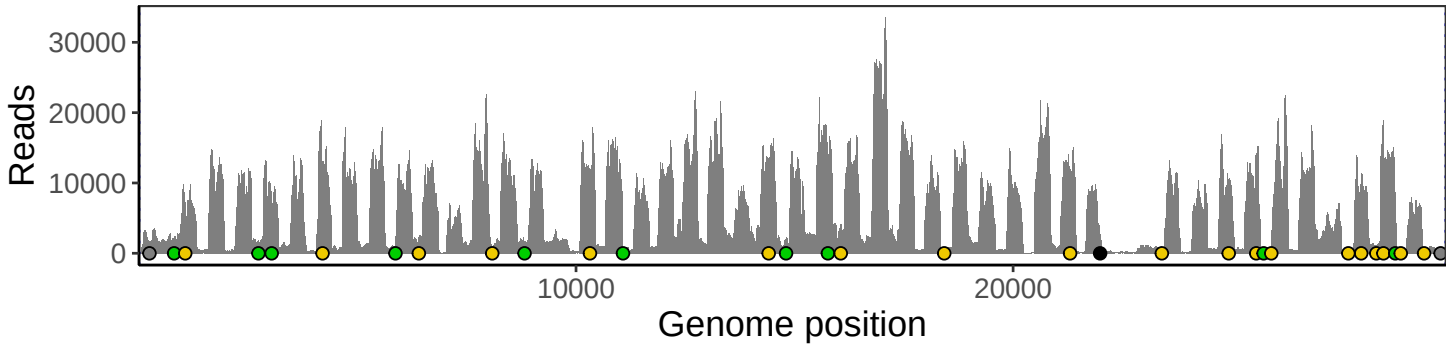
The heat map below shows how variants (reference genome /home/everett/projects/SARS-CoV-2-Philadelphia/Wuhan-Hu-1) are shared across subject samples where the percent variance is colored. Variants are called if a variant position is covered by 5 or more reads, the alternative base is found in > 50% of 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.


Saline
2021−03−15
241 intergenic
1485
805 ORF1ab silent
2037
1059 ORF1ab T265I
5805
2734 ORF1ab silent
1657
3037 ORF1ab silent
7165
4201 ORF1ab M1312I
11502
5869 ORF1ab silent
532
6404 ORF1ab V2047I
2264
8083 ORF1ab M2606I
966
8822 ORF1ab silent
1542
10319 ORF1ab L3352F
11576
11074 ORF1ab silent
598
14408 ORF1ab P314L
12153
14805 ORF1ab silent
2200
15763 ORF1ab silent
13204
16057 ORF1ab P864S
3971
18424 ORF1ab N1653D
879
21304 ORF1ab R2613C
11309
21991 S del 3
2585
23403 S D614G
957
24933 S G1124V
9823
25563 ORF3a Q57H
13793
25731 ORF3a silent
366
25907 ORF3a G172V
360
27670 ORF7a V93F
203
27964 ORF8 S24L
10641
28310 N P13T
1096
28472 N P67S
18176
28744 N silent
7586
28869 N P199L
186
29402 N D377Y
342
29781 intergenic
671
VSP1144−1
Base change
Expected
A
T
C
G
N
Ins/Del
No data

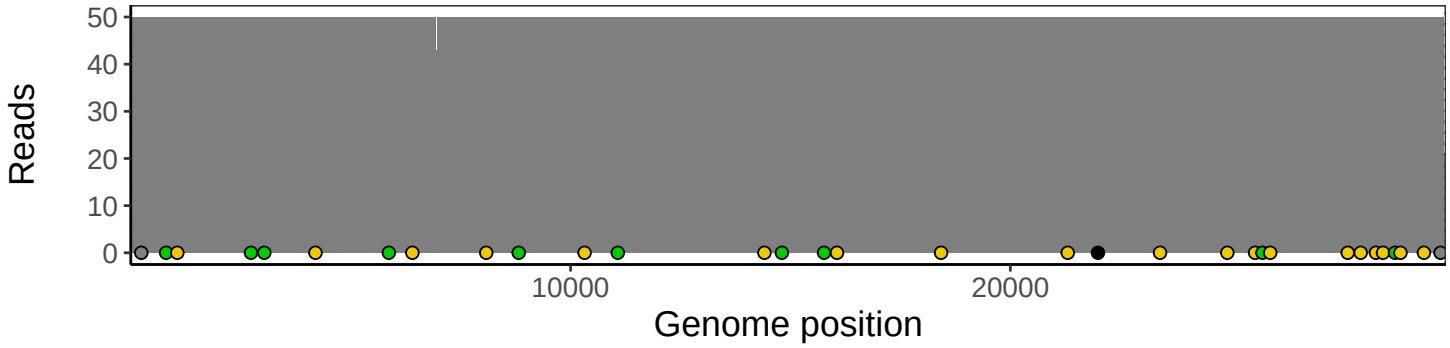
Analyses of individual experiments and composite results

VSP1144-1 | 2021-03-15 | Saline | UPHS-0159 | genomes | single experiment

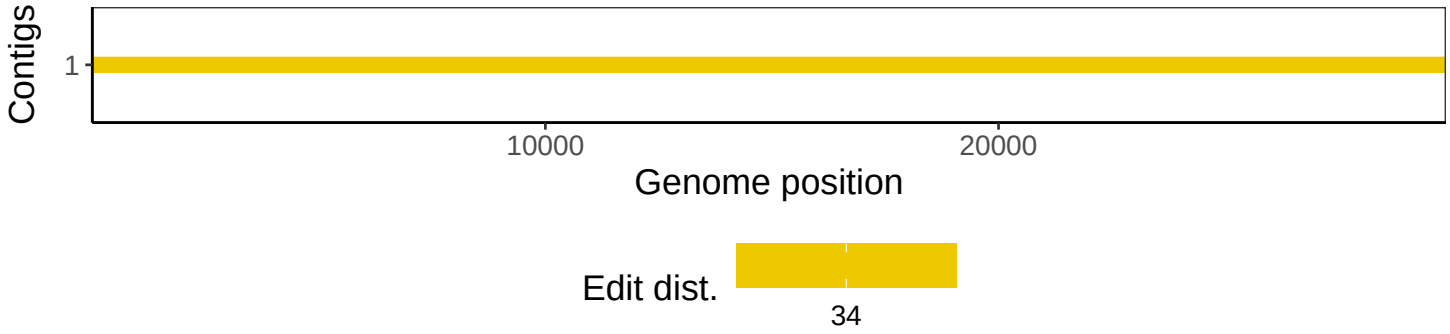
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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