COVID-19 subject UPHS-1637

2021-06-03

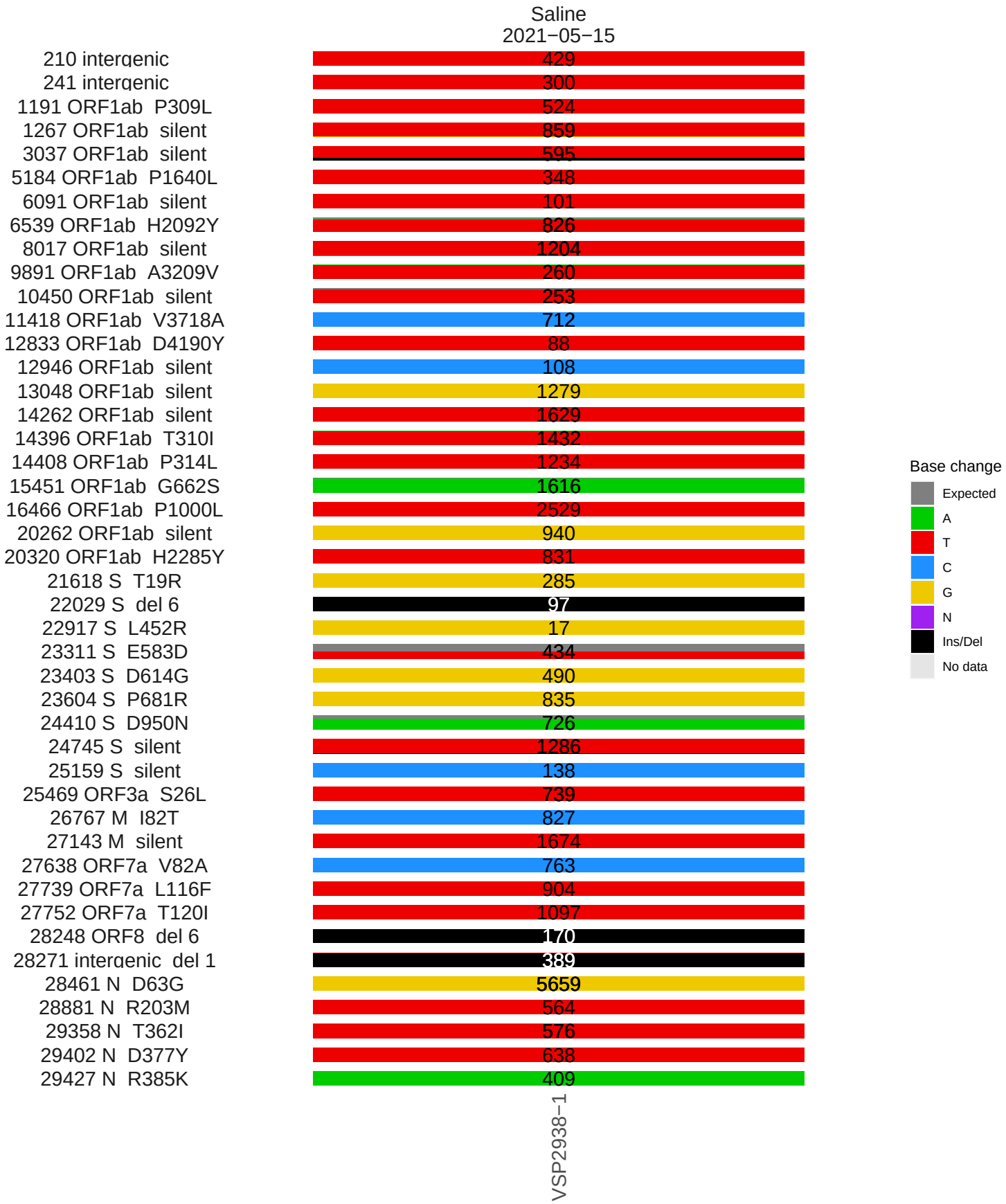
The table below provides a summary of subject samples for which sequencing data is available.
The experiments column shows the number of sequencing experiments performed for each specimen.
Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP2938-1	single experiment	NA	Saline	2021-05-15	21.70	B.1.617	98.5%	98.1%

Variants shared across samples

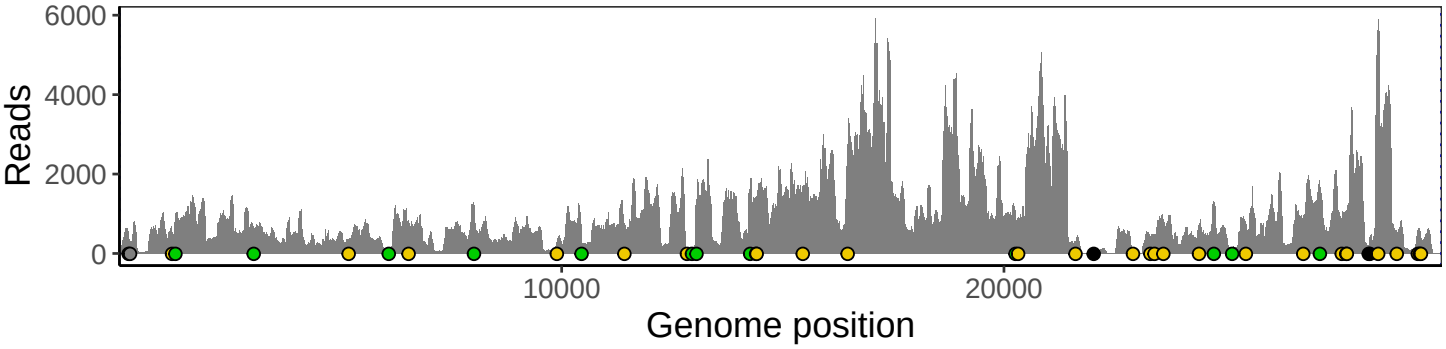
The heat map below shows how variants (reference genome /home/common/SARS-CoV-2-Philadelphia/Wuhan-Hu-1) are shared across subject samples where the percent variance is colored. Variants are called if a variant position is covered by 5 or more reads, the alternative base is found in $> 50\%$ of read 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
2021−05−15
210 intergenic
241 intergenic
1191 ORF1ab P309L
1267 ORF1ab silent
3037 ORF1ab silent
5184 ORF1ab P1640L
6091 ORF1ab silent
6539 ORF1ab H2092Y
8017 ORF1ab silent
9891 ORF1ab A3209V
10450 ORF1ab silent
11418 ORF1ab V3718A
12833 ORF1ab D4190Y
12946 ORF1ab silent
13048 ORF1ab silent
14262 ORF1ab silent
14396 ORF1ab T310I
14408 ORF1ab P314L
15451 ORF1ab G662S
16466 ORF1ab P1000L
20262 ORF1ab silent
20320 ORF1ab H2285Y
21618 S T19R
22029 S del 6
22917 S L452R
23311 S E583D
23403 S D614G
23604 S P681R
24410 S D950N
24745 S silent
25159 S silent
25469 ORF3a S26L
26767 M I82T
27143 M silent
27638 ORF7a V82A
27739 ORF7a L116F
27752 ORF7a T120I
28248 ORF8 del 6
28271 intergenic del 1
28461 N D63G
28881 N R203M
29358 N T362I
29402 N D377Y
29427 N R385K
429
300
524
859
595
348
101
826
1204
260
253
712
88
108
1279
1629
1432
1234
1616
2529
940
831
285
97
17
434
490
835
726
1286
138
739
827
1674
763
904
1097
170
389
5659
564
576
638
409
VSP2938−1
Base change
Expected
A
T
C
G
N
Ins/Del
No data

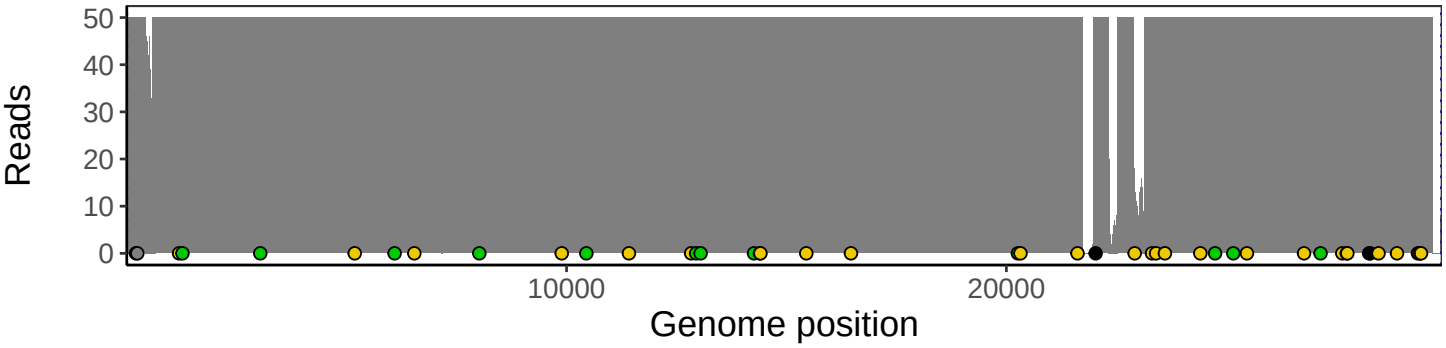
Analyses of individual experiments and composite results

VSP2938-1 | 2021-05-15 | Saline | UPHS-1637 | genomes | single experiment

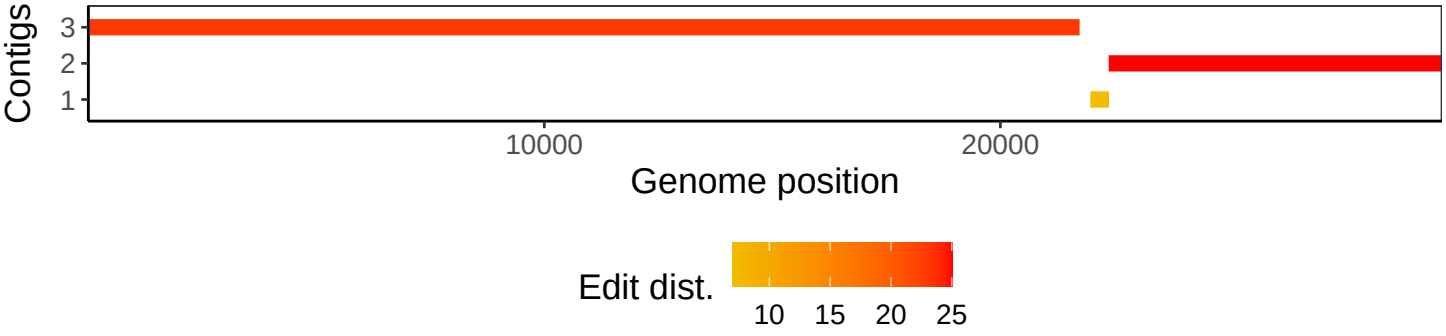
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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	3.3.3
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