

COVID-19 subject UPHS-0734

2021-06-23

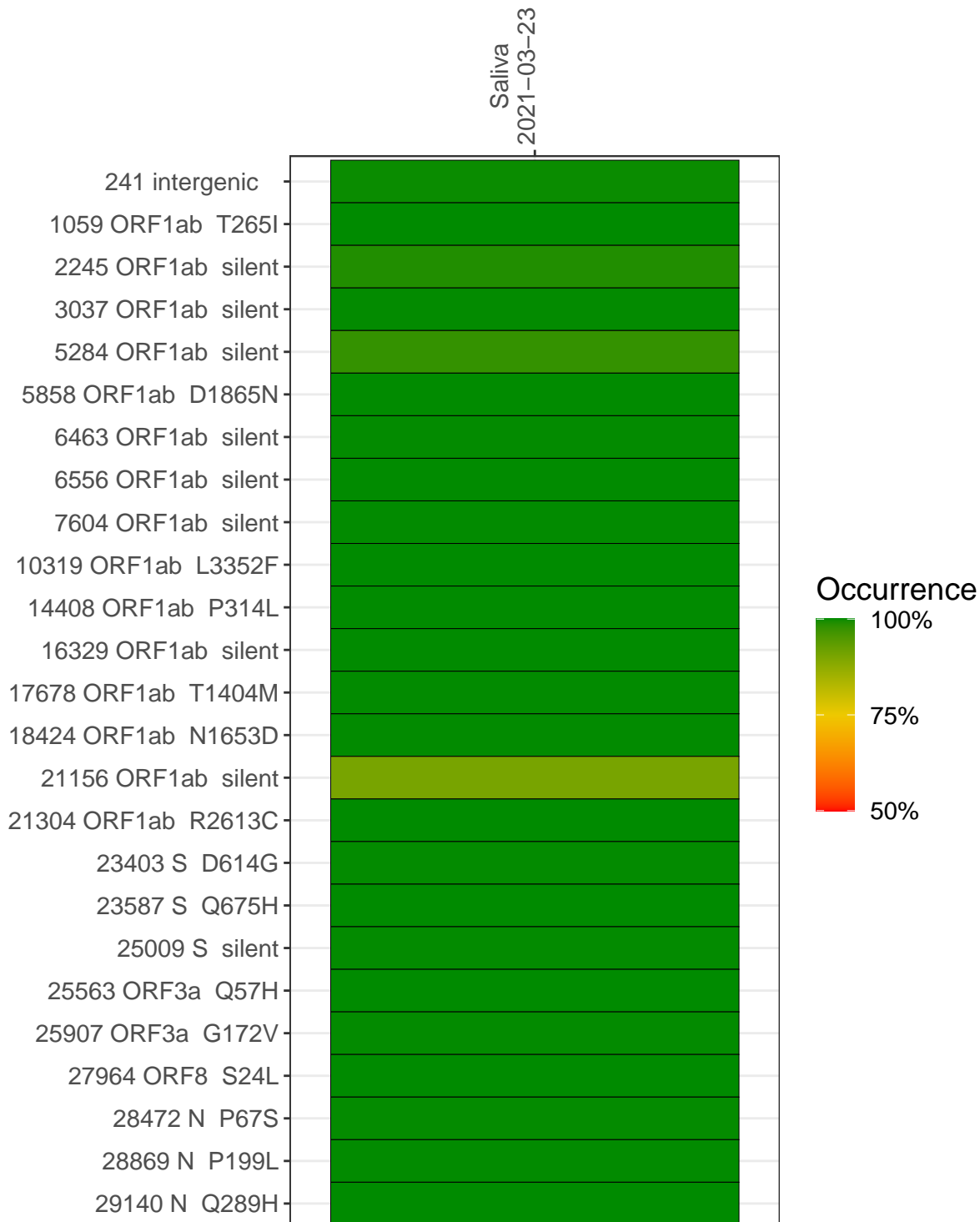
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP1952-1	single experiment	NA	Saliva	2021-03-23	29.65	B.1.2	99.3%	99.2%

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.



Saliva
2021-03-23

241 intergenic	2010
1059 ORF1ab T265I	2548
2245 ORF1ab silent	3805
3037 ORF1ab silent	3619
5284 ORF1ab silent	2349
5858 ORF1ab D1865N	3995
6463 ORF1ab silent	5174
6556 ORF1ab silent	4864
7604 ORF1ab silent	3729
10319 ORF1ab L3352F	3331
14408 ORF1ab P314L	6273
16329 ORF1ab silent	6203
17678 ORF1ab T1404M	5686
18424 ORF1ab N1653D	6410
21156 ORF1ab silent	6729
21304 ORF1ab R2613C	6972
23403 S D614G	5342
23587 S Q675H	2636
25009 S silent	5464
25563 ORF3a Q57H	7518
25907 ORF3a G172V	4607
27964 ORF8 S24L	6618
28472 N P67S	5656
28869 N P199L	2070
29140 N Q289H	1867

Base change

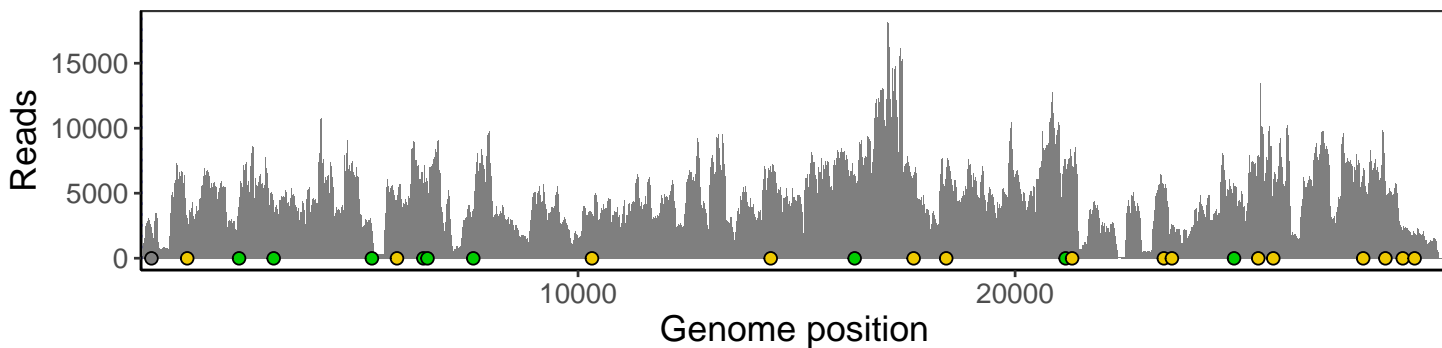


VSP1952-1

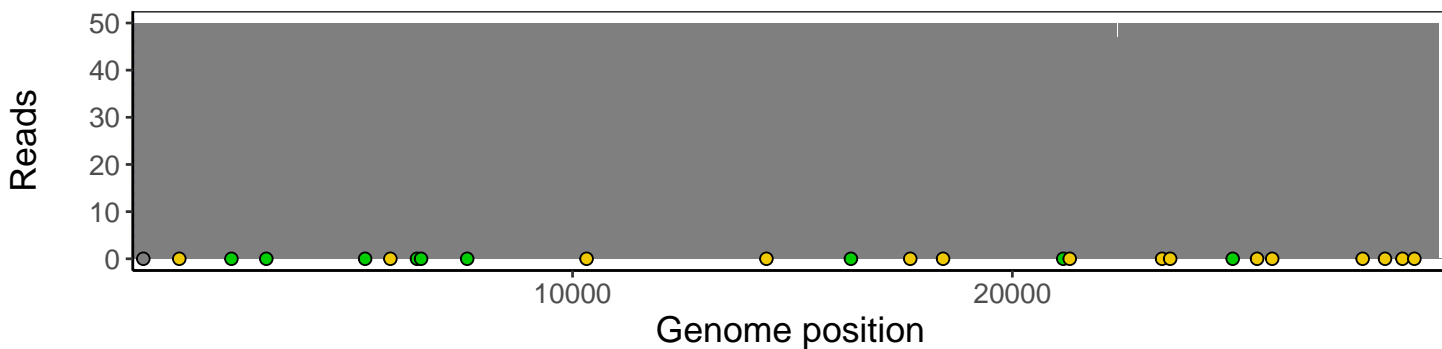
Analyses of individual experiments and composite results

VSP1952-1 | 2021-03-23 | Saliva | UPHS-0734 | genomes | single experiment

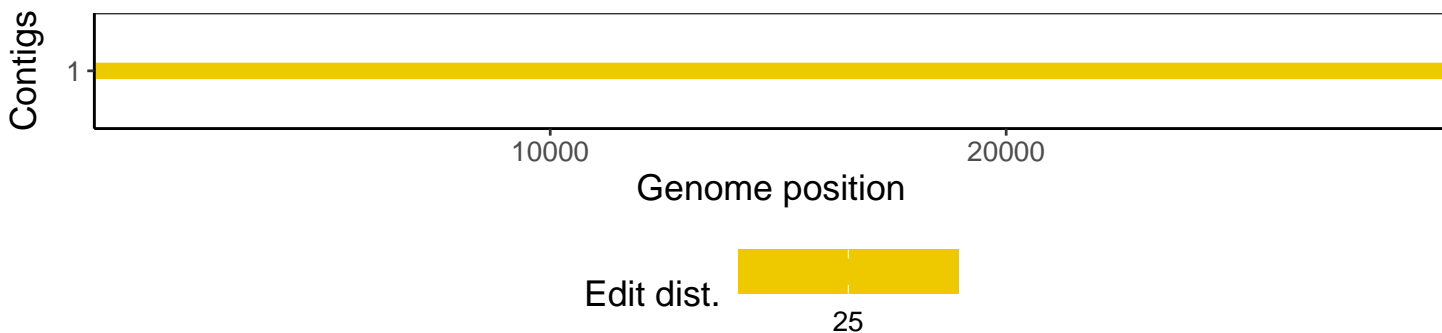
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according to variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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 htlib 1.10
bcftools	1.10.2-34-g1a12af0-dirty Using htlib 1.10.2-57-gf58a6f3
pangolin	3.1.3
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	3.3.3
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