COVID-19 subject UPHS-0747

2021-05-05

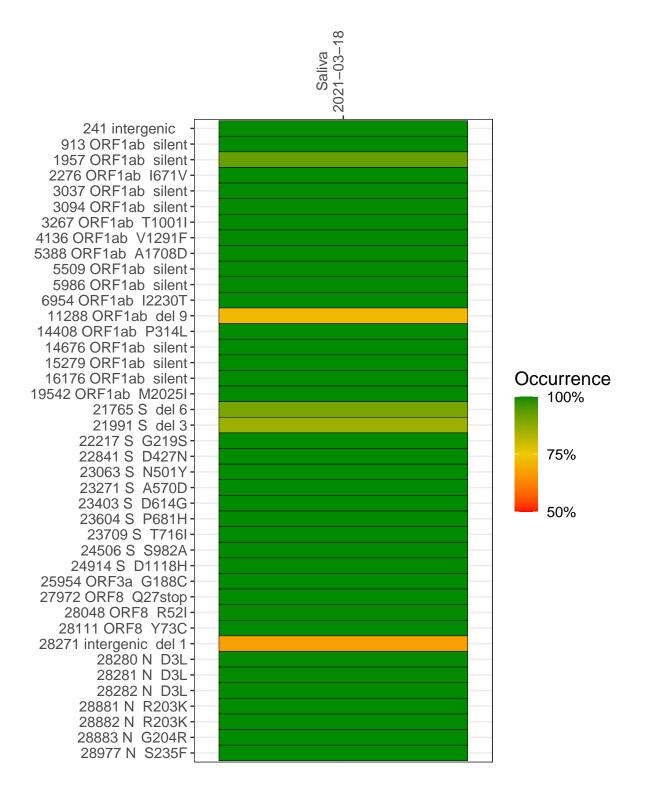
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software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

Table 1. Sample summary.

Experiment	Туре	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP1965-1	single experiment	NA	Saliva	2021-03-18	29.63	B.1.1.7	99.3%	99.3%

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.



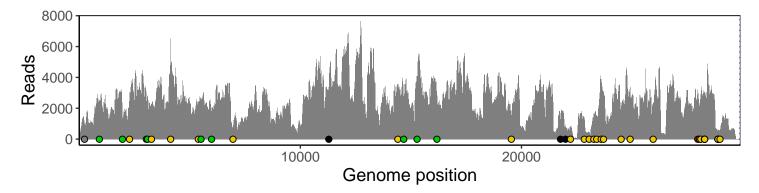
Saliva 2021-03-18

	2021-03-10
241 intergenic	983
913 ORF1ab silent	2261
1957 ORF1ab silent	1372
2276 ORF1ab I671V	2711
3037 ORF1ab silent	2496
3094 ORF1ab silent	2700
3267 ORF1ab T1001I	2292
4136 ORF1ab V1291F	4799
5388 ORF1ab A1708D	2136
5509 ORF1ab silent	2488
5986 ORF1ab silent	1756
6954 ORF1ab I2230T	737
11288 ORF1ab del 9	1780
14408 ORF1ab P314L	3172
14676 ORF1ab silent	2486
15279 ORF1ab silent	3580
16176 ORF1ab silent	2743
19542 ORF1ab M2025l	1546
21765 S del 6	1220
21991 S del 3	735
22217 S G219S	737
22841 S D427N	1214
23063 S N501Y	445
23271 S A570D	1400
23403 S D614G	1497
23604 S P681H	3758
23709 S T716I	2965
24506 S S982A	2168
24914 S D1118H	3185
25954 ORF3a G188C	2285
27972 ORF8 Q27stop	2980
28048 ORF8 R52I	1946
28111 ORF8 Y73C	2845
28271 intergenic del 1	2414
28280 N D3L	1547
28281 N D3L	1547
28282 N D3L	1687
28881 N R203K	387
28882 N R203K	386
28883 N G204R	387
28977 N S235F	524
	7
	965-1
	O)

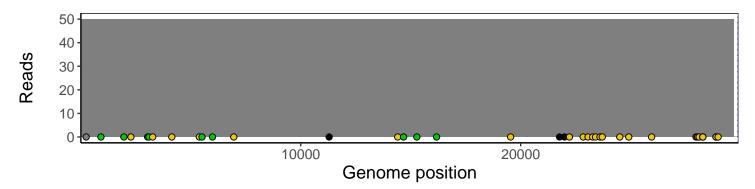
Analyses of individual experiments and composite results

VSP1965-1 | 2021-03-18 | Saliva | UPHS-0747 | genomes | single experiment

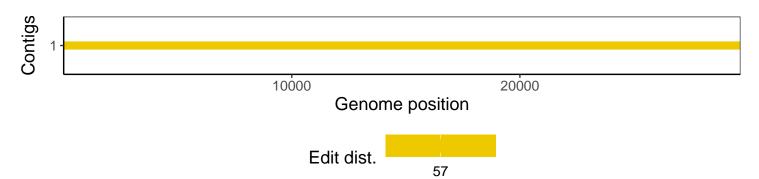
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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
${\it Genomic Alignments}$	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
$\operatorname{GenomeInfoDb}$	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