COVID-19 subject UPHS-0712

2021-04-20

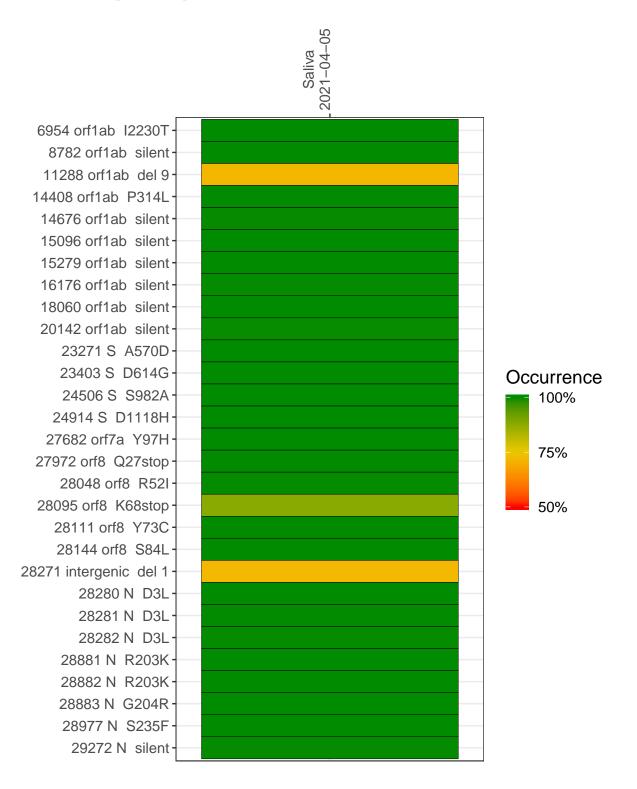
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP1930-1	single experiment	NA	Saliva	2021-04-05	4.42	NA	68.2%	67.8%

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.



Saliva 2021-04-05

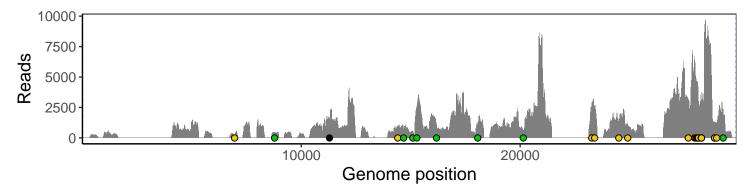
	2021-04-05
6954 orf1ab I2230T	221
8782 orf1ab silent	449
11288 orf1ab del 9	917
14408 orf1ab P314L	915
14676 orf1ab silent	701
15096 orf1ab silent	528
15279 orf1ab silent	2262
16176 orf1ab silent	1079
18060 orf1ab silent	1042
20142 orf1ab silent	553
23271 S A570D	2529
23403 S D614G	2572
24506 S S982A	1594
24914 S D1118H	1171
27682 orf7a Y97H	3135
27972 orf8 Q27stop	6346
28048 orf8 R52I	4239
28095 orf8 K68stop	4607
28111 orf8 Y73C	4243
28144 orf8 S84L	3123
28271 intergenic del 1	2820
28280 N D3L	1976
28281 N D3L	1976
28282 N D3L	2103
28881 N R203K	905
28882 N R203K	899
28883 N G204R	903
28977 N S235F	1190
29272 N silent	2376
	1-0



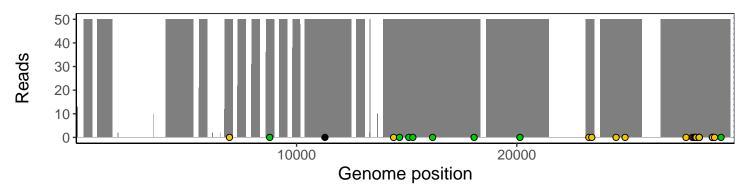
Analyses of individual experiments and composite results

$VSP1930\text{-}1 \mid 2021\text{-}04\text{-}05 \mid Saliva \mid UPHS\text{-}0712 \mid genomes \mid single \ experiment$

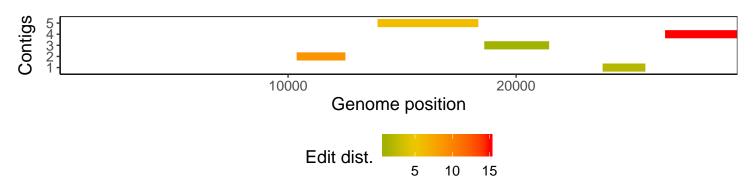
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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