COVID-19 subject UPHS-0720

2021-04-20

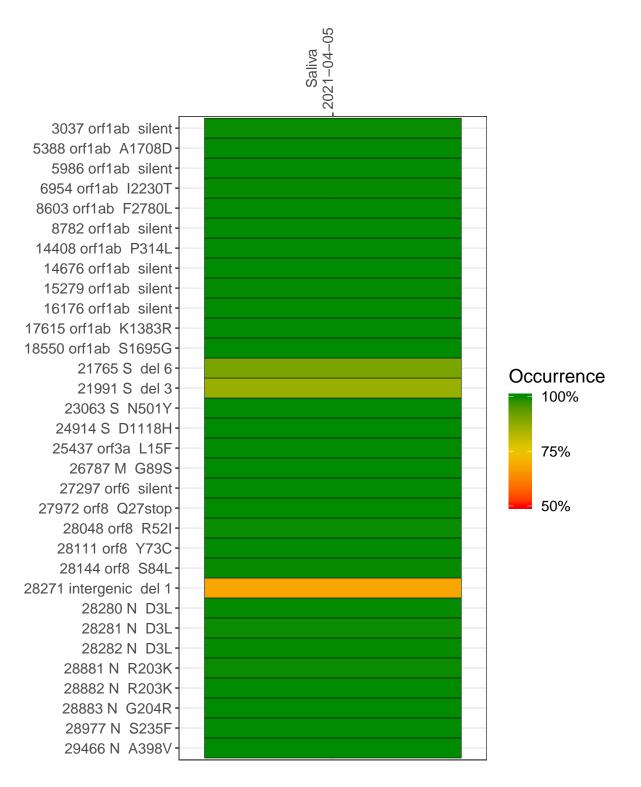
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP1938-1	single experiment	NA	Saliva	2021-04-05	6.75	NA	84.6%	84.4%

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

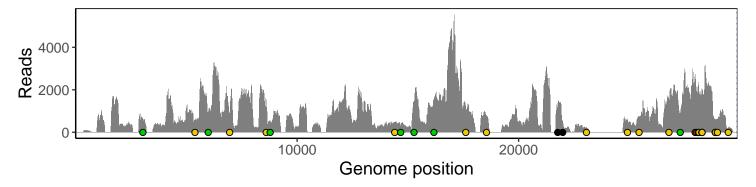
	2021-04-05
3037 orf1ab silent	529
5388 orf1ab A1708D	503
5986 orf1ab silent	906
6954 orf1ab I2230T	879
8603 orf1ab F2780L	1412
8782 orf1ab silent	541
14408 orf1ab P314L	432
14676 orf1ab silent	288
15279 orf1ab silent	1297
16176 orf1ab silent	1552
17615 orf1ab K1383R	1076
18550 orf1ab S1695G	750
21765 S del 6	1012
21991 S del 3	430
23063 S N501Y	232
24914 S D1118H	612
25437 orf3a L15F	744
26787 M G89S	1609
27297 orf6 silent	1073
27972 orf8 Q27stop	2569
28048 orf8 R52I	1641
28111 orf8 Y73C	2161
28144 orf8 S84L	1922
28271 intergenic del 1	1634
28280 N D3L	1064
28281 N D3L	1064
28282 N D3L	1130
28881 N R203K	412
28882 N R203K	410
28883 N G204R	413
28977 N S235F	594
29466 N A398V	202
	7-8
	89



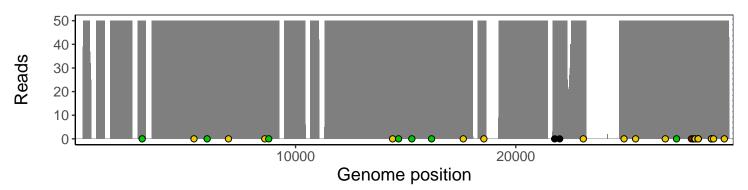
Analyses of individual experiments and composite results

$VSP1938-1 \mid 2021-04-05 \mid Saliva \mid UPHS-0720 \mid genomes \mid single \; experiment$

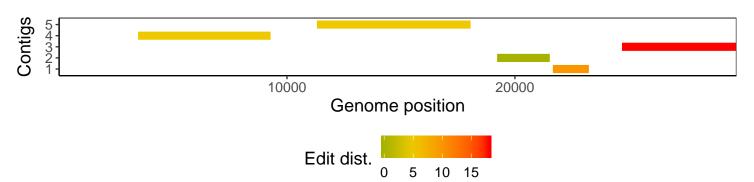
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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