COVID-19 subject 522

2021-06-23

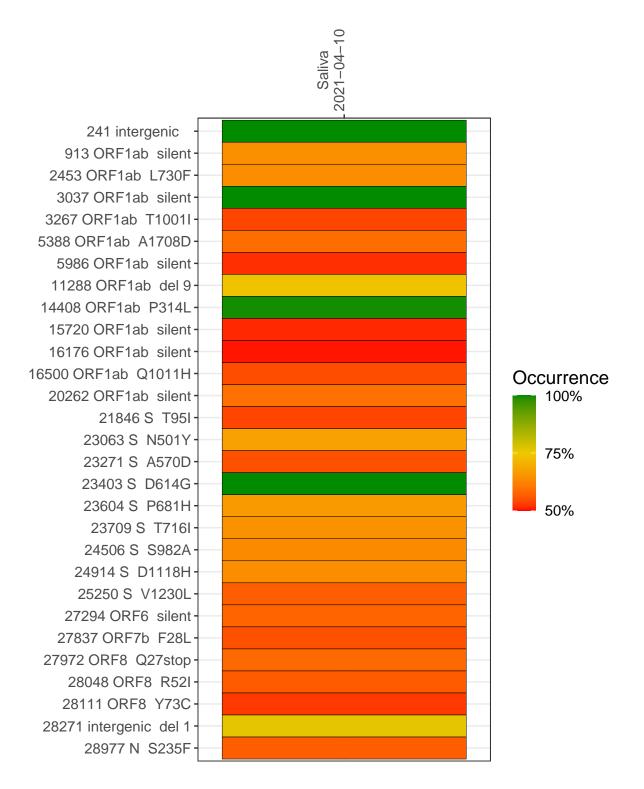
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software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

Table 1. Sample summary.

Experiment	Туре	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP1980-1	single experiment	NA	Saliva	2021-04-10	29.81	B.1.1.7	99.8%	99.7%

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.



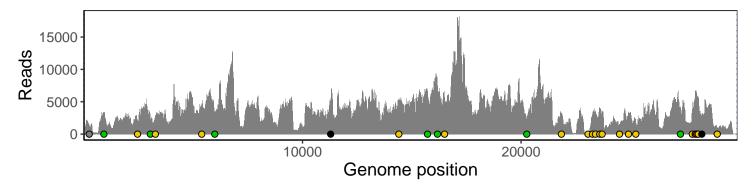
Saliva 2021-04-10

	2021-04-10
241 intergenic	1209
913 ORF1ab silent	3170
2453 ORF1ab L730F	2327
3037 ORF1ab silent	2423
3267 ORF1ab T1001I	2919
5388 ORF1ab A1708D	4151
5986 ORF1ab silent	3344
11288 ORF1ab del 9	3319
14408 ORF1ab P314L	4803
15720 ORF1ab silent	6491
16176 ORF1ab silent	73 97
16500 ORF1ab Q1011H	5470
20262 ORF1ab silent	2225
21846 S T95I	3063
23063 S N501Y	176
23271 S A570D	2941
23403 S D614G	3363
23604 S P681H	3900
23709 S T716I	3465
24506 S S982A	1547
24914 S D1118H	3342
25250 S V1230L	2155
27294 ORF6 silent	3919
27837 ORF7b F28L	3769
27972 ORF8 Q27stop	6391
28048 ORF8 R52I	5747
28111 ORF8 Y73C	5381
28271 intergenic del 1	2723
28977 N S235F	1076
	980-1
	86

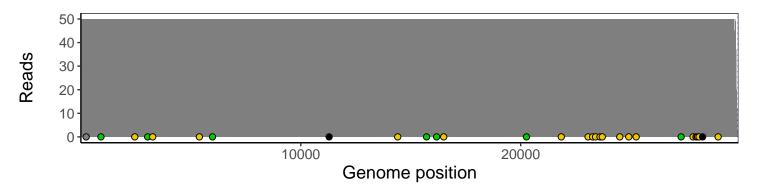
Analyses of individual experiments and composite results

VSP1980-1 | 2021-04-10 | Saliva | 522s | genomes | single experiment

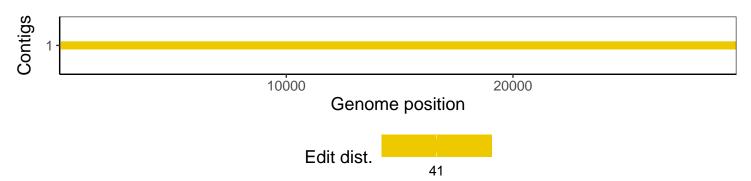
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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	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				
${\bf Summarized Experiment}$	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				