COVID-19 subject UPHS-0183

2021-04-17

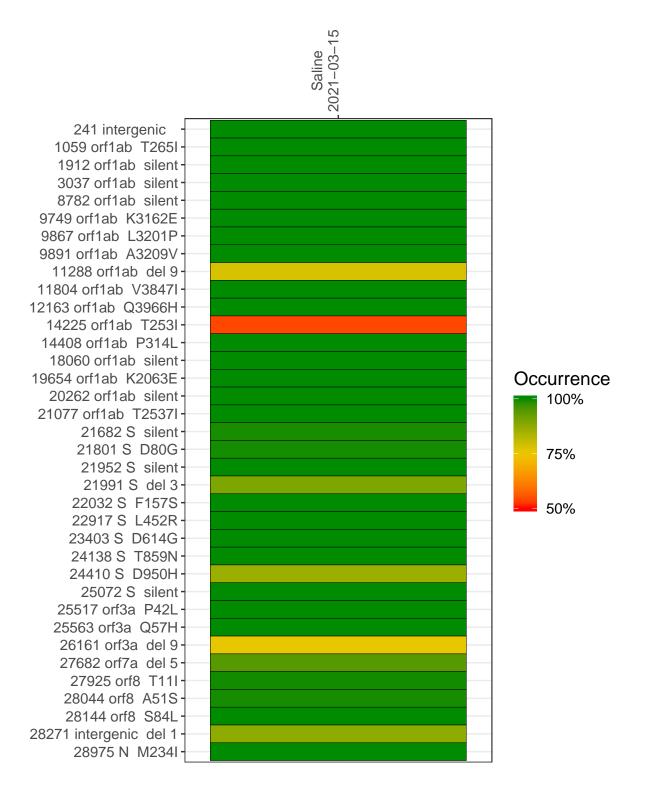
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP1167-1	single experiment	NA	Saline	2021-03-15	21.68	B.1.526.1	99.7%	98.7%

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.



Saline 2021-03-15

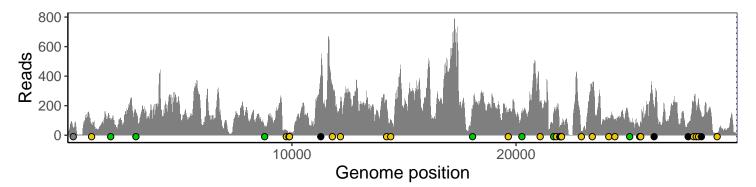
	2021-03-15
241 intergenic	50
1059 orf1ab T265I	60
1912 orf1ab silent	89
3037 orf1ab silent	151
8782 orf1ab silent	57
9749 orf1ab K3162E	34
9867 orf1ab L3201P	20
9891 orf1ab A3209V	21
11288 orf1ab del 9	260
11804 orf1ab V3847I	328
12163 orf1ab Q3966H	218
14225 orf1ab T253l	239
14408 orf1ab P314L	101
18060 orf1ab silent	176
19654 orf1ab K2063E	157
20262 orf1ab silent	159
21077 orf1ab T2537I	160
21682 S silent	140
21801 S D80G	177
21952 S silent	42
21991 S del 3	58
22032 S F157S	52
22917 S L452R	27
23403 S D614G	354
24138 S T859N	118
24410 S D950H	107
25072 S silent	82
25517 orf3a P42L	55
25563 orf3a Q57H	87
26161 orf3a del 9	167
27682 orf7a del 5	71
27925 orf8 T11I	229
28044 orf8 A51S	165
28144 orf8 S84L	231
28271 intergenic del 1	133
28975 N M234I	14
	7
	VSP1167-1
	<u>C</u>
	S)



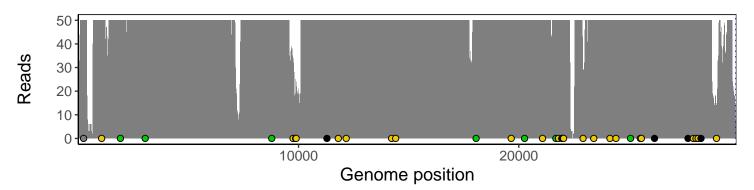
Analyses of individual experiments and composite results

VSP1167-1 | 2021-03-15 | Saline | UPHS-0183 | genomes | single experiment

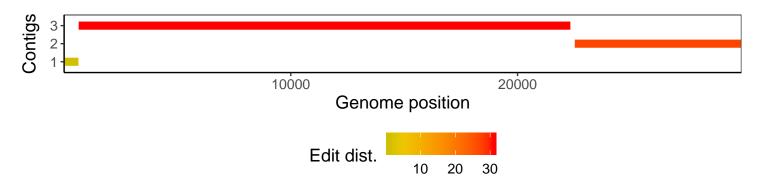
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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