COVID-19 subject UPHS-0103

2021-04-01

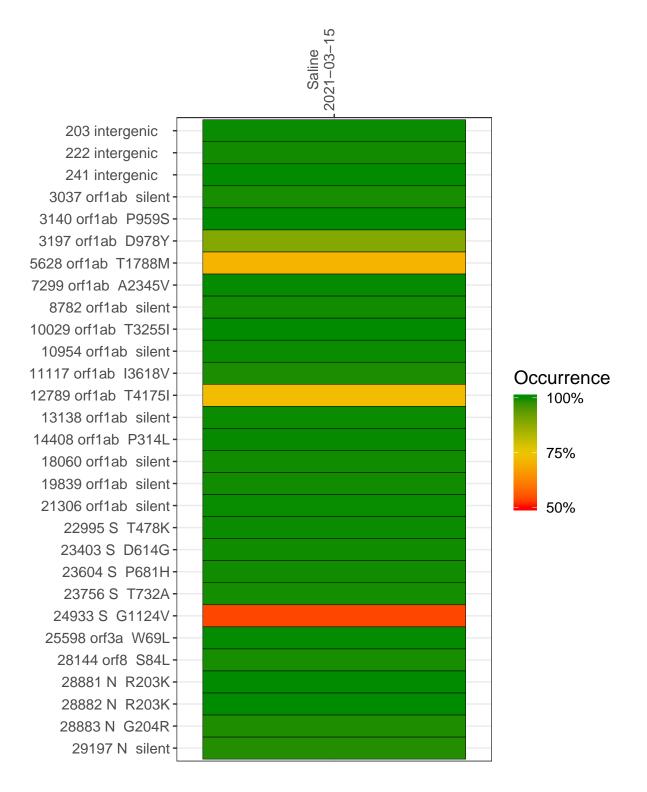
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP1088-1	single experiment	NA	Saline	2021-03-15	29.88	B.1.1.222	99.8%	99.8%

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.



Saline 2021-03-15

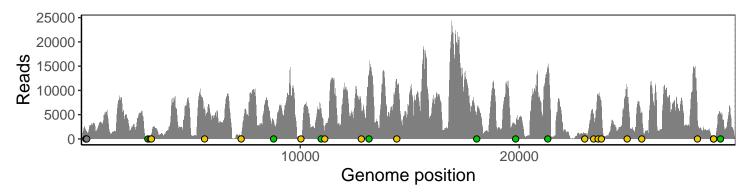
	2021-03-13
203 intergenic	1664
222 intergenic	1610
241 intergenic	1353
3037 orf1ab silent	288
3140 orf1ab P959S	266
3197 orf1ab D978Y	1006
5628 orf1ab T1788M	5944
7299 orf1ab A2345V	7 67
8782 orf1ab silent	3676
10029 orf1ab T3255I	369
10954 orf1ab silent	5695
11117 orf1ab I3618V	2290
12789 orf1ab T4175I	5483
13138 orf1ab silent	13039
14408 orf1ab P314L	10961
18060 orf1ab silent	5165
19839 orf1ab silent	2746
21306 orf1ab silent	13380
22995 S T478K	1162
23403 S D614G	4937
23604 S P681H	8637
23756 S T732A	7070
24933 S G1124V	10401
25598 orf3a W69L	4828
28144 orf8 S84L	3996
28881 N R203K	205
28882 N R203K	204
28883 N G204R	207
29197 N silent	5111
	088-1
	108



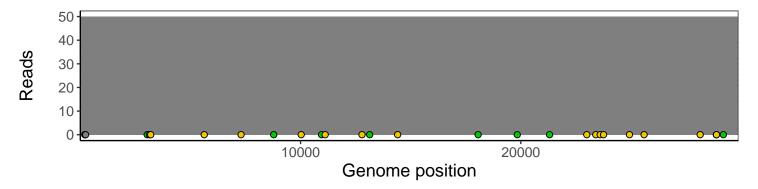
Analyses of individual experiments and composite results

VSP1088-1 | 2021-03-15 | Saline | UPHS-0103 | genomes | single experiment

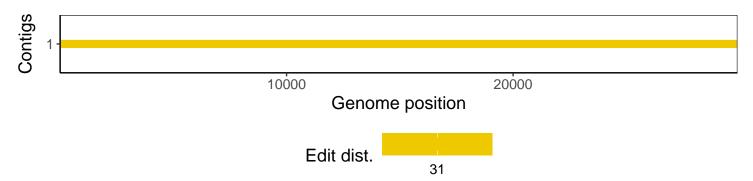
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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	2.3.3
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	3.0.0
tidyverse	1.2.1
ShortRead	1.34.2
$\operatorname{GenomicAlignments}$	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