COVID-19 subject HUP PH-0025

2021-04-01

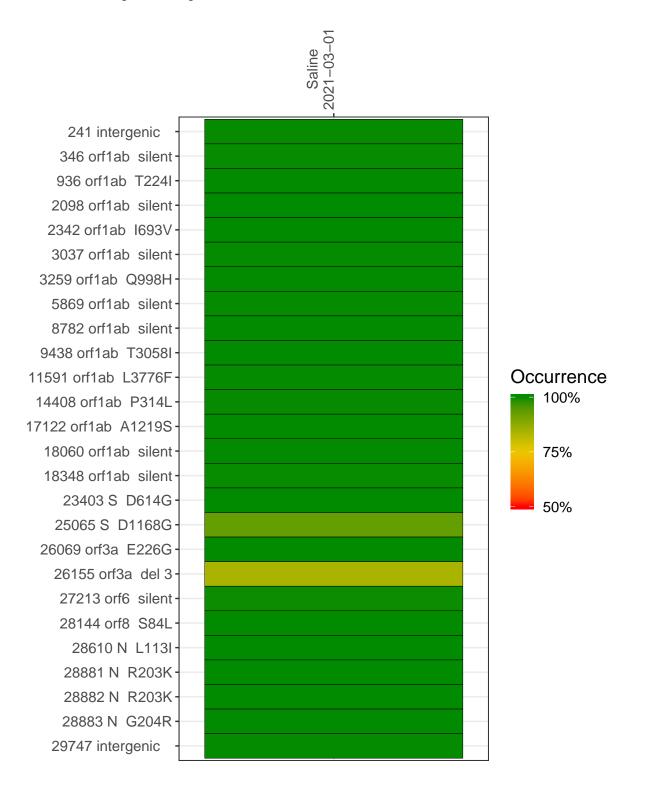
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP0899-1	single experiment	NA	Saline	2021-03-01	29.89	B.1.1.304	99.9%	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 2021-03-01

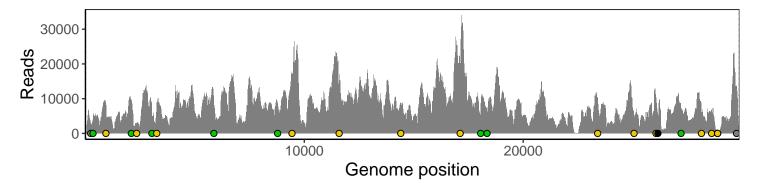
	2021-03-01
241 intergenic	2708
346 orf1ab silent	3997
936 orf1ab T224I	8155
2098 orf1ab silent	9410
2342 orf1ab I693V	3001
3037 orf1ab silent	5136
3259 orf1ab Q998H	8131
5869 orf1ab silent	8457
8782 orf1ab silent	8387
9438 orf1ab T3058I	17341
11591 orf1ab L3776F	15135
14408 orf1ab P314L	8034
17122 orf1ab A1219S	19923
18060 orf1ab silent	8110
18348 orf1ab silent	5528
23403 S D614G	10026
25065 S D1168G	4058
26069 orf3a E226G	10259
26155 orf3a del 3	5373
27213 orf6 silent	6102
28144 orf8 S84L	6267
28610 N L113I	6479
28881 N R203K	629
28882 N R203K	625
28883 N G204R	630
29747 intergenic	13365
	1-669
	9,



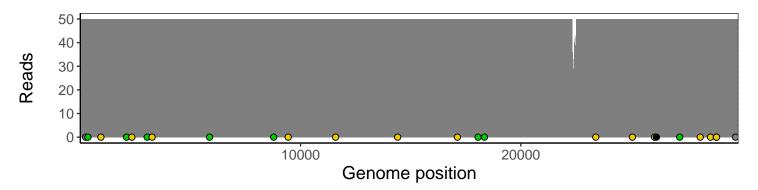
Analyses of individual experiments and composite results

$VSP0899-1 \mid 2021-03-01 \mid Saline \mid HUP\ PH-0025 \mid genomes \mid single\ experiment$

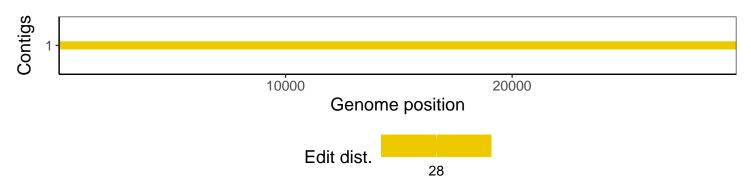
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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	2.3.3
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