

COVID-19 subject HUP PH-0027

2021-03-29

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.

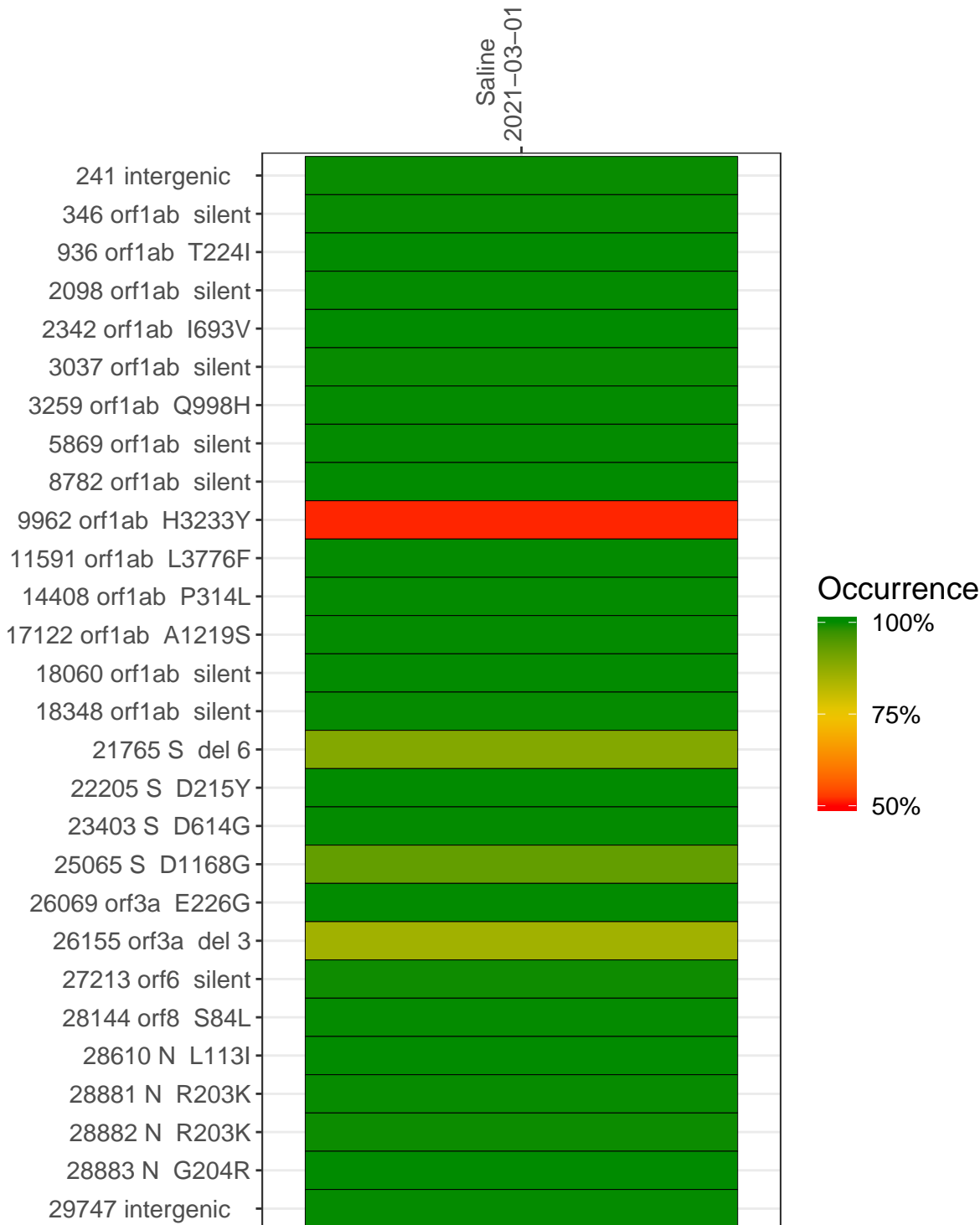
Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software tool (Rambaut et al 2020) for genomes with $> 90\%$ sequence coverage.

Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (≥ 5 reads)
VSP0901-1	single experiment	NA	Saline	2021-03-01	29.84	B.1.1.304	99.8%	99.8%

Variants shared across samples

The heat map below shows how variants (reference genome 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

241 intergenic	3719
346 orf1ab silent	5399
936 orf1ab T224I	9945
2098 orf1ab silent	11190
2342 orf1ab I693V	4207
3037 orf1ab silent	7325
3259 orf1ab Q998H	9382
5869 orf1ab silent	10087
8782 orf1ab silent	11223
9962 orf1ab H3233Y	5629
11591 orf1ab L3776F	17659
14408 orf1ab P314L	11319
17122 orf1ab A1219S	22105
18060 orf1ab silent	10845
18348 orf1ab silent	6181
21765 S del 6	4801
22205 S D215Y	5328
23403 S D614G	11254
25065 S D1168G	5452
26069 orf3a E226G	12849
26155 orf3a del 3	6927
27213 orf6 silent	7456
28144 orf8 S84L	7267
28610 N L113I	7817
28881 N R203K	679
28882 N R203K	677
28883 N G204R	686
29747 intergenic	16263

Base change

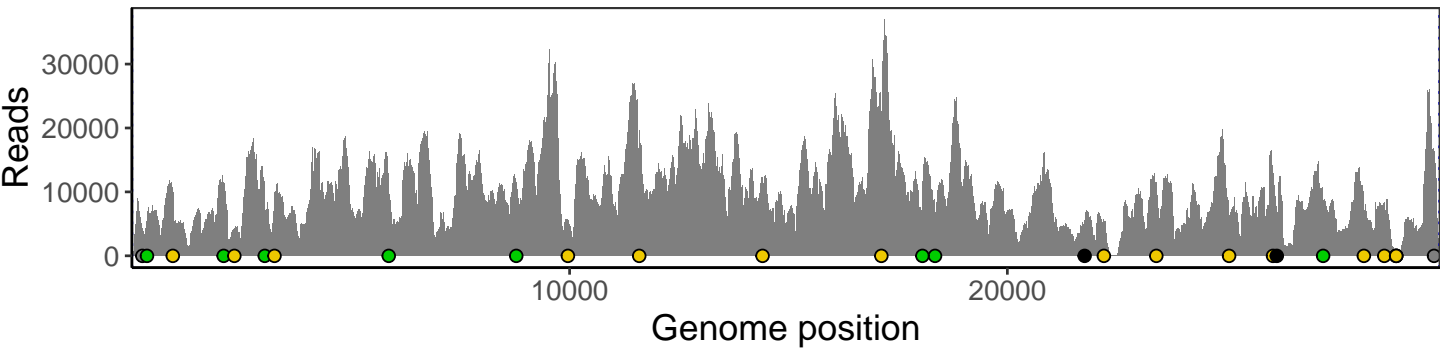


VSP0901-1

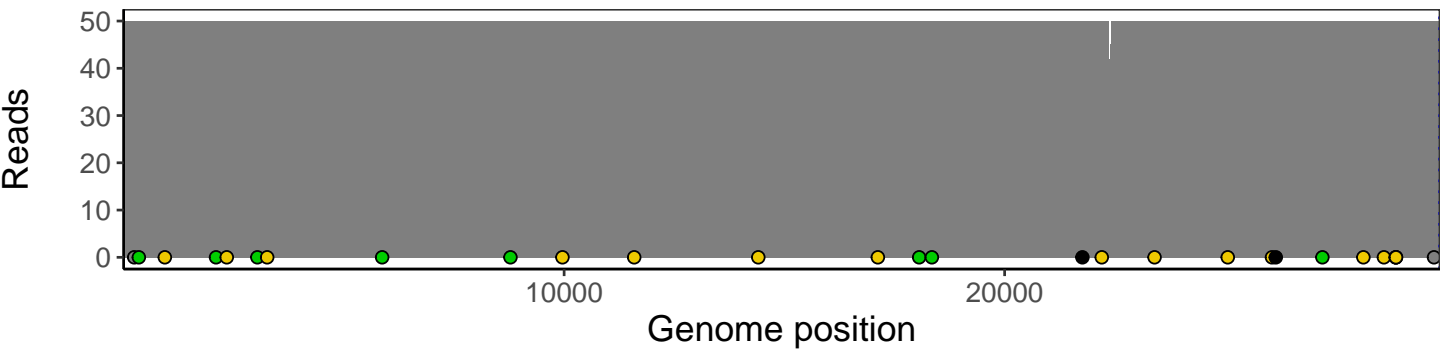
Analyses of individual experiments and composite results

VSP0901-1 | 2021-03-01 | Saline | HUP PH-0027 | genomes | single experiment

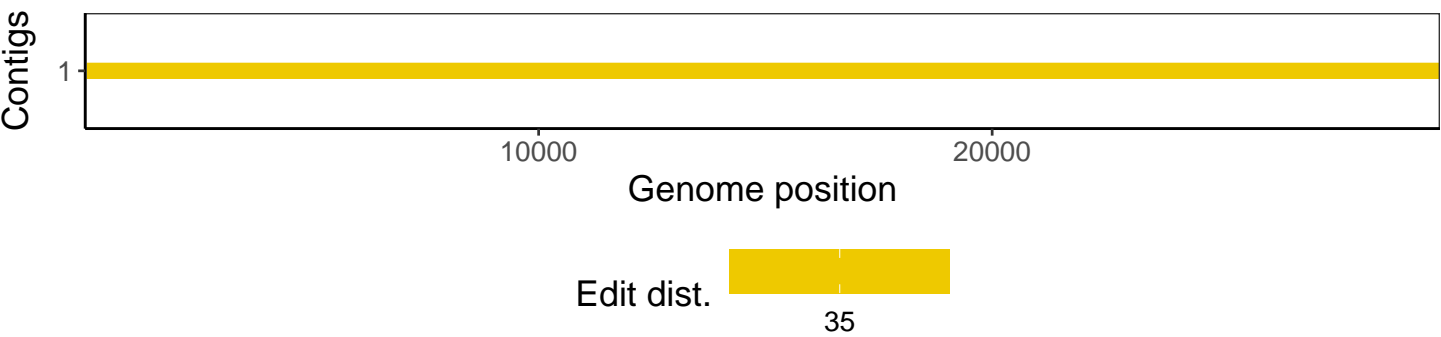
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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 htlib 1.10
bcftools	1.10.2-34-g1a12af0-dirty Using htlib 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
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
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