

COVID-19 subject HUP Q-0047

2021-03-29

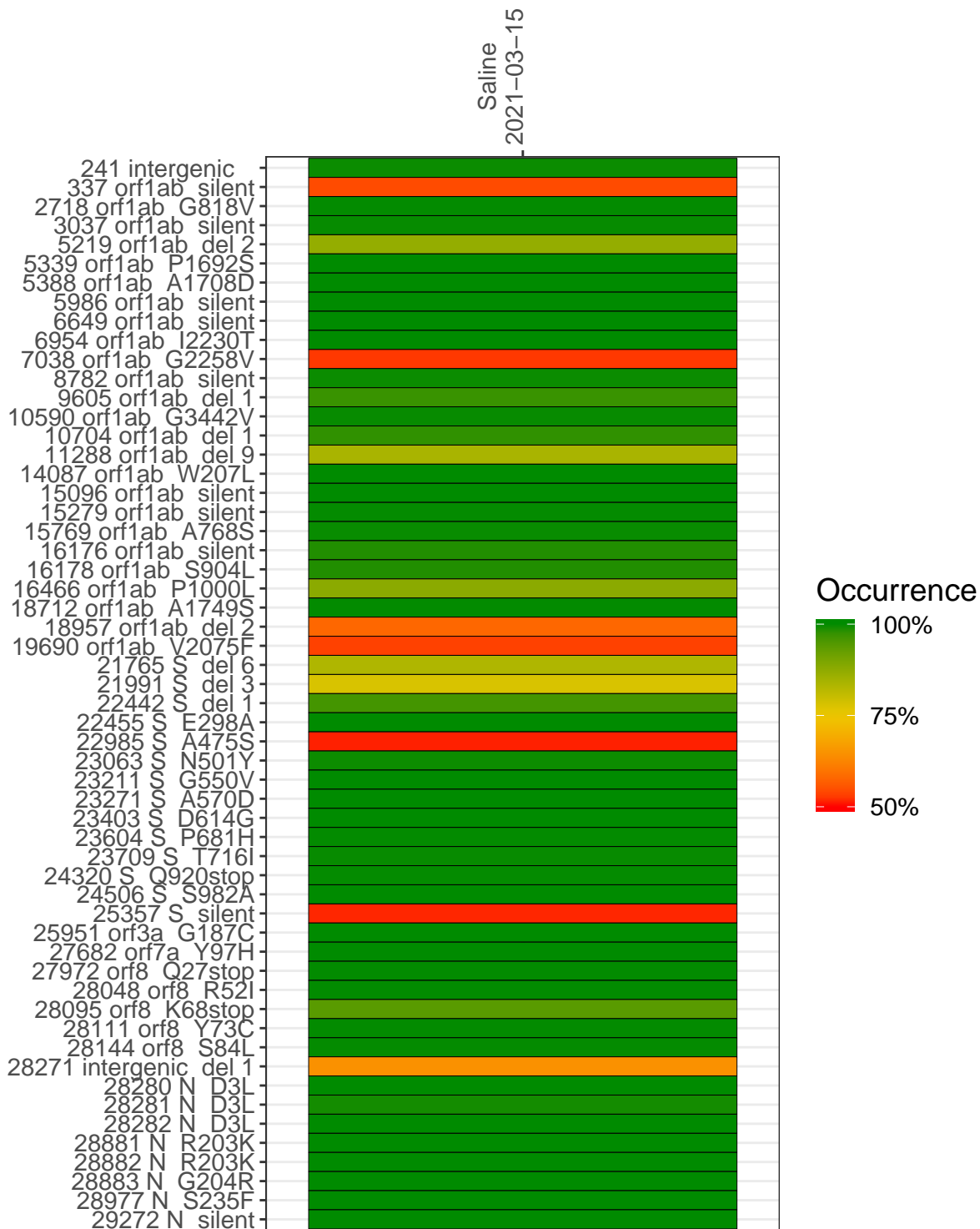
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP1079-1	single experiment	NA	Saline	2021-03-15	3.11	NA	70.3%	70.0%

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	757
337 orf1ab silent	882
2718 orf1ab G818V	1279
3037 orf1ab silent	1500
5219 orf1ab del 2	3116
5339 orf1ab P1692S	634
5388 orf1ab A1708D	643
5986 orf1ab silent	955
6649 orf1ab silent	1937
6954 orf1ab I2230T	60
7038 orf1ab G2258V	97
8782 orf1ab silent	1460
9605 orf1ab del 1	555
10590 orf1ab G3442V	624
10704 orf1ab del 1	323
11288 orf1ab del 9	1111
14087 orf1ab W207L	598
15096 orf1ab silent	1271
15279 orf1ab silent	1147
15769 orf1ab A768S	1494
16176 orf1ab silent	522
16178 orf1ab S904L	527
16466 orf1ab P1000L	434
18712 orf1ab A1749S	1365
18957 orf1ab del 2	901
19690 orf1ab V2075F	887
21765 S del 6	2100
21991 S del 3	732
22442 S del 1	52
22455 S E298A	58
22985 S A475S	369
23063 S N501Y	331
23211 S G550V	439
23271 S A570D	612
23403 S D614G	703
23604 S P681H	1556
23709 S T716I	1757
24320 S Q920stop	1002
24506 S S982A	455
25357 S silent	1345
25951 orf3a G187C	789
27682 orf7a Y97H	622
27972 orf8 Q27stop	5971
28048 orf8 R52I	5037
28095 orf8 K68stop	4141
28111 orf8 Y73C	3707
28144 orf8 S84L	2074
28271 intergenic del 1	1155
28280 N D3L	753
28281 N D3L	753
28282 N D3L	787
28881 N R203K	65
28882 N R203K	64
28883 N G204R	64
28977 N S235F	59
29272 N silent	3049

Base change

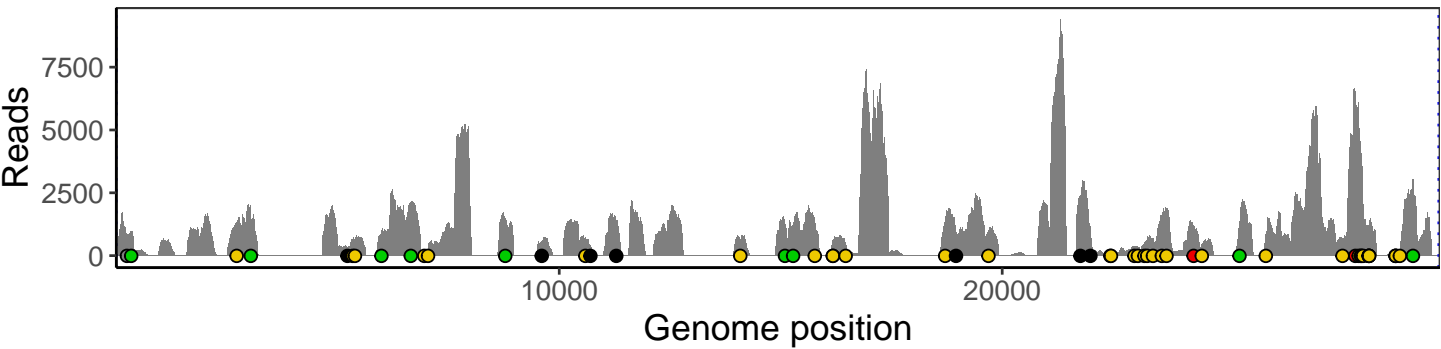


VSP1079-1

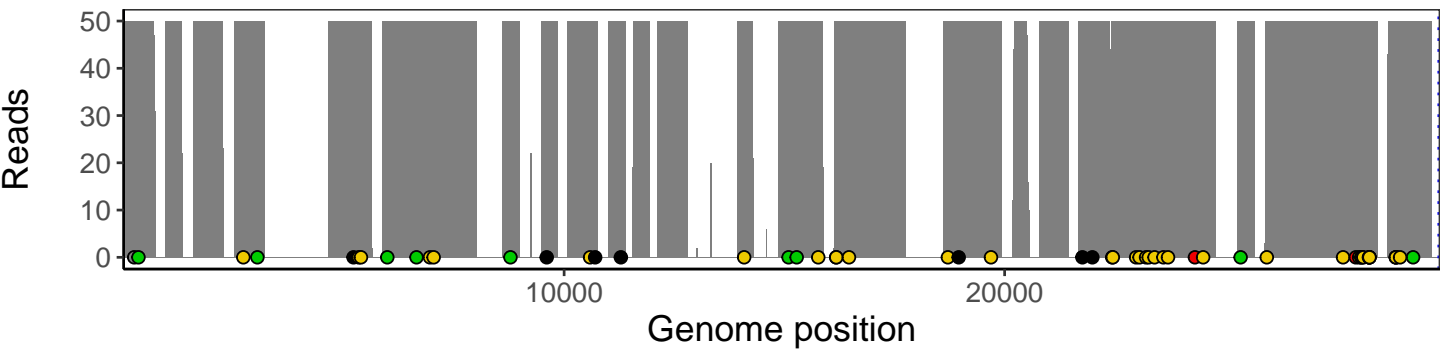
Analyses of individual experiments and composite results

VSP1079-1 | 2021-03-15 | Saline | HUP Q-0047 | genomes | single experiment

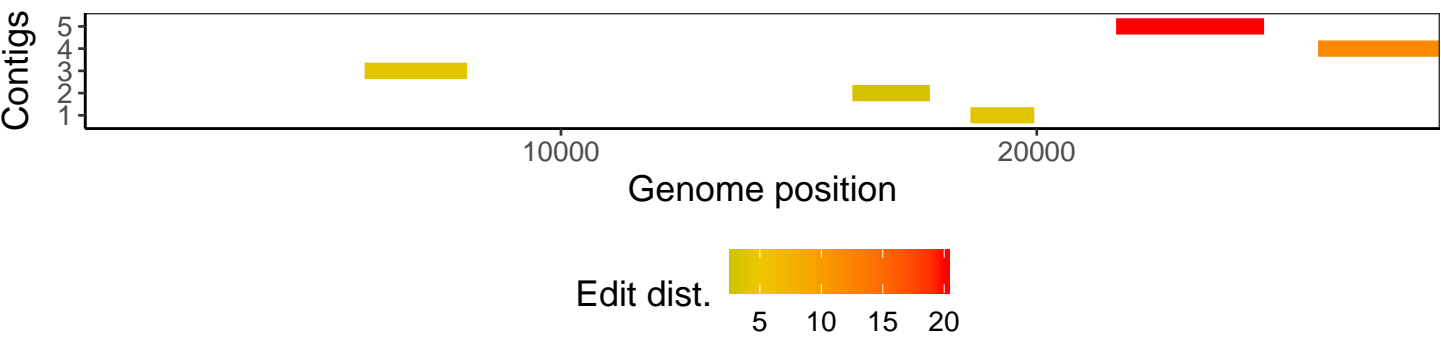
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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