

COVID-19 subject UPHS-0036

2021-04-30

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.

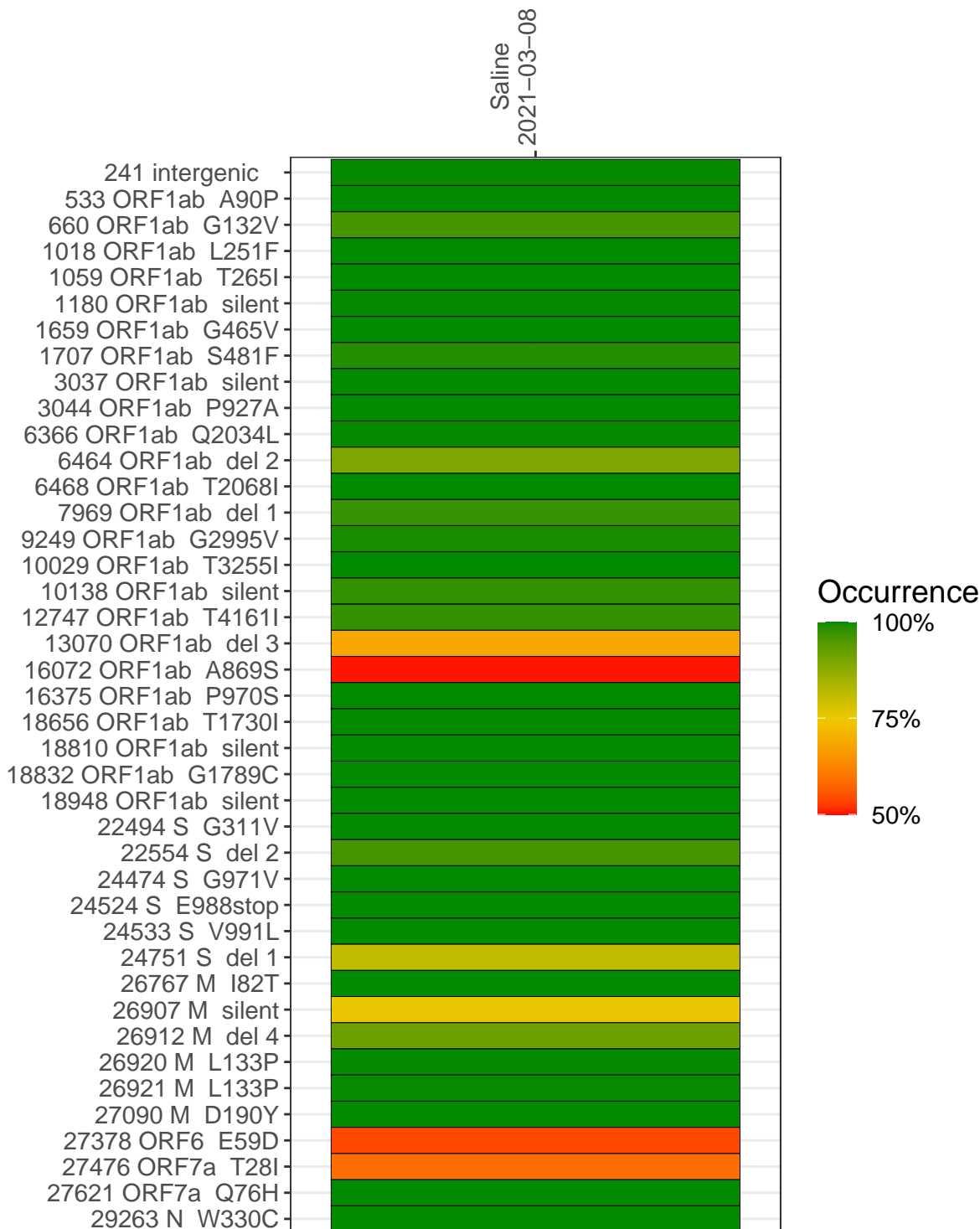
Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software tool (Rambaut et al 2020) for genomes with > 90% sequence coverage.

Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP0968-1	single experiment	NA	Saline	2021-03-08	1.67	NA	38.7%	37.8%

Variants shared across samples

The heat map below shows how variants (reference genome /home/everett/projects/SARS-CoV-2-Philadelphia/Wuhan-Hu-1) 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.



N 2FRFRM M M M N SS ES SS F¹₁F¹₂F¹₃F¹₄F¹₅F¹₆F¹₇F¹₈F¹₉F¹₁₀F¹₁₁F¹₁₂F¹₁₃F¹₁₄F¹₁₅F¹₁₆F¹₁₇F¹₁₈F¹₁₉F¹₂₀F¹₂₁F¹₂₂F¹₂₃F¹₂₄F¹₂₅F¹₂₆F¹₂₇F¹₂₈F¹₂₉F¹₃₀F¹₃₁F¹₃₂F¹₃₃F¹₃₄F¹₃₅F¹₃₆F¹₃₇F¹₃₈F¹₃₉F¹₄₀F¹₄₁F¹₄₂F¹₄₃F¹₄₄F¹₄₅F¹₄₆F¹₄₇F¹₄₈F¹₄₉F¹₅₀F¹₅₁F¹₅₂F¹₅₃F¹₅₄F¹₅₅F¹₅₆F¹₅₇F¹₅₈F¹₅₉F¹₆₀F¹₆₁F¹₆₂F¹₆₃F¹₆₄F¹₆₅F¹₆₆F¹₆₇F¹₆₈F¹₆₉F¹₇₀F¹₇₁F¹₇₂F¹₇₃F¹₇₄F¹₇₅F¹₇₆F¹₇₇F¹₇₈F¹₇₉F¹₈₀F¹₈₁F¹₈₂F¹₈₃F¹₈₄F¹₈₅F¹₈₆F¹₈₇F¹₈₈F¹₈₉F¹₉₀F¹₉₁F¹₉₂F¹₉₃F¹₉₄F¹₉₅F¹₉₆F¹₉₇F¹₉₈F¹₉₉F¹₁₀₀F¹₁₀₁F¹₁₀₂F¹₁₀₃F¹₁₀₄F¹₁₀₅F¹₁₀₆F¹₁₀₇F¹₁₀₈F¹₁₀₉F¹₁₁₀F¹₁₁₁F¹₁₁₂F¹₁₁₃F¹₁₁₄F¹₁₁₅F¹₁₁₆F¹₁₁₇F¹₁₁₈F¹₁₁₉F¹₁₂₀F¹₁₂₁F¹₁₂₂F¹₁₂₃F¹₁₂₄F¹₁₂₅F¹₁₂₆F¹₁₂₇F¹₁₂₈F¹₁₂₉F¹₁₃₀F¹₁₃₁F¹₁₃₂F¹₁₃₃F¹₁₃₄F¹₁₃₅F¹₁₃₆F¹₁₃₇F¹₁₃₈F¹₁₃₉F¹₁₄₀F¹₁₄₁F¹₁₄₂F¹₁₄₃F¹₁₄₄F¹₁₄₅F¹₁₄₆F¹₁₄₇F¹₁₄₈F¹₁₄₉F¹₁₅₀F¹₁₅₁F¹₁₅₂F¹₁₅₃F¹₁₅₄F¹₁₅₅F¹₁₅₆F¹₁₅₇F¹₁₅₈F¹₁₅₉F¹₁₆₀F¹₁₆₁F¹₁₆₂F¹₁₆₃F¹₁₆₄F¹₁₆₅F¹₁₆₆F¹₁₆₇F¹₁₆₈F¹₁₆₉F¹₁₇₀F¹₁₇₁F¹₁₇₂F¹₁₇₃F¹₁₇₄F¹₁₇₅F¹₁₇₆F¹₁₇₇F¹₁₇₈F¹₁₇₉F¹₁₈₀F¹₁₈₁F¹₁₈₂F¹₁₈₃F¹₁₈₄F¹₁₈₅F¹₁₈₆F¹₁₈₇F¹₁₈₈F¹₁₈₉F¹₁₉₀F¹₁₉₁F¹₁₉₂F¹₁₉₃F¹₁₉₄F¹₁₉₅F¹₁₉₆F¹₁₉₇F¹₁₉₈F¹₁₉₉F¹₂₀₀F¹₂₀₁F¹₂₀₂F¹₂₀₃F¹₂₀₄F¹₂₀₅F¹₂₀₆F¹₂₀₇F¹₂₀₈F¹₂₀₉F¹₂₁₀F¹₂₁₁F¹₂₁₂F¹₂₁₃F¹₂₁₄F¹₂₁₅F¹₂₁₆F¹₂₁₇F¹₂₁₈F¹₂₁₉F¹₂₂₀F¹₂₂₁F¹₂₂₂F¹₂₂₃F¹₂₂₄F¹₂₂₅F¹₂₂₆F¹₂₂₇F¹₂₂₈F¹₂₂₉F¹₂₃₀F¹₂₃₁F¹₂₃₂F¹₂₃₃F¹₂₃₄F¹₂₃₅F¹₂₃₆F¹₂₃₇F¹₂₃₈F¹₂₃₉F¹₂₄₀F¹₂₄₁F¹₂₄₂F¹₂₄₃F¹₂₄₄F¹₂₄₅F¹₂₄₆F¹₂₄₇F¹₂₄₈F¹₂₄₉F¹₂₅₀F¹₂₅₁F¹₂₅₂F¹₂₅₃F¹₂₅₄F¹₂₅₅F¹₂₅₆F¹₂₅₇F¹₂₅₈F¹₂₅₉F¹₂₆₀F¹₂₆₁F¹₂₆₂F¹₂₆₃F¹₂₆₄F¹₂₆₅F¹₂₆₆F¹₂₆₇F¹₂₆₈F¹₂₆₉F¹₂₇₀F¹₂₇₁F¹₂₇₂F¹₂₇₃F¹₂₇₄F¹₂₇₅F¹₂₇₆F¹₂₇₇F¹₂₇₈F¹₂₇₉F<

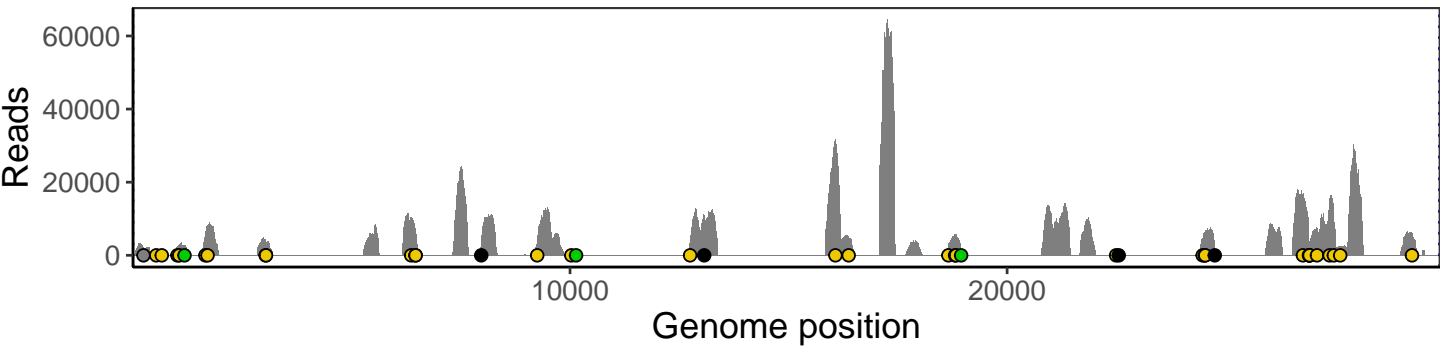


3

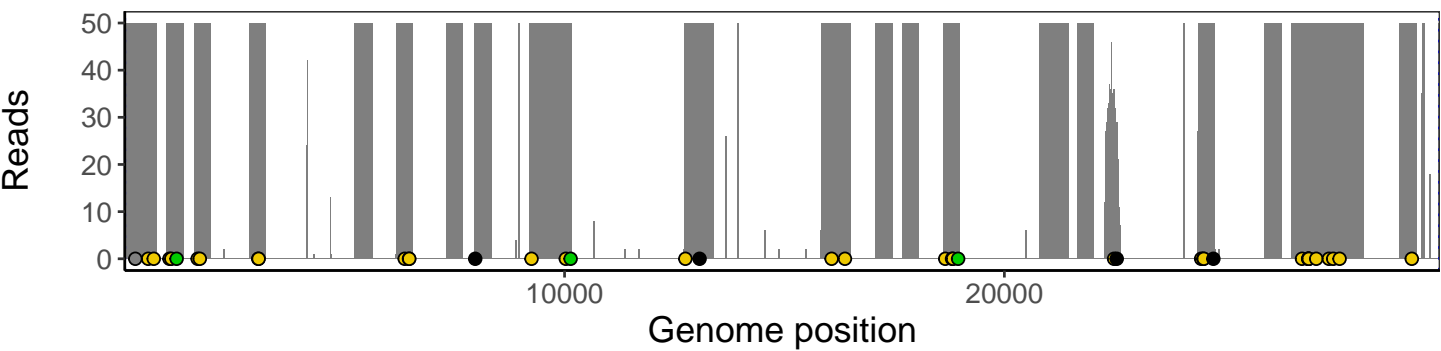
Analyses of individual experiments and composite results

VSP0968-1 | 2021-03-08 | Saline | UPHS-0036 | genomes | single experiment

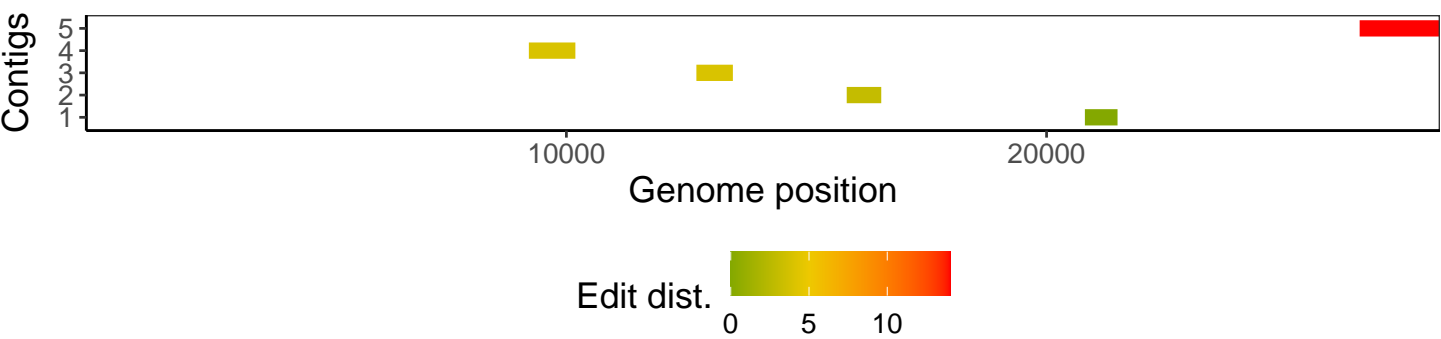
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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	2.3.8
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
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