COVID-19 subject UPHS-0303

2021-04-17

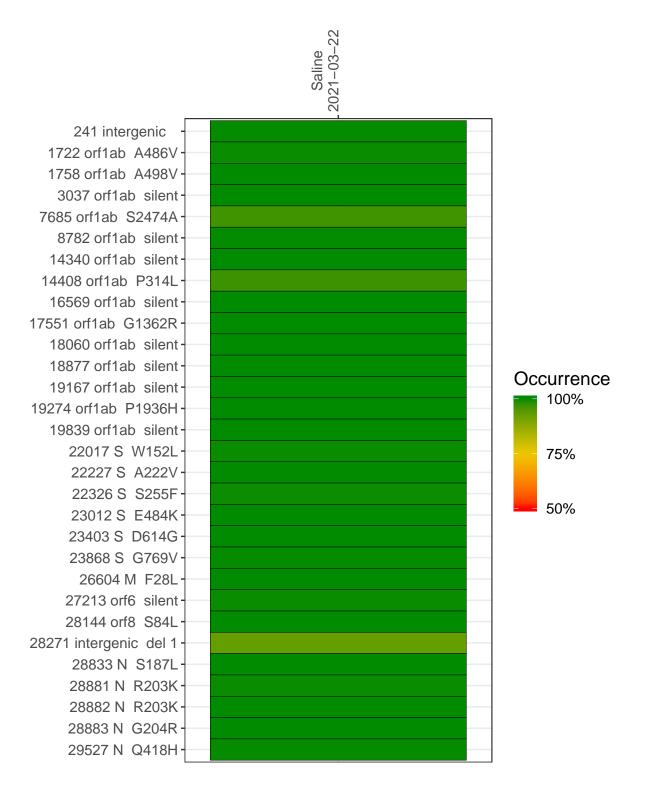
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software tool (Rambaut et al 2020) for genomes with > 90% sequence coverage.

Table 1. Sample summary.

| Experiment | Туре | Genomes | Sample type | Sample date | Largest contig (KD) | Lineage | Reference read coverage | Reference read coverage (>= 5 reads) |
|------------|-------------------|---------|-------------|-------------|------------------------|---------|----------------------------|--------------------------------------|
| VSP1348-1 | single experiment | NA | Saline | 2021-03-22 | 29.65 | R.1 | 99.2% | 99.2% |

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 2021-03-22

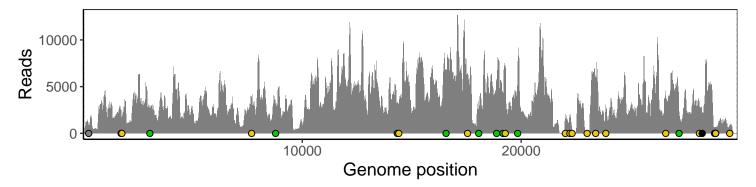
| | 2021-03-22 |
|------------------------|------------|
| 241 intergenic | 881 |
| 1722 orf1ab A486V | 2078 |
| 1758 orf1ab A498V | 1472 |
| 3037 orf1ab silent | 2443 |
| 7685 orf1ab S2474A | 3781 |
| 8782 orf1ab silent | 3099 |
| 14340 orf1ab silent | 3012 |
| 14408 orf1ab P314L | 2870 |
| 16569 orf1ab silent | 6468 |
| 17551 orf1ab G1362R | 5864 |
| 18060 orf1ab silent | 2558 |
| 18877 orf1ab silent | 5494 |
| 19167 orf1ab silent | 4181 |
| 19274 orf1ab P1936H | 6711 |
| 19839 orf1ab silent | 6178 |
| 22017 S W152L | 1427 |
| 22227 S A222V | 1821 |
| 22326 S S255F | 346 |
| 23012 S E484K | 97 |
| 23403 S D614G | 5938 |
| 23868 S G769V | 3049 |
| 26604 M F28L | 4148 |
| 27213 orf6 silent | 2297 |
| 28144 orf8 S84L | 2898 |
| 28271 intergenic del 1 | 2089 |
| 28833 N S187L | 663 |
| 28881 N R203K | 450 |
| 28882 N R203K | 449 |
| 28883 N G204R | 450 |
| 29527 N Q418H | 692 |
| | 348-1 |
| | 348 |



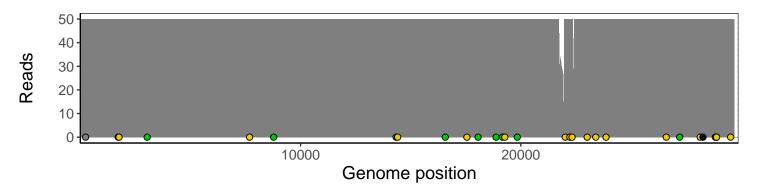
Analyses of individual experiments and composite results

VSP1348-1 | 2021-03-22 | Saline | UPHS-0303 | genomes | single experiment

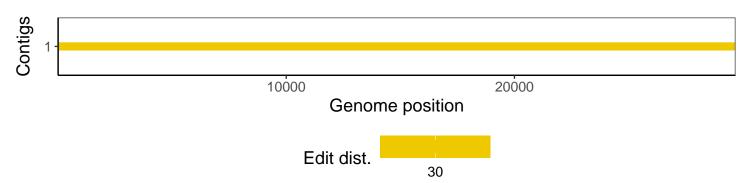
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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 | 2.3.8 |
| 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 |
| ${\it Genomic Alignments}$ | 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 |