

COVID-19 subject 409

2021-01-27

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.

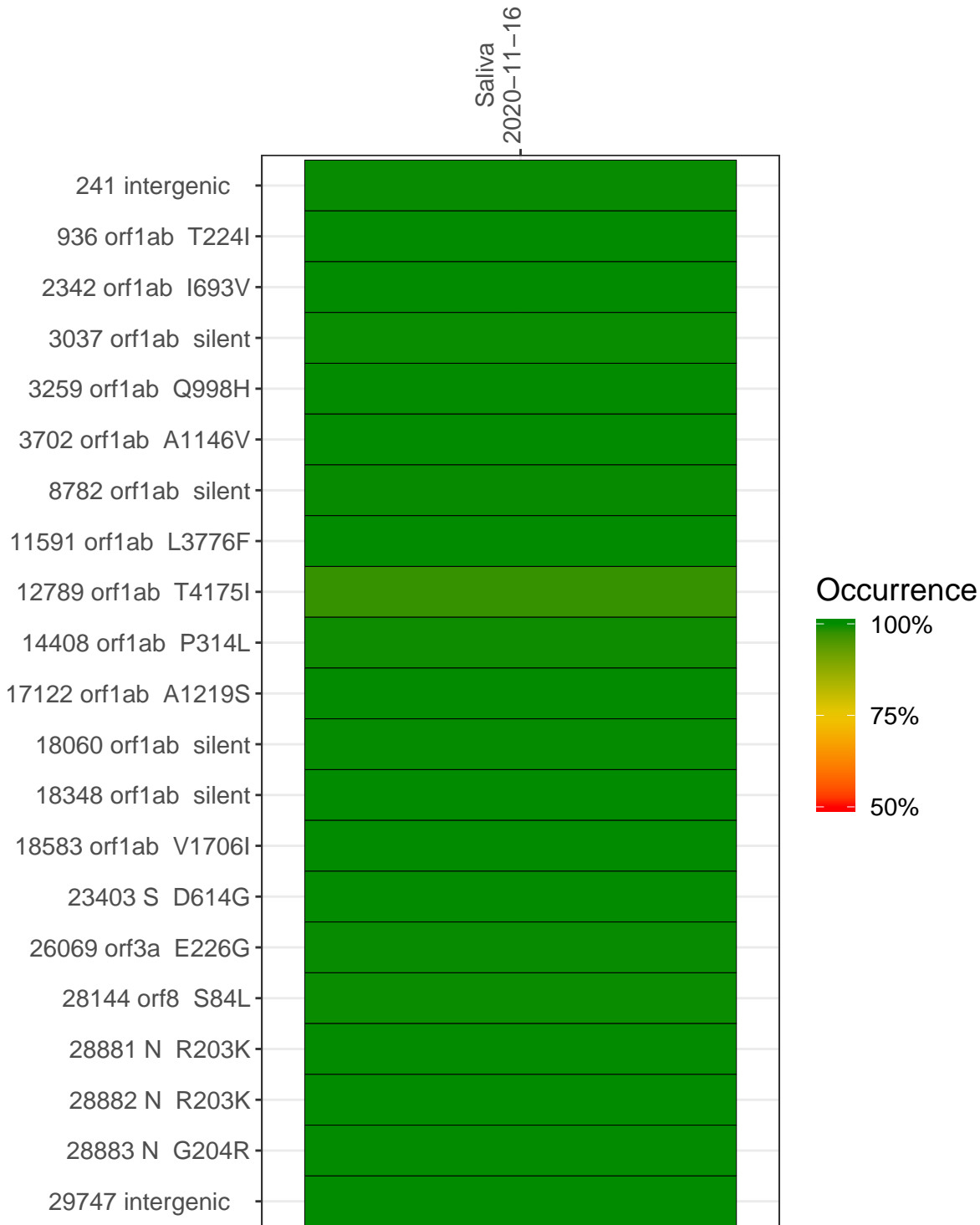
Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software tool (Rambaut et al 2020) for genomes with $> 90\%$ sequence coverage.

Table 1. Sample summary.

| Experiment | Type | Genomes | Sample type | Sample date | Largest contig (KD) | Lineage | Reference read coverage | Reference read coverage (≥ 5 reads) |
|------------|-------------------|---------|-------------|-------------|---------------------|-----------|-------------------------|---|
| VSP0479-2 | single experiment | NA | Saliva | 2020-11-16 | 29.43 | B.1.1.304 | 99.9% | 99.7% |

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.



Saliva
2020-11-16

| | |
|---------------------|-------|
| 241 intergenic | 10340 |
| 936 orf1ab T224I | 14040 |
| 2342 orf1ab I693V | 3805 |
| 3037 orf1ab silent | 5498 |
| 3259 orf1ab Q998H | 10776 |
| 3702 orf1ab A1146V | 4450 |
| 8782 orf1ab silent | 5207 |
| 11591 orf1ab L3776F | 1014 |
| 12789 orf1ab T4175I | 21322 |
| 14408 orf1ab P314L | 8835 |
| 17122 orf1ab A1219S | 24638 |
| 18060 orf1ab silent | 4392 |
| 18348 orf1ab silent | 11220 |
| 18583 orf1ab V1706I | 11224 |
| 23403 S D614G | 19361 |
| 26069 orf3a E226G | 11103 |
| 28144 orf8 S84L | 7058 |
| 28881 N R203K | 2457 |
| 28882 N R203K | 2454 |
| 28883 N G204R | 2464 |
| 29747 intergenic | 136 |

Base change

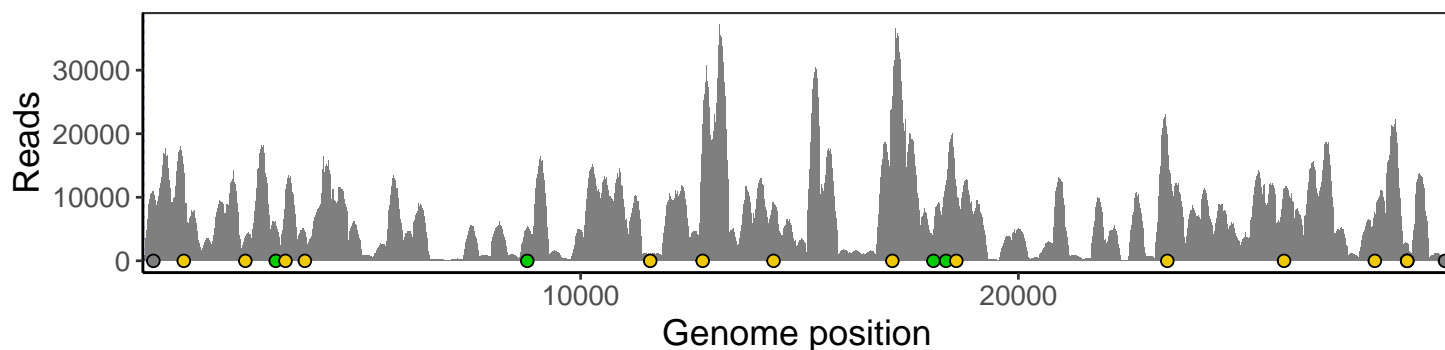


VSP0479-2

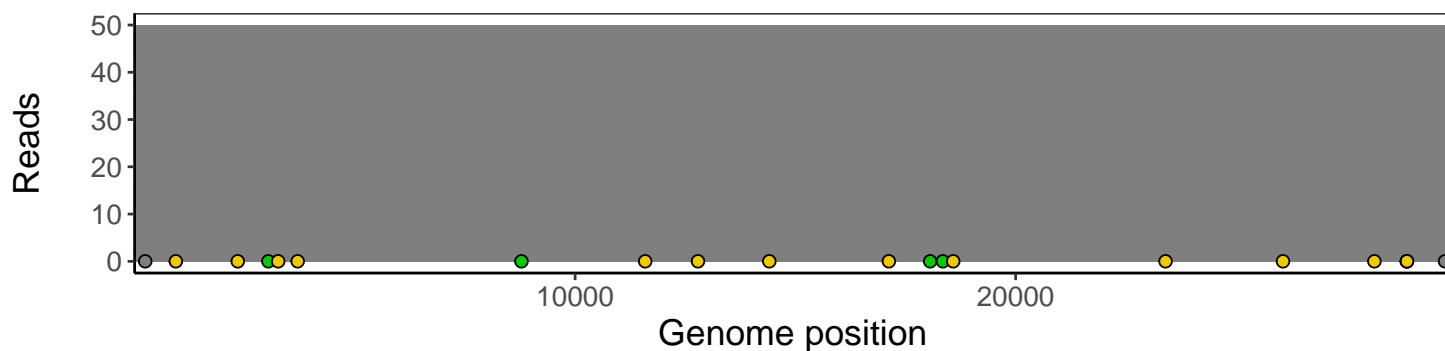
Analyses of individual experiments and composite results

VSP0479-2 | 2020-11-16 | Saliva | 409s-q | genomes | single experiment

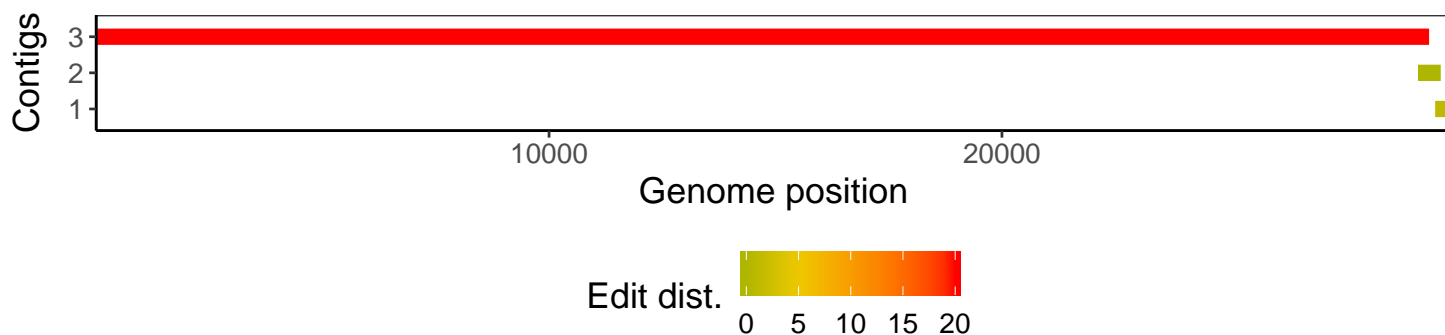
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according to variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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 | 2.1.7 |
| 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 |