# COVID-19 subject HUP Q-0223

*2021-05-21*

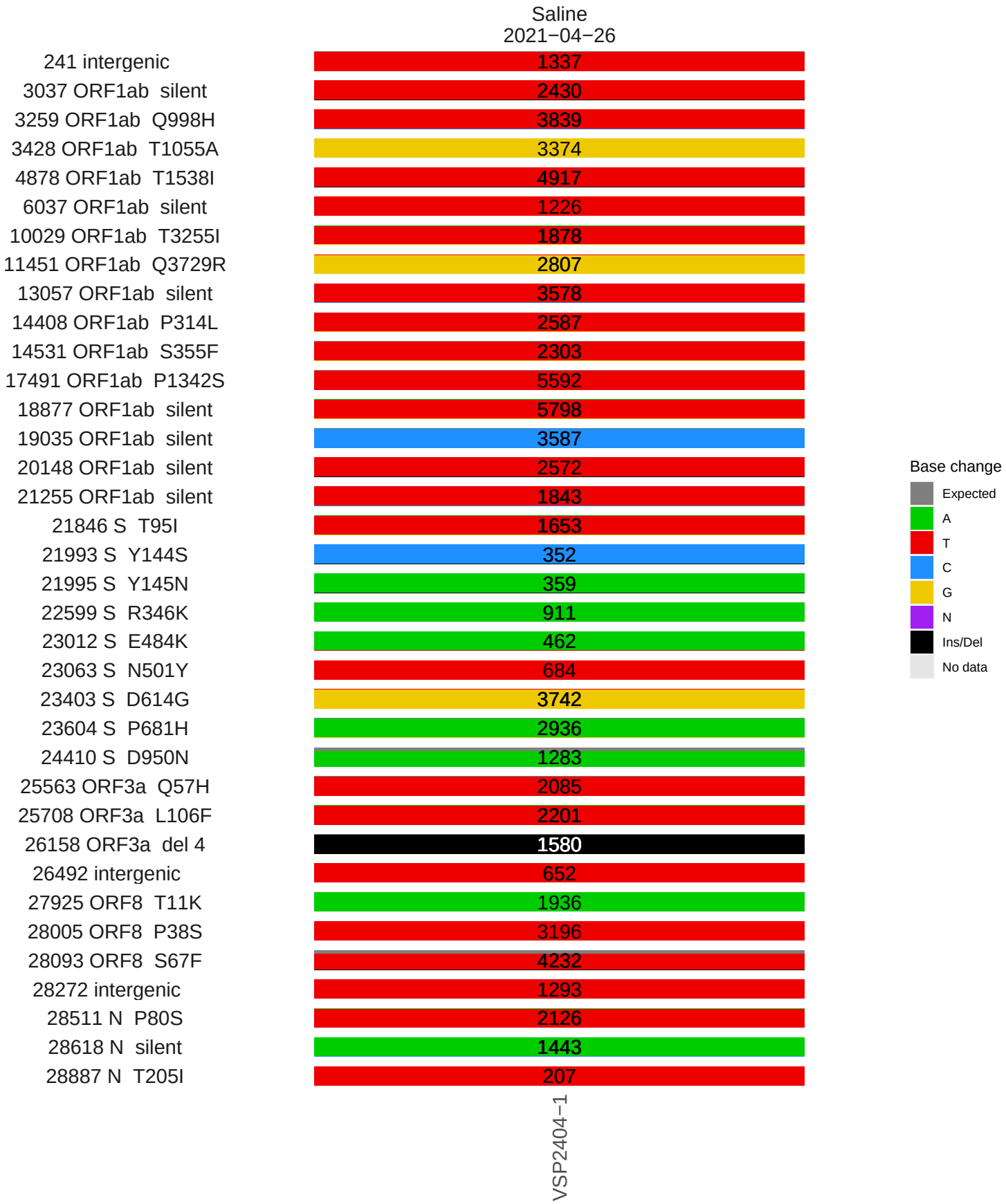
The table below provides a summary of subject samples for which sequencing data is available.
The experiments column shows the number of sequencing experiments performed for each specimen.
Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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) |
|---|---|---|---|---|---|---|---|---|
| VSP2404-1 | single experiment | NA | Saline | 2021-04-26 | 29.83 | B.1 | 99.7% | 99.7% |

**Variants shared across samples**

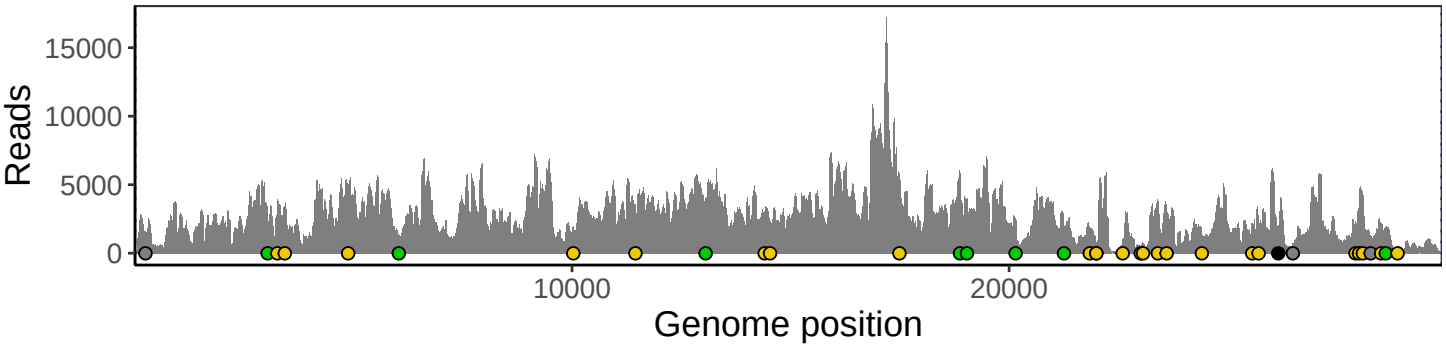
The heat map below shows how variants (reference genome /home/everett/projects/SARS-CoV-2-Philadelphia/Wuhan-Hu-1) are shared across subject samples where the percent variance is colored. Variants are called if a variant position is covered by 5 or more reads, the alternative base is found in > 50% of 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.


Saline
2021−04−26
241 intergenic
3037 ORF1ab silent
3259 ORF1ab Q998H
3428 ORF1ab T1055A
4878 ORF1ab T1538I
6037 ORF1ab silent
10029 ORF1ab T3255I
11451 ORF1ab Q3729R
13057 ORF1ab silent
14408 ORF1ab P314L
14531 ORF1ab S355F
17491 ORF1ab P1342S
18877 ORF1ab silent
19035 ORF1ab silent
20148 ORF1ab silent
21255 ORF1ab silent
21846 S T95I
21993 S Y144S
21995 S Y145N
22599 S R346K
23012 S E484K
23063 S N501Y
23403 S D614G
23604 S P681H
24410 S D950N
25563 ORF3a Q57H
25708 ORF3a L106F
26158 ORF3a del 4
26492 intergenic
27925 ORF8 T11K
28005 ORF8 P38S
28093 ORF8 S67F
28272 intergenic
28511 N P80S
28618 N silent
28887 N T205I
1337
2430
3839
3374
4917
1226
1878
2807
3578
2587
2303
5592
5798
3587
2572
1843
1653
352
359
911
462
684
3742
2936
1283
2085
2201
1580
652
1936
3196
4232
1293
2126
1443
207
VSP2404−1
Base change
Expected
A
T
C
G
N
Ins/Del
No data

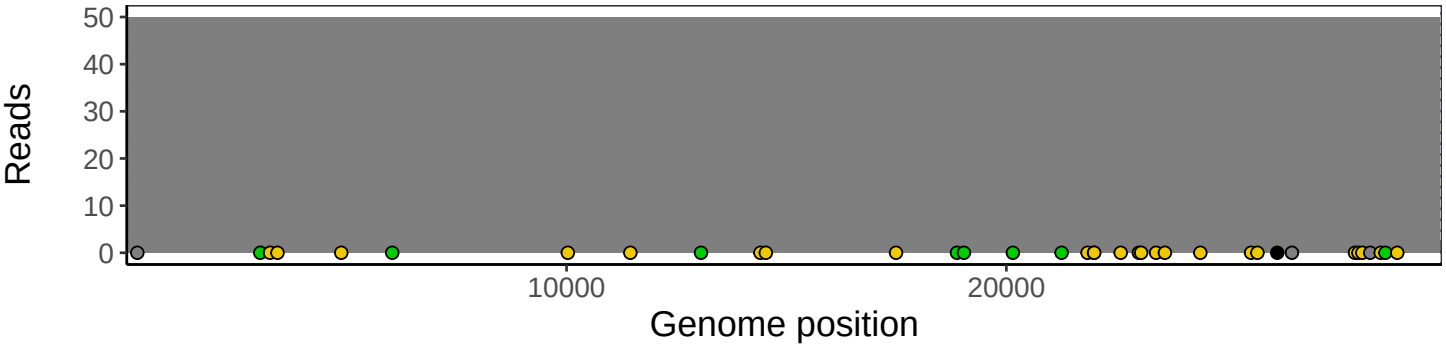
## Analyses of individual experiments and composite results

### VSP2404-1 | 2021-04-26 | Saline | HUP Q-0223 | genomes | single experiment

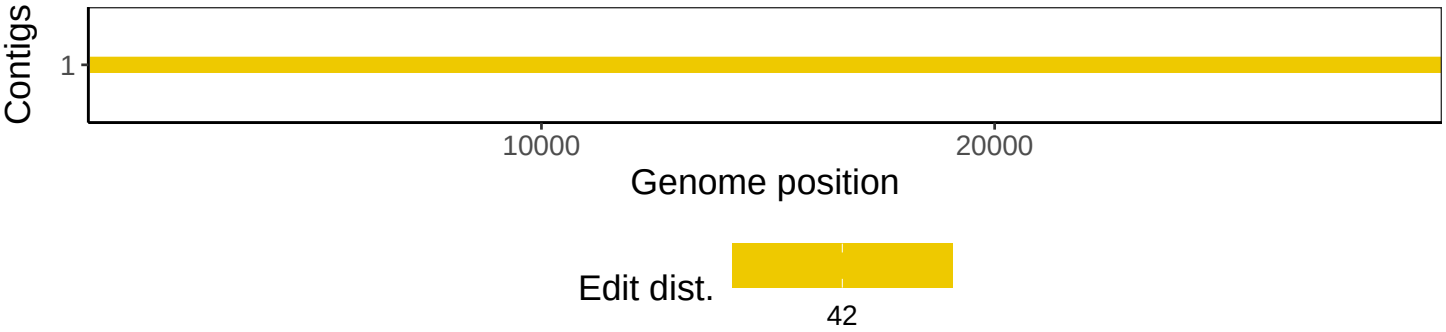
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.




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 | 3.3.3 |
| 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 |