

COVID-19 subject UPHS-0299

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.

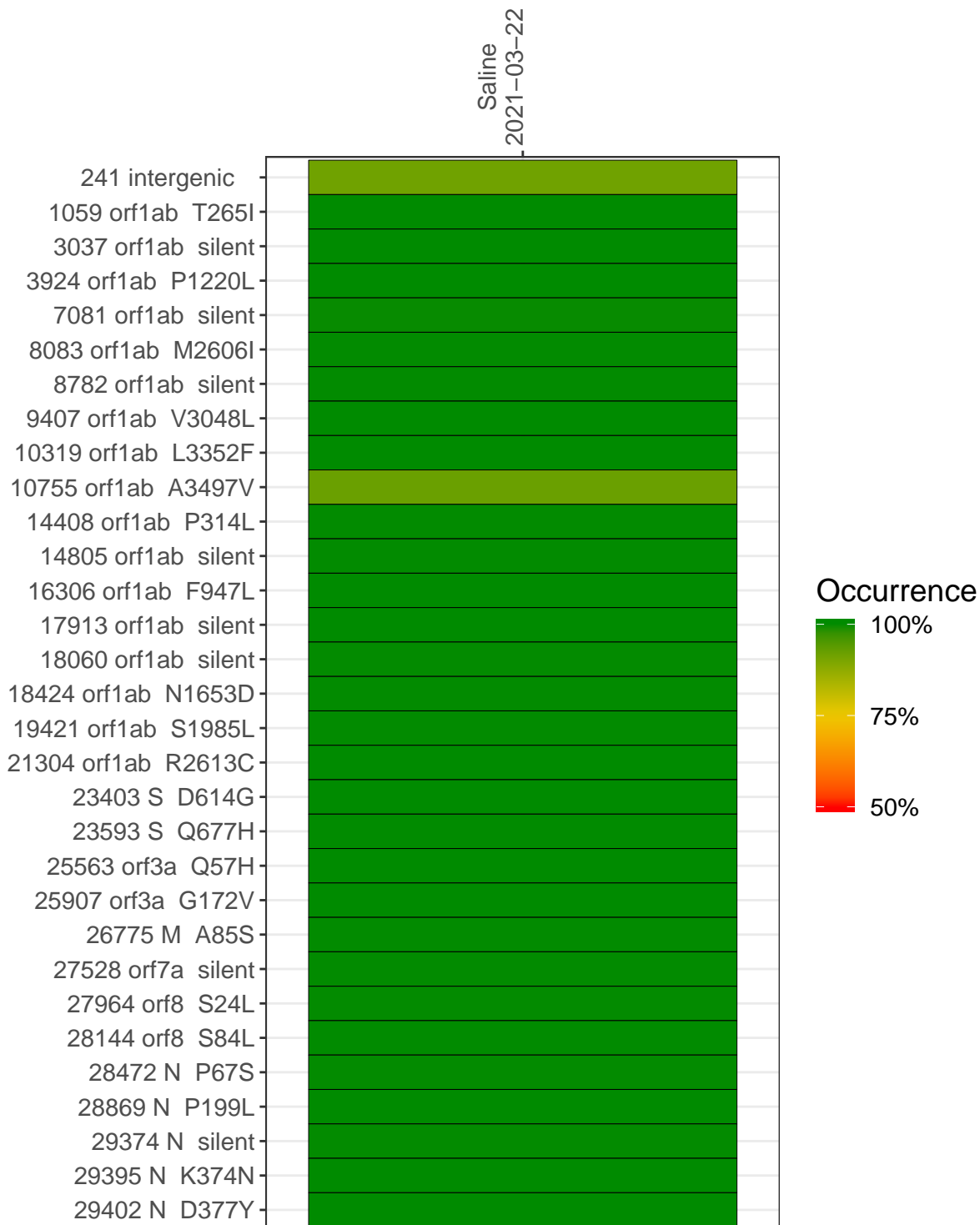
Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software tool (Rambaut et al 2020) for genomes with $> 90\%$ sequence coverage.

Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (≥ 5 reads)
VSP1344-1	single experiment	NA	Saline	2021-03-22	29.65	B.1.2	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	
241 intergenic	1025	
1059 orf1ab T265I	1331	
3037 orf1ab silent	1710	
3924 orf1ab P1220L	1872	
7081 orf1ab silent	1119	
8083 orf1ab M2606I	3558	
8782 orf1ab silent	2674	
9407 orf1ab V3048L	2081	
10319 orf1ab L3352F	3623	
10755 orf1ab A3497V	3138	
14408 orf1ab P314L	4368	
14805 orf1ab silent	5250	
16306 orf1ab F947L	3809	
17913 orf1ab silent	1100	
18060 orf1ab silent	3019	
18424 orf1ab N1653D	7580	
19421 orf1ab S1985L	3320	
21304 orf1ab R2613C	4954	
23403 S D614G	7015	
23593 S Q677H	3094	
25563 orf3a Q57H	6163	
25907 orf3a G172V	1399	
26775 M A85S	5499	
27528 orf7a silent	7006	
27964 orf8 S24L	3902	
28144 orf8 S84L	6809	
28472 N P67S	14840	
28869 N P199L	1102	
29374 N silent	1214	
29395 N K374N	1358	
29402 N D377Y	1454	
	VSP1344-1	

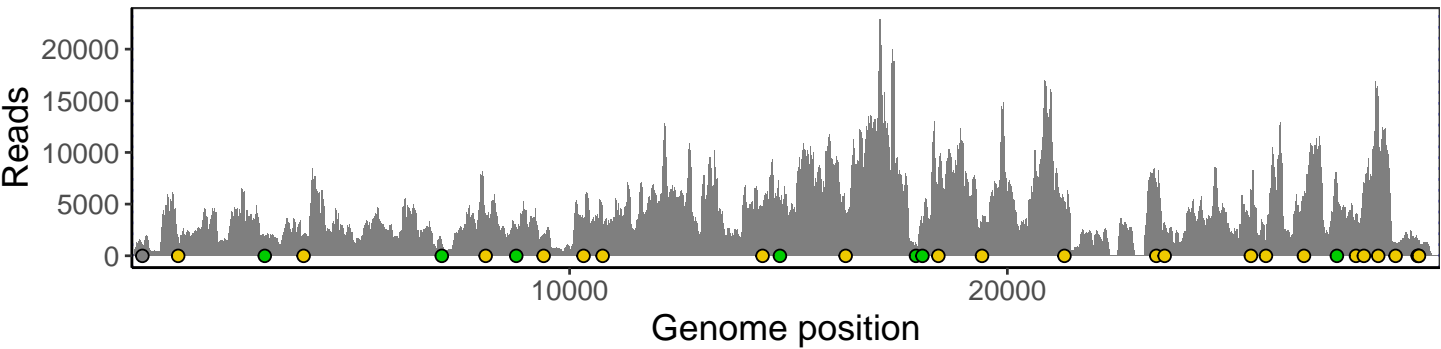
Base change

- Expected
- A
- T
- C
- G
- N
- Ins/Del
- No data

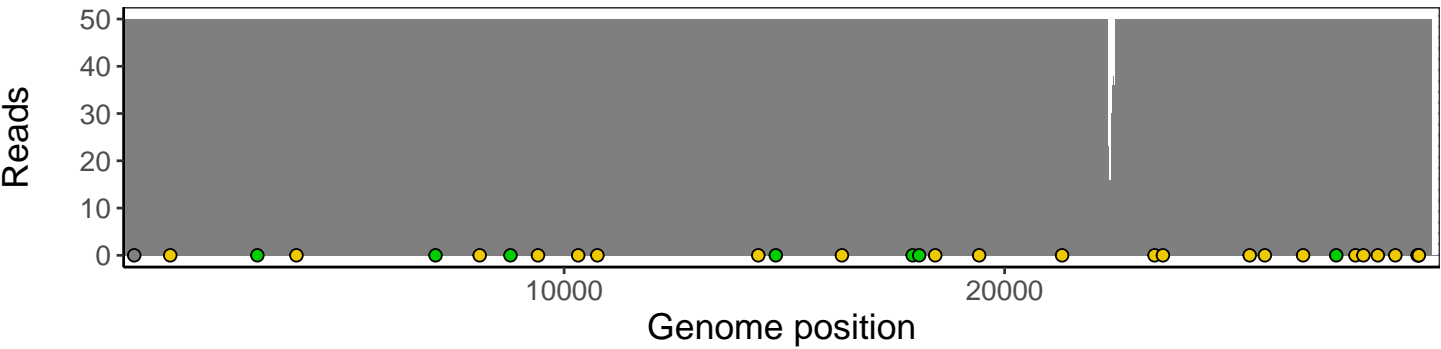
Analyses of individual experiments and composite results

VSP1344-1 | 2021-03-22 | Saline | UPHS-0299 | genomes | single experiment

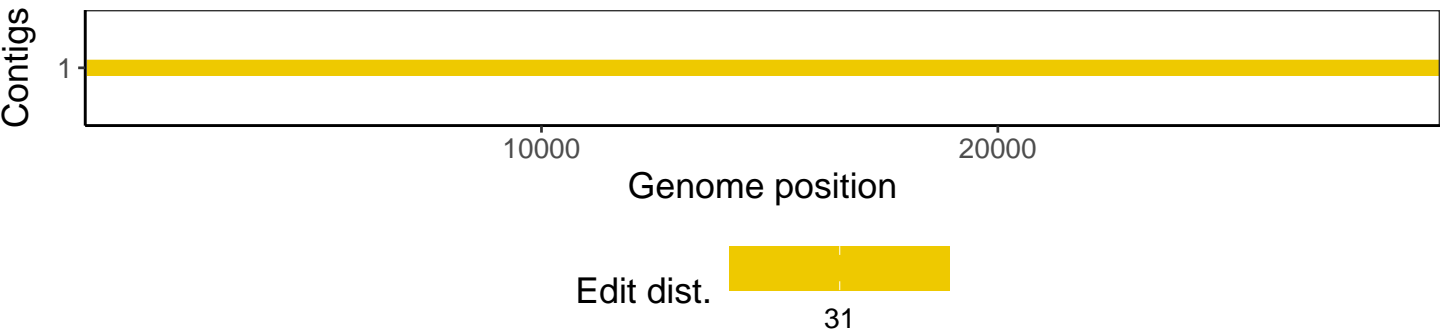
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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