

COVID-19 subject UPHS-0040

2021-04-17

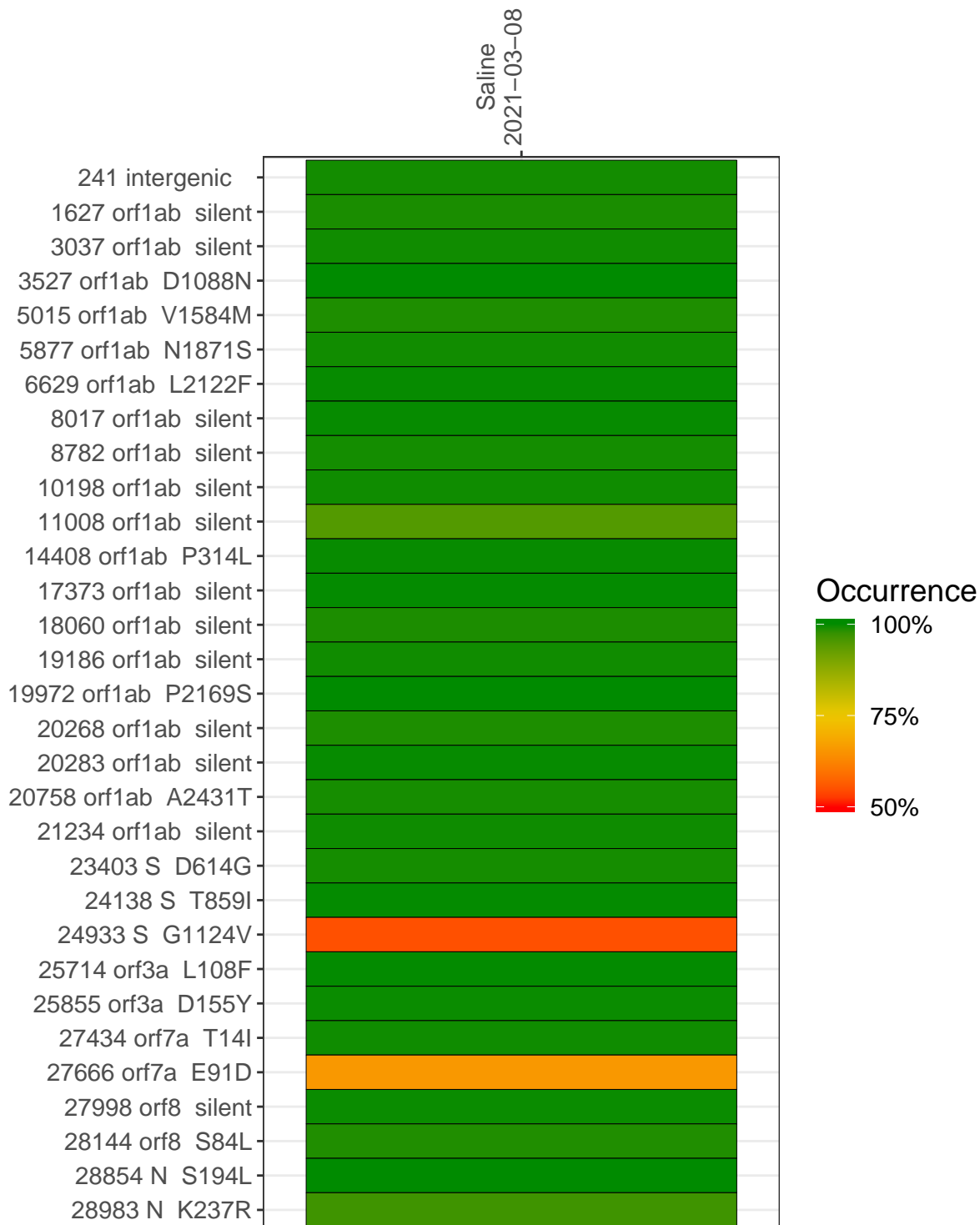
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP0972-1	single experiment	NA	Saline	2021-03-08	29.86	B.1.409	99.9%	99.8%

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 2021-03-08	
241 intergenic	2212	
1627 orf1ab silent	17346	
3037 orf1ab silent	2645	
3527 orf1ab D1088N	324	
5015 orf1ab V1584M	790	
5877 orf1ab N1871S	20310	
6629 orf1ab L2122F	3486	
8017 orf1ab silent	10125	
8782 orf1ab silent	3290	
10198 orf1ab silent	12092	
11008 orf1ab silent	6356	
14408 orf1ab P314L	20726	
17373 orf1ab silent	128525	
18060 orf1ab silent	1658	
19186 orf1ab silent	2753	
19972 orf1ab P2169S	1915	
20268 orf1ab silent	636	
20283 orf1ab silent	709	
20758 orf1ab A2431T	8930	
21234 orf1ab silent	12823	
23403 S D614G	32132	
24138 S T859I	2972	
24933 S G1124V	9838	
25714 orf3a L108F	1641	
25855 orf3a D155Y	2105	
27434 orf7a T14I	6874	
27666 orf7a E91D	4495	
27998 orf8 silent	29355	
28144 orf8 S84L	10627	
28854 N S194L	1105	
28983 N K237R	718	
	VSP0972-1	

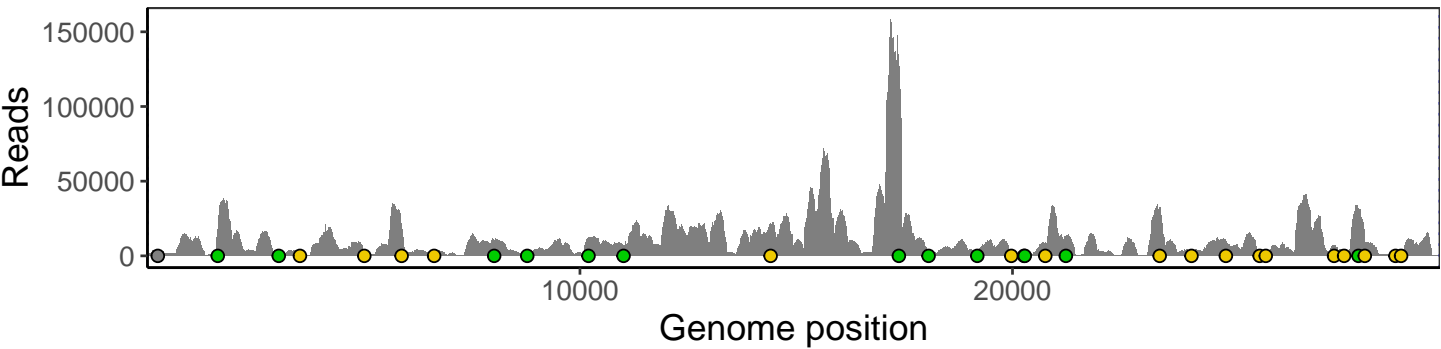
Base change

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

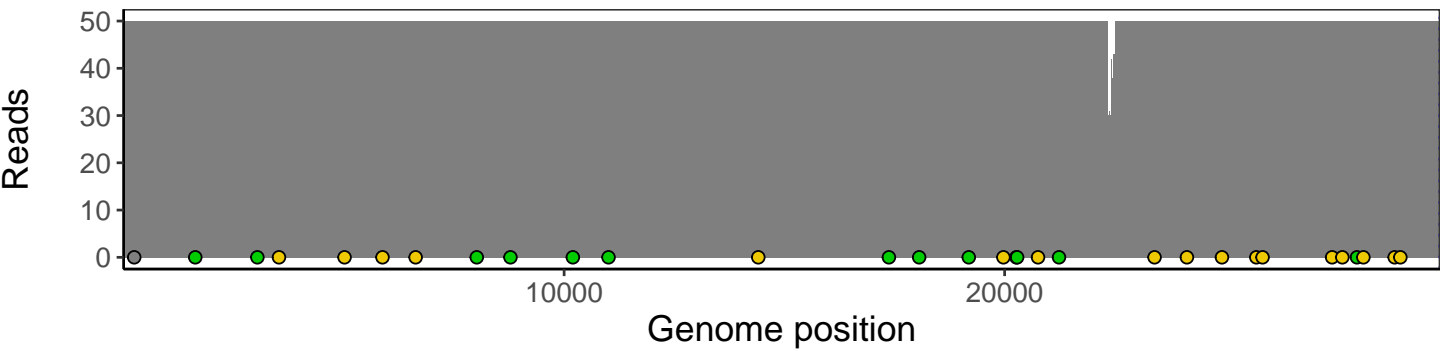
Analyses of individual experiments and composite results

VSP0972-1 | 2021-03-08 | Saline | UPHS-0040 | genomes | single experiment

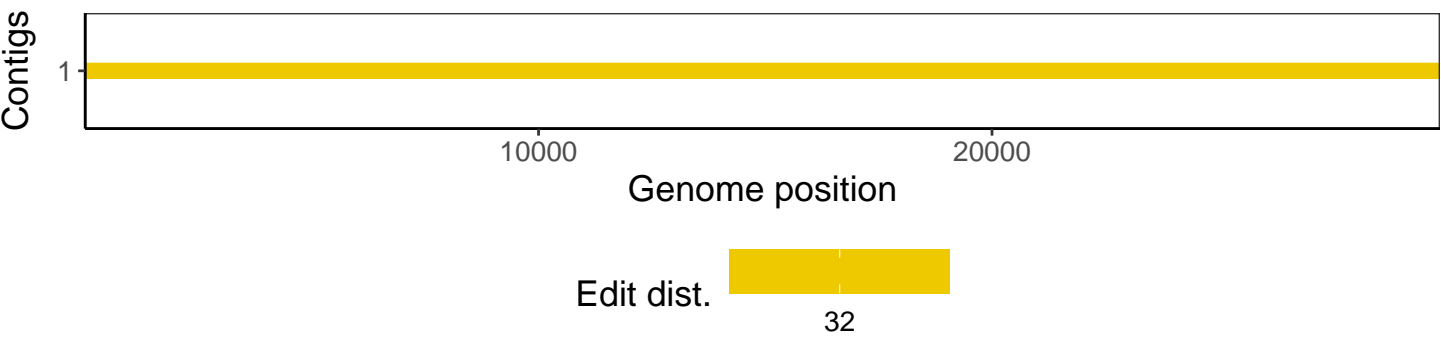
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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