

COVID-19 subject HUP-PH-0016

2021-03-05

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.

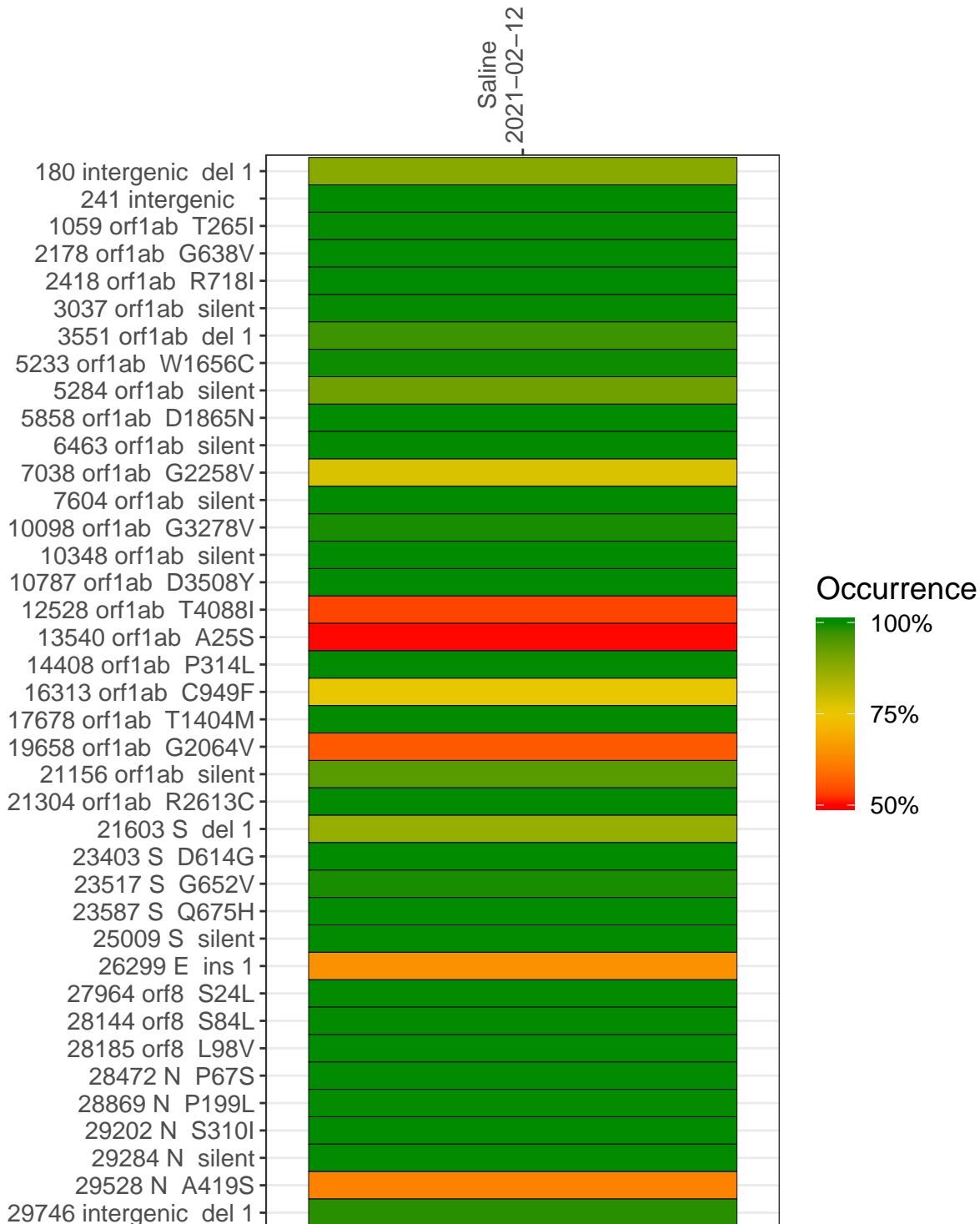
Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software tool (Rambaut et al 2020) for genomes with > 90% sequence coverage.

Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP0829-1	single experiment	NA	Saline	2021-02-12	5.01	NA	84.4%	84.1%

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.



	Saline 2021-02-12	
180 intergenic del 1	2404	
241 intergenic	1168	
1059 orf1ab T265I	4007	
2178 orf1ab G638V	1642	
2418 orf1ab R718I	3219	
3037 orf1ab silent	729	
3551 orf1ab del 1	3325	
5233 orf1ab W1656C	4281	
5284 orf1ab silent	4398	
5858 orf1ab D1865N	33617	
6463 orf1ab silent	17361	
7038 orf1ab G2258V	10375	
7604 orf1ab silent	6674	
10098 orf1ab G3278V	828	
10348 orf1ab silent	13	
10787 orf1ab D3508Y	3792	
12528 orf1ab T4088I	15222	
13540 orf1ab A25S	6236	
14408 orf1ab P314L	10036	
16313 orf1ab C949F	2460	
17678 orf1ab T1404M	13688	
19658 orf1ab G2064V	9681	
21156 orf1ab silent	19772	
21304 orf1ab R2613C	15371	
21603 S del 1	1837	
23403 S D614G	17359	
23517 S G652V	5314	
23587 S Q675H	8780	
25009 S silent	9399	
26299 E ins 1	3997	
27964 orf8 S24L	39260	
28144 orf8 S84L	10515	
28185 orf8 L98V	3541	
28472 N P67S	8645	
28869 N P199L	1972	
29202 N S310I	2846	
29284 N silent	1822	
29528 N A419S	5102	
29746 intergenic del 1	3450	

Base change

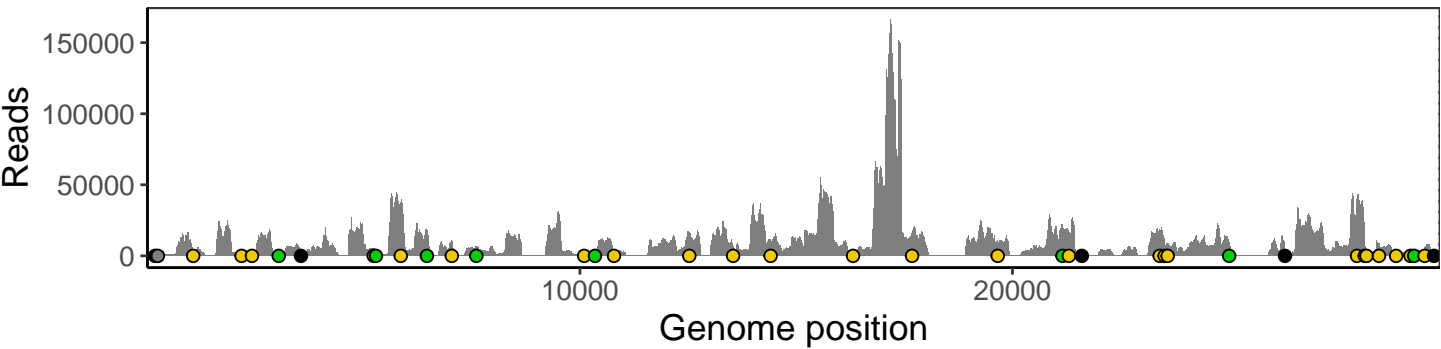
- Expected
- A
- T
- C
- G
- N
- Ins/Del
- No data

VSP0829-1

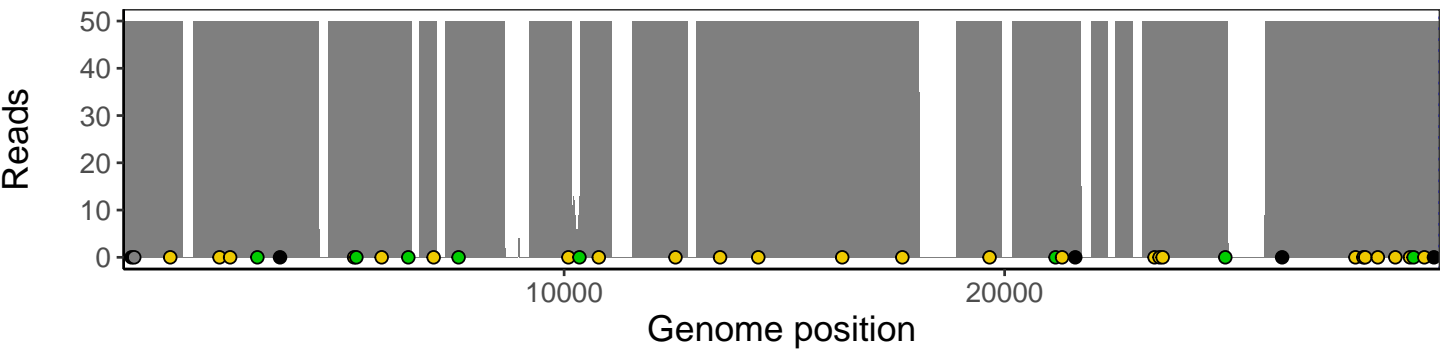
Analyses of individual experiments and composite results

VSP0829-1 | 2021-02-12 | Saline | HUP-PH-0016 | genomes | single experiment

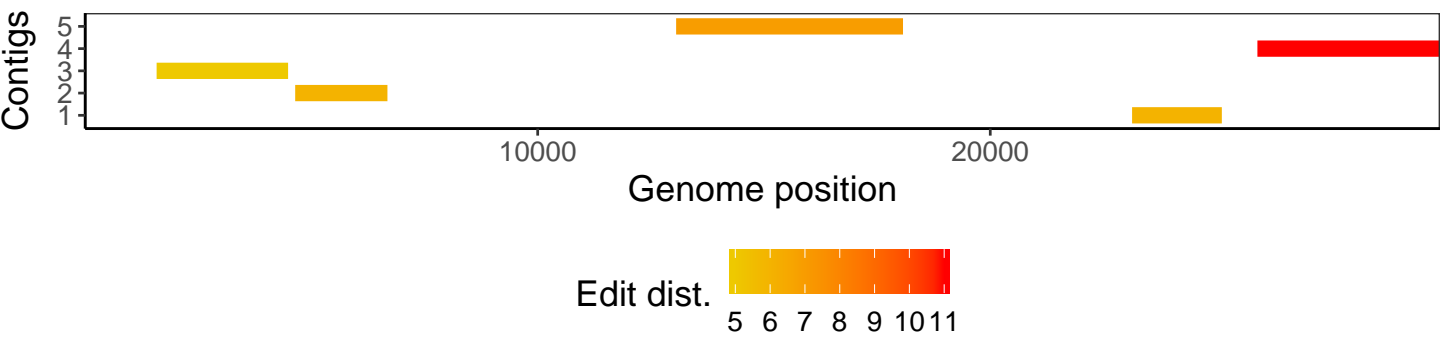
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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	2.3.3
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