COVID-19 subject HUP-PH-0016

2021-05-05

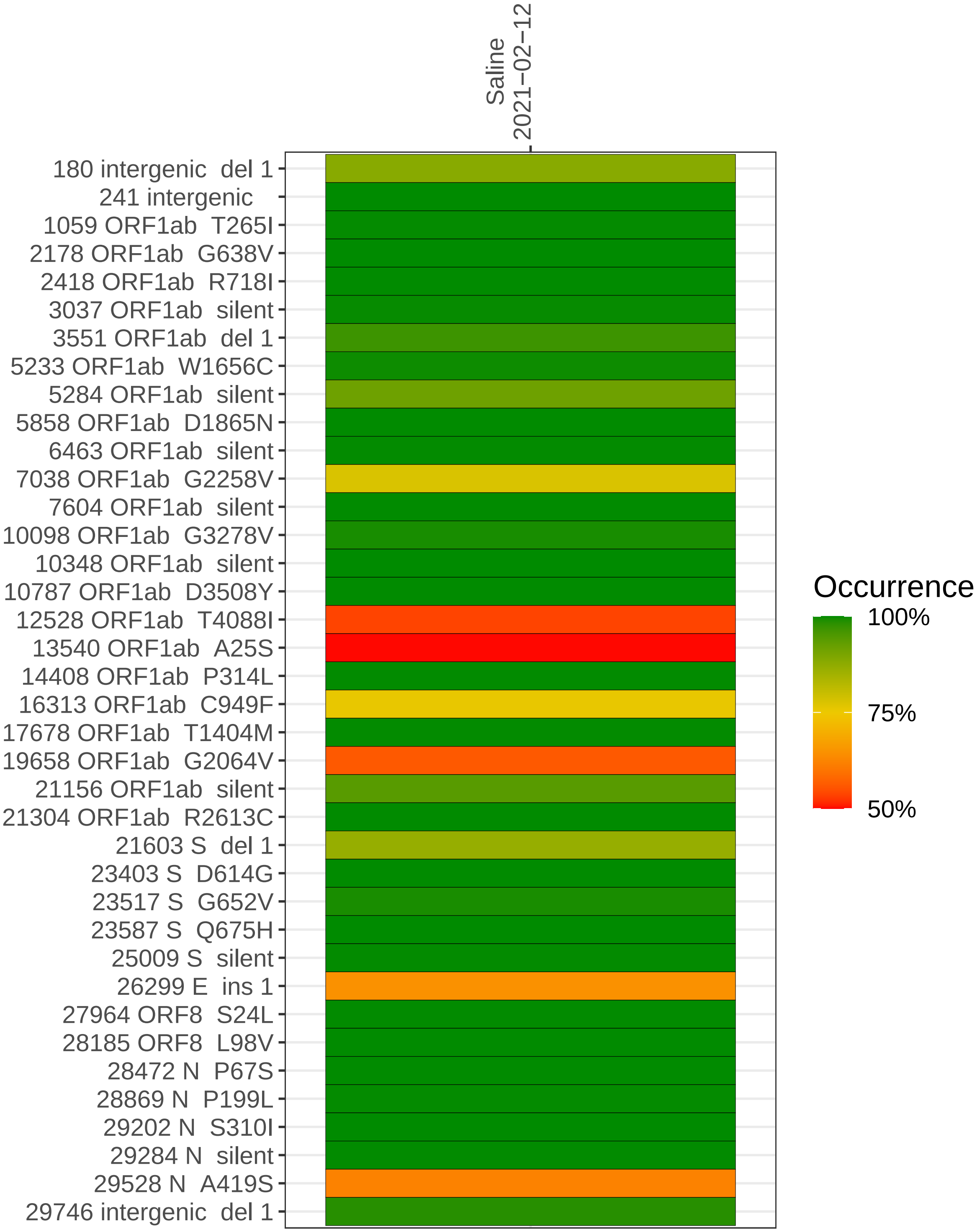
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software tool (Rambaut et al 2020) for genomes with > 90% sequence coverage.

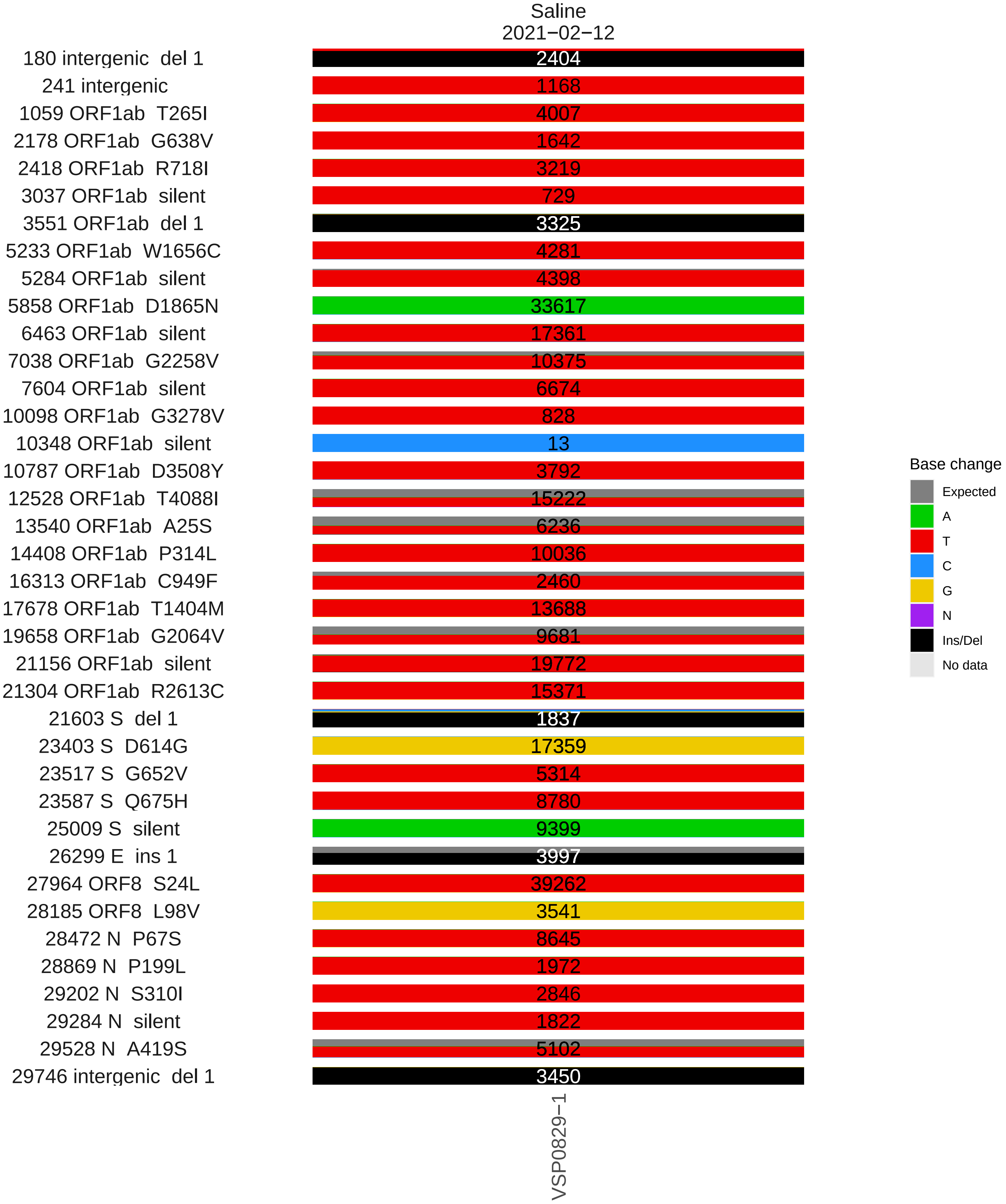
Table 1. Sample summary.

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	84.5%	84.2%

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.

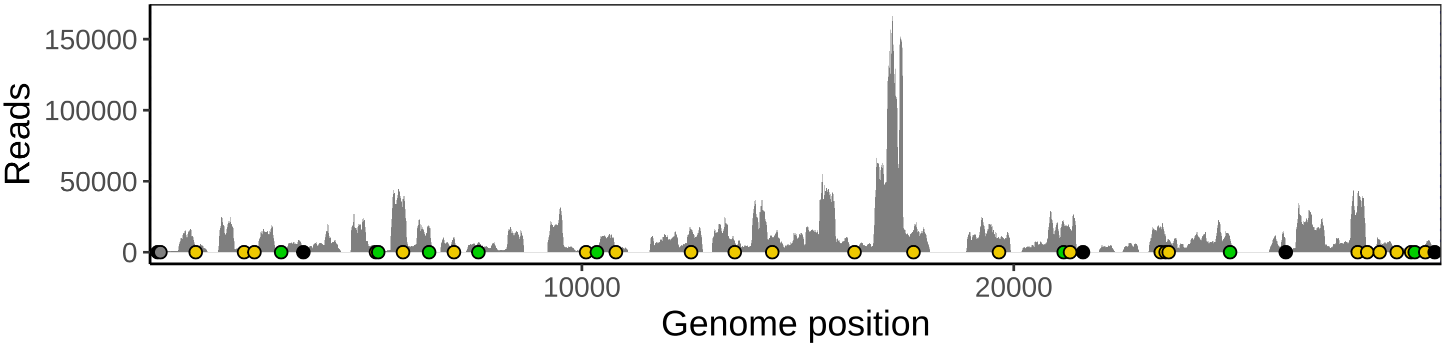



Saline
2021−02−12
180 intergenic del 1
241 intergenic
1059 ORF1ab T265I
2178 ORF1ab G638V
2418 ORF1ab R718I
3037 ORF1ab silent
3551 ORF1ab del 1
5233 ORF1ab W1656C
5284 ORF1ab silent
5858 ORF1ab D1865N
6463 ORF1ab silent
7038 ORF1ab G2258V
7604 ORF1ab silent
10098 ORF1ab G3278V
10348 ORF1ab silent
10787 ORF1ab D3508Y
12528 ORF1ab T4088I
13540 ORF1ab A25S
14408 ORF1ab P314L
16313 ORF1ab C949F
17678 ORF1ab T1404M
19658 ORF1ab G2064V
21156 ORF1ab silent
21304 ORF1ab R2613C
21603 S del 1
23403 S D614G
23517 S G652V
23587 S Q675H
25009 S silent
26299 E ins 1
27964 ORF8 S24L
28185 ORF8 L98V
28472 N P67S
28869 N P199L
29202 N S310I
29284 N silent
29528 N A419S
29746 intergenic del 1
2404
1168
4007
1642
3219
729
3325
4281
4398
33617
17361
10375
6674
828
13
3792
15222
6236
10036
2460
13688
9681
19772
15371
1837
17359
5314
8780
9399
3997
39262
3541
8645
1972
2846
1822
5102
3450
VSP0829−1
Base change
Expected
A
T
C
G
N
Ins/Del
No data

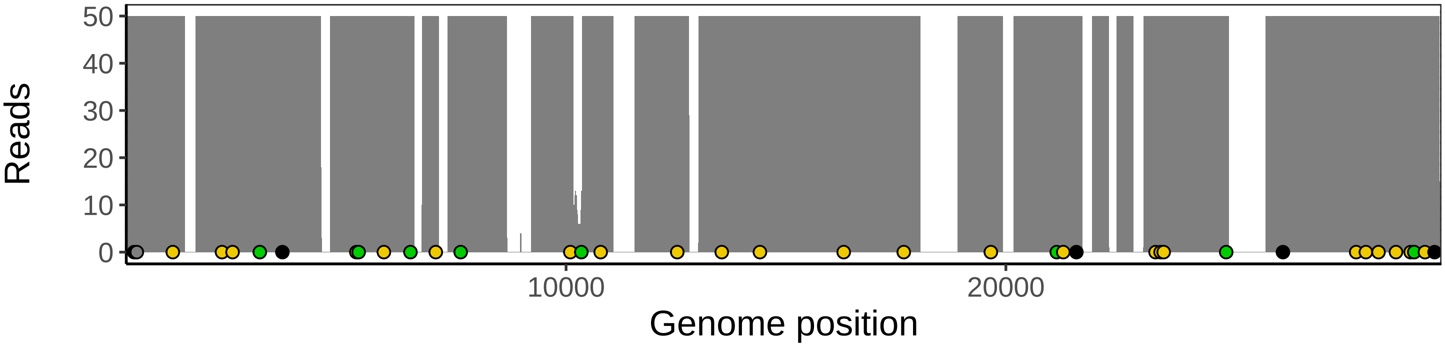
Analyses of individual experiments and composite results

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

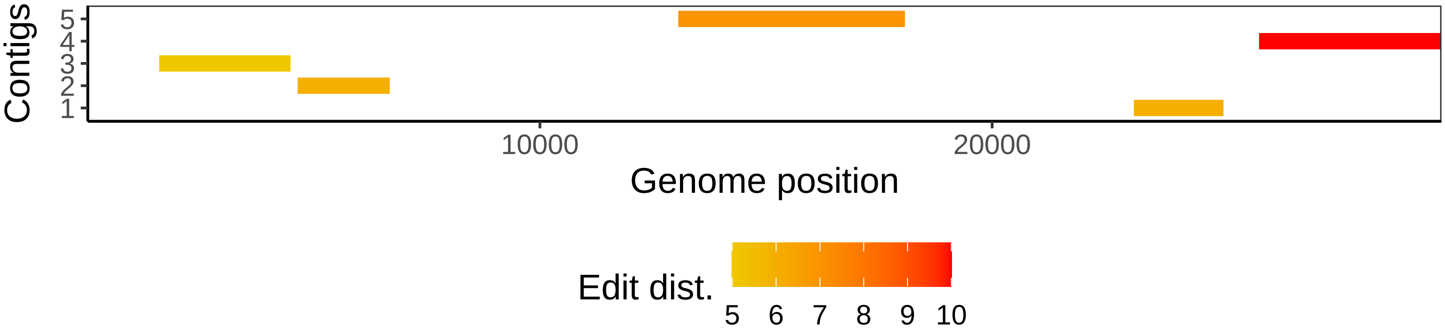
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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 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	3.0.0
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