COVID-19 subject HUP PH-0026

2021-03-29

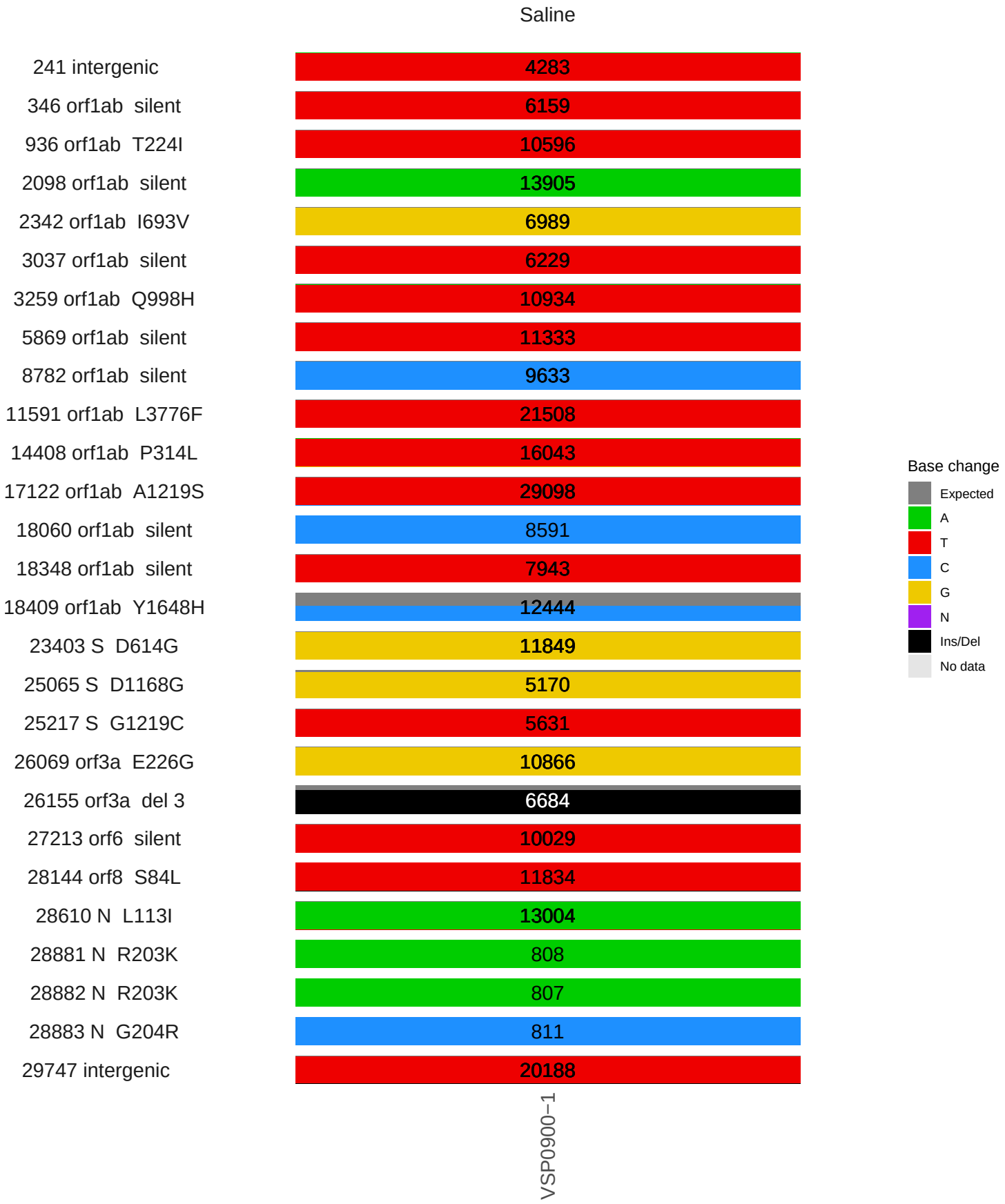
The table below provides a summary of subject samples for which sequencing data is available.
The experiments column shows the number of sequencing experiments performed for each specimen.
Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP0900-1	single experiment	NA	Saline	2021-03-01	29.82	B.1.1.304	99.8%	99.8%

Variants shared across samples

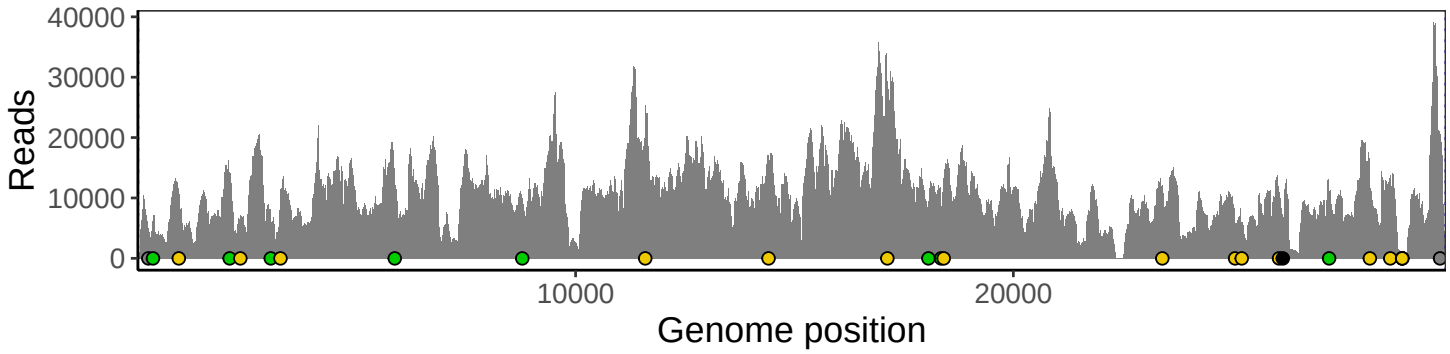
The heat map below shows how variants (reference genome USA-WA1-2020) are shared across subject samples where the percent variance is colored. Variants are called if a variant position is covered by 5 or more reads, the alternative base is found in $> 50\%$ of read pairs and the variant yields a PHRED 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
241 intergenic
346 orf1ab silent
936 orf1ab T224I
2098 orf1ab silent
2342 orf1ab I693V
3037 orf1ab silent
3259 orf1ab Q998H
5869 orf1ab silent
8782 orf1ab silent
11591 orf1ab L3776F
14408 orf1ab P314L
17122 orf1ab A1219S
18060 orf1ab silent
18348 orf1ab silent
18409 orf1ab Y1648H
23403 S D614G
25065 S D1168G
25217 S G1219C
26069 orf3a E226G
26155 orf3a del 3
27213 orf6 silent
28144 orf8 S84L
28610 N L113I
28881 N R203K
28882 N R203K
28883 N G204R
29747 intergenic
4283
6159
10596
13905
6989
6229
10934
11333
9633
21508
16043
29098
8591
7943
12444
11849
5170
5631
10866
6684
10029
11834
13004
808
807
811
20188
VSP0900−1
Base change
Expected
A
T
C
G
N
Ins/Del
No data

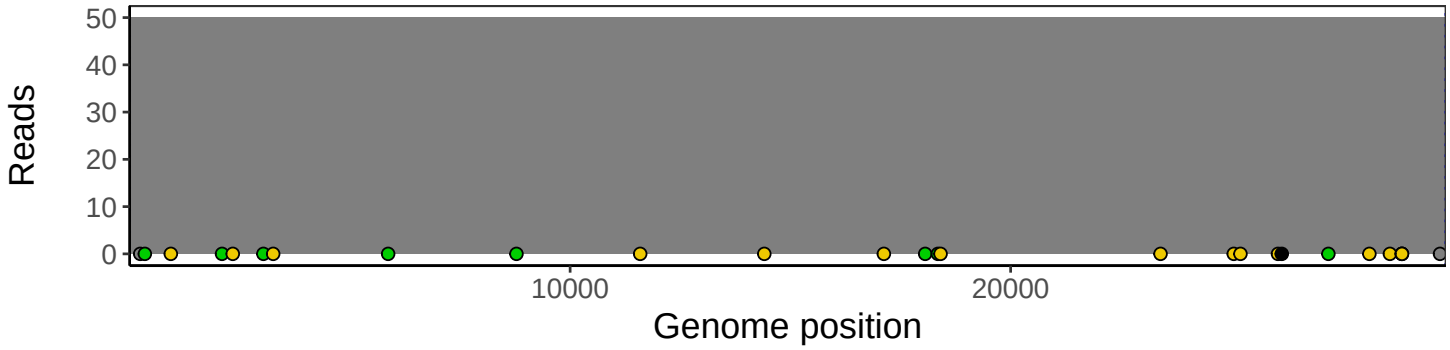
Analyses of individual experiments and composite results

VSP0900-1 | 2021-03-01 | Saline | HUP PH-0026 | genomes | single experiment

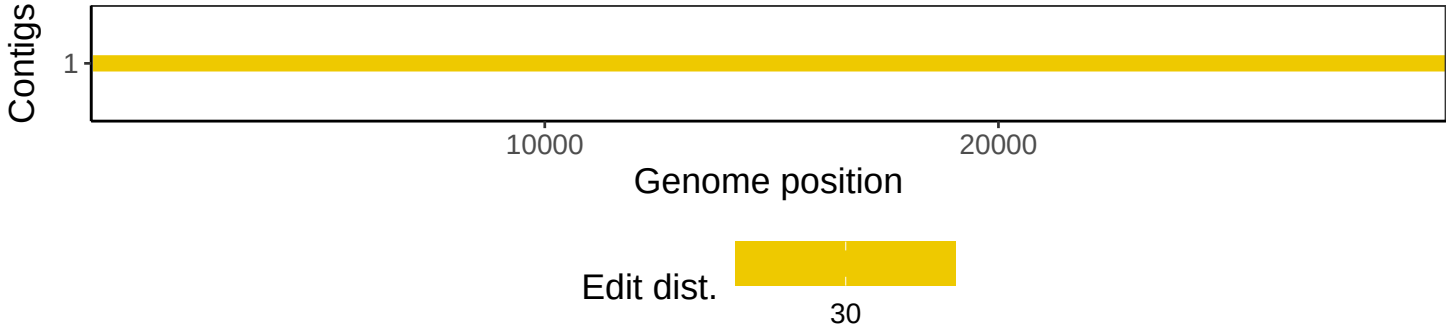
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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	2.3.3
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