COVID-19 subject UPHS-0737

2021-04-20

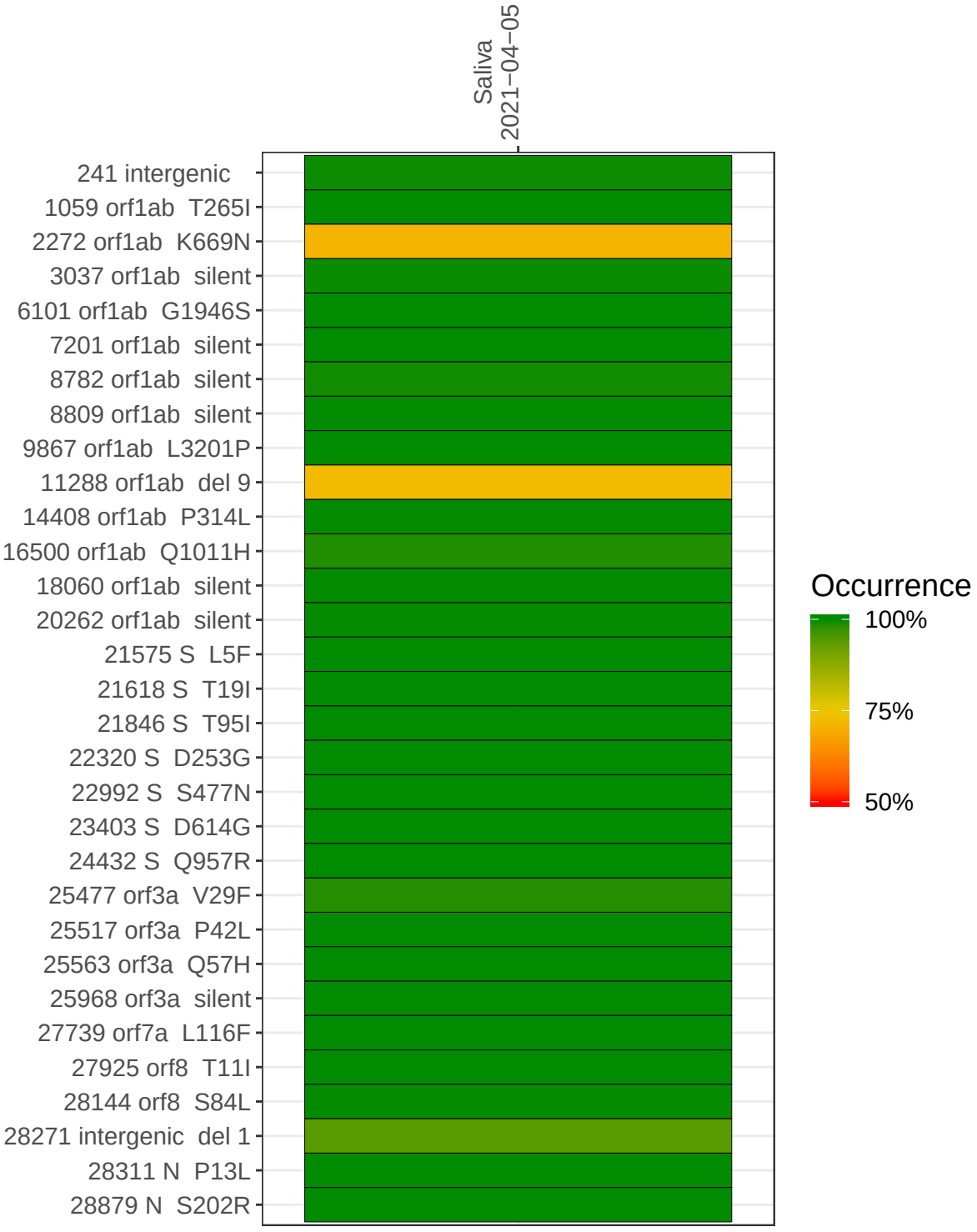
The table below provides a summary of subject samples for which sequencing data is available.
The experiments column shows the number of sequencing experiments performed for each specimen.
Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software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

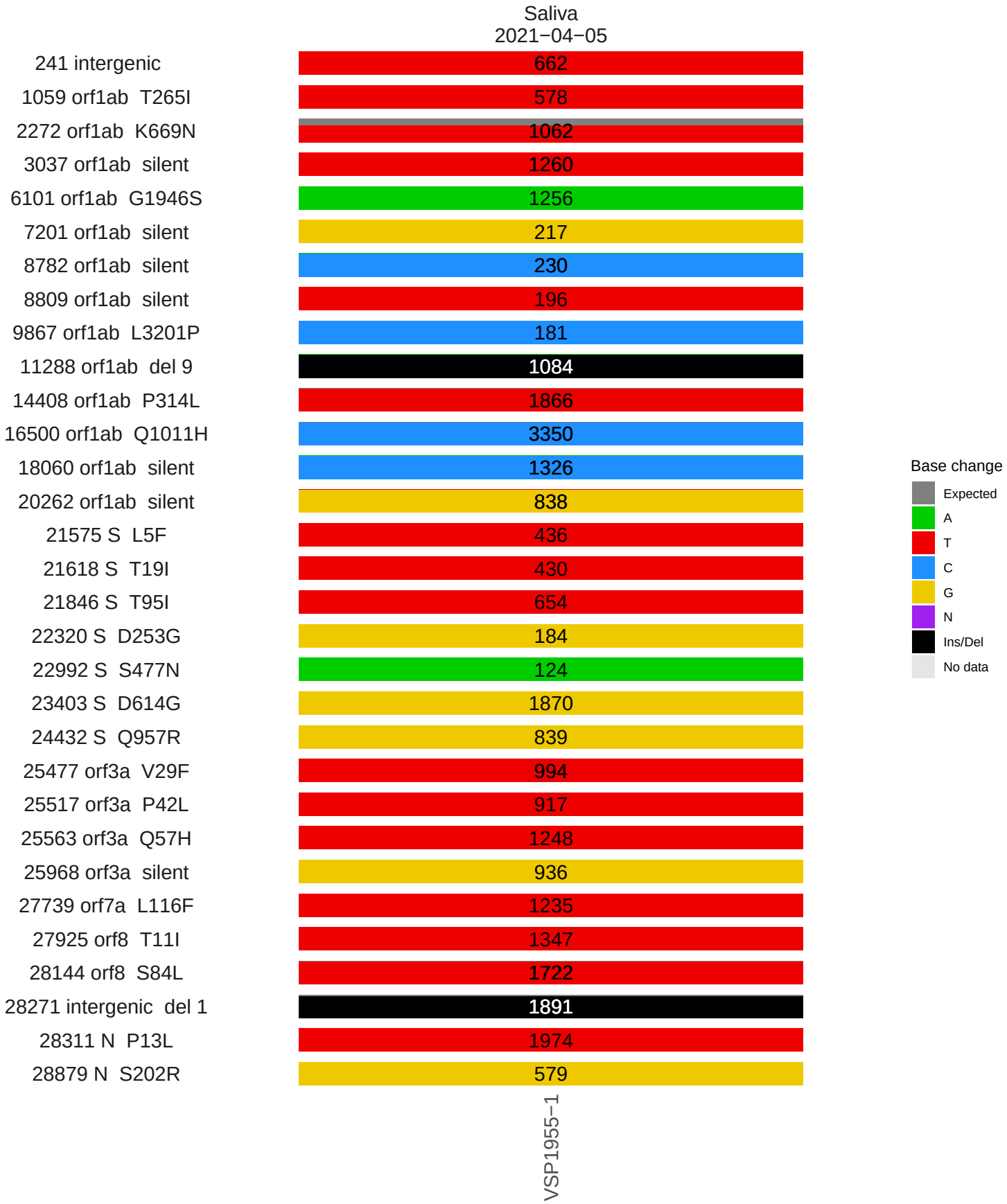
Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP1955-1	single experiment	NA	Saliva	2021-04-05	29.69	B.1.526.2	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.

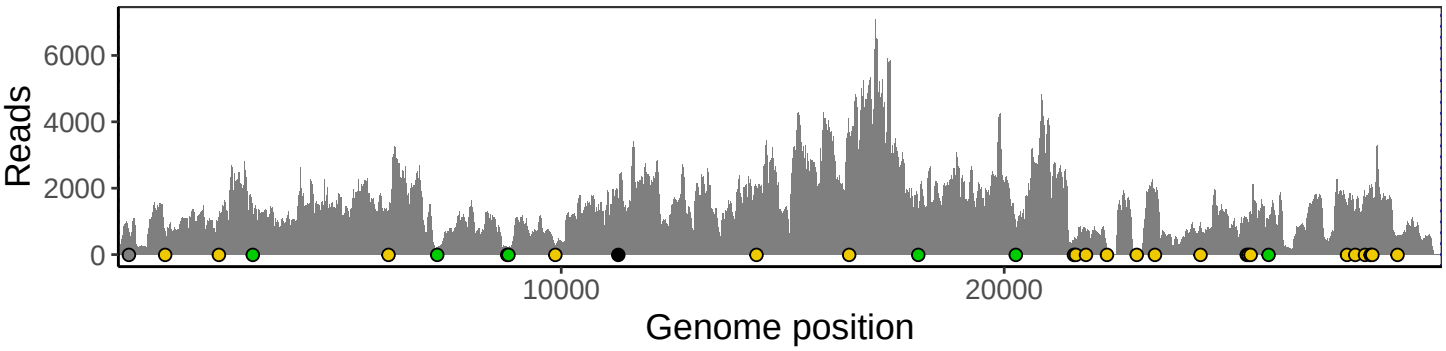



Saliva
2021−04−05
241 intergenic
662
1059 orf1ab T265I
578
2272 orf1ab K669N
1062
3037 orf1ab silent
1260
6101 orf1ab G1946S
1256
7201 orf1ab silent
217
8782 orf1ab silent
230
8809 orf1ab silent
196
9867 orf1ab L3201P
181
11288 orf1ab del 9
1084
14408 orf1ab P314L
1866
16500 orf1ab Q1011H
3350
18060 orf1ab silent
1326
20262 orf1ab silent
838
21575 S L5F
436
21618 S T19I
430
21846 S T95I
654
22320 S D253G
184
22992 S S477N
124
23403 S D614G
1870
24432 S Q957R
839
25477 orf3a V29F
994
25517 orf3a P42L
917
25563 orf3a Q57H
1248
25968 orf3a silent
936
27739 orf7a L116F
1235
27925 orf8 T11I
1347
28144 orf8 S84L
1722
28271 intergenic del 1
1891
28311 N P13L
1974
28879 N S202R
579
VSP1955−1
Base change
Expected
A
T
C
G
N
Ins/Del
No data

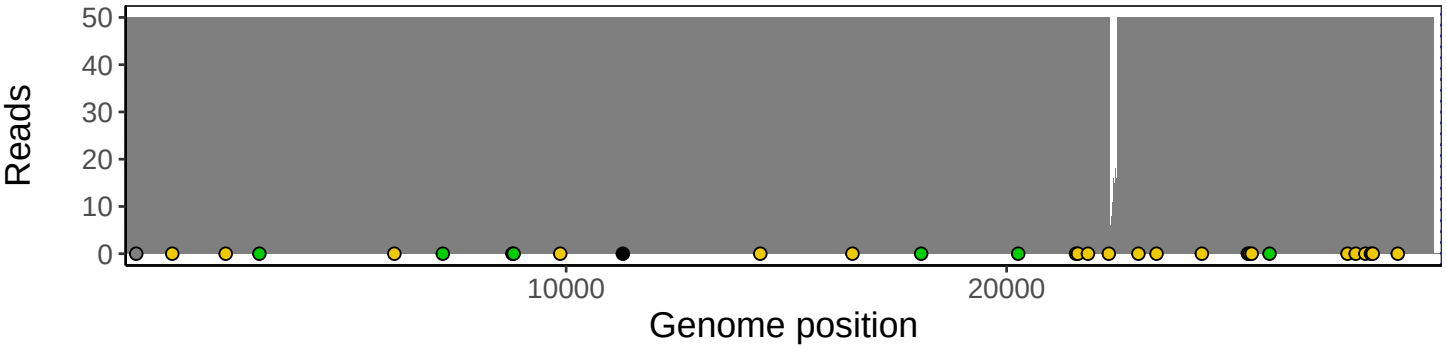
Analyses of individual experiments and composite results

VSP1955-1 | 2021-04-05 | Saliva | UPHS-0737 | genomes | single experiment

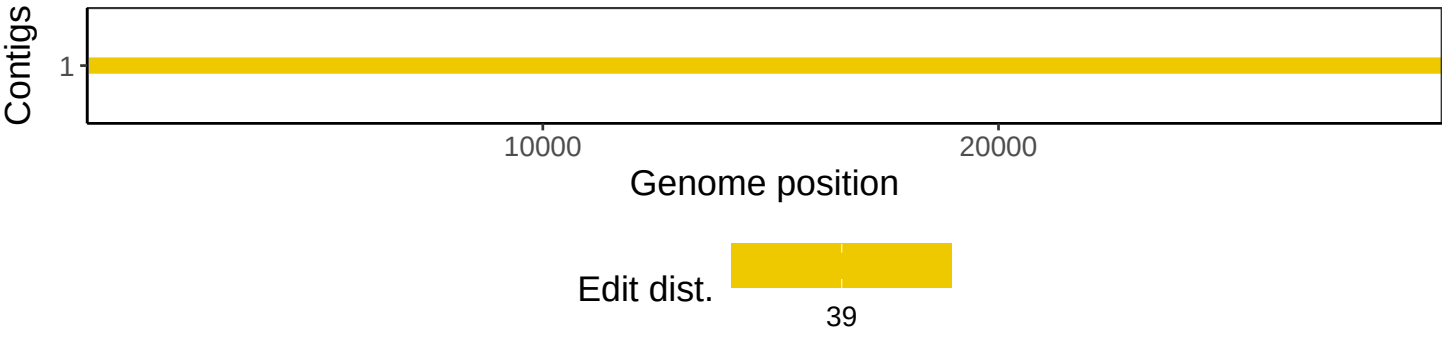
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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