COVID-19 subject UPHS-0704

2021-04-20

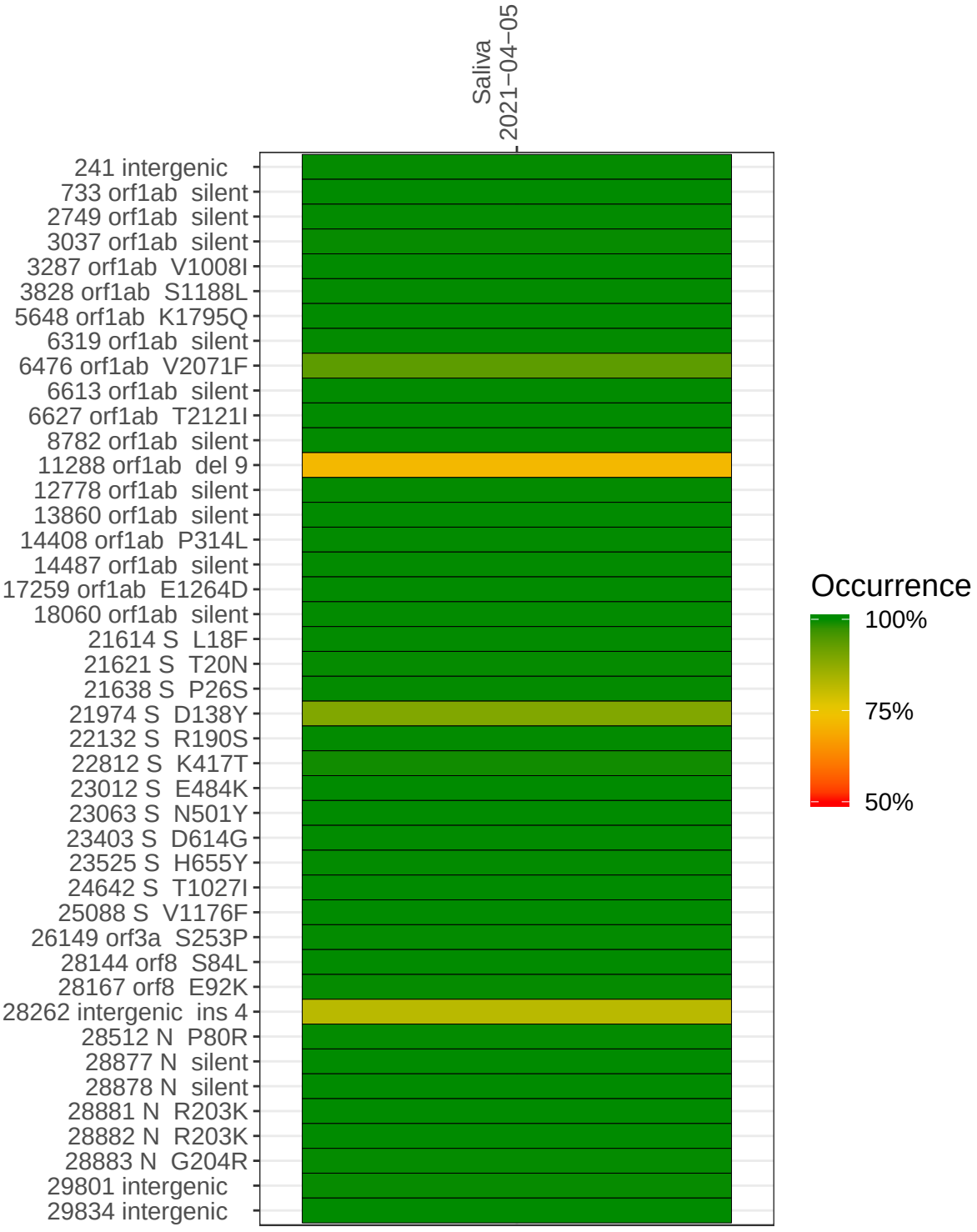
The table below provides a summary of subject samples for which sequencing data is available.
The experiments column shows the number of sequencing experiments performed for each specimen.
Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software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

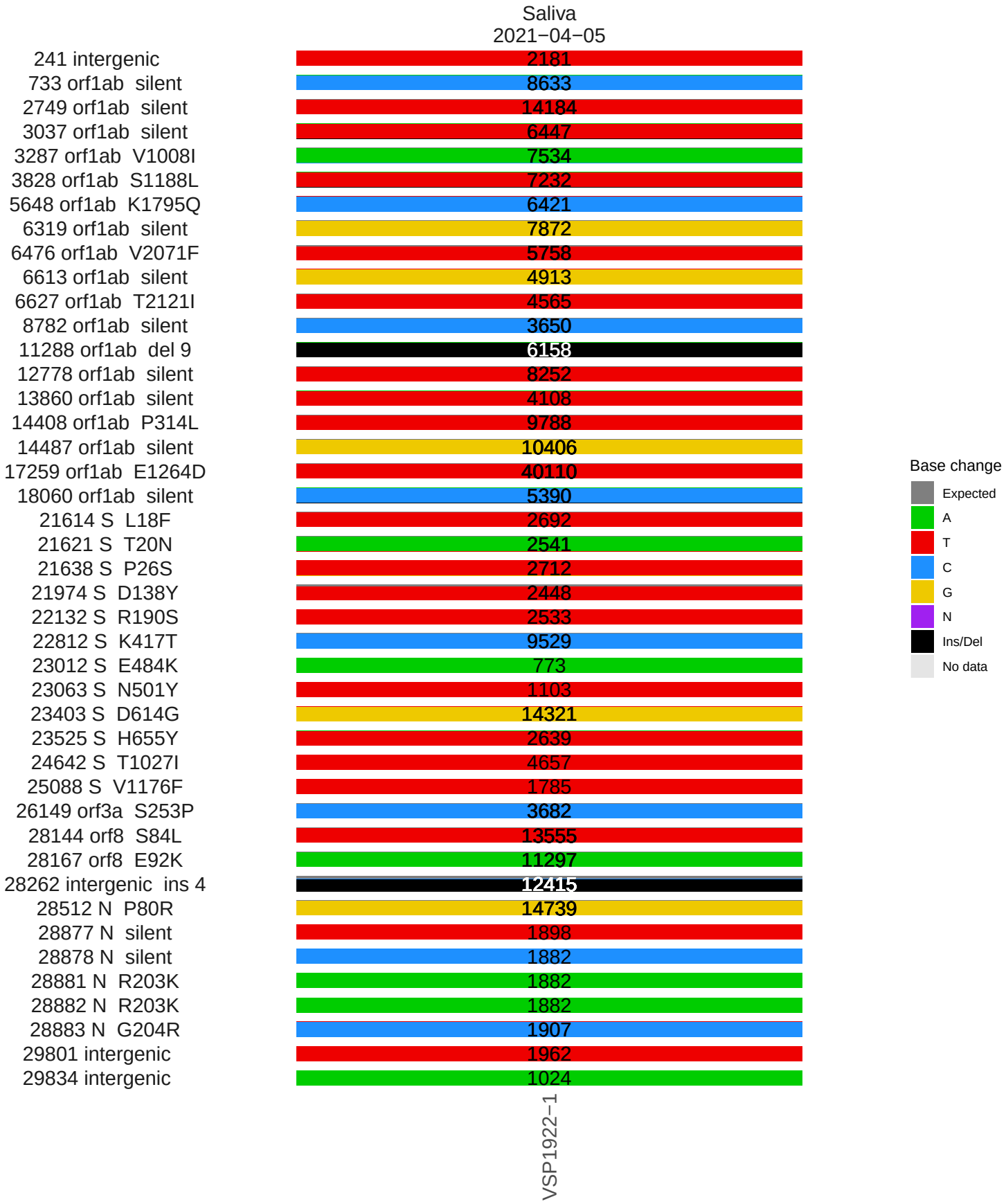
Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP1922-1	single experiment	NA	Saliva	2021-04-05	29.87	P.1	99.8%	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.

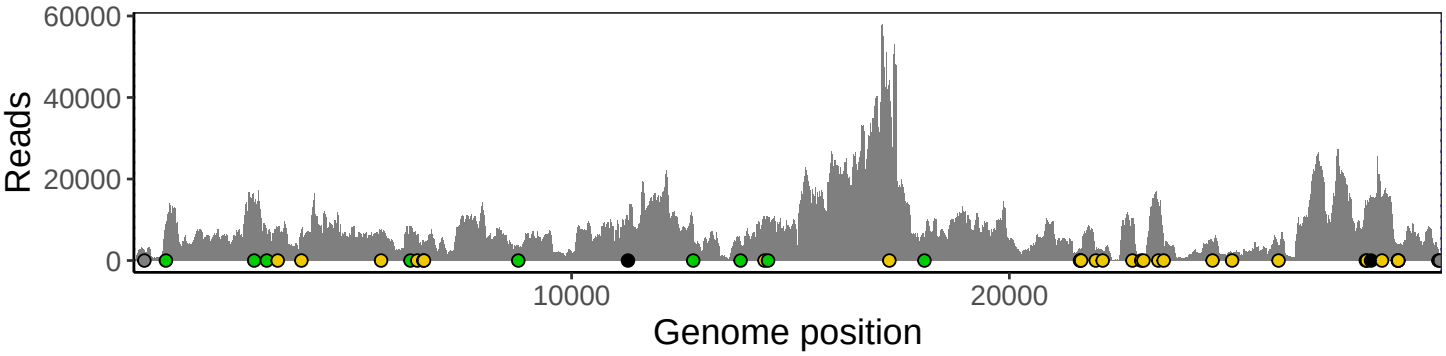



Saliva
2021−04−05
241 intergenic
733 orf1ab silent
2749 orf1ab silent
3037 orf1ab silent
3287 orf1ab V1008I
3828 orf1ab S1188L
5648 orf1ab K1795Q
6319 orf1ab silent
6476 orf1ab V2071F
6613 orf1ab silent
6627 orf1ab T2121I
8782 orf1ab silent
11288 orf1ab del 9
12778 orf1ab silent
13860 orf1ab silent
14408 orf1ab P314L
14487 orf1ab silent
17259 orf1ab E1264D
18060 orf1ab silent
21614 S L18F
21621 S T20N
21638 S P26S
21974 S D138Y
22132 S R190S
22812 S K417T
23012 S E484K
23063 S N501Y
23403 S D614G
23525 S H655Y
24642 S T1027I
25088 S V1176F
26149 orf3a S253P
28144 orf8 S84L
28167 orf8 E92K
28262 intergenic ins 4
28512 N P80R
28877 N silent
28878 N silent
28881 N R203K
28882 N R203K
28883 N G204R
29801 intergenic
29834 intergenic
2181
8633
14184
6447
7534
7232
6421
7872
5758
4913
4565
3650
6158
8252
4108
9788
10406
40110
5390
2692
2541
2712
2448
2533
9529
773
1103
14321
2639
4657
1785
3682
13555
11297
12415
14739
1898
1882
1882
1882
1907
1962
1024
VSP1922−1
Base change
Expected
A
T
C
G
N
Ins/Del
No data

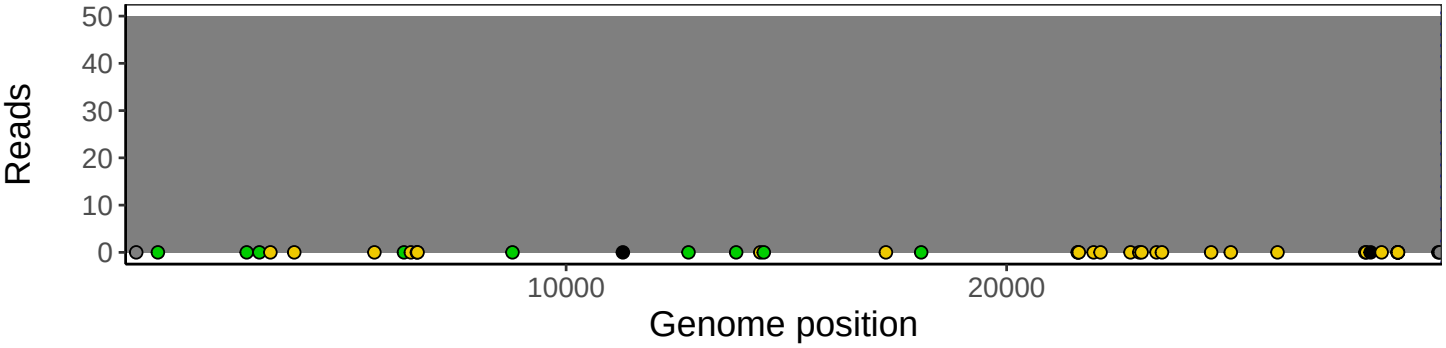
Analyses of individual experiments and composite results

VSP1922-1 | 2021-04-05 | Saliva | UPHS-0704 | genomes | single experiment

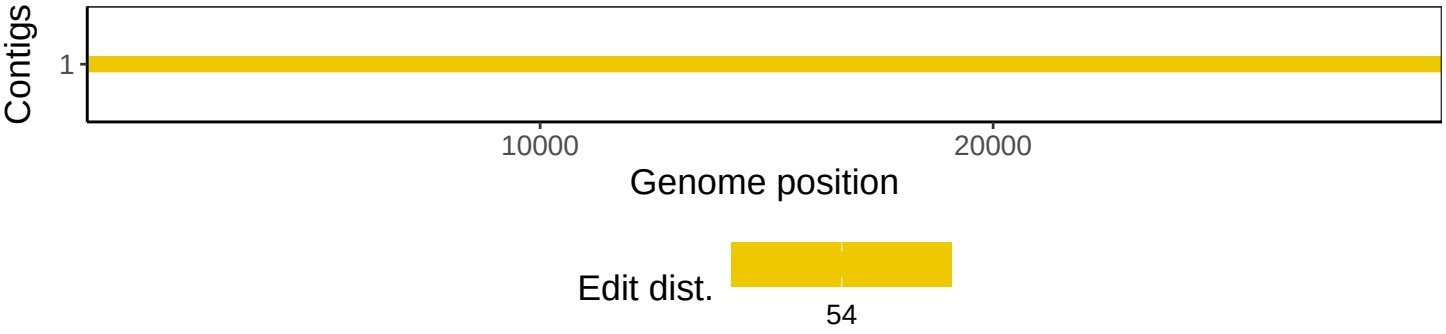
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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