COVID-19 subject UPHS- 0784

2021-04-20

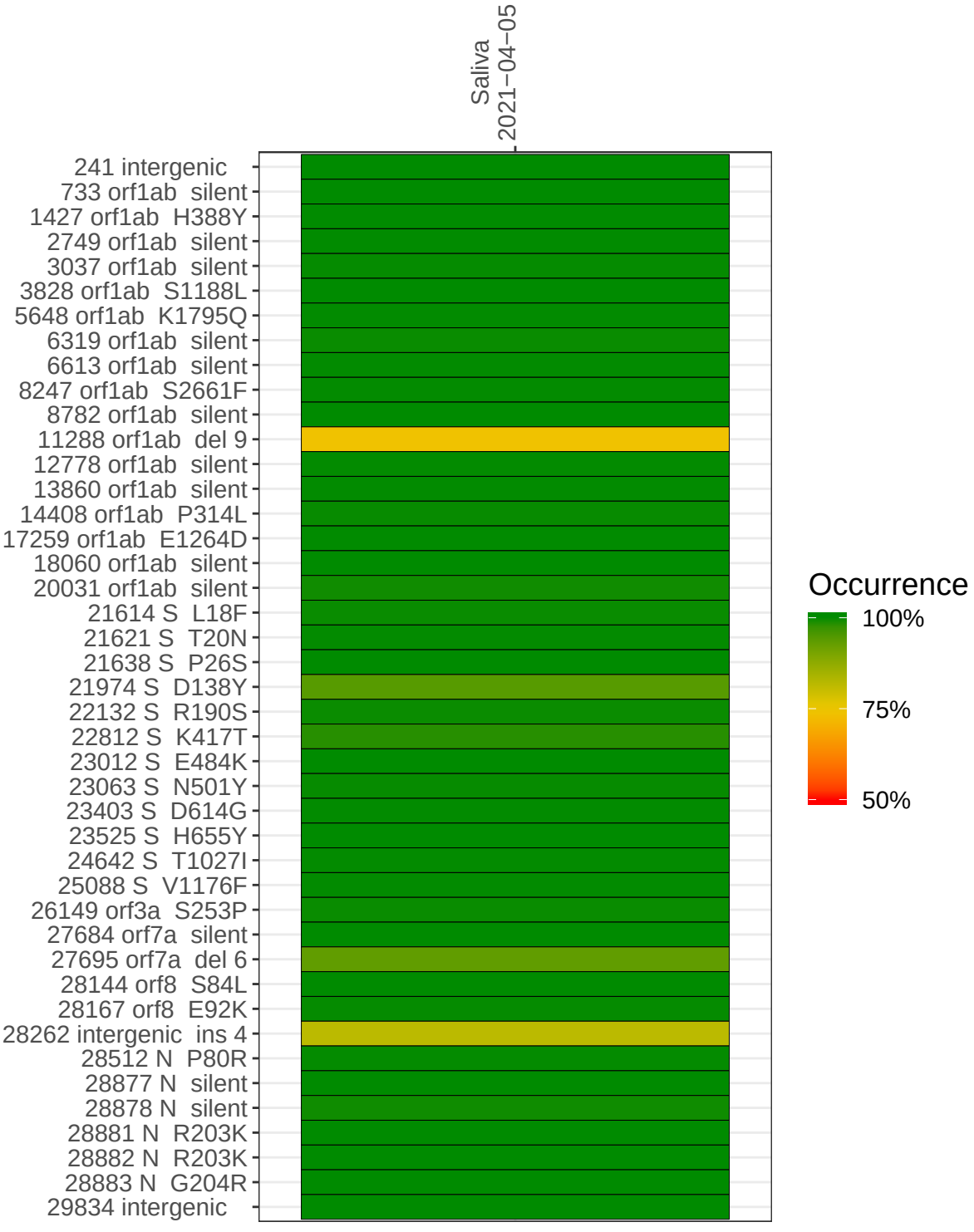
The table below provides a summary of subject samples for which sequencing data is available.
The experiments column shows the number of sequencing experiments performed for each specimen.
Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software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

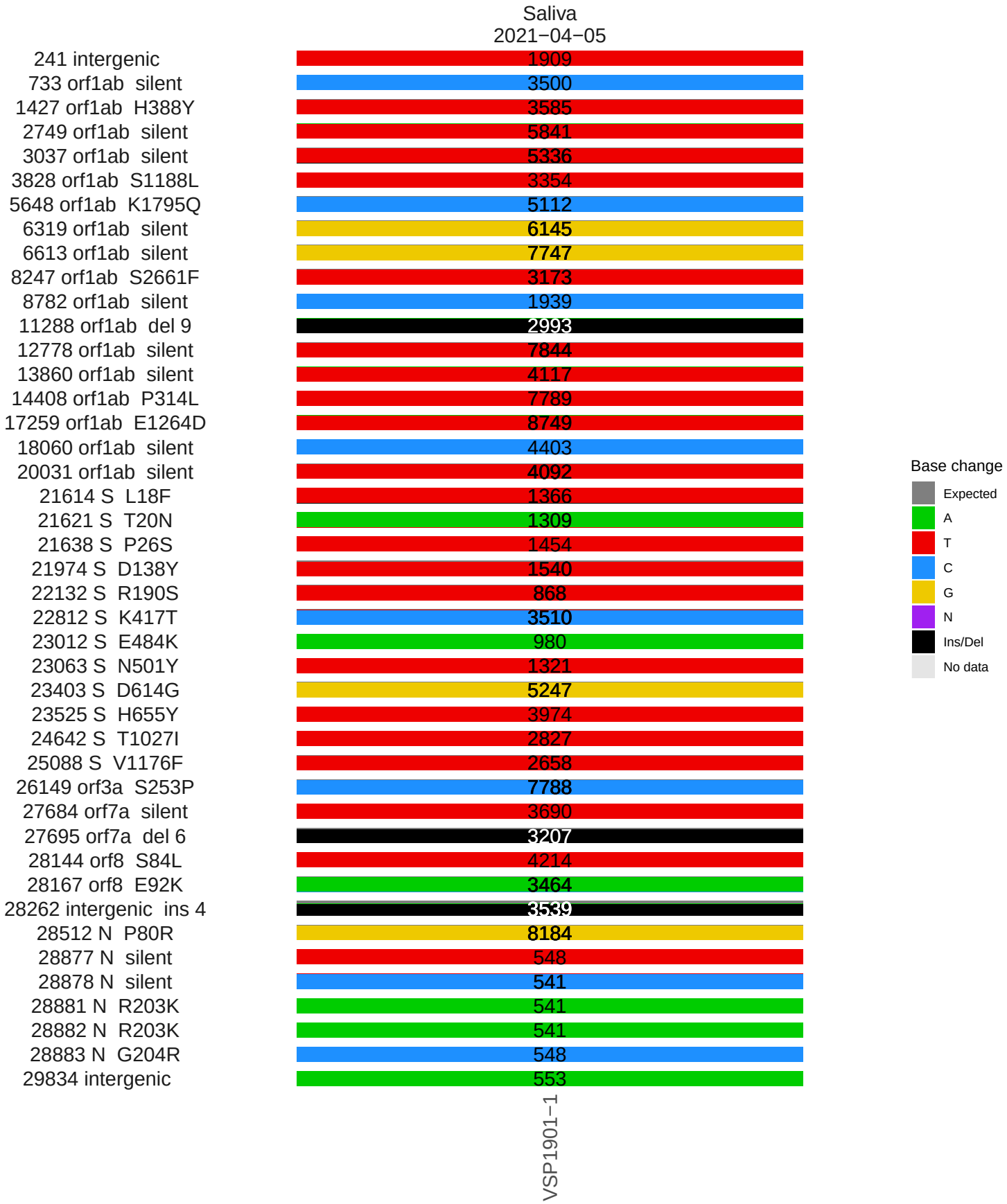
Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP1901-1	single experiment	NA	Saliva	2021-04-05	29.82	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.

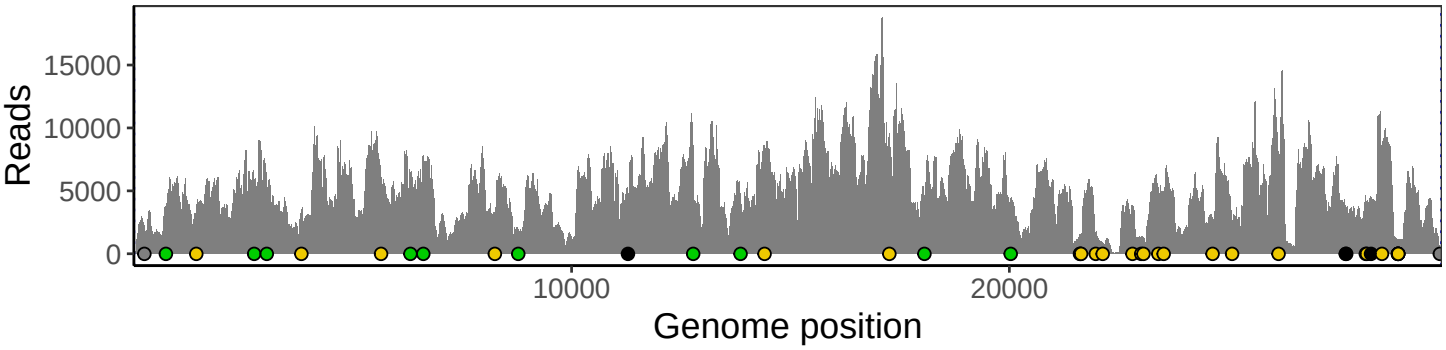



Saliva
2021−04−05
241 intergenic 1909
733 orf1ab silent 3500
1427 orf1ab H388Y 3585
2749 orf1ab silent 5841
3037 orf1ab silent 5336
3828 orf1ab S1188L 3354
5648 orf1ab K1795Q 5112
6319 orf1ab silent 6145
6613 orf1ab silent 7747
8247 orf1ab S2661F 3173
8782 orf1ab silent 1939
11288 orf1ab del 9 2993
12778 orf1ab silent 7844
13860 orf1ab silent 4117
14408 orf1ab P314L 7789
17259 orf1ab E1264D 8749
18060 orf1ab silent 4403
20031 orf1ab silent 4092
21614 S L18F 1366
21621 S T20N 1309
21638 S P26S 1454
21974 S D138Y 1540
22132 S R190S 868
22812 S K417T 3510
23012 S E484K 980
23063 S N501Y 1321
23403 S D614G 5247
23525 S H655Y 3974
24642 S T1027I 2827
25088 S V1176F 2658
26149 orf3a S253P 7788
27684 orf7a silent 3690
27695 orf7a del 6 3207
28144 orf8 S84L 4214
28167 orf8 E92K 3464
28262 intergenic ins 4 3539
28512 N P80R 8184
28877 N silent 548
28878 N silent 541
28881 N R203K 541
28882 N R203K 541
28883 N G204R 548
29834 intergenic 553
VSP1901−1
Base change
Expected
A
T
C
G
N
Ins/Del
No data

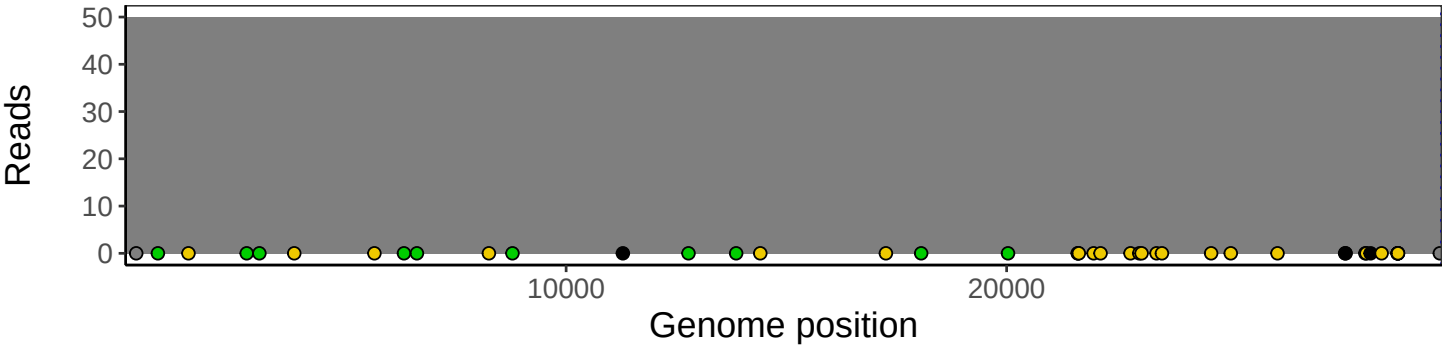
Analyses of individual experiments and composite results

VSP1901-1 | 2021-04-05 | Saliva | UPHS- 0784 | genomes | single experiment

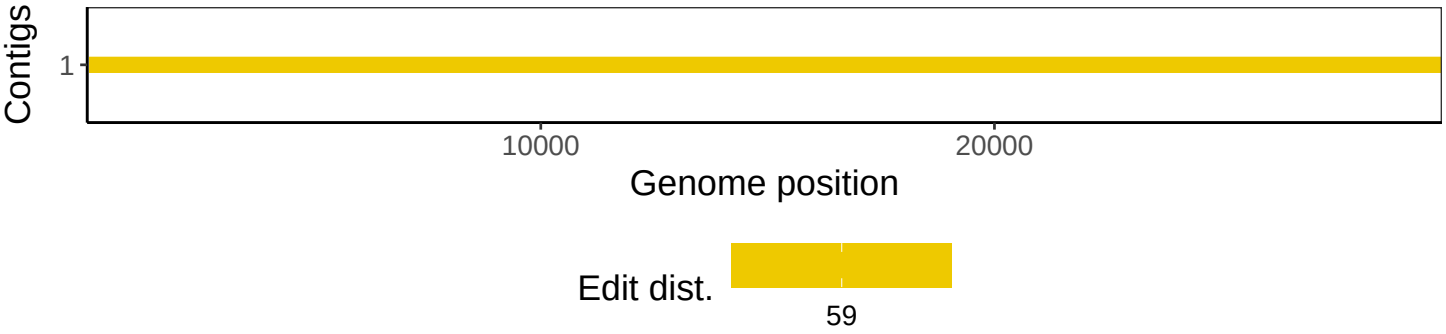
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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