COVID-19 subject UPHS-0744

2021-04-20

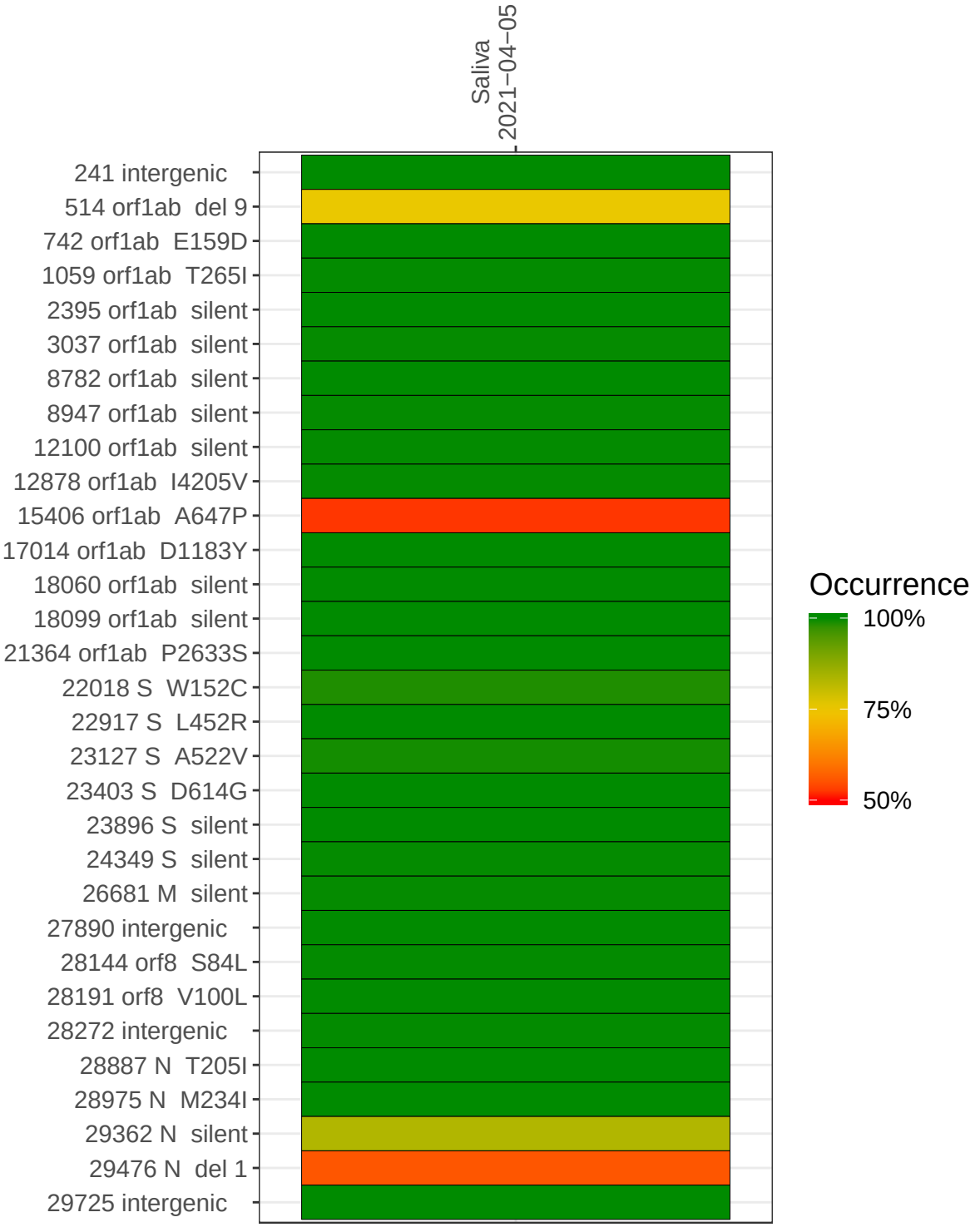
The table below provides a summary of subject samples for which sequencing data is available.
The experiments column shows the number of sequencing experiments performed for each specimen.
Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software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

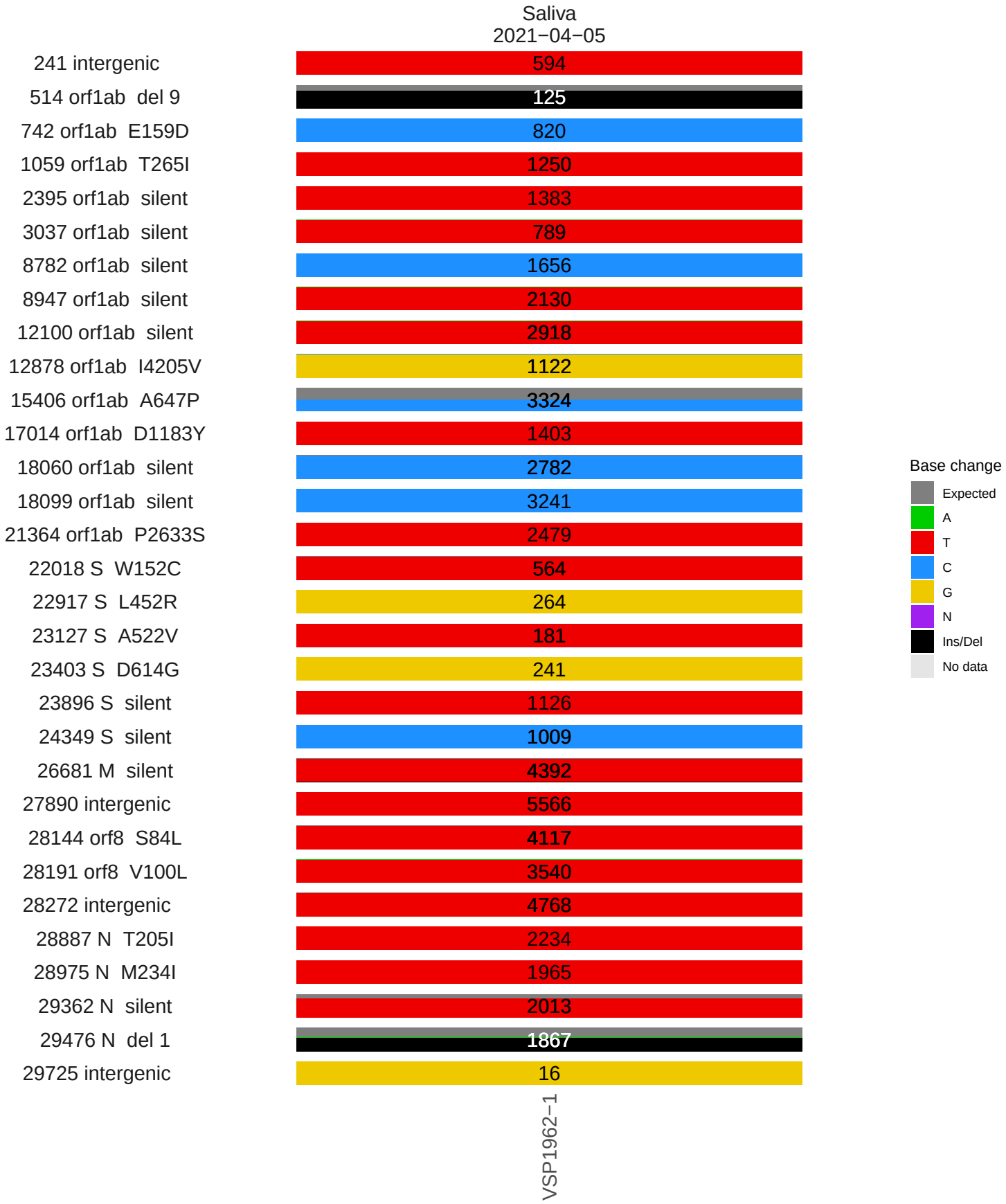
Table 1. Sample summary.

Experiment	Type	Genomes	Sample type	Sample date	Largest contig (KD)	Lineage	Reference read coverage	Reference read coverage (>= 5 reads)
VSP1962-1	single experiment	NA	Saliva	2021-04-05	4.26	NA	84.6%	83.9%

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.

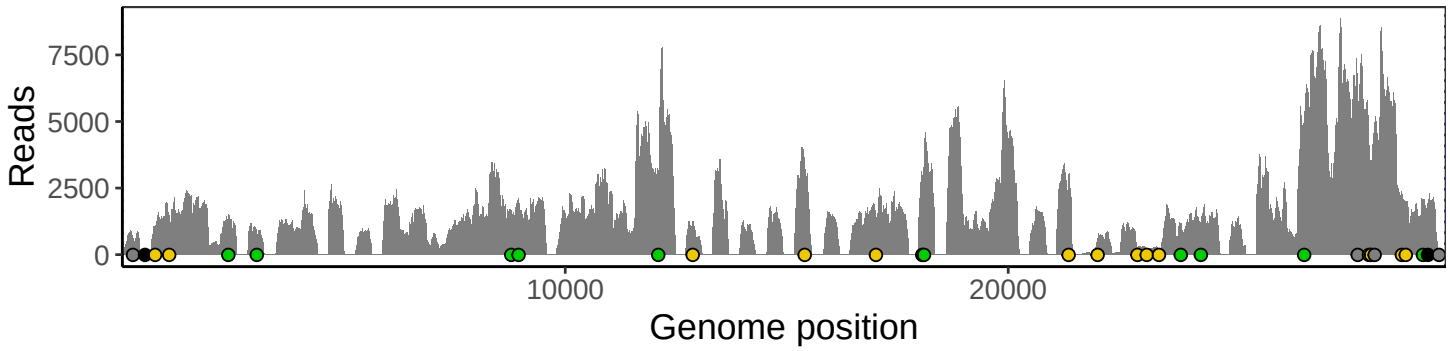



Saliva
2021−04−05
241 intergenic
514 orf1ab del 9
742 orf1ab E159D
1059 orf1ab T265I
2395 orf1ab silent
3037 orf1ab silent
8782 orf1ab silent
8947 orf1ab silent
12100 orf1ab silent
12878 orf1ab I4205V
15406 orf1ab A647P
17014 orf1ab D1183Y
18060 orf1ab silent
18099 orf1ab silent
21364 orf1ab P2633S
22018 S W152C
22917 S L452R
23127 S A522V
23403 S D614G
23896 S silent
24349 S silent
26681 M silent
27890 intergenic
28144 orf8 S84L
28191 orf8 V100L
28272 intergenic
28887 N T205I
28975 N M234I
29362 N silent
29476 N del 1
29725 intergenic
594
125
820
1250
1383
789
1656
2130
2918
1122
3324
1403
2782
3241
2479
564
264
181
241
1126
1009
4392
5566
4117
3540
4768
2234
1965
2013
1867
16
VSP1962−1
Base change
Expected
A
T
C
G
N
Ins/Del
No data

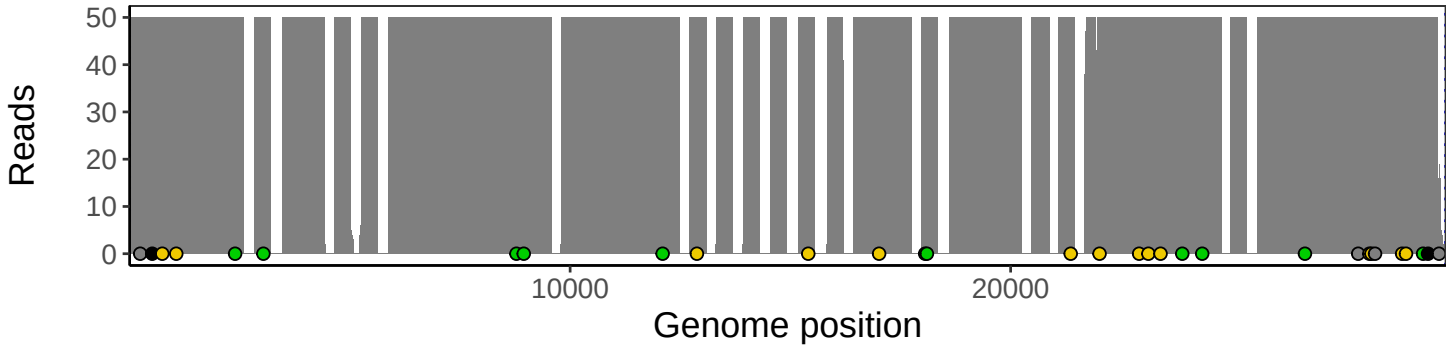
Analyses of individual experiments and composite results

VSP1962-1 | 2021-04-05 | Saliva | UPHS-0744 | genomes | single experiment

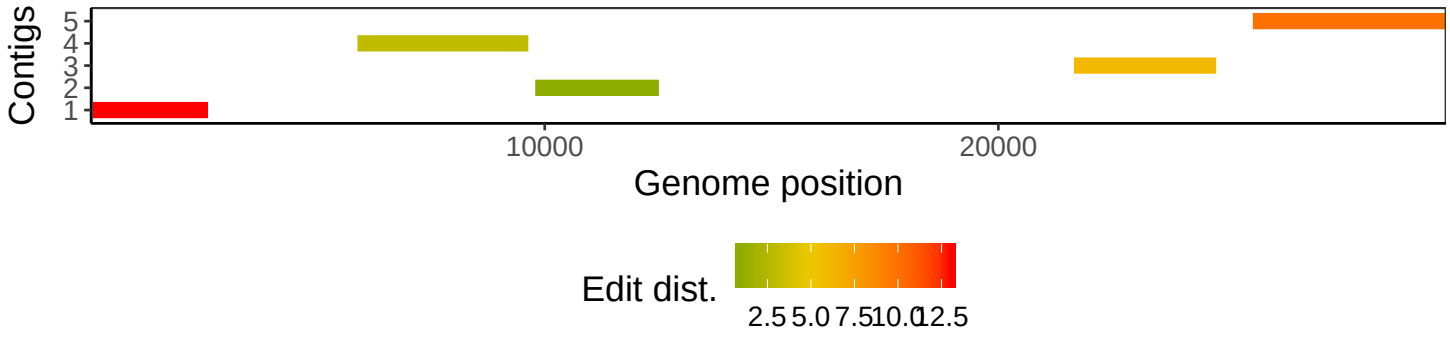
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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