

COVID-19 subject PQ-Seq17

2021-06-23

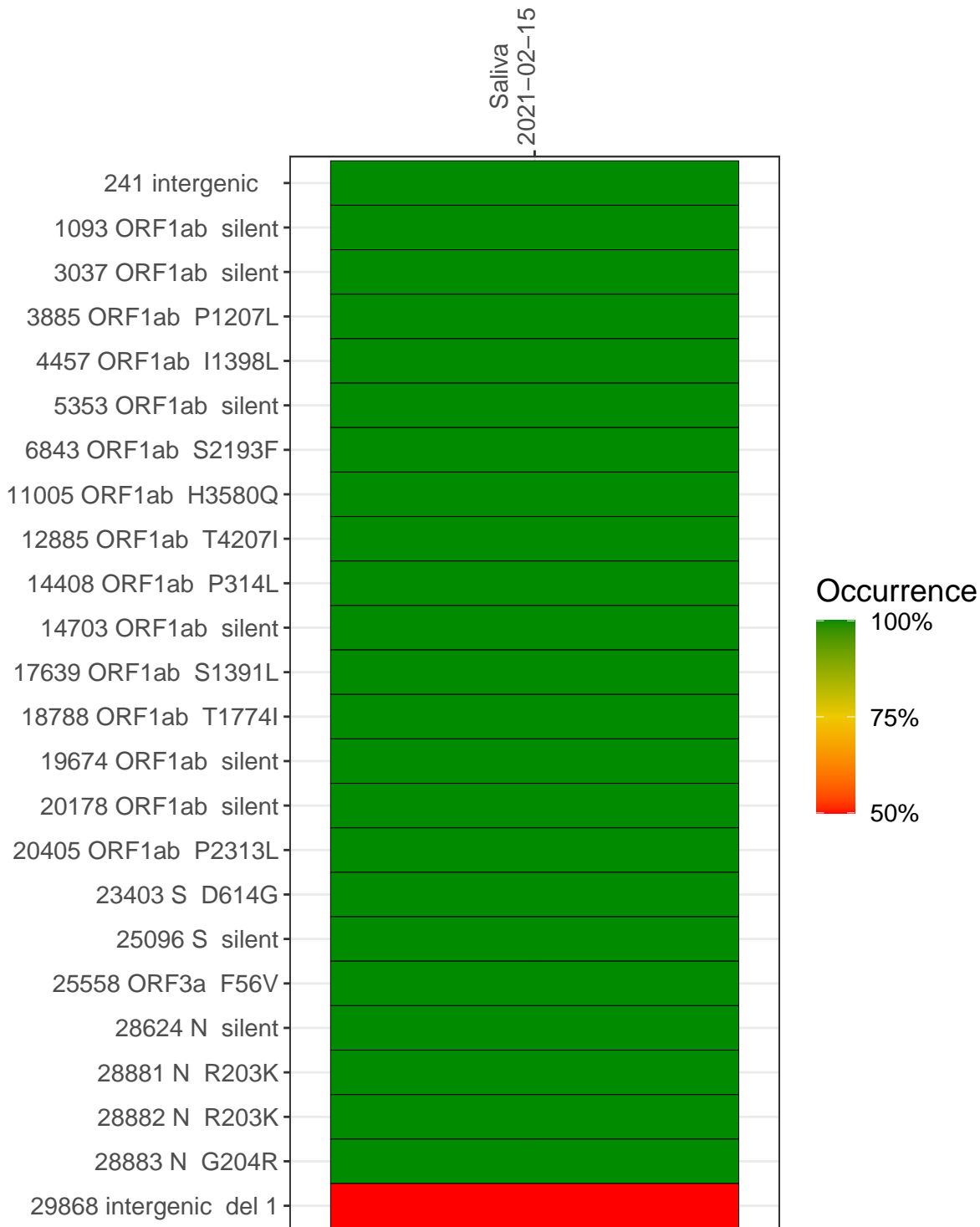
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP0844-1	single experiment	NA	Saliva	2021-02-15	12.21	B.1.1.265	97.2%	96.2%

Variants shared across samples

The heat map below shows how variants (reference genome /home/common/SARS-CoV-2-Philadelphia/Wuhan-Hu-1) 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
2021-02-15

241 intergenic	337
1093 ORF1ab silent	105
3037 ORF1ab silent	101
3885 ORF1ab P1207L	224
4457 ORF1ab I1398L	186
5353 ORF1ab silent	214
6843 ORF1ab S2193F	296
11005 ORF1ab H3580Q	53
12885 ORF1ab T4207I	78
14408 ORF1ab P314L	388
14703 ORF1ab silent	415
17639 ORF1ab S1391L	252
18788 ORF1ab T1774I	268
19674 ORF1ab silent	171
20178 ORF1ab silent	48
20405 ORF1ab P2313L	27
23403 S D614G	422
25096 S silent	46
25558 ORF3a F56V	139
28624 N silent	537
28881 N R203K	26
28882 N R203K	26
28883 N G204R	26
29868 intergenic del 1	59

Base change

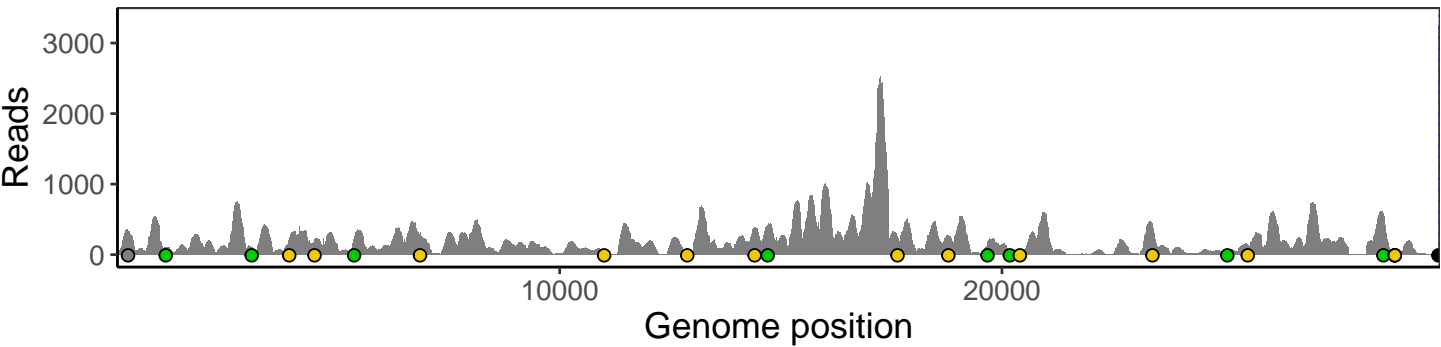


VSP0844-1

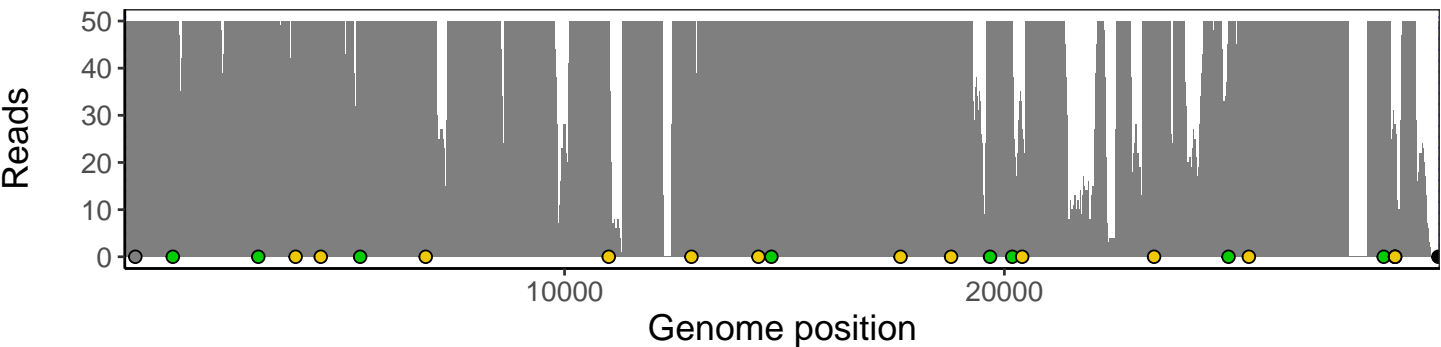
Analyses of individual experiments and composite results

VSP0844-1 | 2021-02-15 | Saliva | PQ-Seq17 | genomes | single experiment

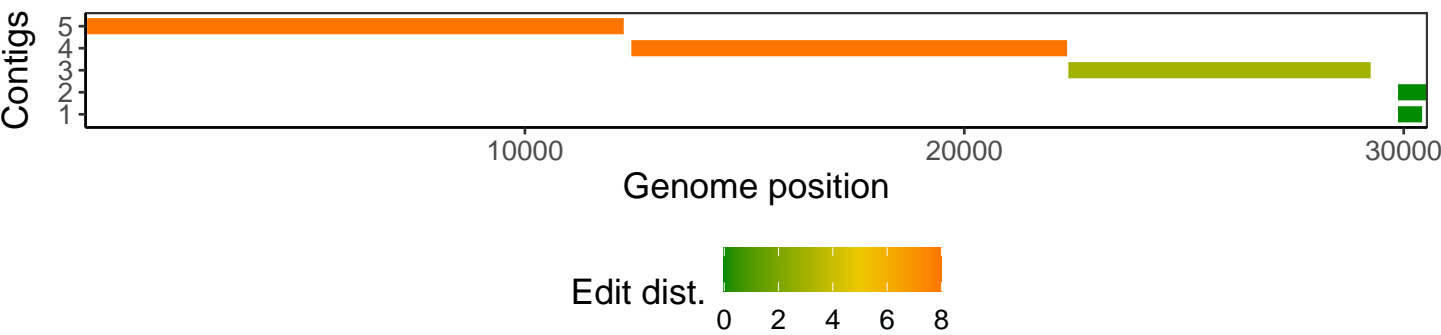
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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 htlib 1.10
bcftools	1.10.2-34-g1a12af0-dirty Using htlib 1.10.2-57-gf58a6f3
pangolin	3.1.3
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	3.3.3
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