

COVID-19 subject PQ-Seq6

2021-04-17

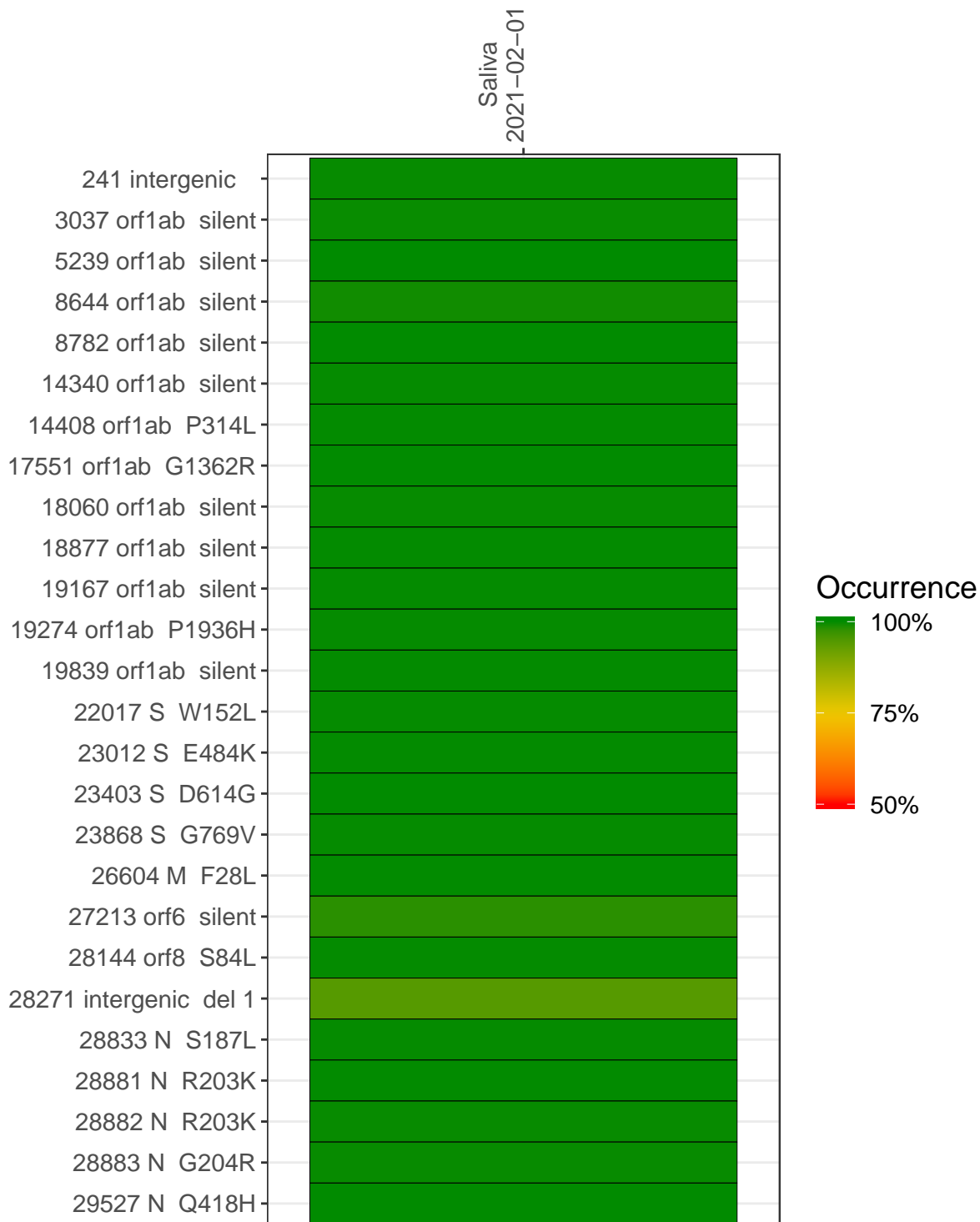
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP0775	composite	NA	Saliva	2021-02-01	21.73	R.1	99.6%	99.1%
VSP0775-1	single experiment	NA	Saliva	2021-02-01	21.70	R.1	99.5%	99.0%
VSP0775-2	single experiment	NA	Saliva	2021-02-01	6.18	NA	93.5%	83.1%
VSP0775-3	single experiment	NA	Saliva	2021-02-01	21.73	R.1	99.1%	99.0%

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 2021-02-01	
241 intergenic	14592	24	2828
3037 orf1ab silent	7636	5	2266
5239 orf1ab silent	1327	136	1635
8644 orf1ab silent	3030	0	2863
8782 orf1ab silent	8641	0	3267
14340 orf1ab silent	8732	17	3454
14408 orf1ab P314L	11689	10	3050
17551 orf1ab G1362R	18412	19	3790
18060 orf1ab silent	5674	46	3662
18877 orf1ab silent	15137	70	5919
19167 orf1ab silent	12499	27	5413
19274 orf1ab P1936H	5670	99	4265
19839 orf1ab silent	5577	8	3291
22017 S W152L	984	6	2441
23012 S E484K	7725	106	2591
23403 S D614G	17574	150	5751
23868 S G769V	4347	38	2694
26604 M F28L	7916	40	5258
27213 orf6 silent	4766	39	4767
28144 orf8 S84L	8209	92	6454
28271 intergenic del 1	13188	59	4398
28833 N S187L	2846	184	979
28881 N R203K	2511	176	716
28882 N R203K	2507	176	716
28883 N G204R	2509	178	717
29527 N Q418H	4528	43	1999
	VSP0775-1	VSP0775-2	VSP0775-3

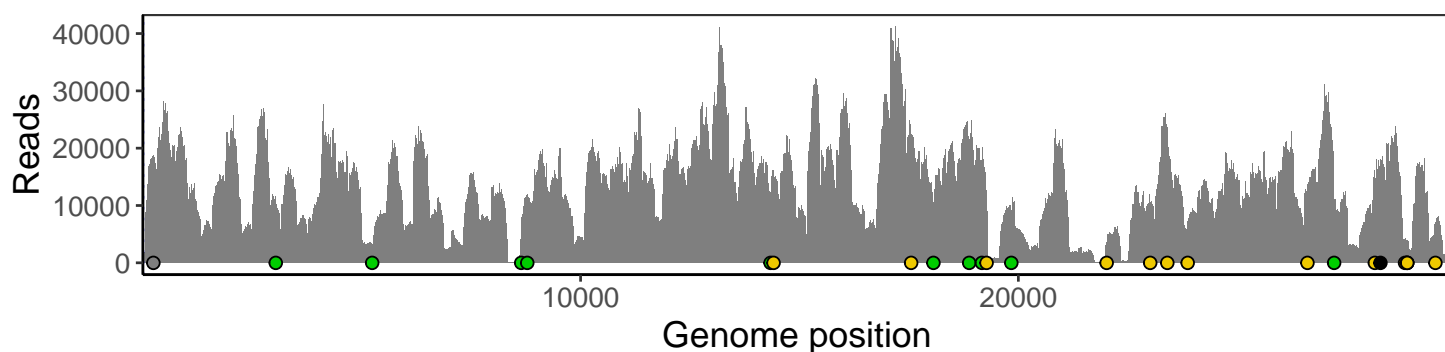
Base change

- Expected
- A
- T
- C
- G
- N
- Ins/Del
- No data

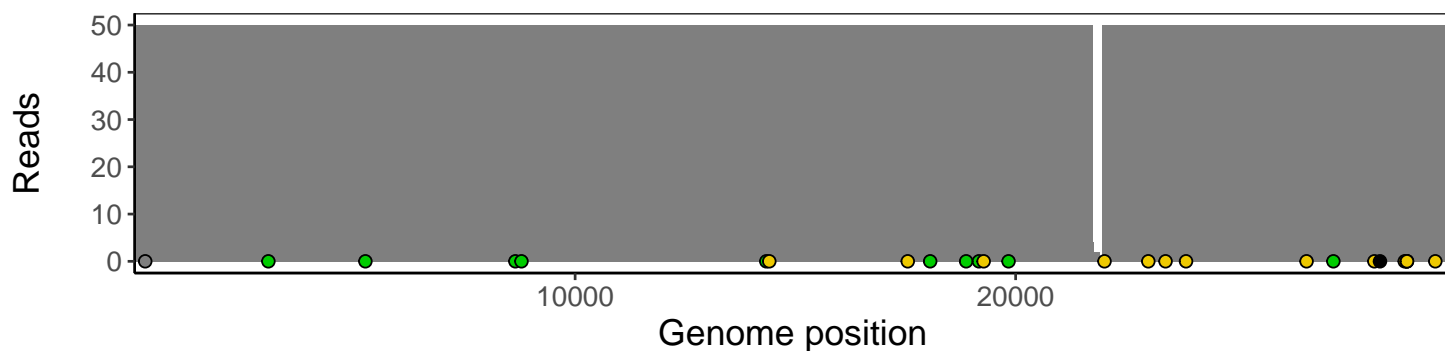
Analyses of individual experiments and composite results

VSP0775 | 2021-02-01 | Saliva | PQ-Seq6 | composite result

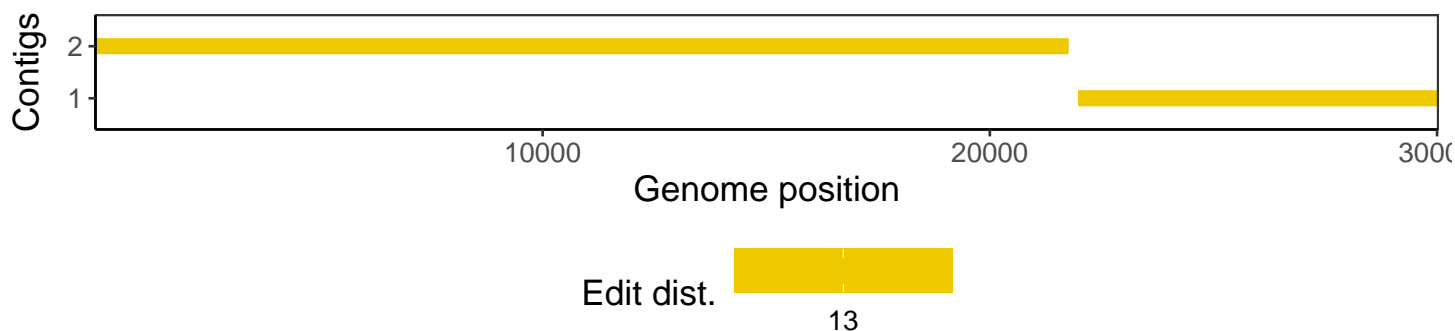
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according to variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



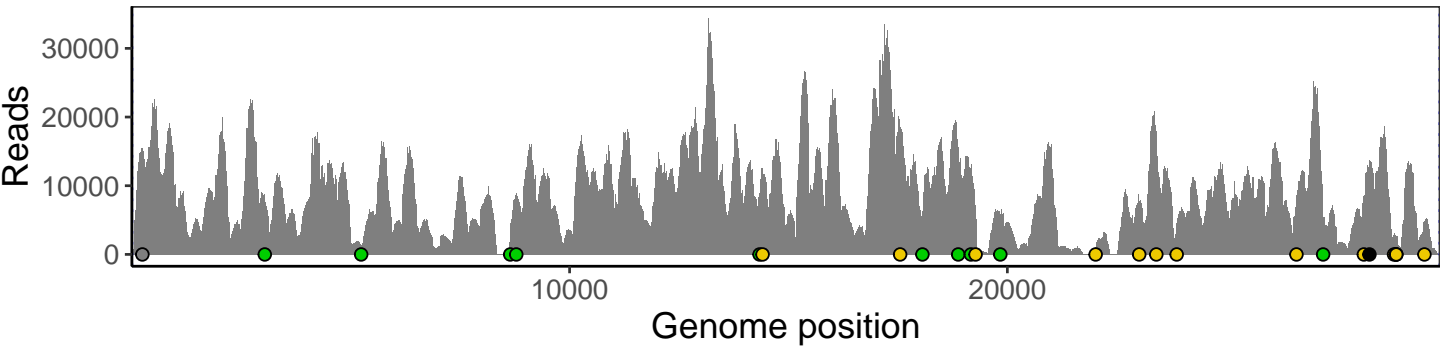
Excerpt from plot above focusing on reads coverage from 0 to 50 NT.



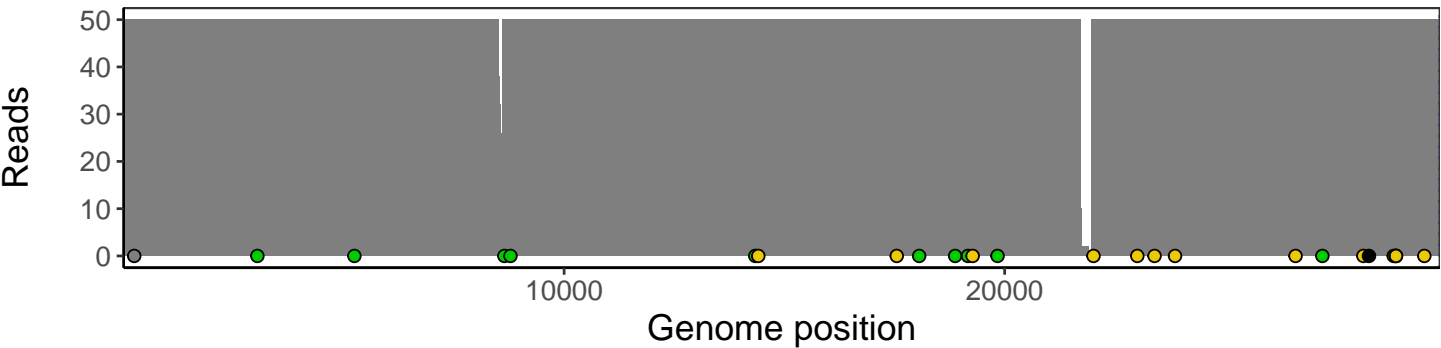
The longest five assembled contigs are shown below colored by their edit distance to the reference genome.



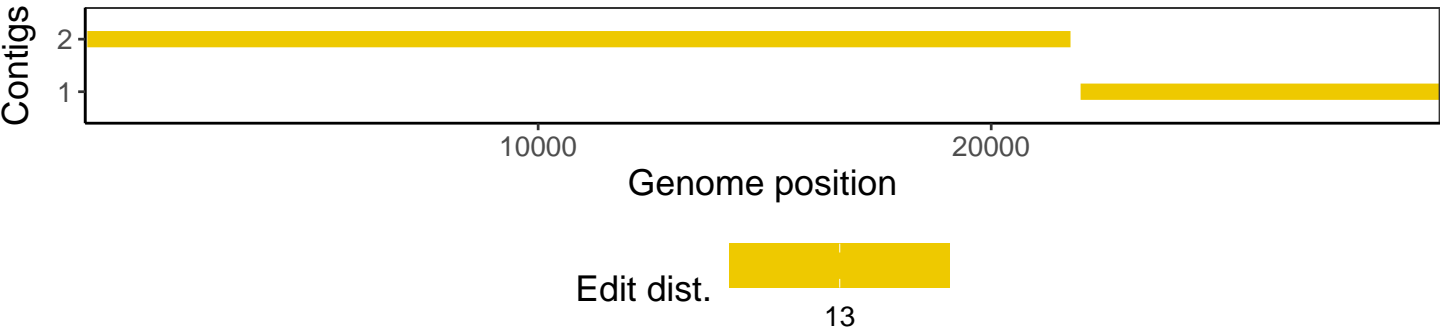
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



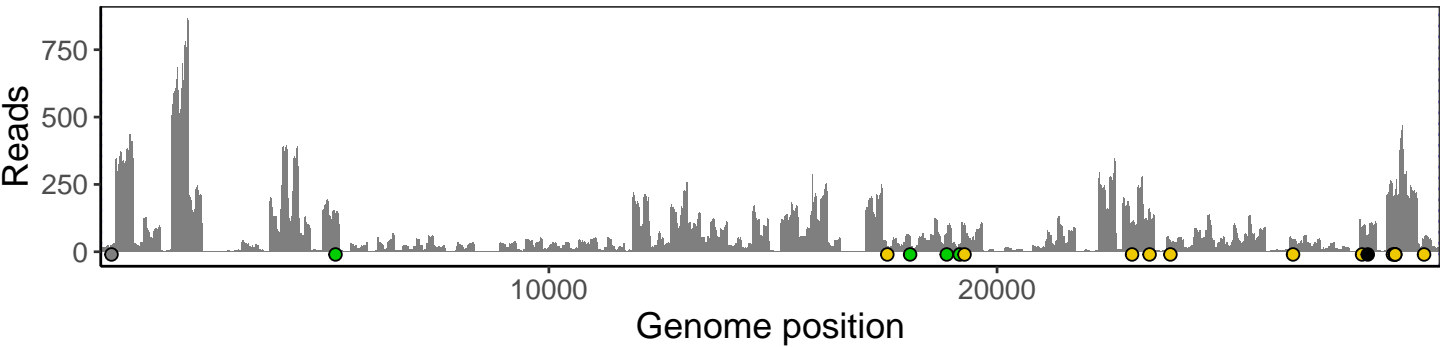
Excerpt from plot above focusing on reads coverage from 0 to 50 NT.



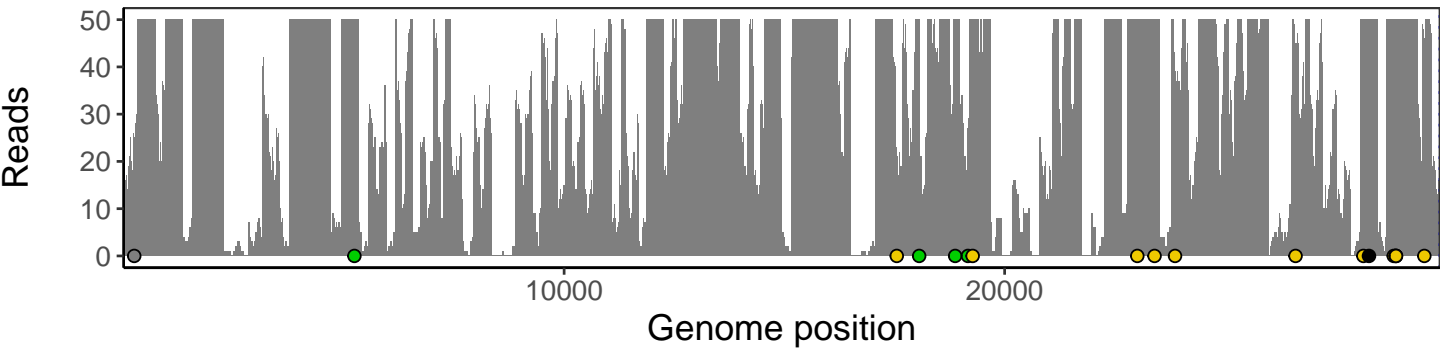
The longest five assembled contigs are shown below colored by their edit distance to the reference genome.



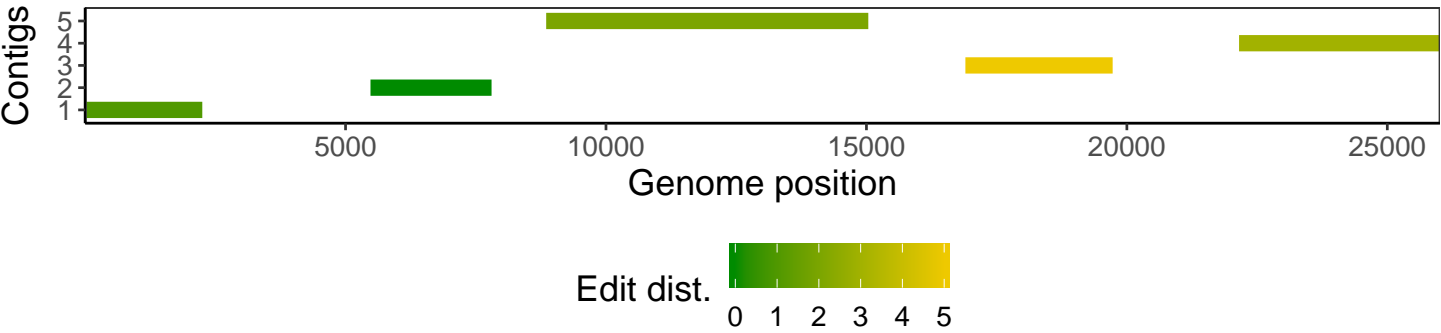
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



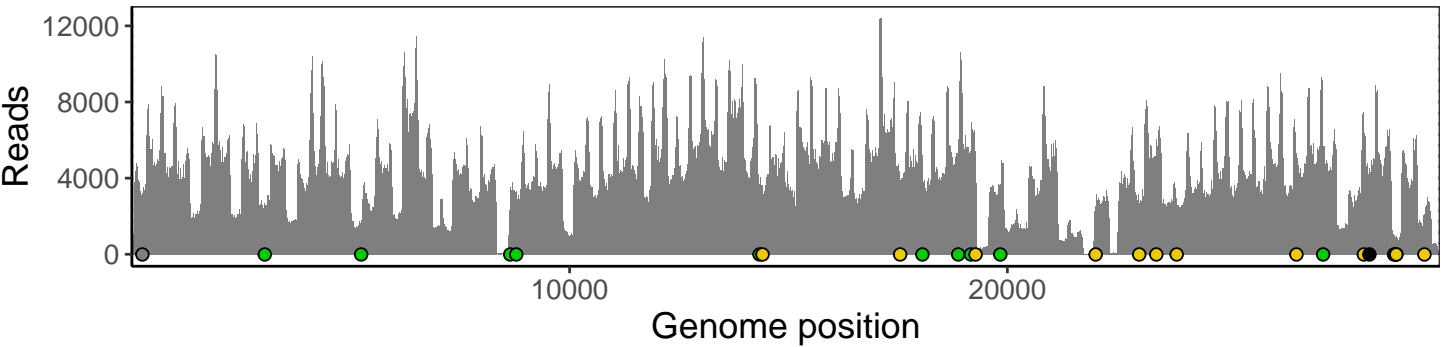
Excerpt from plot above focusing on reads coverage from 0 to 50 NT.



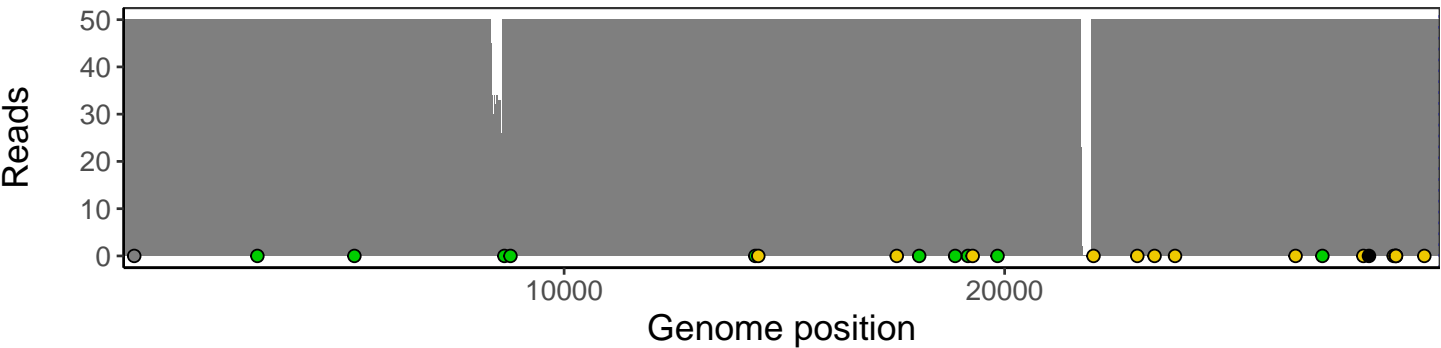
The longest five assembled contigs are shown below colored by their edit distance to the reference genome.



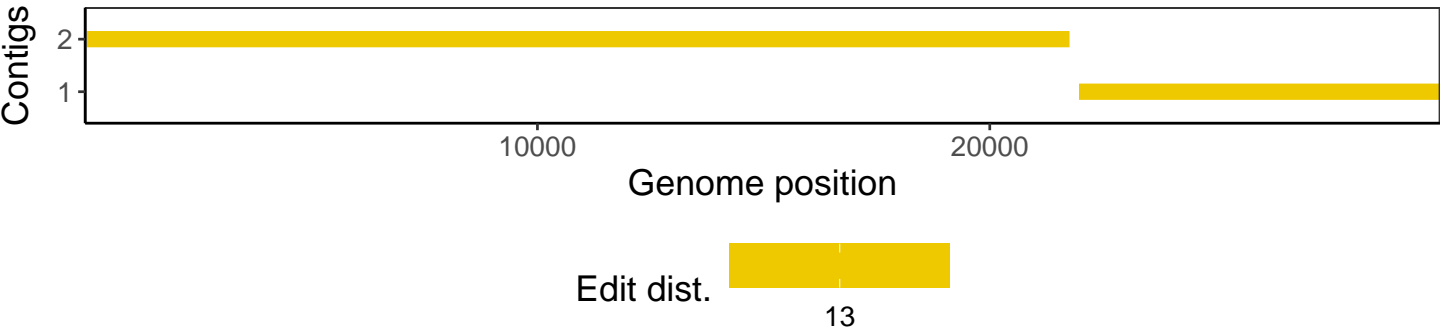
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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	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