

COVID-19 subject molpath-sdrop4

2021-04-17

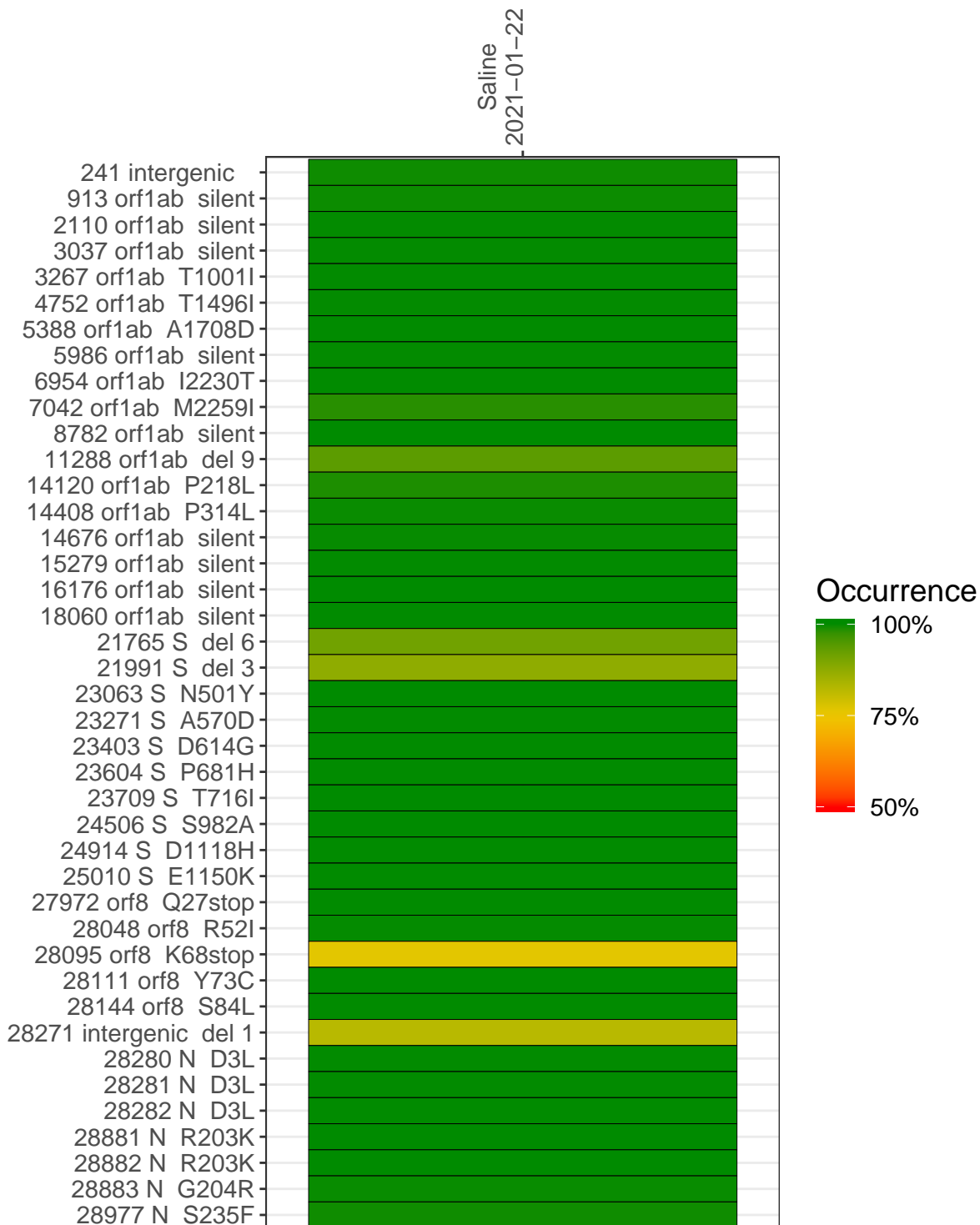
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP0625	composite	NA	Saline	2021-01-22	29.76	B.1.1.7	99.2%	99.2%
VSP0625-1	single experiment	NA	Saline	2021-01-22	29.65	B.1.1.7	99.2%	99.2%
VSP0625-2	single experiment	NA	Saline	2021-01-22	19.57	B.1.1.7	99.2%	99.2%

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.



Saline
2021-01-22

241 intergenic	12844	5305
913 orf1ab silent	20632	7851
2110 orf1ab silent	5960	2203
3037 orf1ab silent	9690	3893
3267 orf1ab T1001I	12700	4750
4752 orf1ab T1496I	16821	6743
5388 orf1ab A1708D	9501	3623
5986 orf1ab silent	4790	1699
6954 orf1ab I2230T	2691	1021
7042 orf1ab M2259I	3196	1331
8782 orf1ab silent	5318	2006
11288 orf1ab del 9	18186	6835
14120 orf1ab P218L	18012	6891
14408 orf1ab P314L	5618	2092
14676 orf1ab silent	15762	6108
15279 orf1ab silent	33500	12578
16176 orf1ab silent	11696	4531
18060 orf1ab silent	6501	2351
21765 S del 6	3671	1511
21991 S del 3	1530	591
23063 S N501Y	3448	1478
23271 S A570D	15187	5942
23403 S D614G	18536	7183
23604 S P681H	6964	2915
23709 S T716I	5330	2405
24506 S S982A	7534	2753
24914 S D1118H	14035	5048
25010 S E1150K	7893	3186
27972 orf8 Q27stop	16254	6241
28048 orf8 R52I	14424	5876
28095 orf8 K68stop	13090	5160
28111 orf8 Y73C	14734	5711
28144 orf8 S84L	15794	5956
28271 intergenic del 1	21732	7562
28280 N D3L	18243	5620
28281 N D3L	18243	5620
28282 N D3L	18373	5708
28881 N R203K	1926	653
28882 N R203K	1926	651
28883 N G204R	1926	656
28977 N S235F	911	441

Base change

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

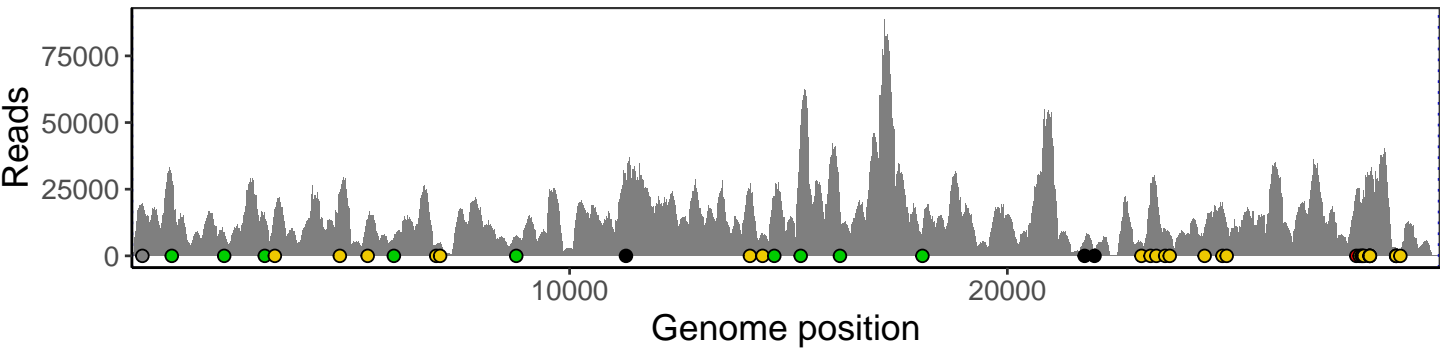
VSP0625-1

VSP0625-2

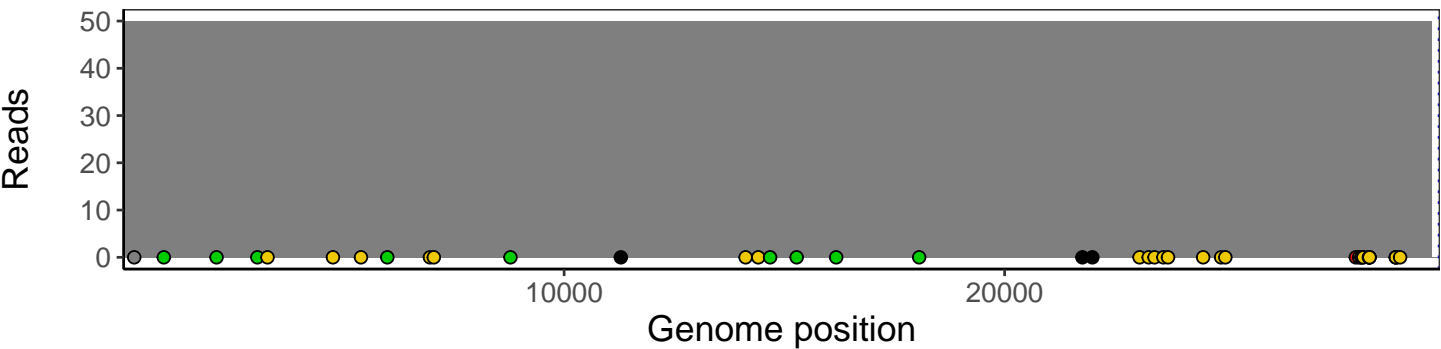
Analyses of individual experiments and composite results

VSP0625 | 2021-01-22 | Saline | molpath-sdrop4 | composite result

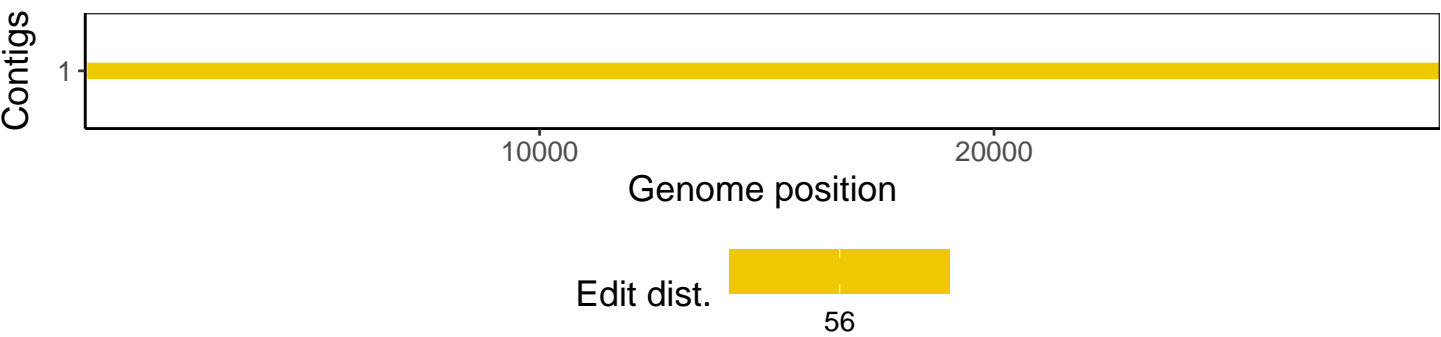
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



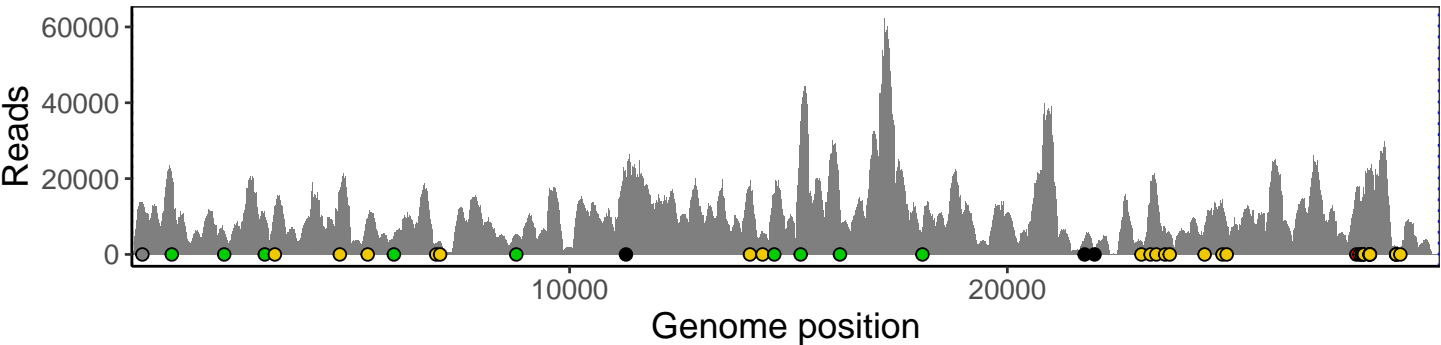
Excerpt from plot above focusing on reads coverage from 0 to 50 NT.



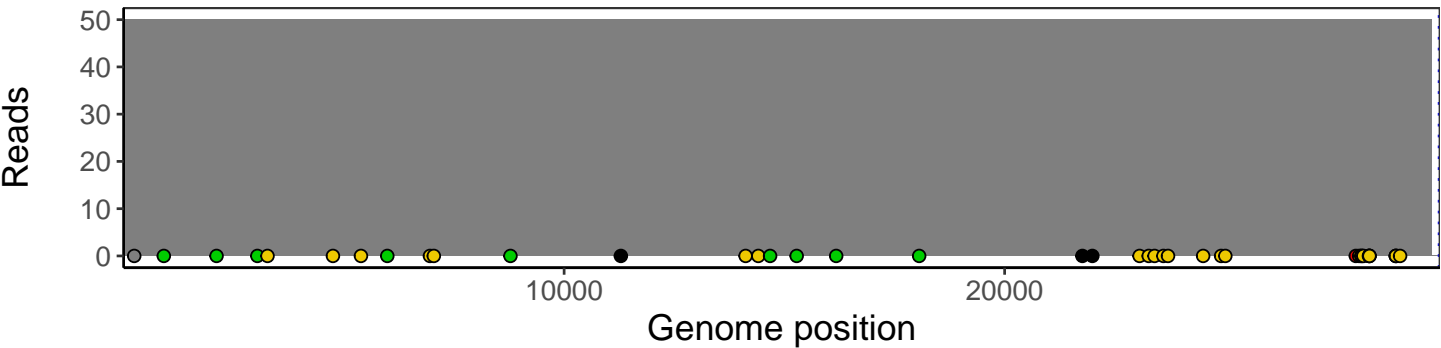
The longest five assembled contigs are shown below colored by their edit distance to the reference genome.



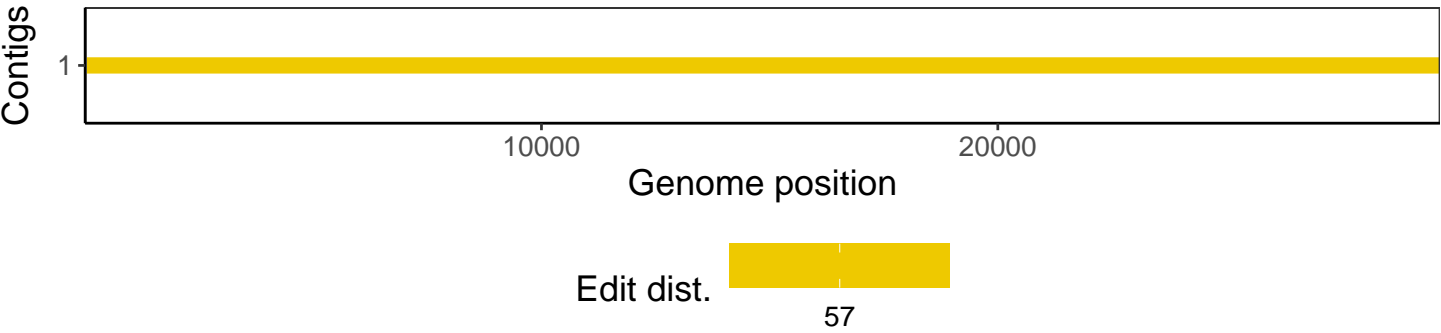
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



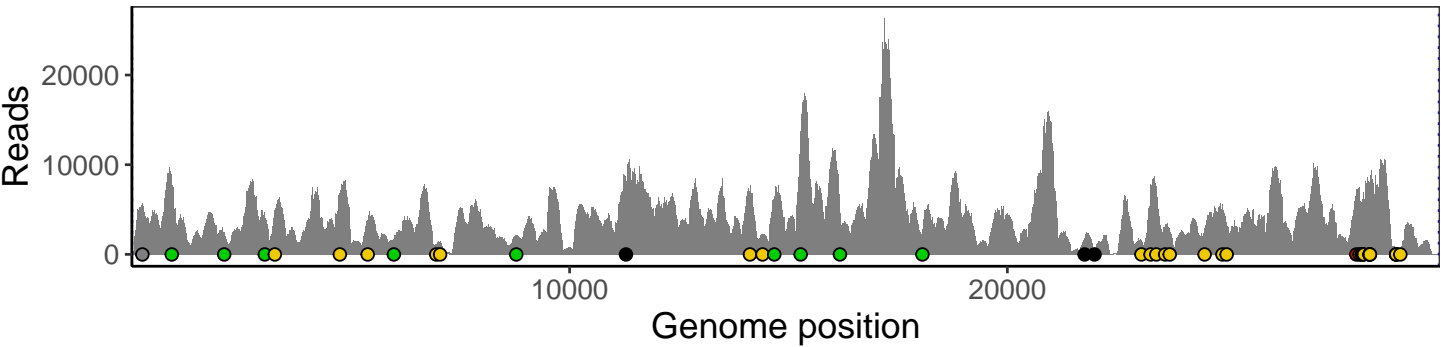
Excerpt from plot above focusing on reads coverage from 0 to 50 NT.



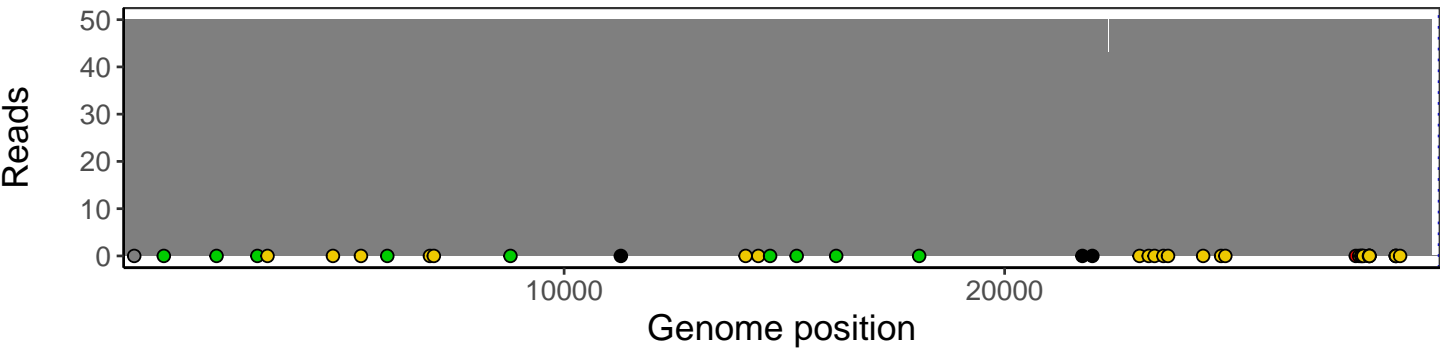
The longest five assembled contigs are shown below colored by their edit distance to the reference genome.



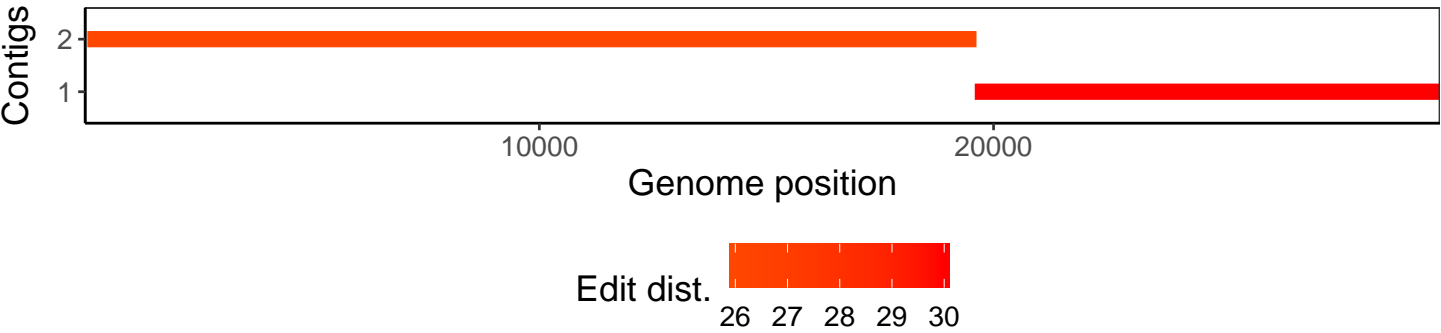
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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