

COVID-19 subject DOH1

2021-06-23

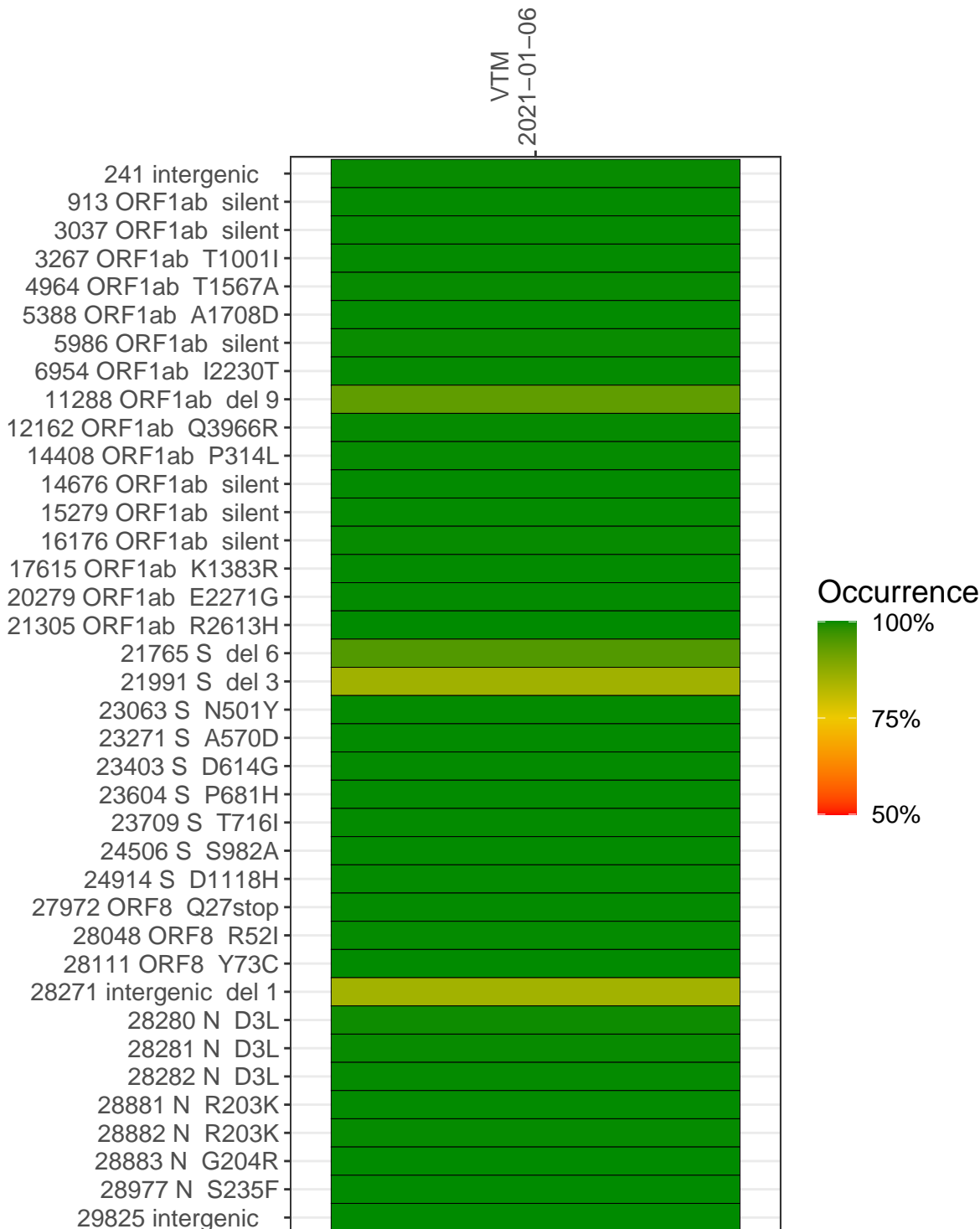
The table below provides a summary of subject samples for which sequencing data is available. The experiments column shows the number of sequencing experiments performed for each specimen. Experiment specific analyses are shown at the end of this report. Lineages are called with the Pangolin software 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)
VSP0563	composite	NA	VTM	2021-01-06	29.90	B.1.1.7	100.0%	99.7%
VSP0563-2	single experiment	NA	VTM	2021-01-06	24.56	B.1.1.7	98.8%	98.1%
VSP0563-3	single experiment	NA	VTM	2021-01-06	29.90	B.1.1.7	100.0%	99.7%

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.



VTM
2021-01-06

241 intergenic	129	6039
913 ORF1ab silent	257	14368
3037 ORF1ab silent	122	5913
3267 ORF1ab T1001I	133	8131
4964 ORF1ab T1567A	107	6914
5388 ORF1ab A1708D	77	4920
5986 ORF1ab silent	71	4521
6954 ORF1ab I2230T	33	2292
11288 ORF1ab del 9	344	17912
12162 ORF1ab Q3966R	200	8792
14408 ORF1ab P314L	192	9923
14676 ORF1ab silent	144	8231
15279 ORF1ab silent	489	19699
16176 ORF1ab silent	160	8842
17615 ORF1ab K1383R	194	9916
20279 ORF1ab E2271G	18	1505
21305 ORF1ab R2613H	36	1826
21765 S del 6	112	6607
21991 S del 3	50	2370
23063 S N501Y	129	7611
23271 S A570D	336	20840
23403 S D614G	469	23880
23604 S P681H	200	12046
23709 S T716I	169	8671
24506 S S982A	109	5124
24914 S D1118H	193	12450
27972 ORF8 Q27stop	305	19001
28048 ORF8 R52I	294	15319
28111 ORF8 Y73C	357	17140
28271 intergenic del 1	341	26412
28280 N D3L	277	22492
28281 N D3L	277	22492
28282 N D3L	281	22685
28881 N R203K	47	4204
28882 N R203K	47	4186
28883 N G204R	47	4187
28977 N S235F	14	1990
29825 intergenic	4	198

Base change

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

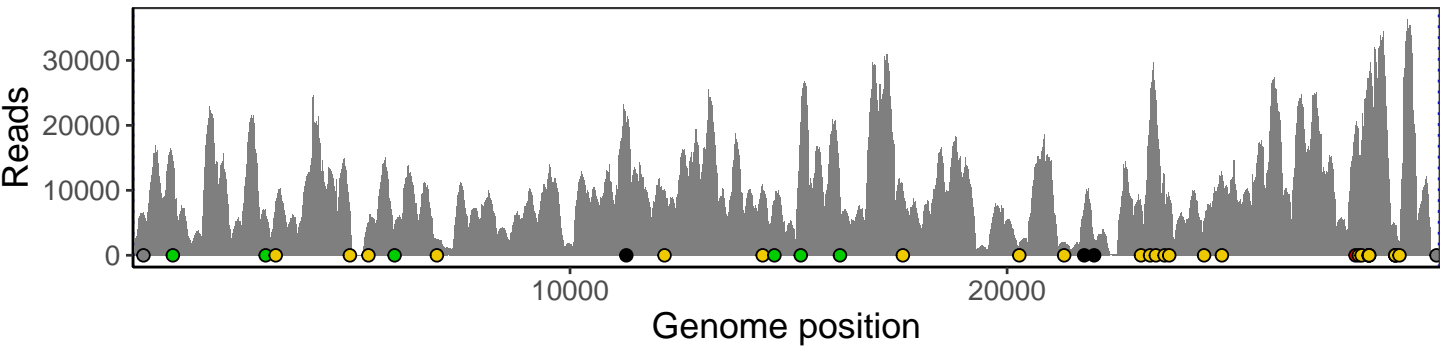
VSP0563-2

VSP0563-3

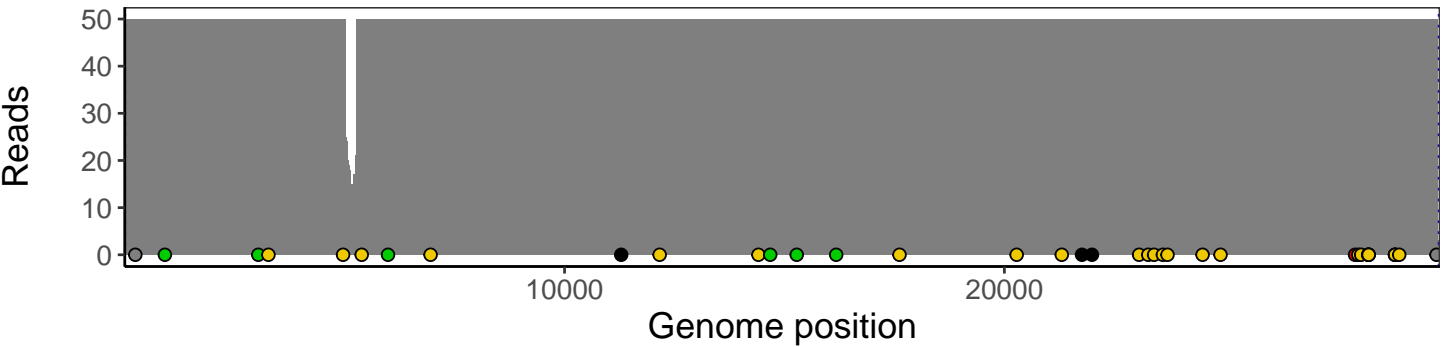
Analyses of individual experiments and composite results

VSP0563 | 2021-01-06 | VTM | DOH1 | composite result

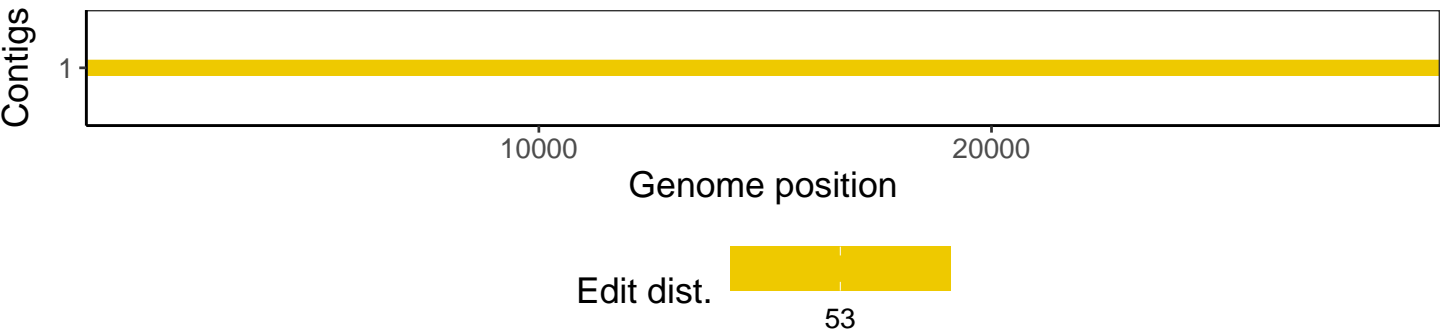
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



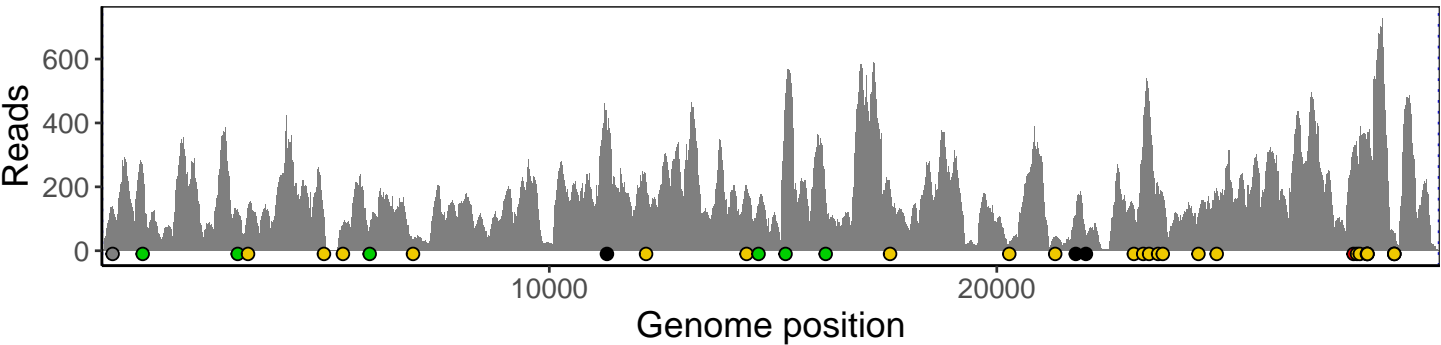
Excerpt from plot above focusing on reads coverage from 0 to 50 NT.



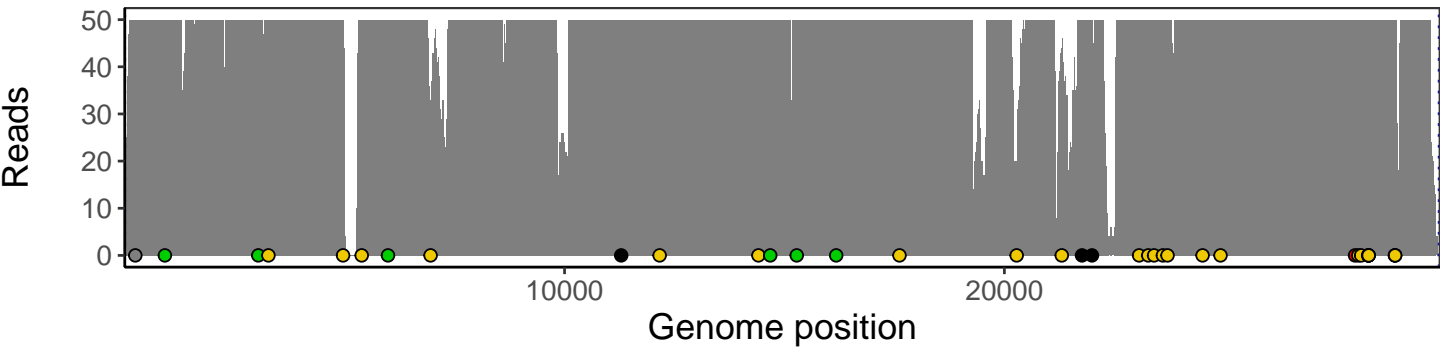
The longest five assembled contigs are shown below colored by their edit distance to the reference genome.



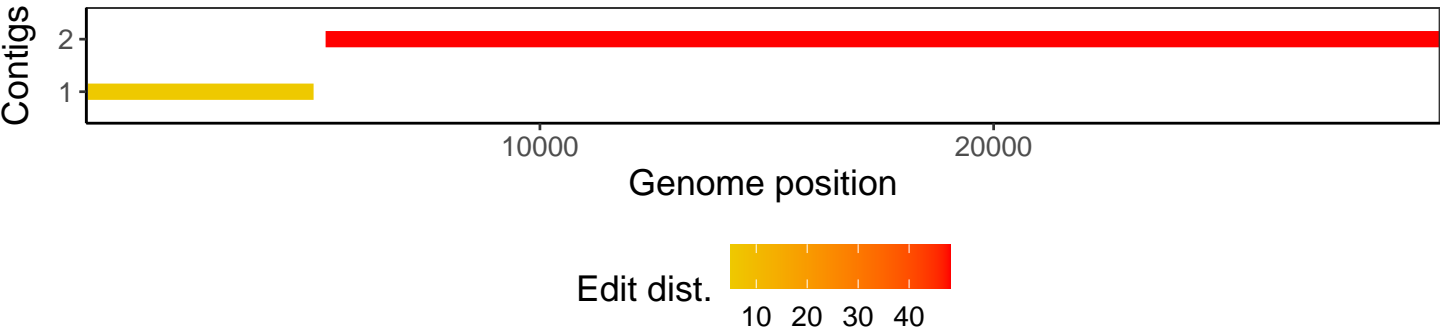
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



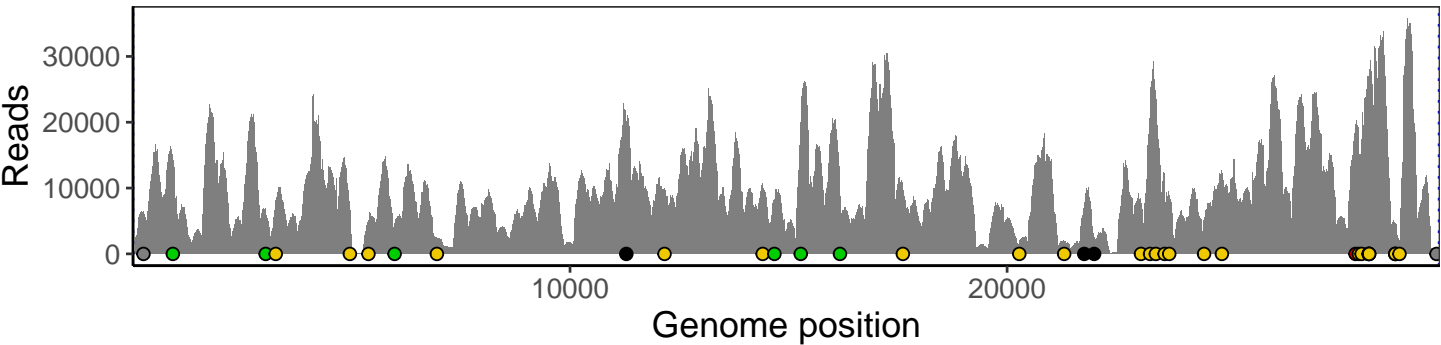
Excerpt from plot above focusing on reads coverage from 0 to 50 NT.



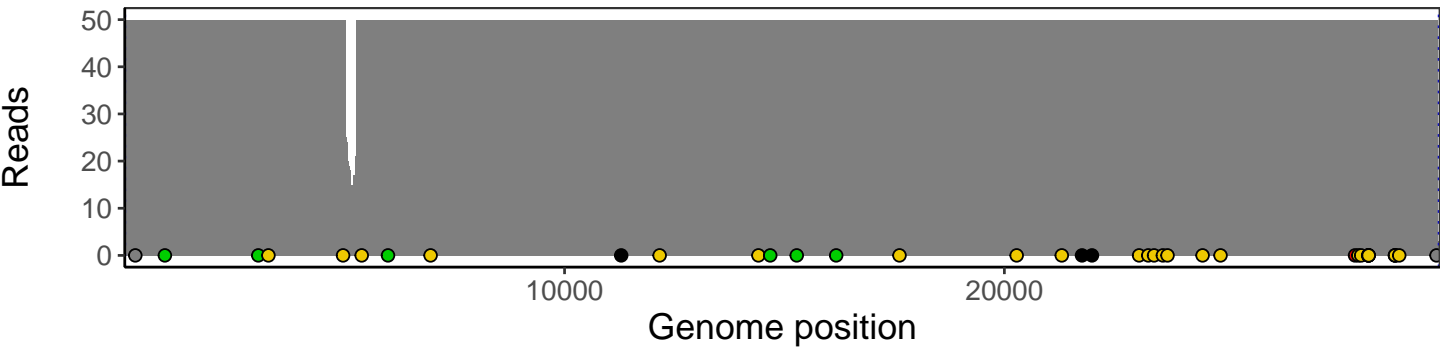
The longest five assembled contigs are shown below colored by their edit distance to the reference genome.



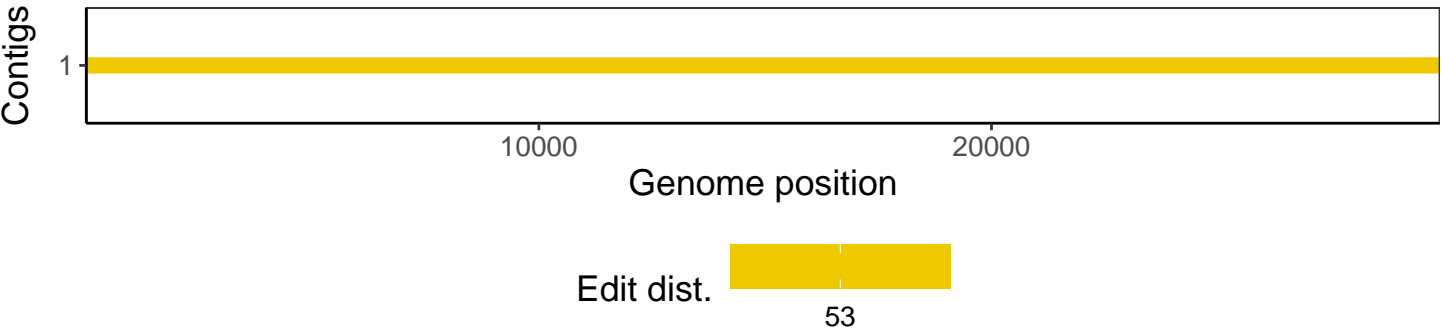
The plot below shows the number of reads covering each nucleotide position in the reference genome. Variants are shown as colored dots along the bottom of the plot and are color coded according by variant types: gray - transgenic, green - silent, gold - missense, red - nonsense, black - indel.



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	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