RNAhybrid fig 2

Nikita Verheyden

2023-04-11

directory

```r
setwd("D:/Krueger_Lab/Publications/miR181_paper/Figure2/RNAhybrid")
```

```r
source("D:/Krueger_Lab/Publications/miR181_paper_v21022023/Figure_theme/theme_paper.R")
library(BSgenome.Mmusculus.UCSC.mm10)
```

```
## Loading required package: BSgenome

## Loading required package: BiocGenerics

##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':
##
##     anyDuplicated, aperm, append, as.data.frame, basename, cbind,
##     colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
##     get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
##     match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##     Position, rank, rbind, Reduce, rownames, sapply, setdiff, sort,
##     table, tapply, union, unique, unsplit, which.max, which.min

## Loading required package: S4Vectors

## Loading required package: stats4

##
## Attaching package: 'S4Vectors'

## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname

## Loading required package: IRanges

##
## Attaching package: 'IRanges'

## The following object is masked from 'package:grDevices':
##
##     windows

## Loading required package: GenomeInfoDb

## Loading required package: GenomicRanges
```

```
## Loading required package: Biostrings

## Loading required package: XVector

##
## Attaching package: 'Biostrings'

## The following object is masked from 'package:base':
##
##     strsplit

## Loading required package: rtracklayer
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:Biostrings':
##
##     collapse, intersect, setdiff, setequal, union

## The following object is masked from 'package:XVector':
##
##     slice

## The following objects are masked from 'package:GenomicRanges':
##
##     intersect, setdiff, union

## The following object is masked from 'package:GenomeInfoDb':
##
##     intersect

## The following objects are masked from 'package:IRanges':
##
##     collapse, desc, intersect, setdiff, slice, union

## The following objects are masked from 'package:S4Vectors':
##
##     first, intersect, rename, setdiff, setequal, union

## The following objects are masked from 'package:BiocGenerics':
##
##     combine, intersect, setdiff, union

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
library(circlize)
```

```
## ========================================
## circlize version 0.4.15
## CRAN page: https://cran.r-project.org/package=circlize
## Github page: https://github.com/jokergoo/circlize
## Documentation: https://jokergoo.github.io/circlize_book/book/
```

```
## 
## If you use it in published research, please cite:
## Gu, Z. circlize implements and enhances circular visualization
##   in R. Bioinformatics 2014.
## 
## This message can be suppressed by:
##   suppressPackageStartupMessages(library(circlize))
## ========================================
library(ComplexHeatmap)
```

```
## Loading required package: grid
```

```
## 
## Attaching package: 'grid'
```

```
## The following object is masked from 'package:Biostrings':
## 
##     pattern
```

```
## ========================================
## ComplexHeatmap version 2.15.2
## Bioconductor page: http://bioconductor.org/packages/ComplexHeatmap/
## Github page: https://github.com/jokergoo/ComplexHeatmap
## Documentation: http://jokergoo.github.io/ComplexHeatmap-reference
## 
## If you use it in published research, please cite either one:
## - Gu, Z. Complex Heatmap Visualization. iMeta 2022.
## - Gu, Z. Complex heatmaps reveal patterns and correlations in multidimensional
##     genomic data. Bioinformatics 2016.
## 
## 
## The new InteractiveComplexHeatmap package can directly export static
## complex heatmaps into an interactive Shiny app with zero effort. Have a try!
## 
## This message can be suppressed by:
##   suppressPackageStartupMessages(library(ComplexHeatmap))
## ========================================
library(seqinr)
```

```
## 
## Attaching package: 'seqinr'
```

```
## The following object is masked from 'package:dplyr':
## 
##     count
```

```
## The following object is masked from 'package:Biostrings':
## 
##     translate
library(GenomicRanges)
library(rGADEM)
```

```
## Loading required package: seqLogo
```

```
## 
## Attaching package: 'seqLogo'
```

```
## The following object is masked from 'package:seqinr':
##
##     consensus

##
## Attaching package: 'rGADEM'

## The following object is masked from 'package:seqLogo':
##
##     consensus

## The following object is masked from 'package:seqinr':
##
##     consensus
```

```r
mir181bs <- makeGRangesFromDataFrame(readRDS("D:/Krueger_Lab/Publications/miR181_paper/Figure1/mir181_b
names(mir181bs) <- 1:length(mir181bs)
```

```r
#colours
farbeneg <- "#b4b4b4"
farbe1 <- "#0073C2FF"
farbe2 <- "#EFC000FF"
farbe3 <- "#CD534CFF"
farbe4 <- "#7AA6DCFF"
farbe5 <- "#868686FF"
farbe6 <- "#003C67FF"
farbe7 <- "#8F7700FF"
farbe8 <- "#3B3B3BFF"
farbe9 <- "#A73030FF"
farbe10 <- "#4A6990FF"
farbe11 <- "#FF6F00FF"
farbe12 <- "#C71000FF"
farbe13 <- "#008EA0FF"
farbe14 <- "#8A4198FF"
farbe15 <- "#5A9599FF"
farbe16 <- "#FF6348FF"


RNApcol <- "#b56504"
RNAncol <- "#027d73"
RPFpcol <- "#c4c404"
RPFncol <- "#8d0391"
```

Get DNA sequences

```r
#resize ranges

mir181bs <- resize(mir181bs, width = width(mir181bs + 50), fix = "start")

# bsGene <- resize(bsGene, width = width(bsGene + 5), fix = "end")

df181 <- mutate(as.data.frame(mir181bs), Sequence = as.character(getSeq(BSgenome.Mmusculus.UCSC.mm10, s

#and turn T into Us
df181$Sequence <- gsub('T', 'U', df181$Sequence)
```

4

```
head(df181)
```

```
##   seqnames   start      end width strand scoreSum  scoreMean scoreMax
## 1     chr1 6245651 6245757   107      +  9.52553   4.762765  6.00678
## 2     chr1 6248341 6248447   107      + 92.68921  23.172303 48.76900
## 3     chr1 6248857 6248963   107      + 14.07133   7.035665  7.04425
## 4     chr1 6248918 6249024   107      + 38.91451  12.971503 20.65080
## 5     chr1 7170481 7170587   107      + 66.92218  13.384436 25.84490
## 6     chr1 9899605 9899711   107      + 25.15963   6.289907  8.61019
##          geneType geneName             geneID region BS_ID         mir_IP
## 1 protein_coding   Rb1cc1 ENSMUSG00000025907    cds     5 mmu-miR-181a-5p
## 2 protein_coding   Rb1cc1 ENSMUSG00000025907    cds     8 mmu-miR-181a-5p
## 3 protein_coding   Rb1cc1 ENSMUSG00000025907    cds    10 mmu-miR-181a-5p
## 4 protein_coding   Rb1cc1 ENSMUSG00000025907    cds    11 mmu-miR-181a-5p
## 5 protein_coding   Pcmtd1 ENSMUSG00000051285   utr3    19 mmu-miR-181a-5p
## 6 protein_coding     Sgk3 ENSMUSG00000025915   utr3    23 mmu-miR-181a-5p
##   n_mir181 n_mir181a n_mir181b n_mir181c n_mir181d               set
## 1        1         1         0         0         0 ago_bs_mir181_chi
## 2        5         5         0         0         0 ago_bs_mir181_chi
## 3        6         6         0         0         0 ago_bs_mir181_chi
## 4        6         6         0         0         0 ago_bs_mir181_chi
## 5        4         4         0         0         0 ago_bs_mir181_chi
## 6        1         1         0         0         0 ago_bs_mir181_chi
##
## 1 UCAUUUUGUGUAUGUAUUUUUUUUUUUUCUAACUAGGACUAAAUUUCUUUUUUUUUUUUUUUAAGGACUAAAUUUCUAUAUUAUAAAACCUUAUUUUUUCA
## 2 UACAACCACUACCUCACCAAAAACUCCUCCUCCACUAACUGUUCAGGACACCUUAUGUCCGGCAGUGUGUCCCUUAGAAGAAUUAUCUCCAGAUAGUAU
## 3 AAAUUAUUUAAAAUGUACAGCUGUAGAAAUAAGAAAUAUUAUUGAAAAAGUAAAAUGUUCUCUAGAAAUAACACUAAAGGAAAAGCAUCAGCAAGAACU
## 4 CUAGAAAUAACACUAAAGGAAAAGCAUCAGCAAGAACUCCAAUCUUUAAAAAAUUGAGUAUGAAUGUAAACUUGAUGCUCUAGUAAAAGACAGUGAAGA
## 5 GAUUUUUAAGAGCUAGUGAAUGUUAAAGUAGGAAGUGGCUACUUGACACAACUAGUUUGUAUGGGCCCUUUGUGUUCUCAGUAUUUAAAGAUAAUGAGG
## 6 GUUUUUUUAAGGUGGUACCCACCAUAAUGUCUCUGUCACAUAUUUAUAUUACAAAUGUUUUCUAUUUAUGUUCCUUUGACUCUUUGAAGGGUUAGUGU
```

Write to .fasta

this is deactivated for now because we only need it once right now just remove the eval if needed

```
candgeneName <- as.list(df181$geneName)
candrname <- as.list(rownames(df181))
condgeneSeq <- as.list(df181$Sequence)


#change to output directory
setwd("D:/Krueger_Lab/Publications/miR181_paper_nongithub/Figure2/RNAhybrid/fastafiles")

for (i in 1:length(candgeneName)) {
  write.fasta(condgeneSeq[i],candrname[i],paste(candrname[i], candgeneName[i], "miR_181", 'fasta', sep =
}
```

Import RNAhybrid results

files that are imported here were run with maximum loop of 3 in both sequences

```
Personalized_Reader <- function(lambda){
 read.table(lambda, sep = ":") %>% select(V1, V5, V6, V7, V9)}
```

```
reslistA <- list.files(path = "D:/Krueger_Lab/Publications/miR181_paper_nongithub/Figure2/RNAhybrid/resu
reslistB <- list.files(path = "D:/Krueger_Lab/Publications/miR181_paper_nongithub/Figure2/RNAhybrid/resu

myfilelistA <- lapply(reslistA, Personalized_Reader)
myfilelistB <- lapply(reslistB, Personalized_Reader)

resframeA <- bind_rows(myfilelistA)
resframeB <- bind_rows(myfilelistB)


colnames(resframeA) <- c("rownumber", "mfs", "pvalue", "start_position", "binding_bases")
colnames(resframeB) <- c("rownumber", "mfs", "pvalue", "start_position", "binding_bases")

head(resframeA)
```

```
##    rownumber    mfs   pvalue start_position
## 1          1 -13.1 1.000000             87
## 2         10 -15.7 0.999882             93
## 3        100 -19.3 0.646155             36
## 4       1000 -21.9 0.197373              4
## 5      10000 -25.4 0.026603             54
## 6      10001 -18.1 0.883059             18
##                             binding_bases
## 1              CUUAU   U   CAG GUU
## 2                     C   UGGCAG
## 3   UCA CUGAUGGU  UGA           AUGUU
## 4              CACC GC    AGUGUUGG   GU
## 5              GCUU CUGAC GC GUUG AAUGU
## 6              ACUUGCUG ACA GC
```

```
head(resframeB)
```

```
##    rownumber    mfs   pvalue start_position
## 1          1 -13.1 1.000000             87
## 2         10 -15.7 0.999882             93
## 3        100 -19.3 0.646155             36
## 4       1000 -21.9 0.197373              4
## 5      10000 -25.4 0.026603             54
## 6      10001 -18.1 0.883059             18
##                             binding_bases
## 1              CUUAU   U   CAG GUU
## 2                     C   UGGCAG
## 3   UCA CUGAUGGU  UGA           AUGUU
## 4              CACC GC    AGUGUUGG   GU
## 5              GCUU CUGAC GC GUUG AAUGU
## 6              ACUUGCUG ACA GC
```

merge with original df

```
# make seperate objects for each mature mirna just to see if they are much different

df181$rownumber <- as.character(rownames(df181))
resframeA$rownumber <- as.character(resframeA$rownumber)
```

```r
resframeB$rownumber <- as.character(resframeB$rownumber)



bsseqHA <- left_join(df181, resframeA, by="rownumber")
bsseqHB <- left_join(df181, resframeB, by="rownumber")

head(bsseqHA)
```

```
##   seqnames   start      end width strand scoreSum scoreMean scoreMax
## 1     chr1 6245651 6245757   107      +  9.52553  4.762765  6.00678
## 2     chr1 6248341 6248447   107      + 92.68921 23.172303 48.76900
## 3     chr1 6248857 6248963   107      + 14.07133  7.035665  7.04425
## 4     chr1 6248918 6249024   107      + 38.91451 12.971503 20.65080
## 5     chr1 7170481 7170587   107      + 66.92218 13.384436 25.84490
## 6     chr1 9899605 9899711   107      + 25.15963  6.289907  8.61019
##         geneType geneName            geneID region BS_ID           mir_IP
## 1 protein_coding   Rb1cc1 ENSMUSG00000025907    cds     5 mmu-miR-181a-5p
## 2 protein_coding   Rb1cc1 ENSMUSG00000025907    cds     8 mmu-miR-181a-5p
## 3 protein_coding   Rb1cc1 ENSMUSG00000025907    cds    10 mmu-miR-181a-5p
## 4 protein_coding   Rb1cc1 ENSMUSG00000025907    cds    11 mmu-miR-181a-5p
## 5 protein_coding   Pcmtd1 ENSMUSG00000051285   utr3    19 mmu-miR-181a-5p
## 6 protein_coding     Sgk3 ENSMUSG00000025915   utr3    23 mmu-miR-181a-5p
##   n_mir181 n_mir181a n_mir181b n_mir181c n_mir181d               set
## 1        1         1         0         0         0 ago_bs_mir181_chi
## 2        5         5         0         0         0 ago_bs_mir181_chi
## 3        6         6         0         0         0 ago_bs_mir181_chi
## 4        6         6         0         0         0 ago_bs_mir181_chi
## 5        4         4         0         0         0 ago_bs_mir181_chi
## 6        1         1         0         0         0 ago_bs_mir181_chi
##
## 1 UCAUUUUGUGUAUGUAUUUUUUUUUUUUCUAACUAGGACUAAAUUUCUUUUUUUUUUUUUUUAAGGACUAAAUUUCUAUAUUAUAAAACCUUAUUUUUUCA
## 2 UACAACCACUACCUCACCAAAAACUCCUCCUCCACUAACUGUUCAGGACACCUUAUGUCCGGCAGUGUGUCCCUUAGAAGAAUUAUCUCCAGAUAGUAU
## 3 AAAUUAUUUAAAAUGUACAGCUGUAGAAAUAAGAAAUAUUUAUUGAAAAAGUAAAAUGUUCUCUAGAAAUAACACUAAAGGAAAAGCAUCAGCAAGAACU
## 4 CUAGAAAUAACACUAAAGGAAAAGCAUCAGCAAGAACUCCAAUCUUUAAAAAAUUGAGUAUGAAUGUAAACUUGAUGCUCUAGUAAAAGACAGUGAAGA
## 5 GAUUUUUAAGAGCUAGUGAAUGUUAAAGUAGGAAGUGGCUACUUGACACAACUAGUUUGUAUGGGCCCUUUGUGUUCUCAGUAUUUAAAGAUAAUGAGG
## 6 GUUUUUUUAAGGUGGUACCCACCAUAAUGUCUCUCUGUCACAUAUUUAUAUUACAAAUGUUUUCUAUUUAUGUUCCUUUGACUCUUUGAAGGGUUAGUGUA
##   rownumber   mfs   pvalue start_position              binding_bases
## 1         1 -13.1 1.000000            87      CUUAU  U  CAG GUU
## 2         2 -23.1 0.101541            51      CUUA   CCGGCAGUGU
## 3         3 -16.9 0.987805             5    AUUUA   UG ACAGC GU GAA
## 4         4 -21.5 0.243804            75 GCUC    A  GACAGUG  GA  AUGU
## 5         5 -17.3 0.968792            10          AGC   UGAAUGUU
## 6         6 -19.8 0.539032            78     ACUC  UGA AG GUU AGUGU
```

```r
head(bsseqHB)
```

```
##   seqnames   start      end width strand scoreSum scoreMean scoreMax
## 1     chr1 6245651 6245757   107      +  9.52553  4.762765  6.00678
## 2     chr1 6248341 6248447   107      + 92.68921 23.172303 48.76900
## 3     chr1 6248857 6248963   107      + 14.07133  7.035665  7.04425
## 4     chr1 6248918 6249024   107      + 38.91451 12.971503 20.65080
## 5     chr1 7170481 7170587   107      + 66.92218 13.384436 25.84490
## 6     chr1 9899605 9899711   107      + 25.15963  6.289907  8.61019
```

```
##            geneType geneName               geneID region BS_ID            mir_IP
## 1 protein_coding   Rb1cc1 ENSMUSG00000025907    cds     5 mmu-miR-181a-5p
## 2 protein_coding   Rb1cc1 ENSMUSG00000025907    cds     8 mmu-miR-181a-5p
## 3 protein_coding   Rb1cc1 ENSMUSG00000025907    cds    10 mmu-miR-181a-5p
## 4 protein_coding   Rb1cc1 ENSMUSG00000025907    cds    11 mmu-miR-181a-5p
## 5 protein_coding   Pcmtd1 ENSMUSG00000051285   utr3    19 mmu-miR-181a-5p
## 6 protein_coding     Sgk3 ENSMUSG00000025915   utr3    23 mmu-miR-181a-5p
##   n_mir181 n_mir181a n_mir181b n_mir181c n_mir181d               set
## 1        1         1         0         0         0 ago_bs_mir181_chi
## 2        5         5         0         0         0 ago_bs_mir181_chi
## 3        6         6         0         0         0 ago_bs_mir181_chi
## 4        6         6         0         0         0 ago_bs_mir181_chi
## 5        4         4         0         0         0 ago_bs_mir181_chi
## 6        1         1         0         0         0 ago_bs_mir181_chi
##
## 1 UCAUUUUGUGUAUGUAUUUUUUUUUUUCUAACUAGGACUAAAUUUCUUUUUUUUUUUUUUUAAGGACUAAAUUUCUAUAUUAUAAAACCUUAUUUUUUCA
## 2 UACAACCACUACCUCACCAAAAACUCCUCCUCCACUAACUGUUCAGGACACCUUAUGUCCGGCAGUGUGUCCCUUAGAAGAAUUAUCUCCAGAUAGUAU
## 3 AAAUUAUUUAAAAUGUACAGCUGUAGAAAUAAGAAAUAUUAUUGAAAAAGUAAAAUGUUCUCUAGAAAUAACACUAAAAGGAAAAGCAUCAGCAAGAACU
## 4 CUAGAAAUAACACUAAAAGGAAAAGCAUCAGCAAGAACUCCAAUCUUUAAAAAAUUGAGUAUGAAUGUAAACUUGAUGCUCUAGUAAAAGACAGUGAAGA
## 5 GAUUUUUUAAGAGCUAGUGAAUGUUAAAGUAGGAAGUGGCUACUUGACACAACUAGUUUGUAUGGGCCCUUUGUGUUCUCAGUAUUUAAAGAUAAUGAG
## 6 GUUUUUUUAAGGUGGUACCCACCAUAAUGUCUCUGUCACAUAUUUAUAUUACAAAUGUUUUCUAUUUAUGUUCCUUUGACUCUUUGAAGGGUUAGUGU
##   rownumber   mfs   pvalue start_position               binding_bases
## 1         1 -13.1 1.000000            87       CUUAU  U  CAG GUU
## 2         2 -23.1 0.101541            51       CUUA   CCGGCAGUGU
## 3         3 -16.9 0.987805             5     AUUUA   UG ACAGC GU GAA
## 4         4 -21.5 0.243804            75 GCUC    A  GACAGUG  GA  AUGU
## 5         5 -17.3 0.968792            10              AGC   UGAAUGUU
## 6         6 -19.8 0.539032            78       ACUC  UGA AG GUU AGUGU
```

adjust binding site info

```r
bsseqHA$binding_bases <- chartr(" ", "N", bsseqHA$binding_bases)
bsseqHB$binding_bases <- chartr(" ", "N", bsseqHB$binding_bases)

head(bsseqHA)
```

```
##   seqnames   start     end width strand scoreSum scoreMean scoreMax
## 1     chr1 6245651 6245757   107      +  9.52553  4.762765  6.00678
## 2     chr1 6248341 6248447   107      + 92.68921 23.172303 48.76900
## 3     chr1 6248857 6248963   107      + 14.07133  7.035665  7.04425
## 4     chr1 6248918 6249024   107      + 38.91451 12.971503 20.65080
## 5     chr1 7170481 7170587   107      + 66.92218 13.384436 25.84490
## 6     chr1 9899605 9899711   107      + 25.15963  6.289907  8.61019
##          geneType geneName               geneID region BS_ID            mir_IP
## 1 protein_coding   Rb1cc1 ENSMUSG00000025907    cds     5 mmu-miR-181a-5p
## 2 protein_coding   Rb1cc1 ENSMUSG00000025907    cds     8 mmu-miR-181a-5p
## 3 protein_coding   Rb1cc1 ENSMUSG00000025907    cds    10 mmu-miR-181a-5p
## 4 protein_coding   Rb1cc1 ENSMUSG00000025907    cds    11 mmu-miR-181a-5p
## 5 protein_coding   Pcmtd1 ENSMUSG00000051285   utr3    19 mmu-miR-181a-5p
## 6 protein_coding     Sgk3 ENSMUSG00000025915   utr3    23 mmu-miR-181a-5p
##   n_mir181 n_mir181a n_mir181b n_mir181c n_mir181d               set
## 1        1         1         0         0         0 ago_bs_mir181_chi
## 2        5         5         0         0         0 ago_bs_mir181_chi
## 3        6         6         0         0         0 ago_bs_mir181_chi
```

```
## 4         6       6       0       0         0 ago_bs_mir181_chi
## 5         4       4       0       0         0 ago_bs_mir181_chi
## 6         1       1       0       0         0 ago_bs_mir181_chi
##
## 1 UCAUUUUGUGUAUGUAUUUUUUUUUUUUCUAACUAGGACUAAAUUUCUUUUUUUUUUUUUUUAAGGACUAAAUUUCUAUAUUAUAAAACCUUAUUUUUUCA
## 2 UACAACCACUACCUCACCAAAAACUCCUCCUCCACUAACUGUUCAGGACACCUUAUGUCCGGCAGUGUGUCCCUUAGAAGAAUUAUCUCCAGAUAGUAU
## 3 AAAUUAUUUAAAAAUGUACAGCUGUAGAAAUAAGAAAUAUUAUUGAAAAAGUAAAAUGUUCUCUAGAAAUAACACUAAAGGAAAAGCAUCAGCAAGAAC
## 4 CUAGAAAUAACACUAAAGGAAAAGCAUCAGCAAGAACUCCAAUCUUUAAAAAUUGAGUAUGAAUGUAAACUUGAUGCUCUAGUAAAAGACAGUGAAGA
## 5 GAUUUUUUAAGAGCUAGUGAAUGUUAAAGUAGGAAGUGGCUACUUGACACAACUAGUUUGUAUGGGCCCUUUGUGUUCUCAGUAUUUAAAGAUAAUGAG
## 6 GUUUUUUUUAAGGUGGUACCCACCAUAAUGUCUCUGUCACAUAUUUAUAUUACAAAUGUUUUCUAUUUAUGUUCCUUUGACUCUUUGAAGGGUUAGUGUA
##   rownumber   mfs   pvalue start_position              binding_bases
## 1         1 -13.1 1.000000            87    NCUUAUNNUNNCAGNGUUNNNNNNN
## 2         2 -23.1 0.101541            51    NCUUANNNCCGGCAGUGUNNNNNNN
## 3         3 -16.9 0.987805             5   NAUUUANNNUGNACAGCNGUNGAANNNN
## 4         4 -21.5 0.243804            75 NGCUCNNNNNANNGACAGUGNNGANNAUGUN
## 5         5 -17.3 0.968792            10    NNNNNNNNNNAGCNNNUGAAUGUUN
## 6         6 -19.8 0.539032            78    NACUCNNUGANAGNGUUNAGUGUN
```

head(bsseqHB)

```
##   seqnames   start      end width strand  scoreSum scoreMean  scoreMax
## 1     chr1 6245651 6245757   107      +   9.52553  4.762765   6.00678
## 2     chr1 6248341 6248447   107      +  92.68921 23.172303  48.76900
## 3     chr1 6248857 6248963   107      +  14.07133  7.035665   7.04425
## 4     chr1 6248918 6249024   107      +  38.91451 12.971503  20.65080
## 5     chr1 7170481 7170587   107      +  66.92218 13.384436  25.84490
## 6     chr1 9899605 9899711   107      +  25.15963  6.289907   8.61019
##         geneType geneName             geneID region BS_ID          mir_IP
## 1 protein_coding   Rb1cc1 ENSMUSG00000025907    cds     5 mmu-miR-181a-5p
## 2 protein_coding   Rb1cc1 ENSMUSG00000025907    cds     8 mmu-miR-181a-5p
## 3 protein_coding   Rb1cc1 ENSMUSG00000025907    cds    10 mmu-miR-181a-5p
## 4 protein_coding   Rb1cc1 ENSMUSG00000025907    cds    11 mmu-miR-181a-5p
## 5 protein_coding   Pcmtd1 ENSMUSG00000051285   utr3    19 mmu-miR-181a-5p
## 6 protein_coding    Sgk3 ENSMUSG00000025915   utr3    23 mmu-miR-181a-5p
##   n_mir181 n_mir181a n_mir181b n_mir181c n_mir181d               set
## 1        1         1         0         0         0 ago_bs_mir181_chi
## 2        5         5         0         0         0 ago_bs_mir181_chi
## 3        6         6         0         0         0 ago_bs_mir181_chi
## 4        6         6         0         0         0 ago_bs_mir181_chi
## 5        4         4         0         0         0 ago_bs_mir181_chi
## 6        1         1         0         0         0 ago_bs_mir181_chi
##
## 1 UCAUUUUGUGUAUGUAUUUUUUUUUUUUCUAACUAGGACUAAAUUUCUUUUUUUUUUUUUUUAAGGACUAAAUUUCUAUAUUAUAAAACCUUAUUUUUUCA
## 2 UACAACCACUACCUCACCAAAAACUCCUCCUCCACUAACUGUUCAGGACACCUUAUGUCCGGCAGUGUGUCCCUUAGAAGAAUUAUCUCCAGAUAGUAU
## 3 AAAUUAUUUAAAAAUGUACAGCUGUAGAAAUAAGAAAUAUUAUUGAAAAAGUAAAAUGUUCUCUAGAAAUAACACUAAAGGAAAAGCAUCAGCAAGAAC
## 4 CUAGAAAUAACACUAAAGGAAAAGCAUCAGCAAGAACUCCAAUCUUUAAAAAUUGAGUAUGAAUGUAAACUUGAUGCUCUAGUAAAAGACAGUGAAGA
## 5 GAUUUUUUAAGAGCUAGUGAAUGUUAAAGUAGGAAGUGGCUACUUGACACAACUAGUUUGUAUGGGCCCUUUGUGUUCUCAGUAUUUAAAGAUAAUGAG
## 6 GUUUUUUUUAAGGUGGUACCCACCAUAAUGUCUCUGUCACAUAUUUAUAUUACAAAUGUUUUCUAUUUAUGUUCCUUUGACUCUUUGAAGGGUUAGUGUA
##   rownumber   mfs   pvalue start_position              binding_bases
## 1         1 -13.1 1.000000            87    NCUUAUNNUNNCAGNGUUNNNNNNN
## 2         2 -23.1 0.101541            51    NCUUANNNCCGGCAGUGUNNNNNNN
## 3         3 -16.9 0.987805             5   NAUUUANNNUGNACAGCNGUNGAANNNN
## 4         4 -21.5 0.243804            75 NGCUCNNNNNANNGACAGUGNNGANNAUGUN
## 5         5 -17.3 0.968792            10    NNNNNNNNNNAGCNNNUGAAUGUUN
## 6         6 -19.8 0.539032            78    NACUCNNUGANAGNGUUNAGUGUN
```
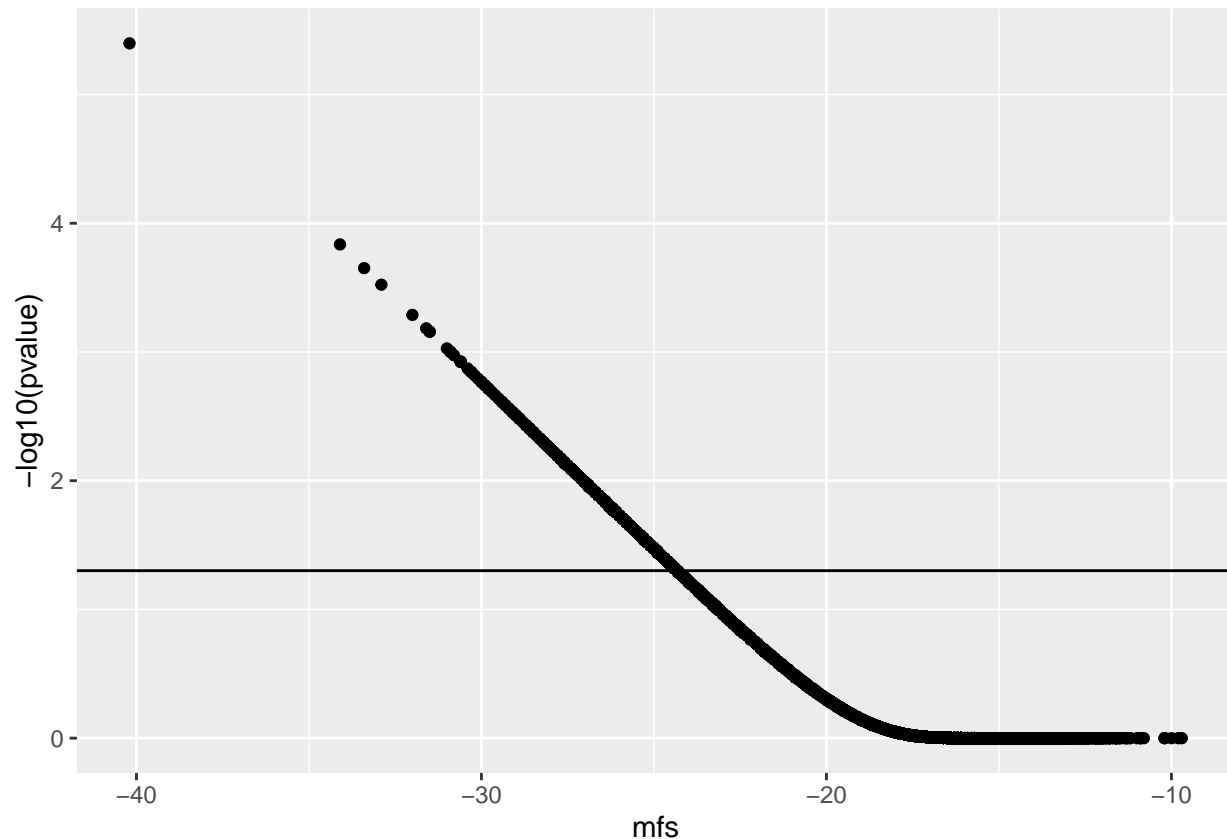
9

Plot pValues and mfs

```
pmfsplotA <- ggplot(bsseqHA, aes(y=-log10(pvalue), x=mfs)) + geom_point() +
  geom_hline(yintercept=-log10(0.05))
```

```
pmfsplotA
```
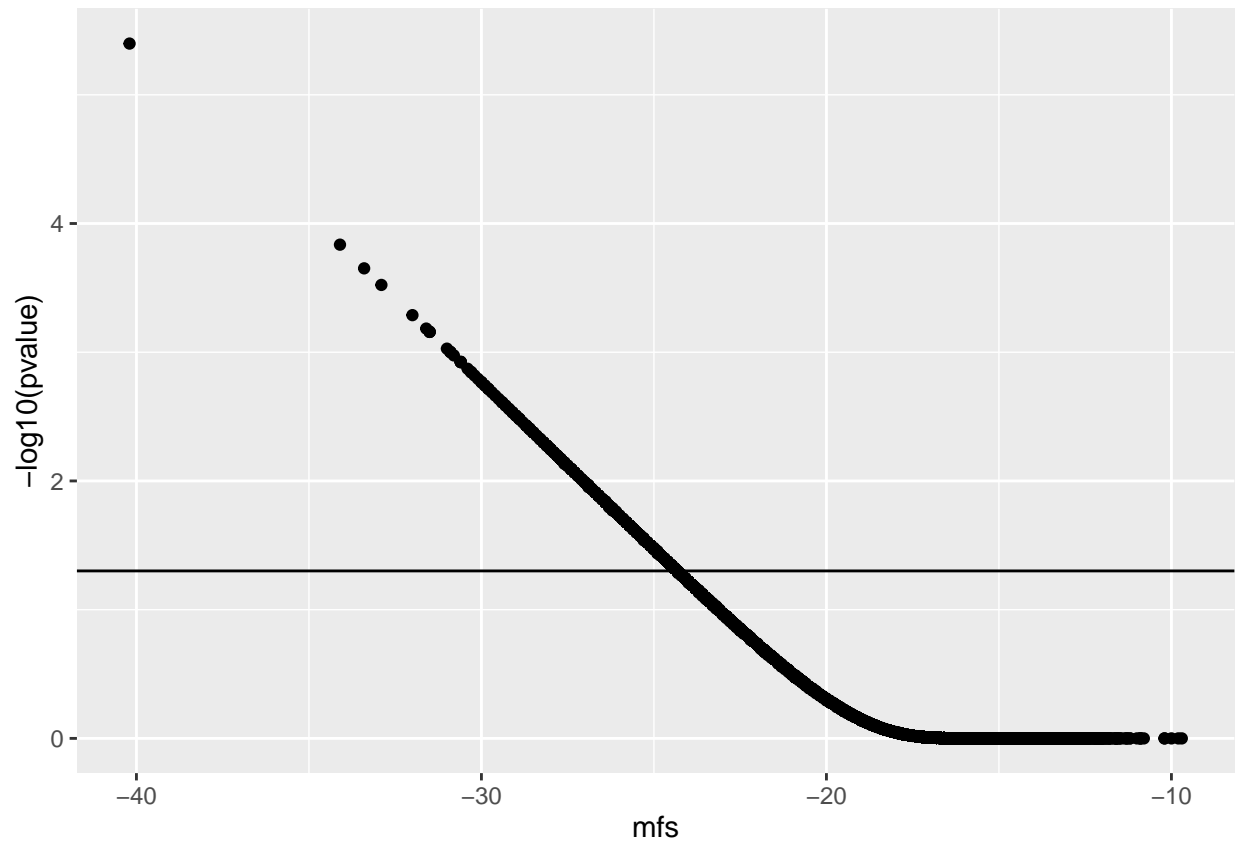


```
pmfsplotB <- ggplot(bsseqHA, aes(y=-log10(pvalue), x=mfs)) + geom_point() +
  geom_hline(yintercept=-log10(0.05))
```

```
pmfsplotB
```

find motivs

make DNAStringSet

```
#make a DNA sting set as well because the motif finder requires it
b_baseSetDNAA <- DNAStringSet(chartr("U", "T", bsseqHA[bsseqHA$pvalue <= 0.05, "binding_bases"]))
b_baseSetDNAB <- DNAStringSet(chartr("U", "T", bsseqHB[bsseqHB$pvalue <= 0.05, "binding_bases"]))

head(b_baseSetDNAA)
```

```
## DNAStringSet object of length 6:
##     width seq
## [1]    31 NCTCNCCGACAGNGTNNNTGANNNNNATGTN
## [2]    26 NGCNNACCGACGGCNNGTTNNNTGTN
## [3]    25 NNTTACCNNNACAGCNNTGAATGTN
## [4]    24 NNTCANCCGACAGTNNTGAATGNN
## [5]    40 NGCTCACTNNNNNNNNNNNNNNNGGCAGCNGTNGANNATGNN
## [6]    40 NGCTCACTNNNNNNNNNNNNNNNGGCAGCNGTNGANNATGNN
```

```
head(b_baseSetDNAB)
```

```
## DNAStringSet object of length 6:
##     width seq
## [1]    31 NCTCNCCGACAGNGTNNNTGANNNNNATGTN
## [2]    26 NGCNNACCGACGGCNNGTTNNNTGTN
## [3]    25 NNTTACCNNNACAGCNNTGAATGTN
```

```
## [4]     24 NNTCANCCGACAGTNNTGAATGNN
## [5]     40 NGCTCACTNNNNNNNNNNNNNNGGCAGCNGTNGANNATGNN
## [6]     40 NGCTCACTNNNNNNNNNNNNNNGGCAGCNGTNGANNATGNN
```

Run rGadem

```
enrichmotA <- GADEM(b_baseSetDNAA, seed = 123)
```
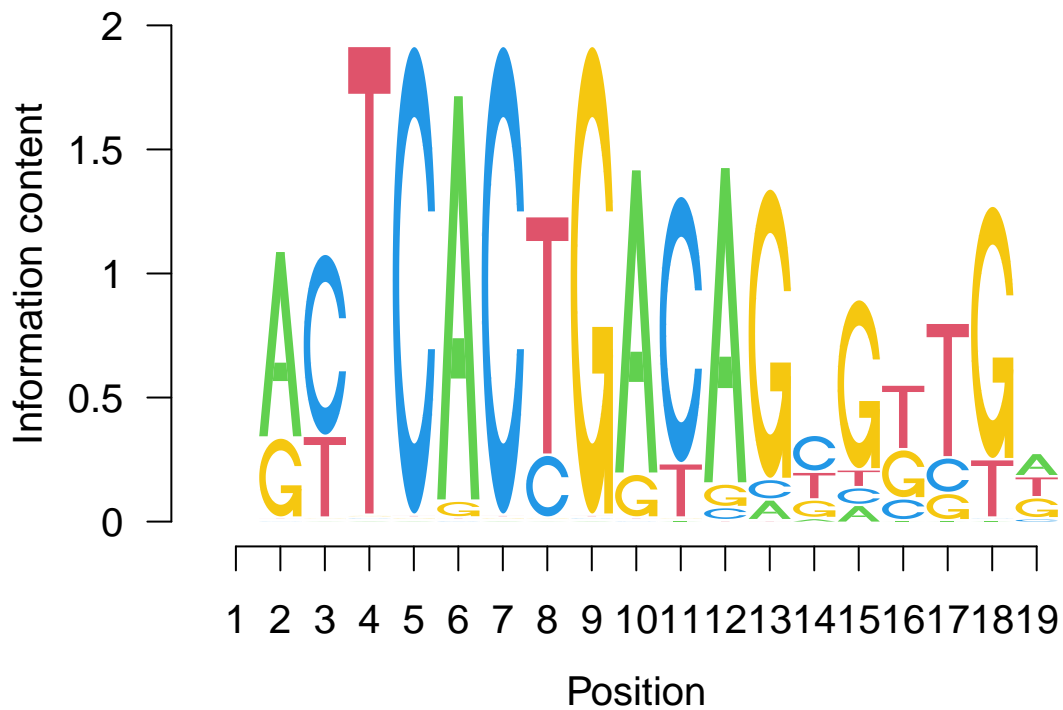
```
## top 3  4, 5-mers: 20 40 60
## top 3  4, 5-mers: 20 40 60
```
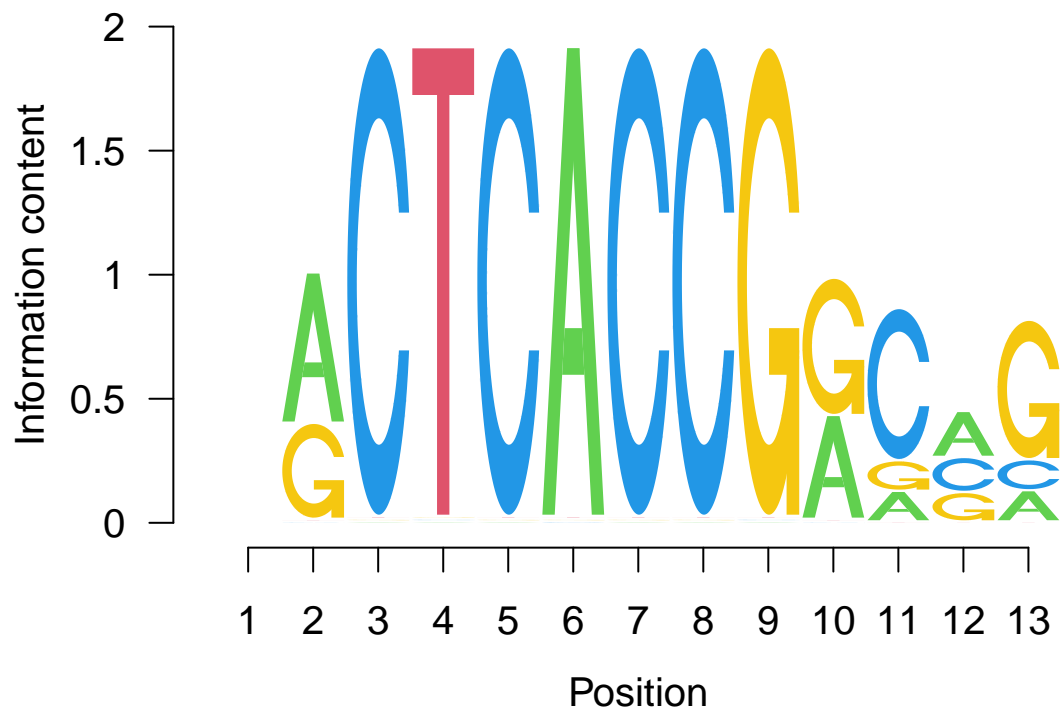
```
enrichmotB <- GADEM(b_baseSetDNAB, seed = 123)
```

```
## top 3  4, 5-mers: 20 40 60
## top 3  4, 5-mers: 20 40 60
```
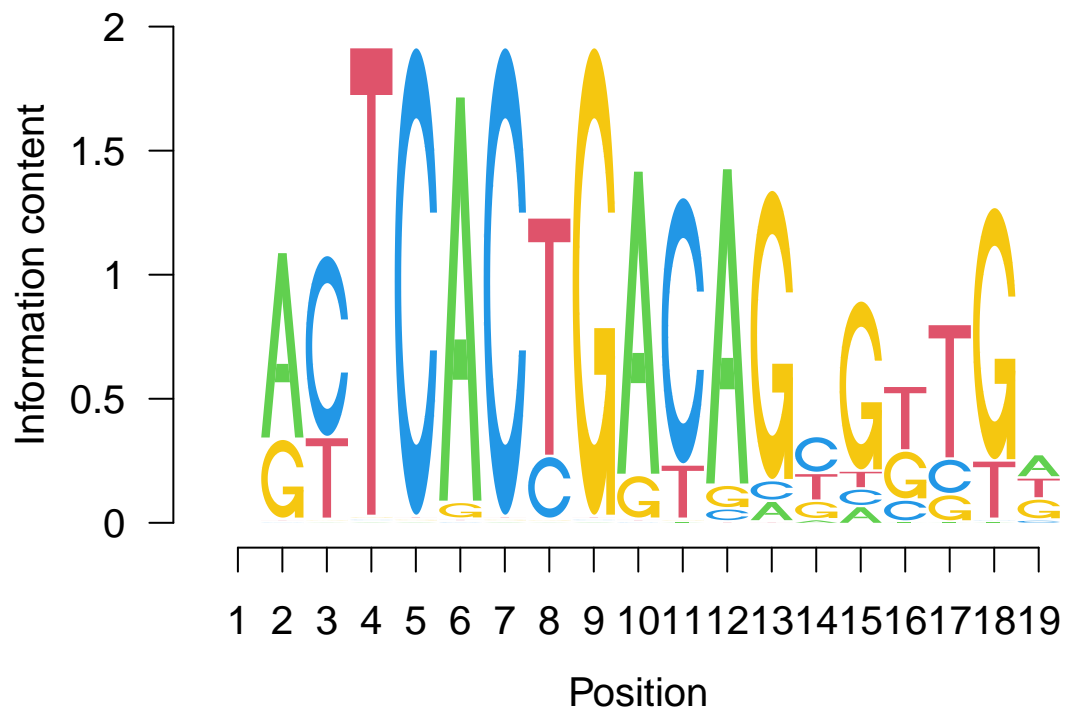
miR181a

```
rGADEM::plot(enrichmotA)
```
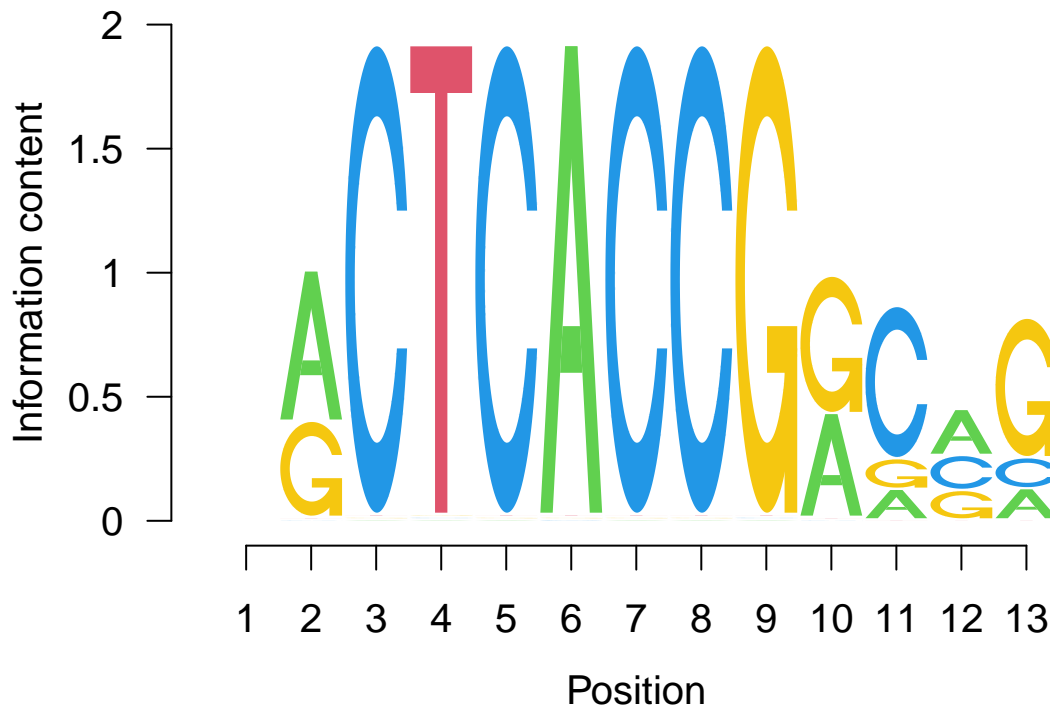
miR181b

```
rGADEM::plot(enrichmotB)
```

Session info

```
sessionInfo()
```

```
## R version 4.2.3 (2023-03-15 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19044)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=German_Germany.utf8  LC_CTYPE=German_Germany.utf8
## [3] LC_MONETARY=German_Germany.utf8 LC_NUMERIC=C
## [5] LC_TIME=German_Germany.utf8
##
## attached base packages:
## [1] grid      stats4    stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] rGADEM_2.46.0                     seqLogo_1.64.0
##  [3] seqinr_4.2-30                     ComplexHeatmap_2.15.2
##  [5] circlize_0.4.15                   ggplot2_3.4.2
##  [7] dplyr_1.1.1                       BSgenome.Mmusculus.UCSC.mm10_1.4.3
##  [9] BSgenome_1.66.3                   rtracklayer_1.58.0
```

```
## [11] Biostrings_2.66.0              XVector_0.38.0
## [13] GenomicRanges_1.50.2          GenomeInfoDb_1.34.9
## [15] IRanges_2.32.0                S4Vectors_0.36.2
## [17] BiocGenerics_0.44.0
##
## loaded via a namespace (and not attached):
##  [1] MatrixGenerics_1.10.0    Biobase_2.58.0
##  [3] foreach_1.5.2            highr_0.10
##  [5] GenomeInfoDbData_1.2.9   Rsamtools_2.14.0
##  [7] yaml_2.3.7               pillar_1.9.0
##  [9] lattice_0.20-45          glue_1.6.2
## [11] digest_0.6.31           RColorBrewer_1.1-3
## [13] colorspace_2.1-0        htmltools_0.5.4
## [15] Matrix_1.5-3            XML_3.99-0.14
## [17] pkgconfig_2.0.3         GetoptLong_1.0.5
## [19] zlibbioc_1.44.0         scales_1.2.1
## [21] BiocParallel_1.32.6     tibble_3.2.1
## [23] farver_2.1.1            generics_0.1.3
## [25] withr_2.5.0             SummarizedExperiment_1.28.0
## [27] cli_3.6.0               magrittr_2.0.3
## [29] crayon_1.5.2            evaluate_0.20
## [31] fansi_1.0.4             doParallel_1.0.17
## [33] MASS_7.3-58.2           tools_4.2.3
## [35] GlobalOptions_0.1.2     BiocIO_1.8.0
## [37] lifecycle_1.0.3         matrixStats_0.63.0
## [39] munsell_0.5.0           cluster_2.1.4
## [41] DelayedArray_0.23.2     ade4_1.7-22
## [43] compiler_4.2.3          rlang_1.1.0
## [45] RCurl_1.98-1.12         iterators_1.0.14
## [47] rstudioapi_0.14         rjson_0.2.21
## [49] bitops_1.0-7            labeling_0.4.2
## [51] rmarkdown_2.21          restfulr_0.0.15
## [53] gtable_0.3.3            codetools_0.2-19
## [55] R6_2.5.1                GenomicAlignments_1.34.1
## [57] knitr_1.42              fastmap_1.1.1
## [59] utf8_1.2.3              clue_0.3-64
## [61] shape_1.4.6             parallel_4.2.3
## [63] Rcpp_1.0.10             vctrs_0.6.1
## [65] png_0.1-8               tidyselect_1.2.0
## [67] xfun_0.37
```