basic ChIPseq analysis

TomanB

14 5 2020

load packages

```
library(BiocManager)
```

```
## Bioconductor version 3.10 (BiocManager 1.30.10), ?BiocManager::install for help
```

```
## Bioconductor version '3.10' is out-of-date; the current release version '3.11'
##   is available with R version '4.0'; see https://bioconductor.org/install
```

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.3.0     v purrr   0.3.3
## v tibble  3.0.0     v dplyr   0.8.5
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(Rsamtools)
```

```
## Loading required package: GenomeInfoDb
```

```
## Loading required package: BiocGenerics
```

```
## Loading required package: parallel
```

```
##
## Attaching package: 'BiocGenerics'
```

```
## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB
```

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

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

```
## The following objects are masked from 'package:base':
##
```

```
##     anyDuplicated, 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, which.max, which.min

## Loading required package: S4Vectors

## Loading required package: stats4

##
## Attaching package: 'S4Vectors'

## The following objects are masked from 'package:dplyr':
##
##     first, rename

## The following object is masked from 'package:tidyr':
##
##     expand

## The following object is masked from 'package:base':
##
##     expand.grid

## Loading required package: IRanges

##
## Attaching package: 'IRanges'

## The following objects are masked from 'package:dplyr':
##
##     collapse, desc, slice

## The following object is masked from 'package:purrr':
##
##     reduce

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

## Loading required package: GenomicRanges

## Loading required package: Biostrings

## Loading required package: XVector

##
## Attaching package: 'XVector'

## The following object is masked from 'package:purrr':
##
##     compact

##
## Attaching package: 'Biostrings'

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

```
library(GenomicRanges)
library(rtracklayer)
library(GenomicFeatures)
```

```
## Loading required package: AnnotationDbi
```

```
## Loading required package: Biobase
```

```
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```
##
## Attaching package: 'AnnotationDbi'
```

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

```
library(ggraph)
library(clusterProfiler)
```

```
##
```

```
## Registered S3 method overwritten by 'enrichplot':
##   method               from
##   fortify.enrichResult DOSE
```

```
## clusterProfiler v3.14.3  For help: https://guangchuangyu.github.io/software/clusterProfiler
##
## If you use clusterProfiler in published research, please cite:
## Guangchuang Yu, Li-Gen Wang, Yanyan Han, Qing-Yu He. clusterProfiler: an R package for comparing biol
```

```
##
## Attaching package: 'clusterProfiler'
```

```
## The following object is masked from 'package:purrr':
##
##     simplify
```

```
library(ChIPpeakAnno)
```

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

```
## Loading required package: VennDiagram
```

```
## Loading required package: futile.logger
```

```
library(org.Hs.eg.db)
```

```
##
```

```
library(TxDb.Hsapiens.UCSC.hg38.knownGene)
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
```

```
##     smiths
```

```
library(BSgenome.Hsapiens.UCSC.hg38)
```

```
## Loading required package: BSgenome
```

```
library(TFBSTools)
```

```
## Warning: namespace 'VGAM' is not available and has been replaced
## by .GlobalEnv when processing object ''

## Warning: namespace 'VGAM' is not available and has been replaced
## by .GlobalEnv when processing object ''

## Warning: namespace 'VGAM' is not available and has been replaced
## by .GlobalEnv when processing object ''

## Warning: namespace 'VGAM' is not available and has been replaced
## by .GlobalEnv when processing object ''

## Warning: namespace 'VGAM' is not available and has been replaced
## by .GlobalEnv when processing object ''

## No methods found in package 'IRanges' for request: 'score' when loading 'TFBSTools'

## Warning: namespace 'VGAM' is not available and has been replaced
## by .GlobalEnv when processing object ''

## Warning: namespace 'VGAM' is not available and has been replaced
## by .GlobalEnv when processing object ''
```

```
library(Biostrings)
library(ChIPseeker)
```

```
## ChIPseeker v1.22.1  For help: https://guangchuangyu.github.io/software/ChIPseeker
##
## If you use ChIPseeker in published research, please cite:
## Guangchuang Yu, Li-Gen Wang, Qing-Yu He. ChIPseeker: an R/Bioconductor package for ChIP peak annotati
```

```
library(biomaRt)
library(phylotools)
```

```
## Loading required package: ape

##
## Attaching package: 'ape'

## The following object is masked from 'package:VennDiagram':
##
##     rotate

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

```
library(msigdbr)
library(motifRG)
```

```
## Loading required package: seqLogo

##
```

```
## Attaching package: 'seqLogo'

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

## Loading required package: BSgenome.Hsapiens.UCSC.hg19

##
## Attaching package: 'BSgenome.Hsapiens.UCSC.hg19'

## The following object is masked from 'package:BSgenome.Hsapiens.UCSC.hg38':
##
##     Hsapiens

##
## Attaching package: 'motifRG'

## The following object is masked from 'package:biomaRt':
##
##     getSequence
```

```
library(motifStack)
```

```
## Loading required package: grImport2

## Loading required package: MotIV

##
## Attaching package: 'MotIV'

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

## The following object is masked from 'package:dplyr':
##
##     filter

## The following object is masked from 'package:stats':
##
##     filter

## Loading required package: ade4

##
## Attaching package: 'ade4'

## The following object is masked from 'package:BSgenome':
##
##     score

## The following object is masked from 'package:rtracklayer':
##
##     score

## The following object is masked from 'package:Biostrings':
##
##     score

## The following object is masked from 'package:GenomicRanges':
##
##     score
```

```
## The following object is masked from 'package:BiocGenerics':
##
##     score
```

```
library(JASPAR2018)


### set working directory

setwd("D:/01_Private Dateien/Biochemie Studium/Masterstudium/01 - Molecular Biosciences - Major Cancer

# load txdb gene model

txdb <- TxDb.Hsapiens.UCSC.hg38.knownGene

# load files in narrowpeak = BED format (can be downloaded from ENCODE project database)
extraCols_narrowPeak <- c(signalValue = "numeric", pValue = "numeric", qValue = "numeric", peak = "integ

# Snyder Stanford convervative IDR peaks example file: ENCFF838NIW.bed <- ChIPseq data of MYC transcrip
ChIP_bed <- import("D:/01_Private Dateien/Biochemie Studium/Masterstudium/01 - Molecular Biosciences -

### annotate the peaks with precompiled ensembl annotation
data(TSS.human.GRCh38)
ucsc.hg38.knownGene <- genes(TxDb.Hsapiens.UCSC.hg38.knownGene)

# use biomaRt for annotation with ensmbl
ensembl = useMart("ensembl", dataset="hsapiens_gene_ensembl")

# annotate GRanges
ChIP.anno <- annotatePeakInBatch(ChIP_bed, AnnotationData = ucsc.hg38.knownGene)
ChIP.anno <- addGeneIDs(annotatedPeak = ChIP.anno, orgAnn = "org.Hs.eg.db", feature_id_type = "entrez_id

## calculate percentage of peaks in promoters etc.
ChIP_aCR <- assignChromosomeRegion(ChIP.anno, nucleotideLevel=FALSE, precedence=c("Promoters", "immediat
                                   "fiveUTRs", "threeUTRs", "Exons", "Introns"), TxDb=TxDb.Hsapiens.UCSC
```

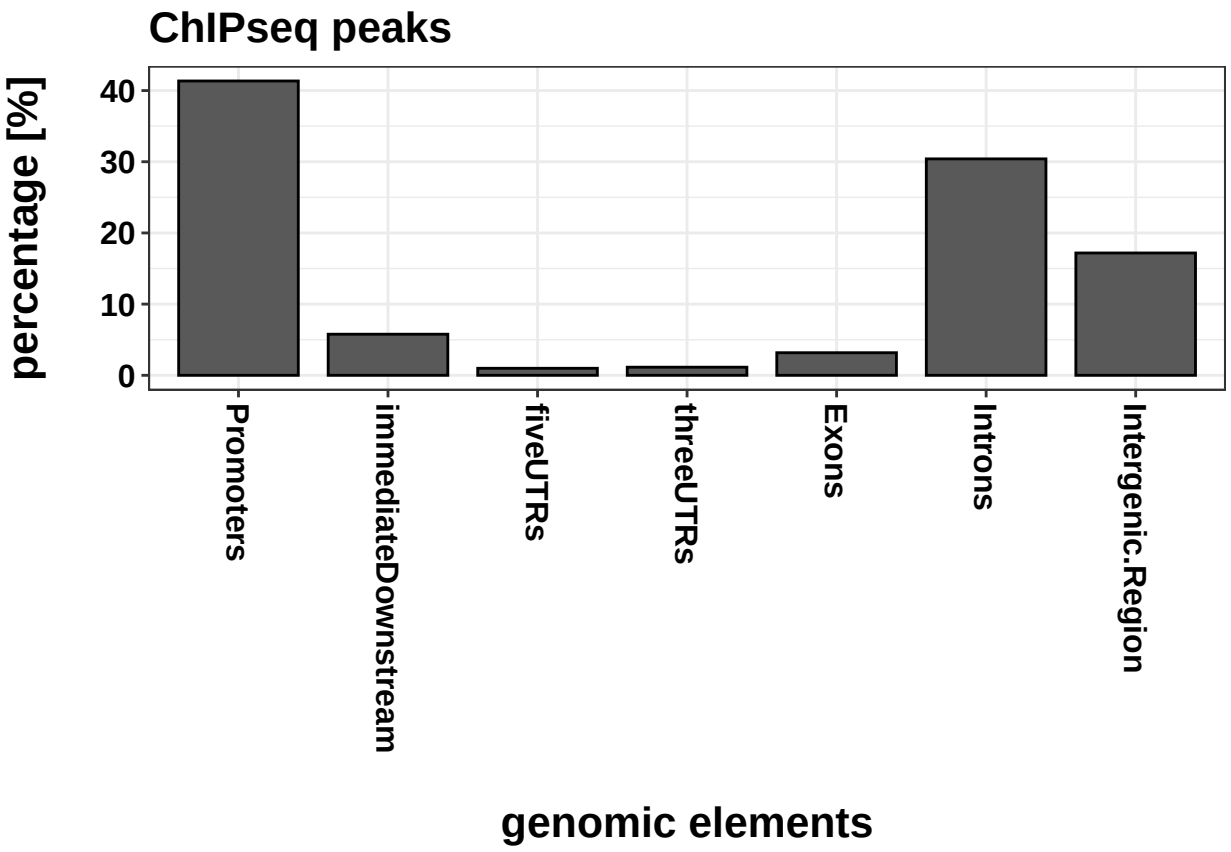
```
## Warning in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE): GRanges object contains 153 out-of-bo
##   chr1_GL383518v1_alt, chr1_KQ458384v1_alt, chr2_GL383522v1_alt,
##   chr4_GL000257v2_alt, chr5_GL339449v2_alt, chr5_KI270795v1_alt,
##   chr5_KI270898v1_alt, chr5_KV575244v1_fix, chr6_KI270797v1_alt,
##   chr6_KI270798v1_alt, chr6_KI270801v1_alt, chr7_GL383534v2_alt,
##   chr7_KI270803v1_alt, chr7_KI270806v1_alt, chr7_KI270809v1_alt,
##   chr7_KZ208912v1_fix, chr9_GL383540v1_alt, chr9_GL383541v1_alt,
##   chr11_KI270902v1_alt, chr12_GL383551v1_alt, chr12_GL383553v2_alt,
##   chr12_KI270834v1_alt, chr14_KI270847v1_alt, chr15_KI270848v1_alt,
##   chr15_KI270850v1_alt, chr15_KI270851v1_alt, chr15_KI270906v1_alt,
##   chr16_GL383556v1_alt, chr16_KI270854v1_alt, chr17_JH159146v1_alt,
##   chr17_JH159147v1_alt, chr17_KI270857v1_alt, chr17_KI270860v1_alt,
##   chr17_KV575245v1_fix, chr17_KV766196v1_fix, chr17_KV766198v1_alt,
##   chr19_GL383575v2_alt, chr19_GL383576v1_alt, chr19_KI270866v1_alt,
##   chr19_KI270884v1_alt, chr19_KI270885v1_alt, chr19_KI270889v1_alt,
##   chr19_KI270890v1_alt, chr19_KI270891v1_alt, chr19_KI270915v1_alt,
##   chr19_KI270916v1_alt, chr19_KI270919v1_alt, chr19_KI270922v1_alt,
##   chr19_KI270923v1_alt, chr19_KI270929v1_alt, chr19_KI270930v1_alt,
##   chr19_KI270931v1_alt, chr19_KI270932v1_alt, chr19_KI270933v1_alt,
```

```
##   chr19_KV575259v1_alt, chr20_KI270869v1_alt, chr22_KI270876v1_alt,
##   chr22_KI270879v1_alt, chr1_GL383519v1_alt, chr1_KN538360v1_fix,
##   chr2_GL582966v2_alt, chr3_GL383526v1_alt, chr3_KV766192v1_fix,
##   chr5_KI270791v1_alt, chr6_GL000251v2_alt, chr6_GL000254v2_alt,
##   chr6_GL000255v2_alt, chr6_KZ208911v1_fix, chr7_KI270899v1_alt,
##   chr8_KI270815v1_alt, chr8_KI270900v1_alt, chr11_JH159136v1_alt,
##   chr11_KI270832v1_alt, chr11_KQ759759v1_fix, chr11_KZ559109v1_fix,
##   chr12_GL877876v1_alt, chr12_KI270835v1_alt, chr12_KI270904v1_alt,
##   chr12_KZ208916v1_fix, chr14_KZ208920v1_fix, chr15_GL383555v2_alt,
##   chr17_GL383564v2_alt, chr17_KI270861v1_alt, chr19_GL383574v1_alt,
##   chr19_KI270865v1_alt, chr19_KI270868v1_alt, chr20_KI270870v1_alt,
##   chr20_KI270871v1_alt, chr21_GL383580v2_alt, chr21_KI270873v1_alt,
##   chr22_KI270877v1_alt, chr22_KN196485v1_alt, and chrUn_KI270750v1. Note
##   that ranges located on a sequence whose length is unknown (NA) or on a
##   circular sequence are not considered out-of-bound (use seqlengths() and
##   isCircular() to get the lengths and circularity flags of the underlying
##   sequences). You can use trim() to trim these ranges. See
##   ?`trim,GenomicRanges-method` for more information.
## Warning in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE): GRanges object contains 153 out-of-b
##   chr1_GL383518v1_alt, chr1_KQ458384v1_alt, chr2_GL383522v1_alt,
##   chr4_GL000257v2_alt, chr5_GL339449v2_alt, chr5_KI270795v1_alt,
##   chr5_KI270898v1_alt, chr5_KV575244v1_fix, chr6_KI270797v1_alt,
##   chr6_KI270798v1_alt, chr6_KI270801v1_alt, chr7_GL383534v2_alt,
##   chr7_KI270803v1_alt, chr7_KI270806v1_alt, chr7_KI270809v1_alt,
##   chr7_KZ208912v1_fix, chr9_GL383540v1_alt, chr9_GL383541v1_alt,
##   chr11_KI270902v1_alt, chr12_GL383551v1_alt, chr12_GL383553v2_alt,
##   chr12_KI270834v1_alt, chr14_KI270847v1_alt, chr15_KI270848v1_alt,
##   chr15_KI270850v1_alt, chr15_KI270851v1_alt, chr15_KI270906v1_alt,
##   chr16_GL383556v1_alt, chr16_KI270854v1_alt, chr17_JH159146v1_alt,
##   chr17_JH159147v1_alt, chr17_KI270857v1_alt, chr17_KI270860v1_alt,
##   chr17_KV575245v1_fix, chr17_KV766196v1_fix, chr17_KV766198v1_alt,
##   chr19_GL383575v2_alt, chr19_GL383576v1_alt, chr19_KI270866v1_alt,
##   chr19_KI270884v1_alt, chr19_KI270885v1_alt, chr19_KI270889v1_alt,
##   chr19_KI270890v1_alt, chr19_KI270891v1_alt, chr19_KI270915v1_alt,
##   chr19_KI270916v1_alt, chr19_KI270919v1_alt, chr19_KI270922v1_alt,
##   chr19_KI270923v1_alt, chr19_KI270929v1_alt, chr19_KI270930v1_alt,
##   chr19_KI270931v1_alt, chr19_KI270932v1_alt, chr19_KI270933v1_alt,
##   chr19_KV575259v1_alt, chr20_KI270869v1_alt, chr22_KI270876v1_alt,
##   chr22_KI270879v1_alt, chr1_GL383519v1_alt, chr1_KN538360v1_fix,
##   chr2_GL582966v2_alt, chr3_GL383526v1_alt, chr3_KV766192v1_fix,
##   chr5_KI270791v1_alt, chr6_GL000251v2_alt, chr6_GL000254v2_alt,
##   chr6_GL000255v2_alt, chr6_KZ208911v1_fix, chr7_KI270899v1_alt,
##   chr8_KI270815v1_alt, chr8_KI270900v1_alt, chr11_JH159136v1_alt,
##   chr11_KI270832v1_alt, chr11_KQ759759v1_fix, chr11_KZ559109v1_fix,
##   chr12_GL877876v1_alt, chr12_KI270835v1_alt, chr12_KI270904v1_alt,
##   chr12_KZ208916v1_fix, chr14_KZ208920v1_fix, chr15_GL383555v2_alt,
##   chr17_GL383564v2_alt, chr17_KI270861v1_alt, chr19_GL383574v1_alt,
##   chr19_KI270865v1_alt, chr19_KI270868v1_alt, chr20_KI270870v1_alt,
##   chr20_KI270871v1_alt, chr21_GL383580v2_alt, chr21_KI270873v1_alt,
##   chr22_KI270877v1_alt, chr22_KN196485v1_alt, and chrUn_KI270750v1. Note
##   that ranges located on a sequence whose length is unknown (NA) or on a
##   circular sequence are not considered out-of-bound (use seqlengths() and
##   isCircular() to get the lengths and circularity flags of the underlying
```

```
##   sequences). You can use trim() to trim these ranges. See
##   ?`trim,GenomicRanges-method` for more information.
```

```
# create dataframes
ChIP_aCR_p <- ChIP_aCR$percentage
ChIP_aCR_p <- as.data.frame(ChIP_aCR_p)
```

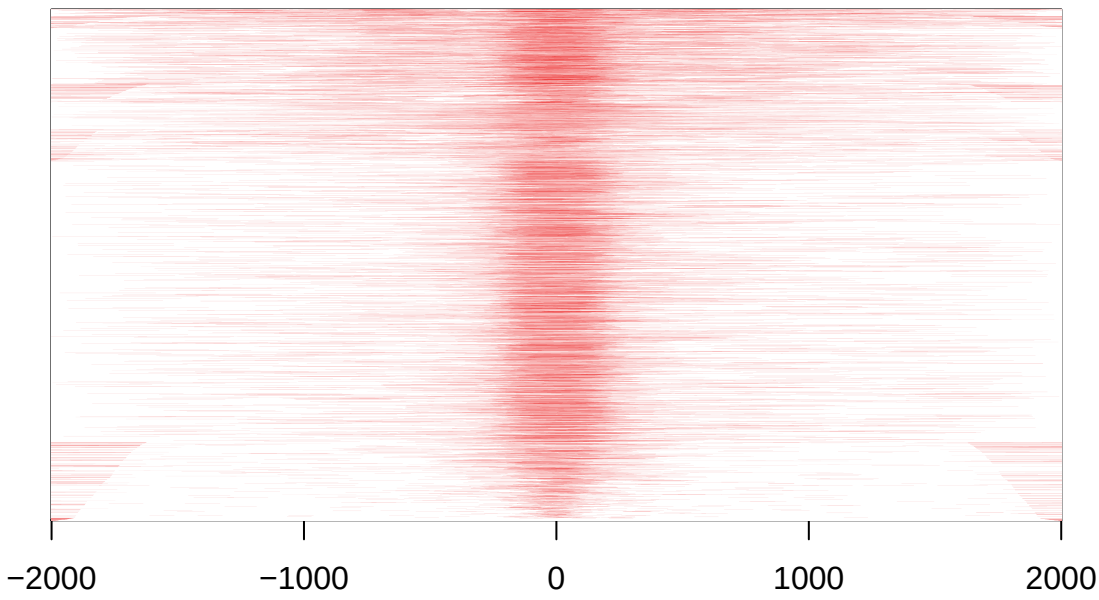
```
ggplot(ChIP_aCR_p, aes(x=`subjectHits`, y=`Freq`)) + geom_bar(position= "dodge", colour="black", width=
  show.legend = TRUE) + ylab("percentage [%]\n") + xlab("\ngenomic elements") + theme_bw() + labs(title
  theme(plot.title = element_text(color="black", size=16, face= "bold"),
        axis.title.x = element_text(color="black", size=16, face= "bold"),
        axis.text.x = element_text(angle = -90, vjust = 0.5, hjust = 0.0, color="black", size=12, face=
        axis.text.y = element_text(color="black", size=12, face = "bold"),
        axis.title.y = element_text(color="black", size=16, face="bold"))
```



peak heatmap

```
peakHeatmap(ChIP.anno, TxDb=txdb, upstream=2000, downstream=2000, color="brown2", title = "peak heatmap
```

```
## >> preparing promoter regions...  2020-05-14 15:17:38
## >> preparing tag matrix...        2020-05-14 15:17:39
## >> generating figure...       2020-05-14 15:17:51
```

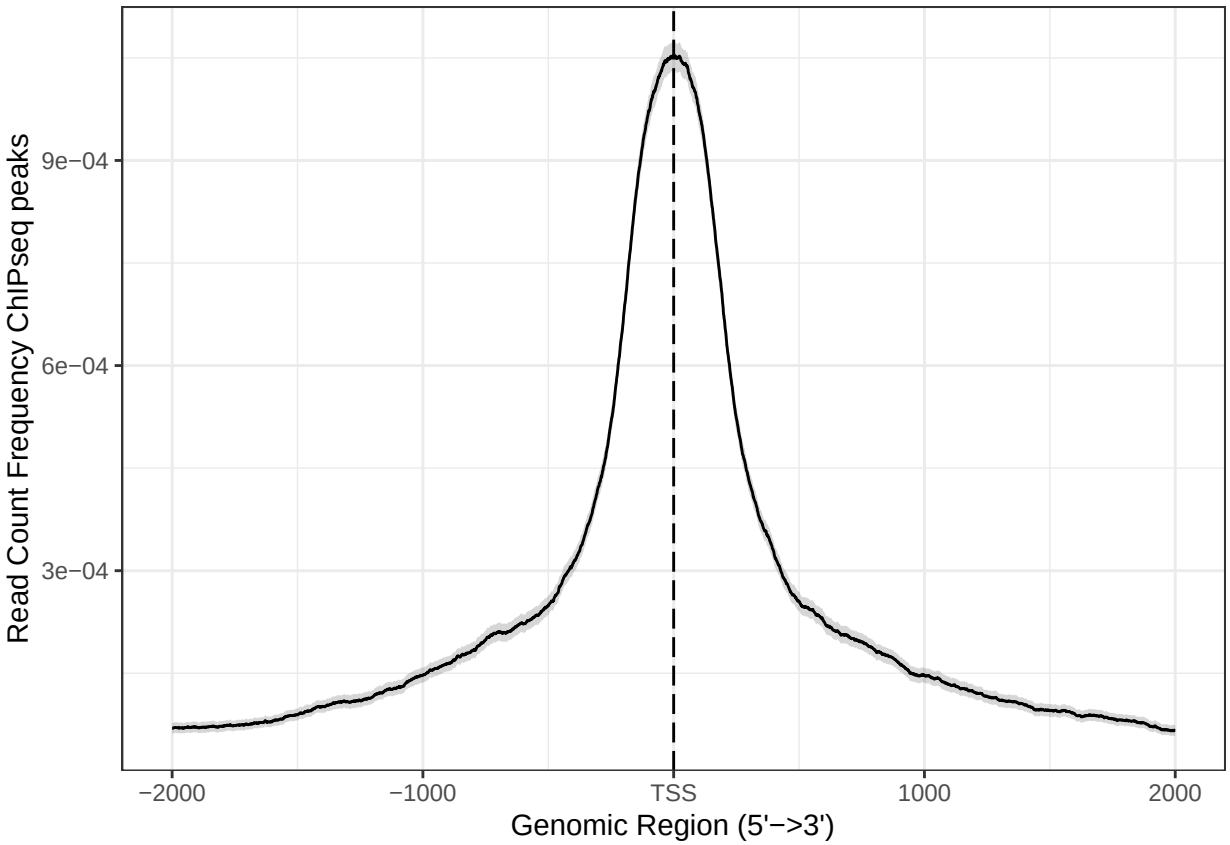



```
## >> done...            2020-05-14 15:18:01
```

density plot of ChIPseq peak profile

```
plotAvgProf2(ChIP.anno, TxDb=txdb, upstream=2000, downstream=2000, conf = 0.95, resample = 1000,
             xlab="Genomic Region (5'->3')", ylab = "Read Count Frequency ChIPseq peaks")
```

```
## >> preparing promoter regions...  2020-05-14 15:18:05
## >> preparing tag matrix...        2020-05-14 15:18:05
## >> plotting figure...             2020-05-14 15:18:11
## >> Running bootstrapping for tag matrix...        2020-05-14 15:21:08
```



get promoters via from reference genome hg38

```
genes <- genes(txdb)
promoters <- promoters(genes, upstream=2000, downstream=200) #  define promoter range around TSS
promoters
```

```
## GRanges object with 25750 ranges and 1 metadata column:
##             seqnames              ranges strand |     gene_id
##                <Rle>           <IRanges>  <Rle> | <character>
##           1    chr19   58362552-58364751      - |           1
##          10     chr8   18389282-18391481      + |          10
##         100    chr20   44652034-44654233      - |         100
##        1000    chr18   28176931-28179130      - |        1000
##   100009613    chr11   70075234-70077433      - |   100009613
##         ...      ...                 ...    ... .         ...
##        9991     chr9 112333468-112335667      - |        9991
##        9992    chr21   34362024-34364223      + |        9992
##        9993    chr22   19122255-19124454      - |        9993
##        9994     chr6   89827894-89830093      + |        9994
##        9997    chr22   50526262-50528461      - |        9997
##   -------
##   seqinfo: 595 sequences (1 circular) from hg38 genome
```

```
# subset annotaded ChIPseq peaks within promoters
ChIP_prom <- subsetByOverlaps(ChIP.anno, promoters)
ChIP_prom
```

```
## GRanges object with 8063 ranges and 16 metadata columns:
```

```
##                   seqnames              ranges strand |        name     score
##                      <Rle>           <IRanges>  <Rle> | <character> <numeric>
##   X00007.101927612     chr8 124472531-124472910      * |        <NA>       781
##   X00009.105377348     chr4 102827580-102827959      * |        <NA>      1000
##       X00015.51078     chr2 241636406-241636785      * |        <NA>       674
##       X00018.60386    chr17   75289551-75289930      * |        <NA>      1000
##        X00020.6161     chr3   12841698-12842077      * |        <NA>       961
##                ...      ...                 ...    ... .         ...       ...
##       X27653.92579    chr17   44070519-44070842      * |        <NA>      1000
##       X27654.55226    chr11   34105471-34105830      * |        <NA>      1000
##        X27655.6449    chr19     2783355-2783741      * |        <NA>      1000
##        X27658.6397    chr17   77088514-77088859      * |        <NA>      1000
##       X27659.11325    chr17   63773667-63774105      * |        <NA>      1000
##                   signalValue    pValue    qValue        peak     feature
##                     <numeric> <numeric> <numeric> <character> <character>
##   X00007.101927612     5.39297        -1  -0.14152       00007   101927612
##   X00009.105377348     5.39875        -1  -0.14256       00009   105377348
##       X00015.51078     5.43734        -1  -0.15267       00015       51078
##       X00018.60386     5.45674        -1  -0.14767       00018       60386
##        X00020.6161     5.47562        -1  -0.14897       00020        6161
##                ...         ...       ...       ...         ...         ...
##       X27653.92579   372.85637        -1   4.57055       27653       92579
##       X27654.55226   378.62518        -1   4.57055       27654       55226
##        X27655.6449   382.89774        -1   4.57055       27655        6449
##        X27658.6397   402.89835        -1   4.57055       27658        6397
##       X27659.11325   425.27089        -1   4.57055       27659       11325
##                   start_position end_position feature_strand insideFeature
##                        <integer>    <integer>    <character>   <character>
##   X00007.101927612      124462485    124474582              -        inside
##   X00009.105377348      102828055    102844075              +      upstream
##       X00015.51078      241584405    241637158              -        inside
##       X00018.60386       75272981     75289510              -      upstream
##        X00020.6161       12834485     12841582              -      upstream
##                ...            ...          ...            ...           ...
##       X27653.92579       44070735     44076344              +  overlapStart
##       X27654.55226       34105617     34146908              +  overlapStart
##        X27655.6449        2754715      2783282              -      upstream
##        X27658.6397       77088749     77217101              +  overlapStart
##       X27659.11325       63773603     63819317              +        inside
##                   distancetoFeature shortestDistance fromOverlappingOrNearest
##                           <numeric>        <integer>              <character>
##   X00007.101927612              2051             1672          NearestLocation
##   X00009.105377348              -475               96          NearestLocation
##       X00015.51078               752              373          NearestLocation
##       X00018.60386               -41               41          NearestLocation
##        X00020.6161              -116              116          NearestLocation
##                ...               ...              ...                      ...
##       X27653.92579              -216              107          NearestLocation
##       X27654.55226              -146              146          NearestLocation
##        X27655.6449               -73               73          NearestLocation
##        X27658.6397              -235              110          NearestLocation
##       X27659.11325                64               64          NearestLocation
##                          ensembl      symbol
##                      <character> <character>
```

```
##   X00007.101927612 ENSG00000245149  RNF139-AS1
##   X00009.105377348 ENSG00000246560  UBE2D3-AS1
##       X00015.51078 ENSG00000176946       THAP4
##       X00018.60386 ENSG00000125454    SLC25A19
##        X00020.6161 ENSG00000144713       RPL32
##                ...             ...         ...
##       X27653.92579 ENSG00000141349       G6PC3
##       X27654.55226 ENSG00000135372       NAT10
##        X27655.6449 ENSG00000104969        SGTA
##        X27658.6397 ENSG00000129657     SEC14L1
##       X27659.11325 ENSG00000198231       DDX42
##   -------
##   seqinfo: 25 sequences from an unspecified genome; no seqlengths
```

```
### get dataframe of GRanges object of ChIPseq data
#ChIP_prom_seq <- getAllPeakSequence(ChIP_prom, upstream = 100, downstream = 100, genome = Hsapiens)

ChIP_prom_df <- as.data.frame(ChIP_prom)
ChIP_prom_df <- tibble::rownames_to_column(ChIP_prom_df, "Ranges")

ChIP_targetgenes <- ChIP_prom_df[,c("ensembl","symbol","Ranges","signalValue")]

# extract fasta sequences
ChIP_seq <- getSeq(BSgenome.Hsapiens.UCSC.hg38, ChIP_prom)

# import fasta sequences
fastalist <- readDNAStringSet("D:/01_Private Dateien/Biochemie Studium/Masterstudium/01 - Molecular Bios

#count G and C nucleotides in sequence and divide by all nucleotides etc.
ChIP_prom_df_G <- str_count(ChIP_prom_df$sequence, "G")
ChIP_prom_df_C <- str_count(ChIP_prom_df$sequence, "C")
ChIP_prom_df_GCAT <- str_count(ChIP_prom_df$sequence, "")
ChIP_prom_df_GC <- ((ChIP_prom_df_C + ChIP_prom_df_G) / ChIP_prom_df_GCAT) * 100

# add % GC-content as column to dataframe
ChIP_prom_df["GC_content"] <- ChIP_prom_df_GC
colnames(ChIP_prom_df)
```

```
##  [1] "Ranges"                  "seqnames"
##  [3] "start"                   "end"
##  [5] "width"                   "strand"
##  [7] "name"                    "score"
##  [9] "signalValue"             "pValue"
## [11] "qValue"                  "peak"
## [13] "feature"                 "start_position"
## [15] "end_position"            "feature_strand"
## [17] "insideFeature"           "distancetoFeature"
## [19] "shortestDistance"        "fromOverlappingOrNearest"
## [21] "ensembl"                 "symbol"
## [23] "GC_content"
```

```
# save as txt
write_tsv(ChIP_prom_df, "ChIPseq_prom_df.txt")

### GO enrichment analysis with Clusterprofiler
```

```
ChIP_prom_df <- read_tsv(file = "D:/01_Private Dateien/Biochemie Studium/Masterstudium/01 - Molecular Bi
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   Ranges = col_character(),
##   seqnames = col_character(),
##   strand = col_character(),
##   name = col_logical(),
##   peak = col_character(),
##   feature_strand = col_character(),
##   insideFeature = col_character(),
##   fromOverlappingOrNearest = col_character(),
##   ensembl = col_character(),
##   symbol = col_character(),
##   sequence = col_character()
## )
```

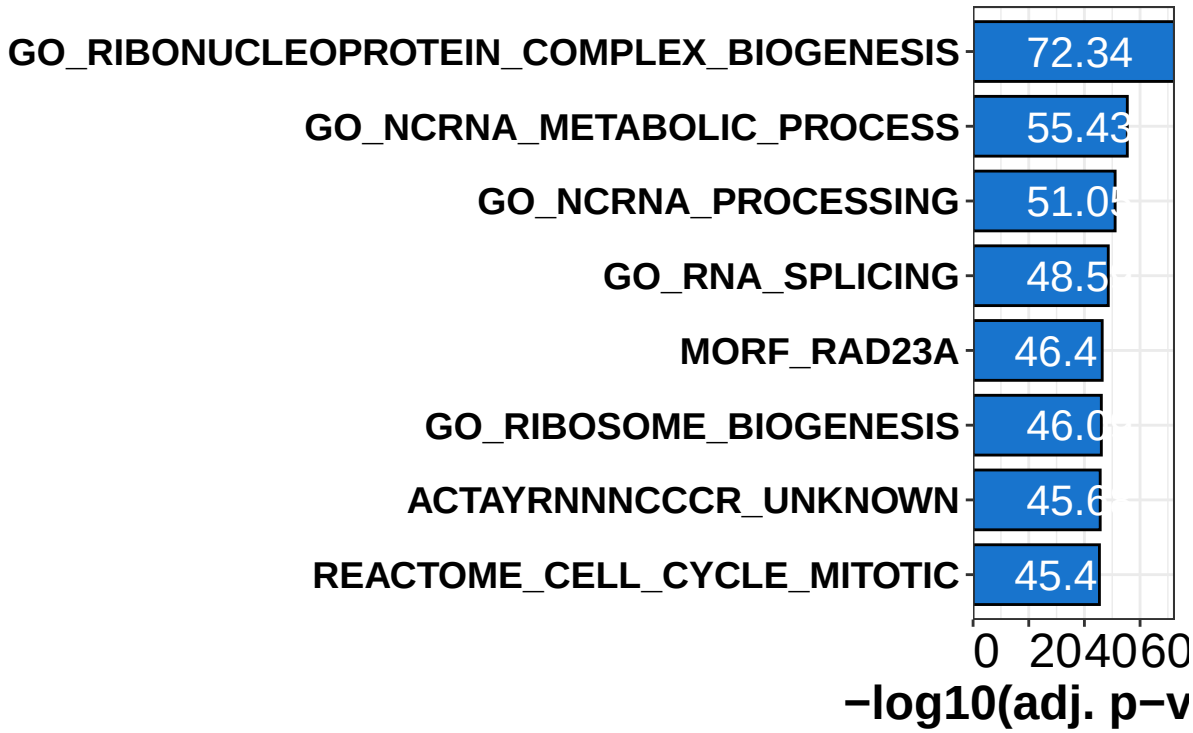
```
## See spec(...) for full column specifications.
```

```
## perform enrichment analysis
# use known genes from hg38 reference genome as background genes -> annotate via org.Hs.eg.db to get ge
backgroundgenes <- ucsc.hg38.knownGene$gene_id
backgroundgenes.anno <- addGeneIDs(annotatedPeak=backgroundgenes,
                                   orgAnn="org.Hs.eg.db", feature_id_type="entrez_id", c("ensembl","syml

# create term2gene dataframe , category = "H" == hallmark gene sets , "C2" == curated gene sets , "C3" =
m_t2g <- msigdbr(species = "Homo sapiens") %>% dplyr::select(gs_name, gene_symbol)
head(m_t2g)
```

```
## # A tibble: 6 x 2
##   gs_name        gene_symbol
##   <chr>          <chr>
## 1 AAACCAC_MIR140 ABCC4
## 2 AAACCAC_MIR140 ABRAXAS2
## 3 AAACCAC_MIR140 ACTN4
## 4 AAACCAC_MIR140 ACVR1
## 5 AAACCAC_MIR140 ADAM9
## 6 AAACCAC_MIR140 ADAMTS5
```

```
em <- enricher(ChIP_prom_df$symbol, TERM2GENE=m_t2g, pvalueCutoff = 0.05, pAdjustMethod = "BH", univers
enricher_result <- em@result

# save txt file
write_tsv(enricher_result, "ChIPseq_enricher_result.txt")

### plot GO enrich results
enricher_result = enricher_result[order(enricher_result$p.adjust), c(1:9)]
enricher_result <- enricher_result[c(1:8),]
nLog10_adjP <- -log10(enricher_result$p.adjust)
enricher_result["log10(p.adjust)"] <- nLog10_adjP
ordeR <- c(8:1)
enricher_result["order"] <- ordeR

ggplot(enricher_result, aes(x=reorder(`Description`, `order`), y=`log10(p.adjust)`)) +
```

```
  geom_bar(position= "dodge", fill="dodgerblue3", width=.8, stat = "identity", show.legend = TRUE, colou
  theme_bw() + labs(title = "") + ylab("-log10(adj. p-value)\n") + xlab("") +
  scale_y_continuous(expand = c(0,0), limits = c()) +
  geom_text(aes(y=`p.adjust`, label=round(`log10(p.adjust)`, digits = 2)), vjust=0.5, hjust=-0.5, size=5
            position=position_stack(vjust = 0.5), colour="white") +
  theme(plot.title = element_text(color="firebrick3", size=18, face= "bold", hjust=1.0),
        axis.title.x = element_text(color="black", size=18, face= "bold", hjust=0.5),
        axis.text.x = element_text(angle = 0, vjust = 1, hjust = 0, color="black", size=18),
        axis.text.y = element_text(color="black", size=14, face = "bold"),
        axis.title.y = element_text(color="black", size=18, face="bold")) + coord_flip()
```




```
### motif analysis
ChIP_prom_f <- ChIP_prom[order(ChIP_prom$signalValue, decreasing = TRUE)]


top500_peaks <- head(ChIP_prom_f, n=500)

# reduce peaks to merge nearby located peaks
#reduced_peaks <- reduce(top500_peaks)

# resize peak sequences (get only +/- 100 bp from center)
resized_peaks <- resize(top500_peaks, width=200, fix="center")

# get peak sequences
peak_seq <- getSeq(BSgenome.Hsapiens.UCSC.hg38, resized_peaks)
```

```r
set.seed(1) # shuffle sequences of peak_seq and use as background for analysis
control_seq <- DNAStringSet(lapply(peak_seq, sample))

### motif analysis
set.seed(1)

#calculate motif enrichments
motifs <- findMotifFgBg(fg.seq = peak_seq, bg.seq = control_seq, max.motif = 1, enriched.only = TRUE, ma
```

```
## CACGTG 1.205683e-07
##  Refine  CACGTG 12.07778 : 11.49449 11.53701 12.05504 11.36509 11.985 TRUE 274 37 237 36
## CACGHG 13.9901 : 14.28315 13.26456 13.9274 14.21191 13.04699 838 241

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
```

```
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## New motif:  CACGHG

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData
```

```
## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData
```

```
## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData

## Warning:   RangedData objects are deprecated. Please migrate your code to use
##   GRanges or GRangesList objects instead. See IMPORTANT NOTE in
##   ?RangedData
```

```r
# refine motif edges
refined_motifs <- lapply(motifs$motifs, function(x){motifRG::refinePWMMotifExtend(motifs = x@match$patt
```

```
## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 1 4682.157
```

```
## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 2 5222.436

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 3 5268.468

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 4 5296.964

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 5 5309.644

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 6 5317.644

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 7 5319.291
```

```
## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 8 5321.355

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 9 5322.441

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 10 5323.886

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 11 5323.951

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Extend score 5323.886 0

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 1 5551.524
```

```
## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 2 5592.383

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 3 5609.687

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 4 5619.869

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 5 5629.428

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 6 5636.112

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 7 5641.691
```

```
## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 8 5643.621

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 9 5646.148

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 10 5648.253

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 11 5649.432

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 12 5650.841

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 13 5652.244
```

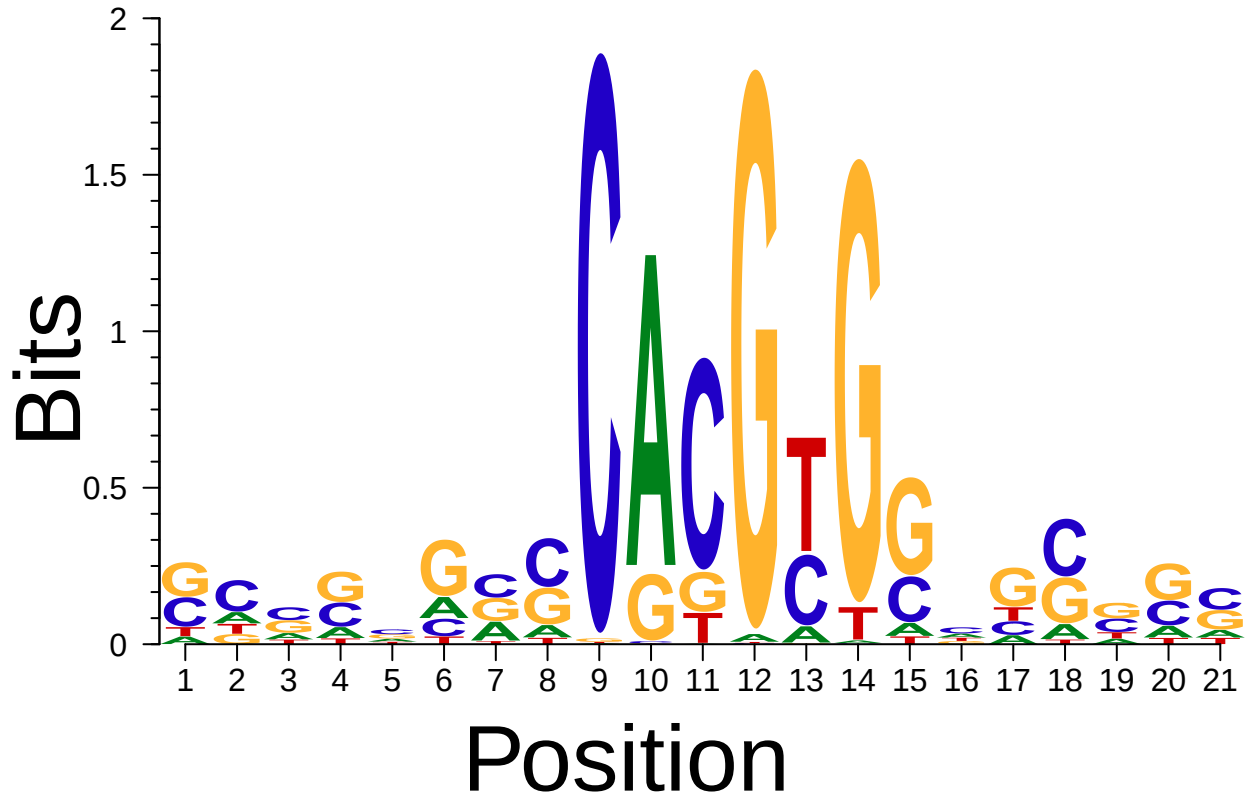
```
## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 14 5653.414

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 15 5654.601

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 16 5654.835

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Extend score 5654.601 5323.886

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 1 5911.556

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 2 5976.197
```

```
## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 3 5995.1

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 4 6004.062

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 5 6009.01

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 6 6012.161

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 7 6013.874

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 8 6015.061
```

```
## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## Warning in .Call2("PWM_score_starting_at", pwm, subject, starting.at,
## base_codes, : 'subject' contains letters not in [ACGT] ==> assigned weight 0 to
## them

## 9 6015.248
```

plot seqlogo

```
plotMotifLogo(refined_motifs[[1]]$model$prob, motifName = "", font = "Helvetica-Bold", xlcex = 2.9, ylc
```



motif annotation

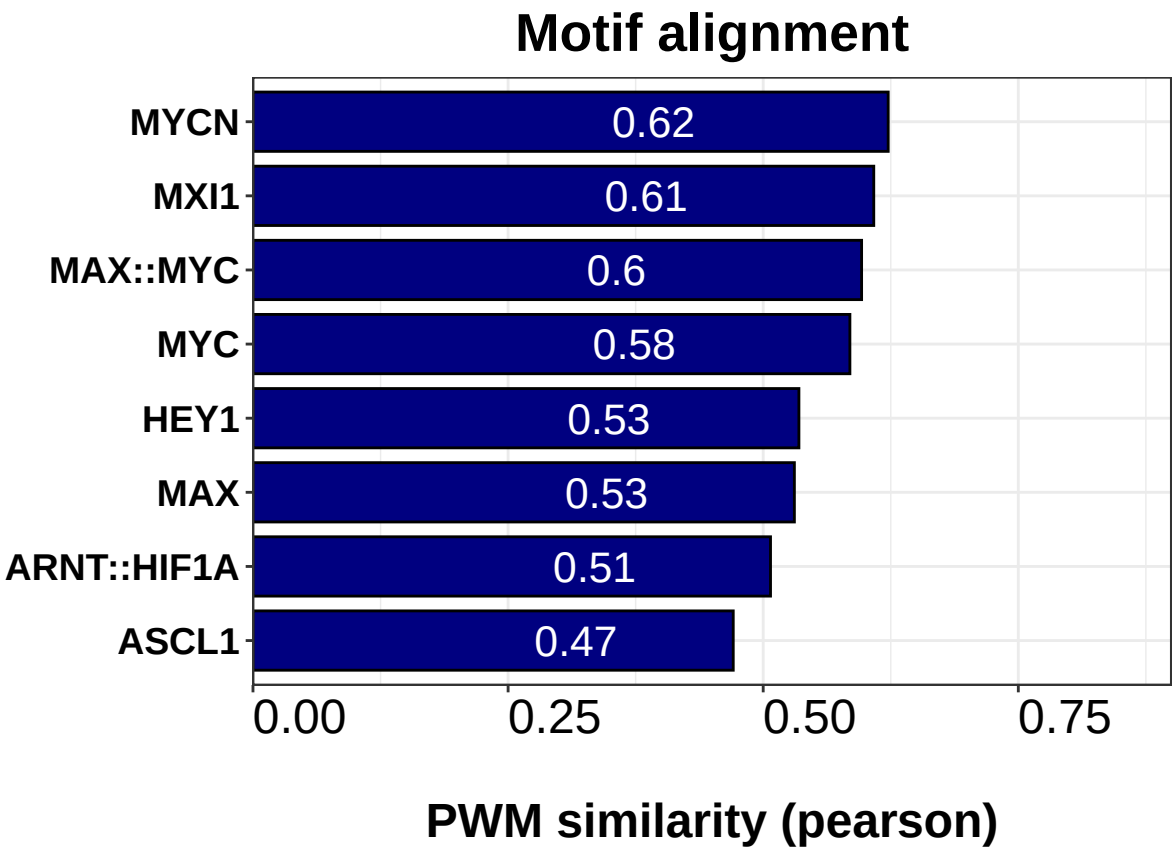
```
unknown_motif <- refined_motifs[[1]]$model$prob

# create PWM (position weight matrix)
unknown_pwm <- PWMatrix(ID = "unk", profileMatrix = unknown_motif)

# get PWM library from JASPAR core motif database
pwm_library <- getMatrixSet(JASPAR2018, opts = list(collection = "CORE", species = "Homo sapiens", matr

# find motifs that are simmilar to the motif (pearson correlation)
pwm_sim <- PWMSimilarity(pwm_library, unknown_pwm, method = "Pearson")
```

```
pwm_library_list = lapply(pwm_library, function(x){data.frame(ID = ID(x), name = name(x))})
pwm_library_dt = dplyr::bind_rows(pwm_library_list)
pwm_library_dt$similarity = pwm_sim[pwm_library_dt$ID]
pwm_library_dt = pwm_library_dt[order(-pwm_library_dt$similarity),]
pwm_library_dt$order = order(pwm_library_dt$similarity)
head(pwm_library_dt)
```

```
##           ID     name similarity order
## 389 MA0104.4     MYCN  0.6224866   452
## 386 MA1108.1     MXI1  0.6083362   451
## 7   MA0059.1 MAX::MYC  0.5964263   450
## 387 MA0147.3      MYC  0.5849530   449
## 299 MA0823.1     HEY1  0.5349980   448
## 301 MA0058.3      MAX  0.5305468   447
```

batplot of pearson correlation values of known motifs from JASPAR2018 data

```
ggplot(pwm_library_dt[1:8,], aes(x=reorder(`name`, `order`), y=`similarity`)) +
  geom_bar(position= "dodge", fill="navy", width=.8, stat = "identity", show.legend = TRUE, colour="bla
  theme_bw() + labs(title = "Motif alignment") + ylab("\nPWM similarity (pearson)") + xlab("") +
  scale_y_continuous(expand = c(0,0), limits = c(0,0.9)) +
  geom_text(aes(y=`similarity`, label=round(`similarity`, digits = 2)), vjust=0.5, hjust=-0.5, size=5.5
  theme(plot.title = element_text(color="black", size=20, face= "bold", hjust=0.5),
        axis.title.x = element_text(color="black", size=18, face= "bold", hjust=0.5),
        axis.text.x = element_text(angle = 0, vjust = 1, hjust = 0, color="black", size=18),
        axis.text.y = element_text(color="black", size=14, face = "bold"),
        axis.title.y = element_text(color="black", size=18, face="bold")) + coord_flip()
```


Motif alignment
MYCN
MXI1
MAX::MYC
MYC
HEY1
MAX
ARNT::HIF1A
ASCL1
0.62
0.61
0.6
0.58
0.53
0.53
0.51
0.47
0.00
0.25
0.50
0.75
PWM similarity (pearson)