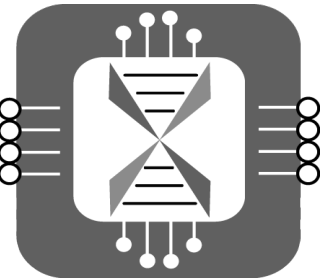
GENOMATION

A toolkit to summarize, annotate and visualize genomic intervals

*Presented by Katarzyna Wreczycka

BIOINFORMATICS PLATFORM
MDC MAX DELBRÜCK CENTER FOR MOLECULAR MEDICINE BERLIN-BUCH
BIMSB THE BERLIN INSTITUTE FOR MEDICAL SYSTEMS BIOLOGY

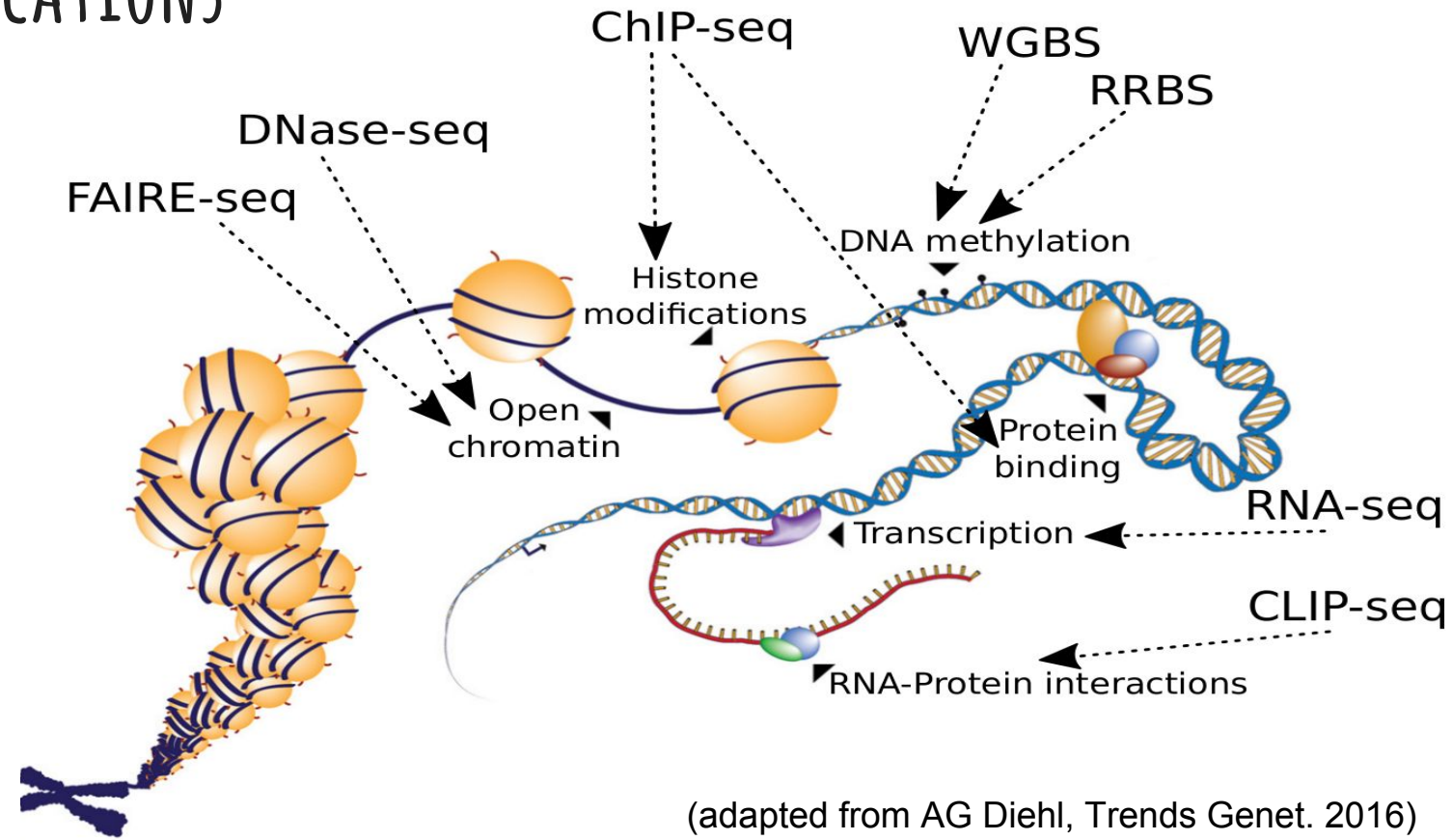


Citation (from within R, enter `citation("genomation")`):

Akalin A, Franke V, Vlahovicek K, Mason C and Schubeler D (2014). “genomation: a toolkit to summarize, annotate and visualize genomic intervals.” *Bioinformatics*.

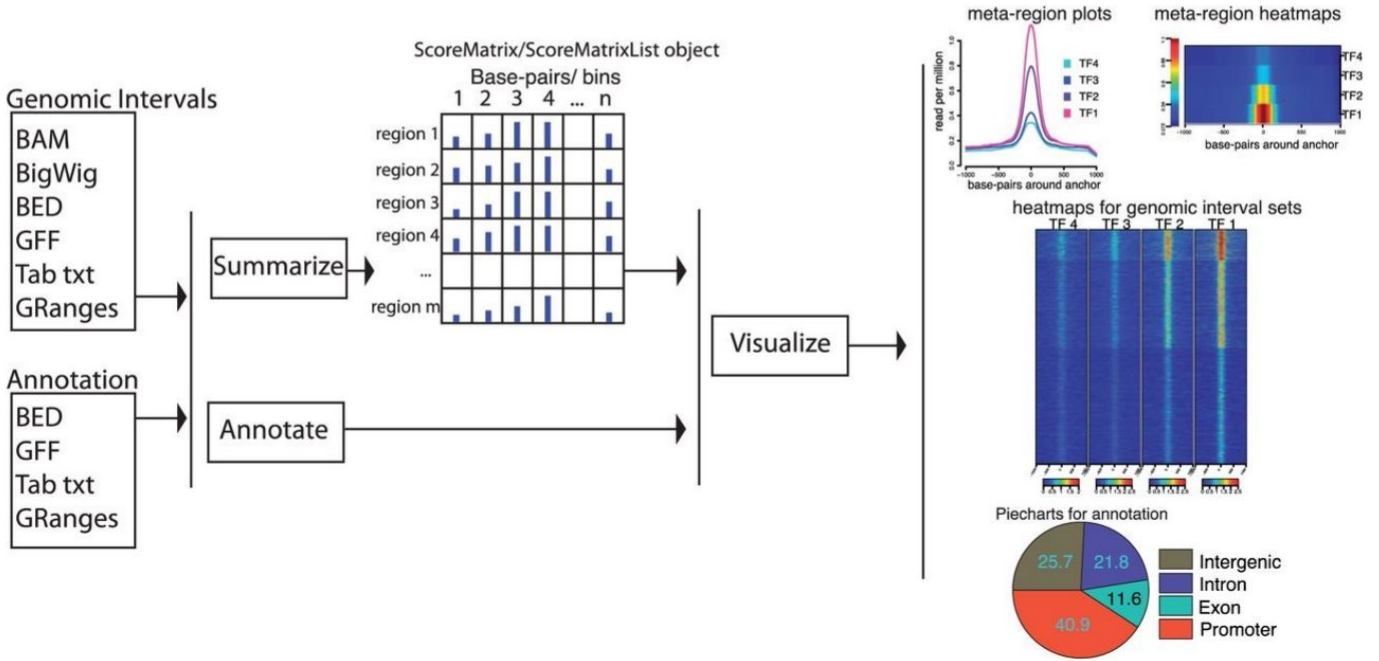
*Created by Altuna Akalin and Vedran Franke

Applications



(adapted from AG Diehl, Trends Genet. 2016)

Overview of genomation features



(Altuna Akalin et al. Bioinformatics 2015)

Installation

Via Bioconductor:

```
source("http://bioconductor.org/biocLite.R")
biocLite("genomation")
```

From GitHub:

```
#' Install dependecies
install.packages( c("data.table","plyr","reshape2","ggplot2","gridBase","devtools"))
source("http://bioconductor.org/biocLite.R")
biocLite(c("GenomicRanges","rtracklayer","impute","Rsamtools"))

#' install the packages
library(devtools)
install_github("BIMSBbioinfo/genomation",build_vignettes=FALSE)
```

Data import

Genomic intervals can be read from various file formats.

To read files:

- readGeneric
- readBed
- readNarrowPeak
- readBroadPeak

```
ctcf.peaks = readGeneric(file.path(genomationDataPath,
                                   'wgEncodeBroadHistoneH1hescCtcfStdPk.broadPeak.gz'))
```

```
> ctcf.peaks
GRanges object with 1681 ranges and 0 metadata columns:
         seqnames              ranges strand
            <Rle>           <IRanges>  <Rle>
     [1]    chr21   [9484643, 9485642]      *
     [2]    chr21   [9589776, 9590775]      *
     [3]    chr21   [9647485, 9648484]      *
     [4]    chr21   [9695637, 9696636]      *
     [5]    chr21   [9699131, 9700130]      *
     ...      ...                 ...    ...
  [1677]    chr21 [47971254, 47972253]      *
  [1678]    chr21 [48054588, 48055587]      *
  [1679]    chr21 [48059377, 48060376]      *
  [1680]    chr21 [48062985, 48063984]      *
  [1681]    chr21 [48080784, 48081783]      *
  -------
  seqinfo: 24 sequences from an unspecified genome; no seqlengths
```

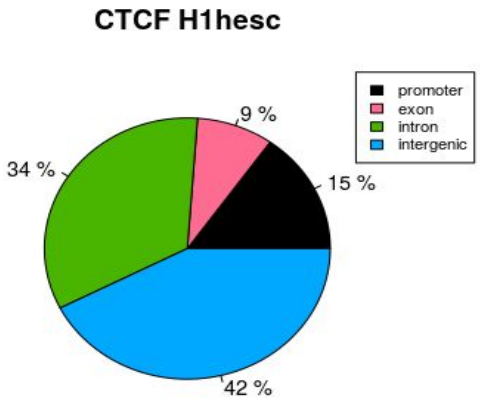
Annotate Genomic intervals

Annotate with gene structures..

```
transcriptFeat <- readTranscriptFeatures("./refseq.hg19.bed",
                                         up.flank = 2000, down.flank = 2000)
ann.ctcf <- annotateWithGeneParts(ctcf.peaks, transcriptFeat)
getTargetAnnotationStats(ann.ctcf, percentage=TRUE, precedence=TRUE)
```

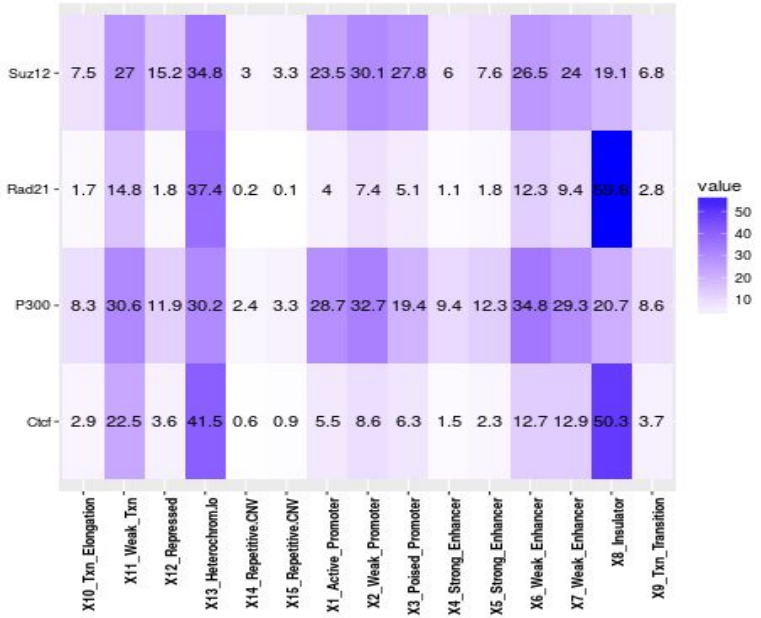
```
> getTargetAnnotationStats(ann.ctcf, percentage=TRUE, precedence=TRUE)
  promoter       exon     intron intergenic
     15.17       8.69      33.67      42.47
```

```
plotTargetAnnotation(ann.ctcf, main="CTCF H1hesc")
```



..or with any feature

```
peak2ann.l=annotateWithFeatures(broadpeak.list,
                                chrHMM.list)
heatTargetAnnotation(peak2ann.l)
```



Summarize genomic intervals over target

Target:

- GRanges
- BAM
- BigWig

To summarize:

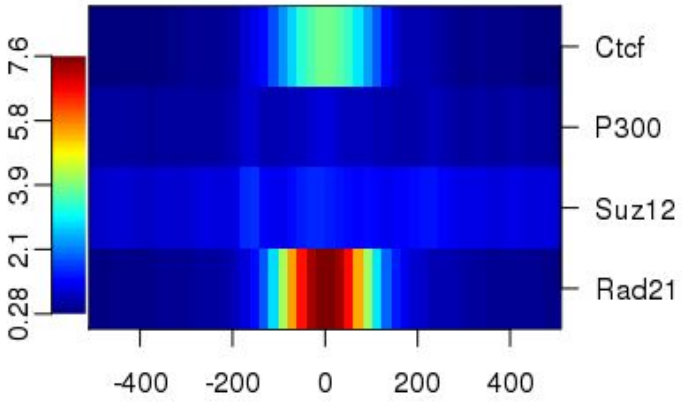
- ScoreMatrix
- ScoreMatrixBin
- ScoreMatrixList

```
sm = ScoreMatrix(bam.files[1], ctcf.peaks,
                 type='bam')

sml = ScoreMatrixList(bam.files, ctcf.peaks,
                      type='bam', bin.num=50, cores=2)
```

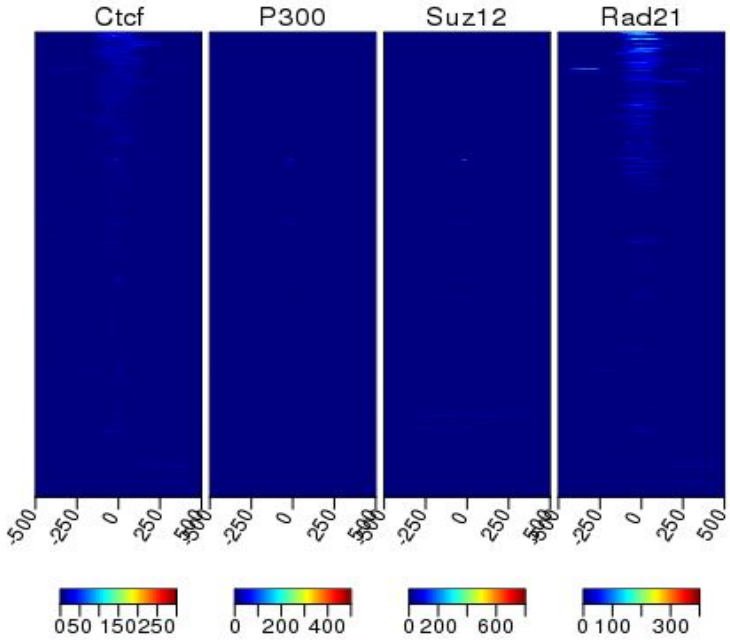
```
> sm
  scoreMatrix with dims: 1681 1000
> sml
scoreMatrixlist of length:4

1. scoreMatrix with dims: 1681 50
2. scoreMatrix with dims: 1681 50
3. scoreMatrix with dims: 1681 50
4. scoreMatrix with dims: 1681 50
```

Visualize summary matrices as heatmaps and meta-regions plots

```
heatMeta(sml, xcoords=c(-500, 500))
```




```
multiHeatMatrix(sml, xcoords=c(-500, 500))
```

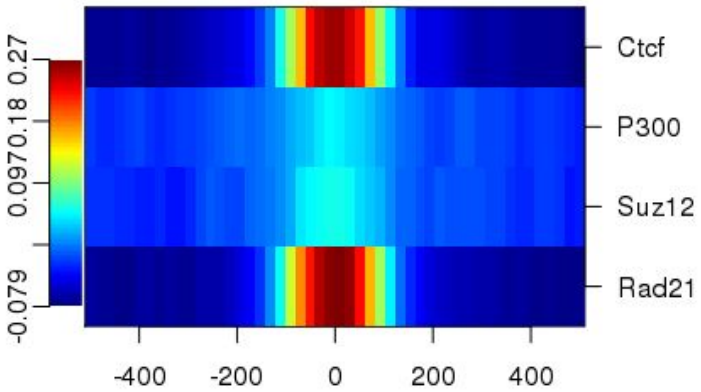


Scale heatmaps

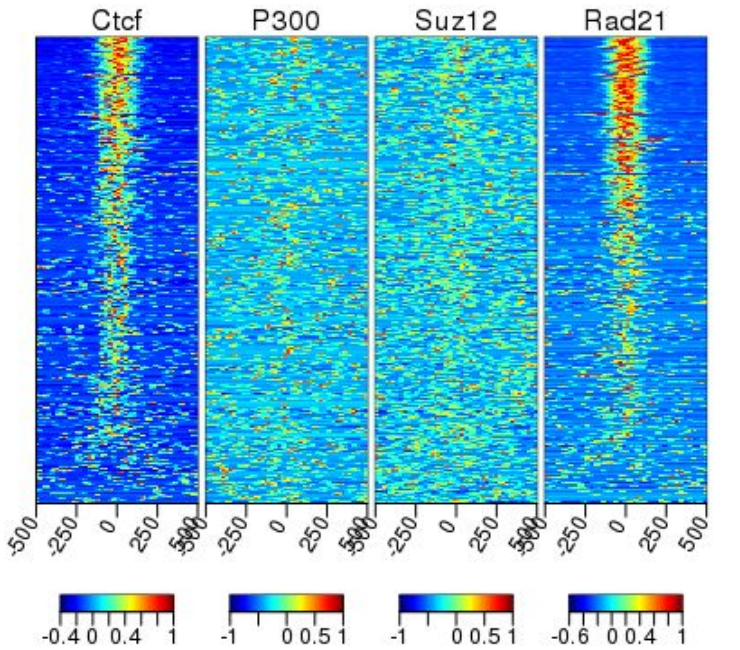
Scale each ScoreMatrix in the ScoreMatrixList object by rows and/or columns

```
sml.scaled = scaleScoreMatrixList(sml,
                                  scalefun = function(x)
                                             (x - mean(x))/(max(x)-min(x)+1),
                                  rows=TRUE)
```

```
heatMeta(sml.scaled, xcoords=c(-500, 500))
```



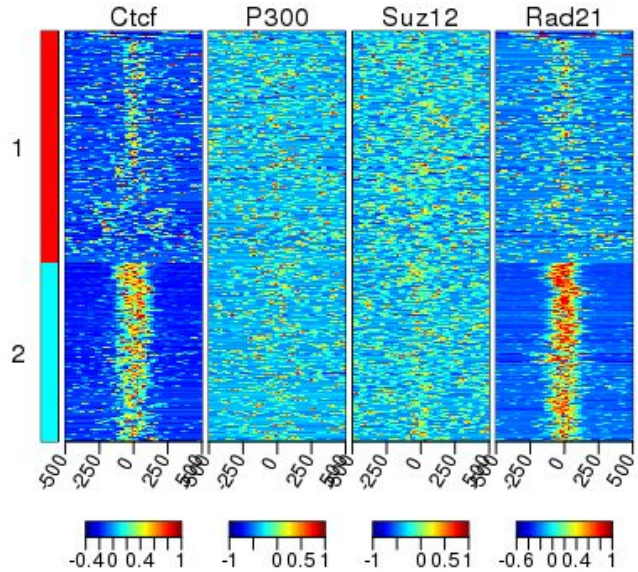

```
multiHeatMatrix(sml.scaled, xcoords=c(-500, 500))
```



Cluster heatmaps

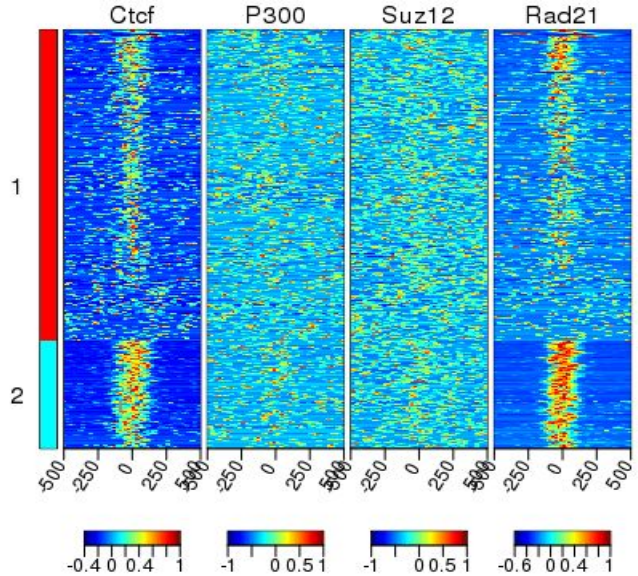
K-means with k=2

```
cl1 <- function(x) kmeans(x, centers=2)$cluster
multiHeatMatrix(sml.scaled, xcoords=c(-500, 500), clustfun = cl1)
```



Hierarchical clustering with Ward's method for agglomeration into 2 cluster

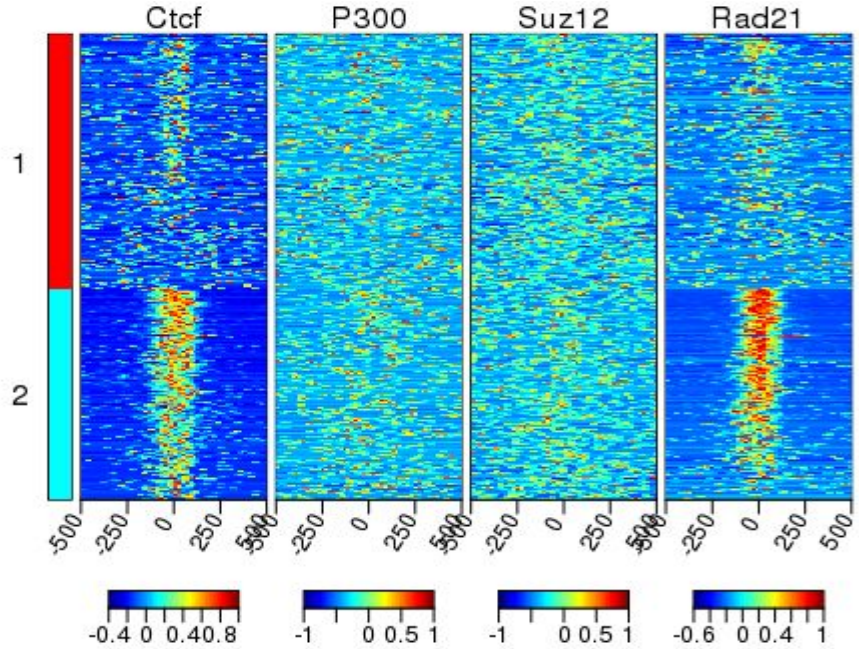
```
cl2 <- function(x) cutree(hclust(dist(x), method="ward"), k=2)
multiHeatMatrix(sml.scaled, xcoords=c(-500, 500), clustfun = cl2)
```



Cluster heatmaps

Define which matrices are used for clustering

```
multiHeatMatrix(sml.scaled, xcoords=c(-500, 500), clustfun = cl1, clust.matrix = 1)
```

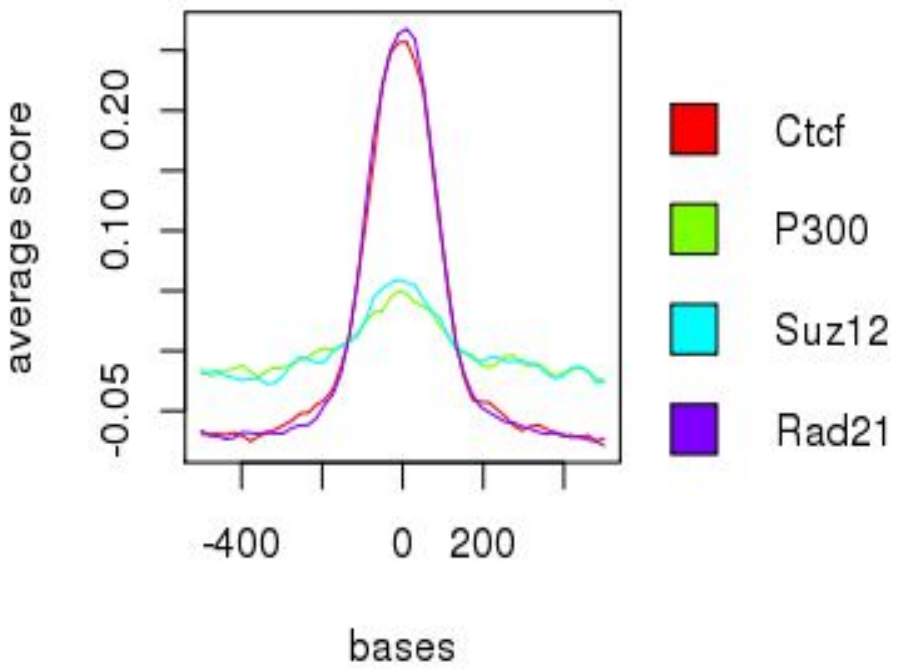


META-PLOTS

```
plotMeta(mat=sml.scaled, profile.names=names(sml.scaled),
         xcoords=c(-500, 500),
         winsorize=c(0,99),
         centralTend="mean")
```

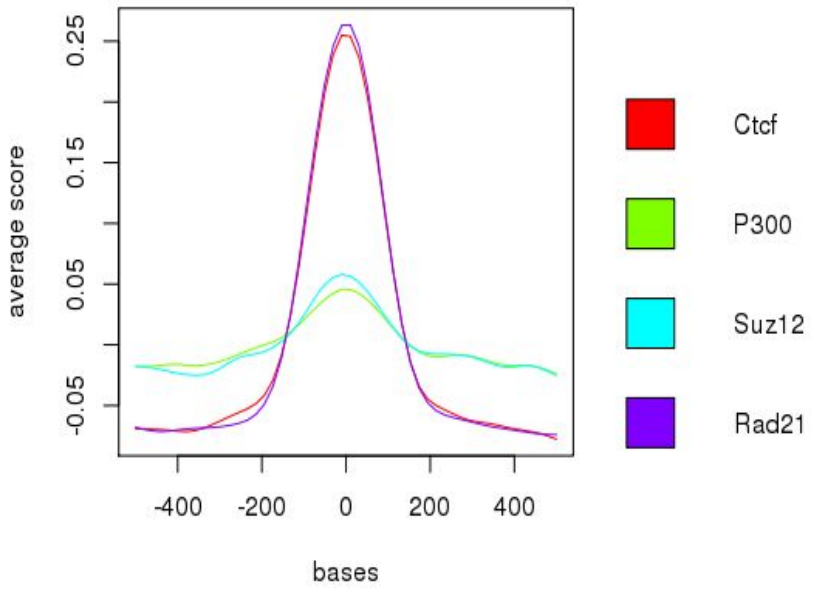
centralTend:

- mean
- median



Smooth meta-plots

```
plotMeta(mat=sml.scaled, profile.names=names(sml.scaled),
         xcoords=c(-500, 500),
         winsorize=c(0,99),
         centralTend="mean",
         smoothfun=function(x) stats::smooth.spline(x, spar=0.5))
```

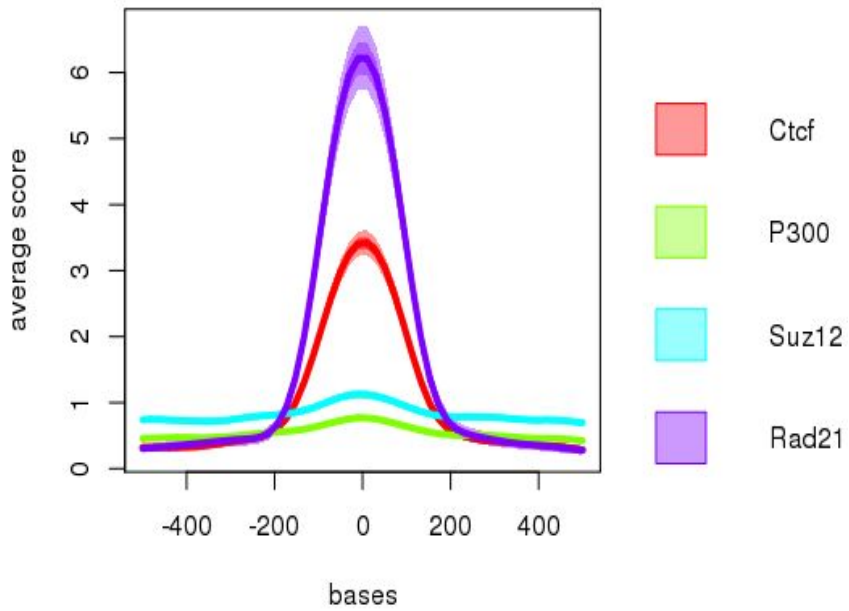


Add dispersion to meta-plots

```
plotMeta(mat=sml, profile.names=names(sml),
         xcoords=c(-500, 500),
         winsorize=c(0,99),
         centralTend="mean",
         smoothfun=function(x) stats::smooth.spline(x, spar=0.5)
         dispersion="se", lwd=4)
```

Dispersion:

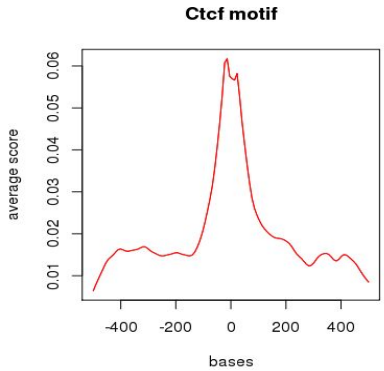
- "se" shows standard error of the mean and 95 percent confidence interval for the mean
- "sd" shows standard deviation and 2*(standard deviation)
- "IQR" shows 1st and 3rd quartile and confidence interval around the median based on the median +/- 1.57 * IQR/sqrt(n) (notches)

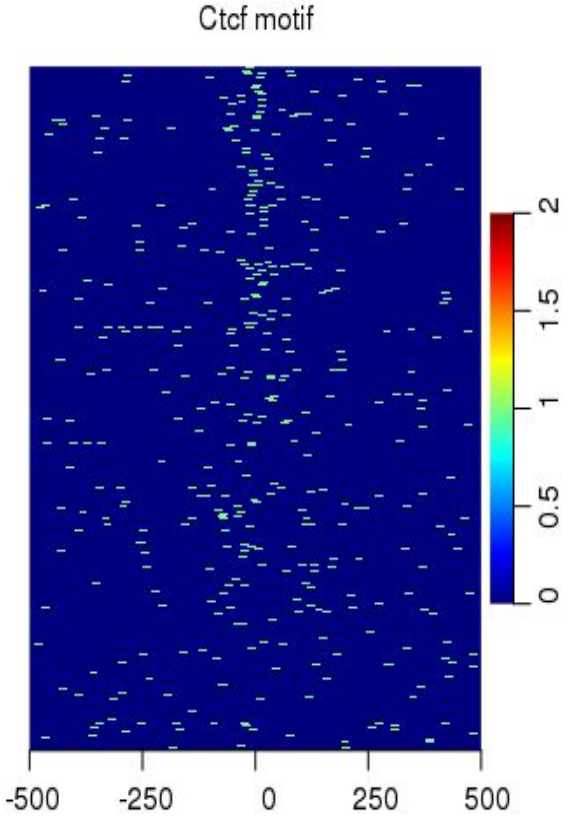


MOTIF ENRICHMENT OVER GENOMIC INTERVALS

Get scores of k-mers or PWM matrix over windows

```
library(BSgenome.Hsapiens.UCSC.hg19)
hg19 = BSgenome.Hsapiens.UCSC.hg19

p = patternMatrix(pattern=ctcf.pwm, windows=ctcf.peaks, genome=hg19, min.score=0.8)

heatMatrix(p, xcoords=c(-500, 500), main="Ctcf motif")
plotMeta(mat=p, xcoords=c(-500, 500), smoothfun=function(x) stats::lowess(x, f = 1/10),
         line.col="red", main="Ctcf motif")
```

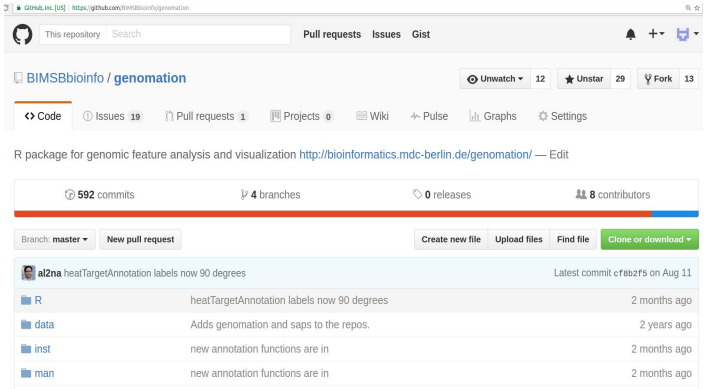




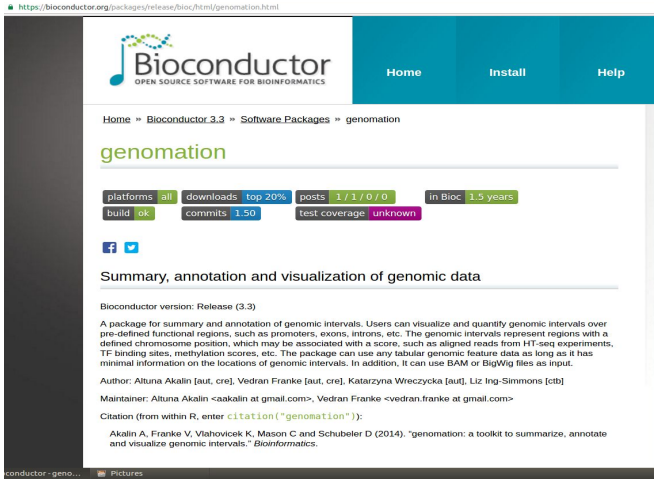
Usefull for ChIP-seq

For more info

- github.com/BIMSBbioinfo/genomation

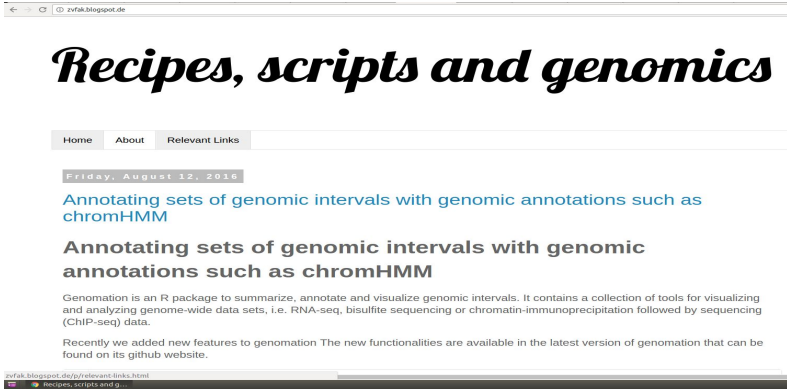


- bioconductor.org



- groups.google.com/forum/#!forum/genomation

- zvfak.blogspot.de (Altuna's blog)



Thank you for your attention!

katarzyna.wreczycka@mdc-berlin.de