genomation

a toolkit to summarize, annotate and visualize genomic intervals

Altuna Akalın[1]

FMI

Friedrich Miescher Institute for Biomedical Research

February 24, 2014

[1]* presented by. Package developed by Altuna Akalın and Vedran Franke

Quick introduction

- The **genomation** is an **R** package that expedites genomic interval summary and annotation. It has the following features
  1. Annotation of genomic intervals: e.g. see what % of your intervals overlap with exon/intron/promoters

Quick introduction

- The **genomation** is an **R** package that expedites genomic interval summary and annotation. It has the following features
  1. Annotation of genomic intervals: e.g. see what % of your intervals overlap with exon/intron/promoters
  2. Summary of genomic scores or read coverages over pre-defined regions
     - e.g. extract the conservation profile over ChIP-seq binding sites (equi-width regions) or CpG islands (nonequi-width regions)

Quick introduction

- The **genomation** is an **R** package that expedites genomic interval summary and annotation. It has the following features
  1. Annotation of genomic intervals: e.g. see what % of your intervals overlap with exon/intron/promoters
  2. Summary of genomic scores or read coverages over pre-defined regions
     - e.g. extract the conservation profile over ChIP-seq binding sites (equi-width regions) or CpG islands (nonequi-width regions)
  3. Visualize genomic interval summaries as meta-region plots or heatmaps.

Quick introduction

- The **genomation** is an **R** package that expedites genomic interval summary and annotation. It has the following features
  1. Annotation of genomic intervals: e.g. see what % of your intervals overlap with exon/intron/promoters
  2. Summary of genomic scores or read coverages over pre-defined regions
     - e.g. extract the conservation profile over ChIP-seq binding sites (equi-width regions) or CpG islands (nonequi-width regions)
  3. Visualize genomic interval summaries as meta-region plots or heatmaps.
  4. Work with multiple file formats
     - e.g. BAM, BED, bigWig, GFF and generic tabular text files containing chromosome location information.

Quick introduction

- The **genomation** is an **R** package that expedites genomic interval summary and annotation. It has the following features
  1. Annotation of genomic intervals: e.g. see what % of your intervals overlap with exon/intron/promoters
  2. Summary of genomic scores or read coverages over pre-defined regions
     - e.g. extract the conservation profile over ChIP-seq binding sites (equi-width regions) or CpG islands (nonequi-width regions)
  3. Visualize genomic interval summaries as meta-region plots or heatmaps.
  4. Work with multiple file formats
     - e.g. BAM, BED, bigWig, GFF and generic tabular text files containing chromosome location information.
  5. do all these in **R** :)

Genomic interval summaries are widely used

- Summaries of genomic intervals are one of the useful ways to communicate high-dimensional data
- Traditionally, regions of interest are picked and distribution of genomic intervals are summarized on those regions

Genomic interval summaries are widely used: Examples from literature

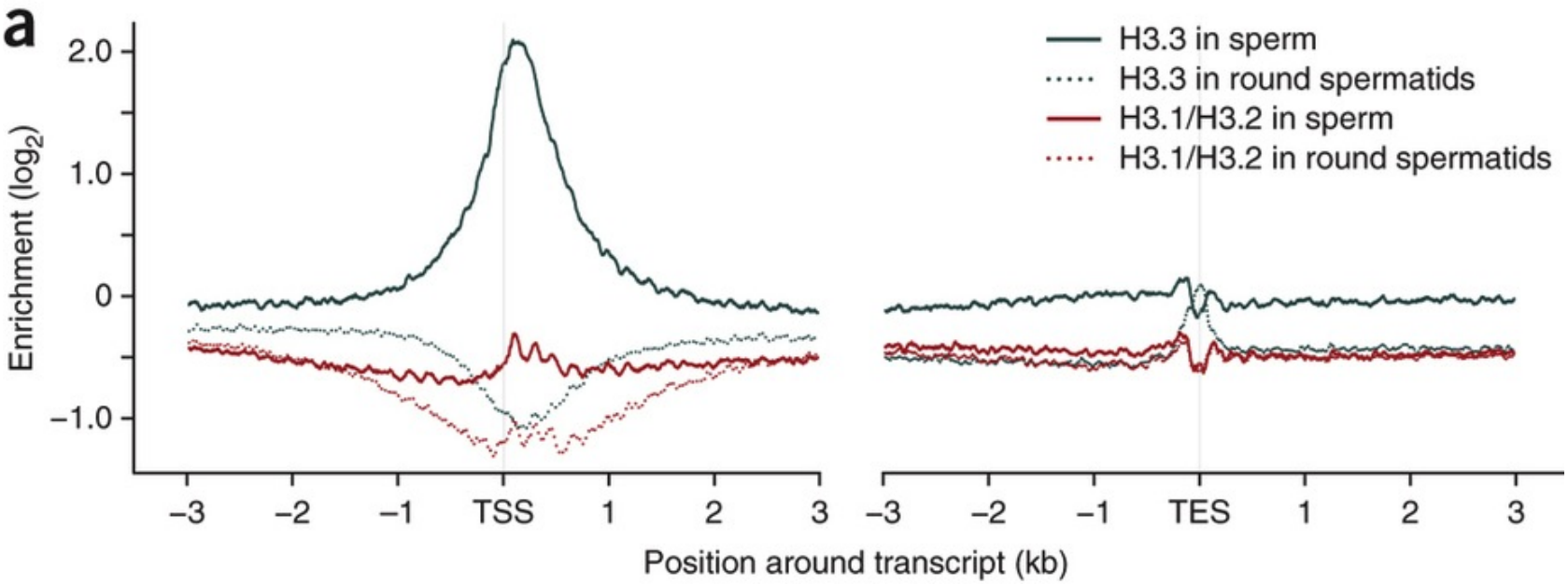


Figure : Erkek, S., et al. (2013). Molecular determinants of nucleosome retention at CpG-rich sequences in mouse spermatozoa. Nature Structural & Molecular Biology

Genomic interval summaries are widely used: Examples from literature

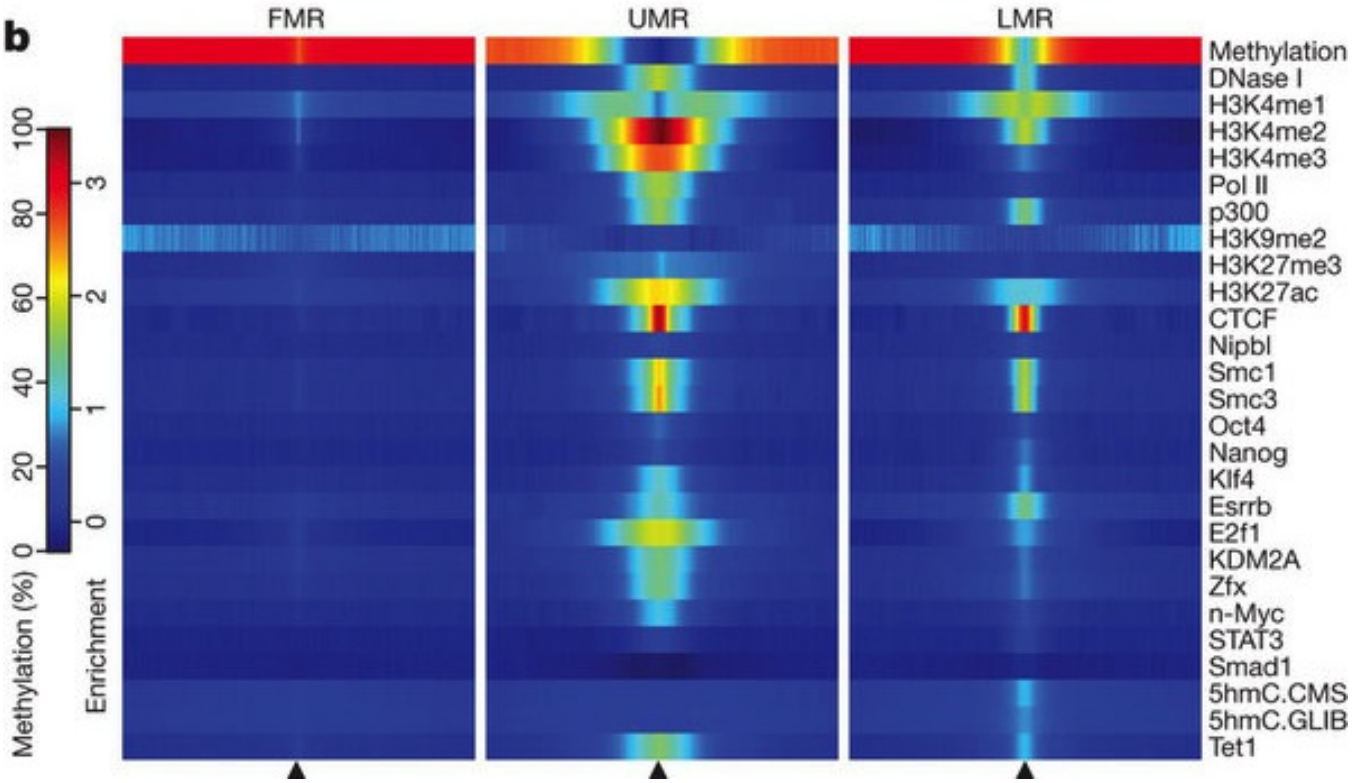
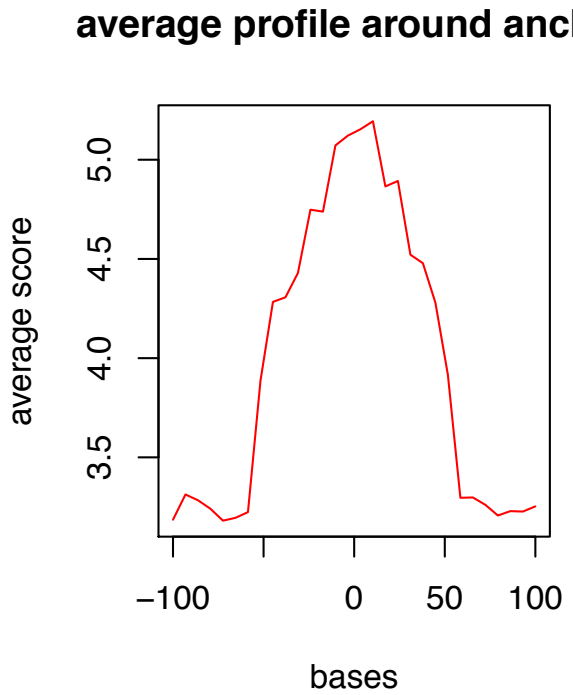


Figure : Stadler, M., Murr, R., Burger, L., et al. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature

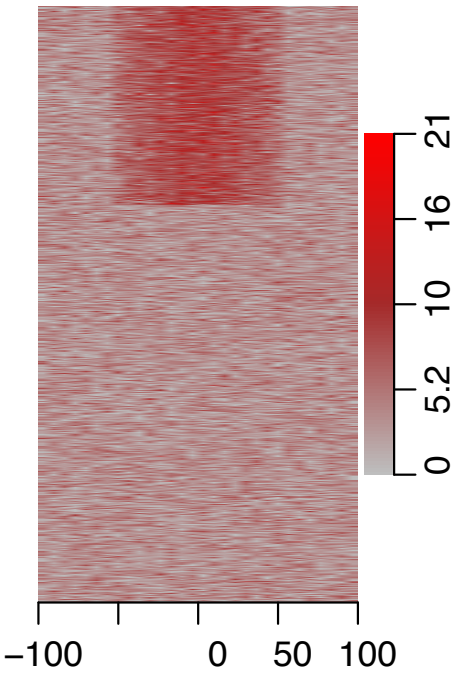
Utility and futility of average profiles

Does this mean all of the windows (viewpoints) have a similar enrichment profile?



Utility and futility of average profiles

Only 1/3 of windows have such enrichment. Be careful when you are interpreting the average profiles.



Genomic interval summaries are widely used: Examples from literature

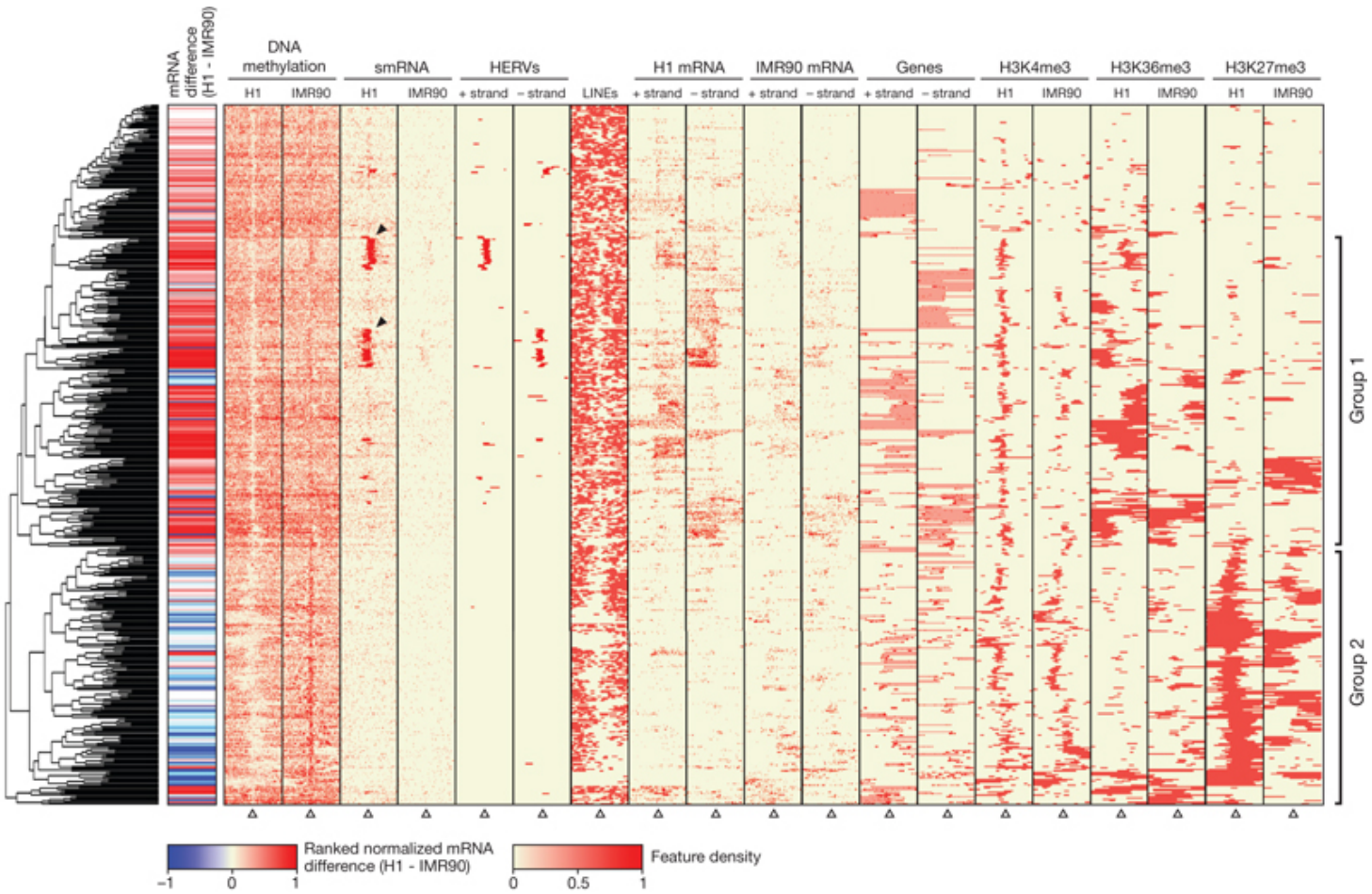


Figure : Lister, R., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature

Genomic interval summaries are widely used: Examples from literature

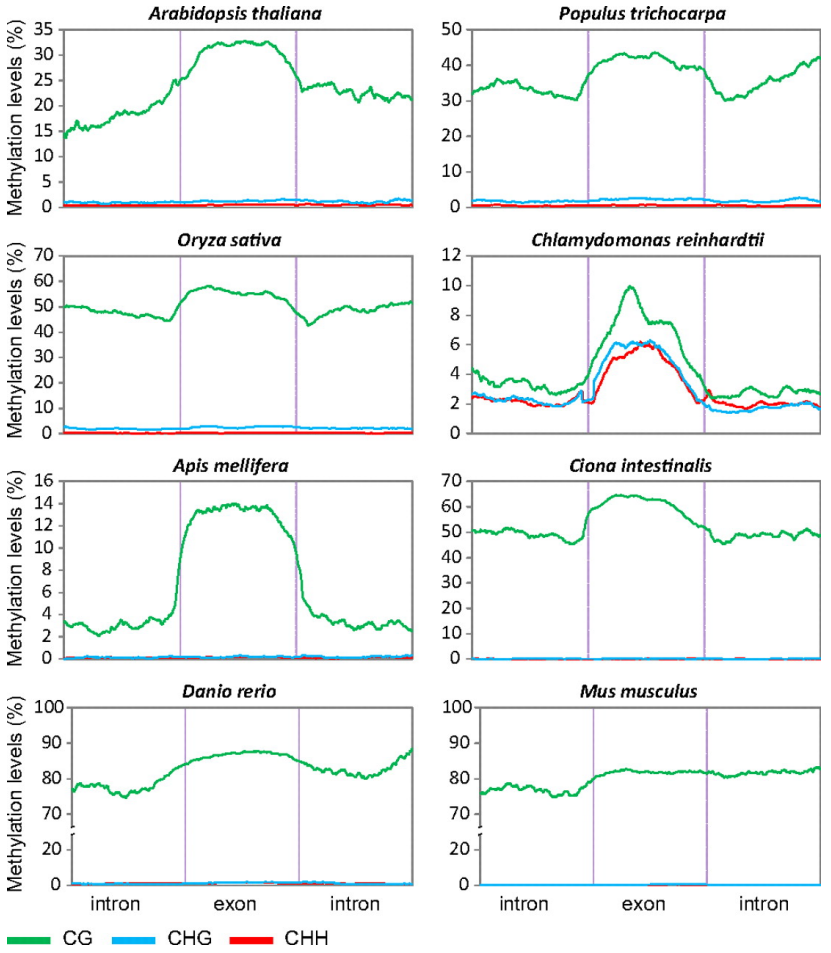
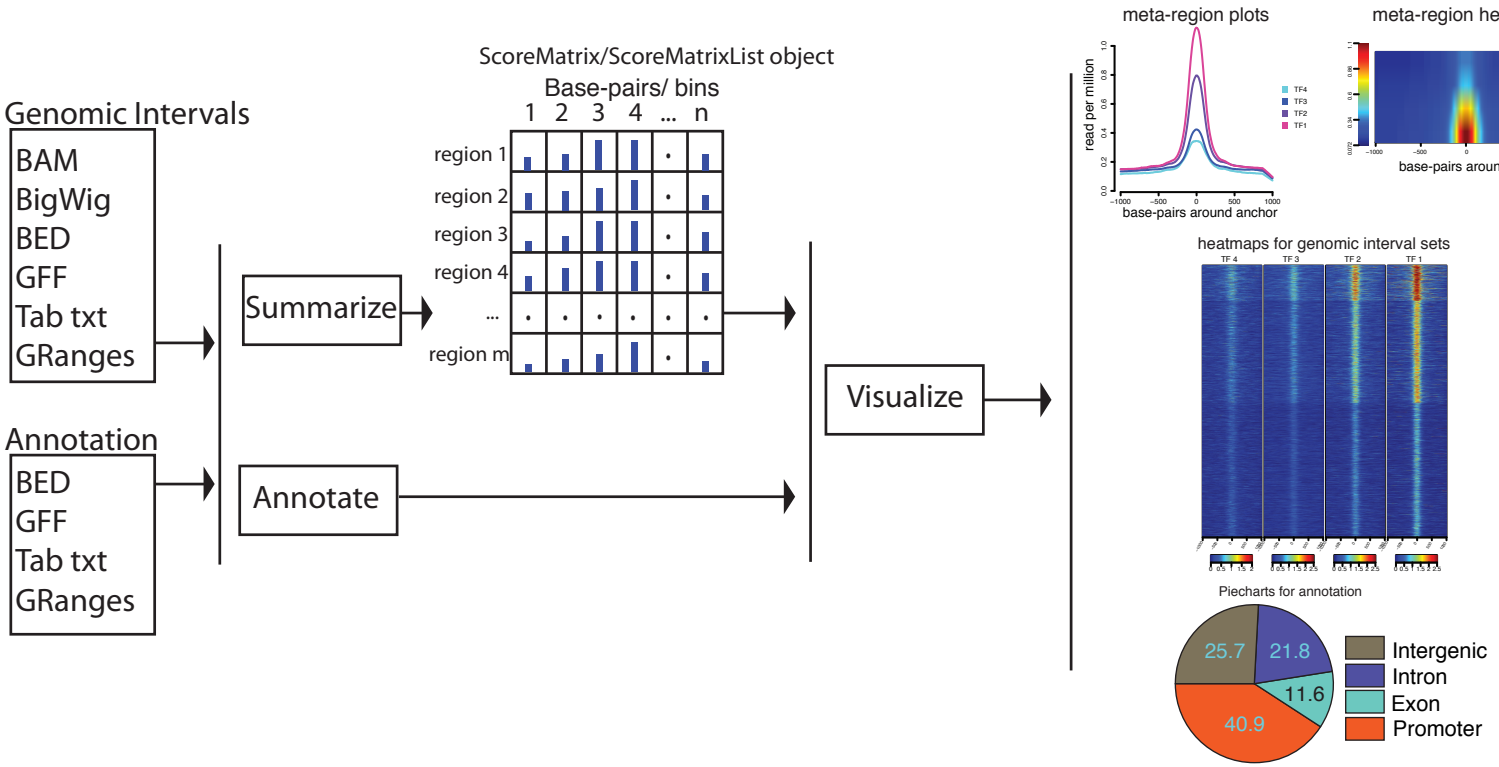


Figure : Feng, S. et al. (2010). Conservation and divergence of methylation patterning in plants and animals. PNAS

Issues to keep in mind when developing summary methods

- Genomic data comes in many formats, we need a method that is able to **work with multiple flat file formats**
- We need a method that is not specialized on one type of data set such as read counts, it should also work on **other scoring schemes**(e.g. conservation scores) easily.
- Regions of interest are not always equi-width, you should be able to **normalize for length differences by binning**.
- **Multiple visualization options** and fast heatmap generation should be available
- **Clustering of regions** based on multiple summaries (e.g. binding for different TFs on the same set of regions) on the heatmap
- **Ease of use**, it should not take hours of coding to generate and visualize summaries.

Overview of genomation features



installation of the package and the example data

We can install the package and the data using *install_github()* function from the *devtools* package.

```
#install dependencies
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("genomation", username = "al2na")

# install the data package
# needed for examples
install_github("genomationData", username = "al2na")
```

Data import

Various file formats can be used in genomation. You can read in annotation or your genomic intervals of interest.

```
library(genomation)
tab.file1 <- system.file("extdata/tab1.bed", package = "genomation")
readGeneric(tab.file1)

## GRanges with 6 ranges and 0 metadata columns:
##       seqnames              ranges strand
##          <Rle>           <IRanges>  <Rle>
##   [1]    chr21 [9437272, 9439473]      *
##   [2]    chr21 [9483485, 9484663]      *
##   [3]    chr21 [9647866, 9648116]      *
##   [4]    chr21 [9708935, 9709231]      *
##   [5]    chr21 [9825442, 9826296]      *
##   [6]    chr21 [9909011, 9909218]      *
##   ---
##   seqlengths:
##    chr21
##       NA
```

Extraction of data over pre-defined genomic regions

ScoreMatrix() and **ScoreMatrixBin()** are functions used to extract data over predefined windows.

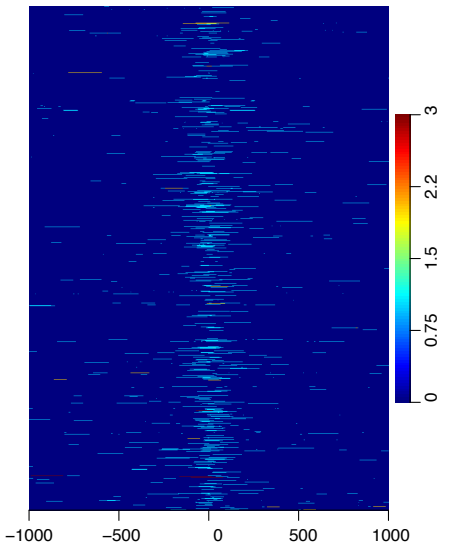
- ScoreMatrix is used when all of the windows have the same width (e.g. region around TSS)
- ScoreMatrixBin is designed for use with windows of unequal width (e.g. enrichment of methylation over exons).

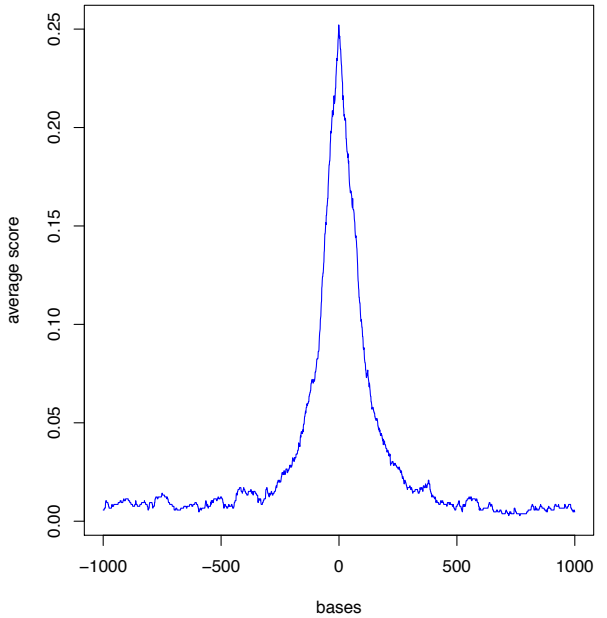
```
data(cage)
data(promoters)
sm <- ScoreMatrix(target = cage, windows = promoters)
sm

##  scoreMatrix with dims:  1055 2001
```

Visualizing ScoreMatrix: summary of genomic invervals over pre-defined regions

plotMeta(),**heatMeta()**, **heatMatrix()** and **multiHeatMatrix()** are the visualization functions.

```
oldmar <- par()$mar
par(oma = c(0, 0, 0, 0))
heatMatrix(sm, xcoords = c(-1000, 1000))
plotMeta(sm, xcoords = c(-1000, 1000),line.col="blue")
par(oma = oldmar)
```





Working with BAM files

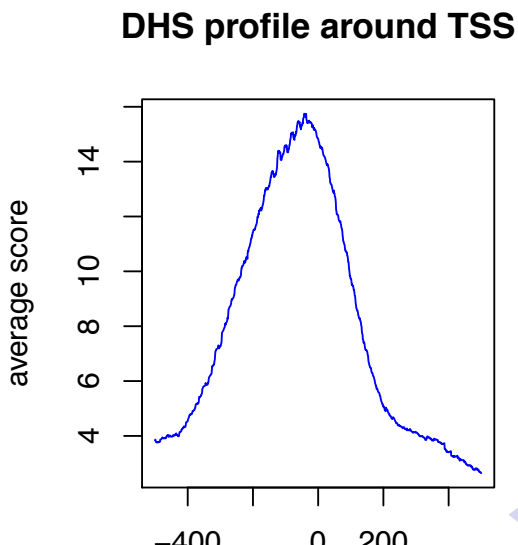
BAM files can also be used in ScoreMatrix() and ScoreMatrixBin() functions

```
bam.file = system.file('tests/test.bam', package='genomation')
windows = GRanges(rep(c(1,2),each=2),
          IRanges(rep(c(1,2), times=2), width=5))
scores3 = ScoreMatrix(target=bam.file,windows=windows, type='bam')
```

Working with bigWig files

ScoreMatrix() and **ScoreMatrixBin()** are functions can handle bigWig files. Here we use ENCODE DHS scores, downloaded from `http://goo.gl/fEVu0g`

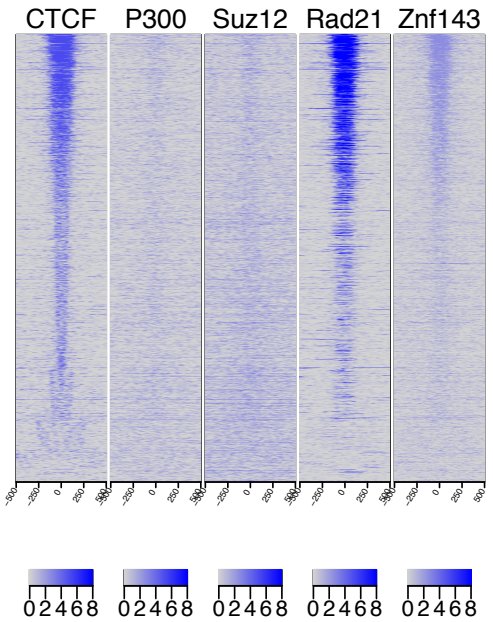
```
my.bed12.file=system.file("extdata/chr21.refseq.hg19.bed",
                          package = "genomation")
feats=readTranscriptFeatures(my.bed12.file,up.flank=500,down.flank=500)
sm=ScoreMatrix(target="wgEncodeUwDnaseA549RawRep1.bw",
     windows=feats$promoters,type='bigWig',strand.aware=TRUE)
plotMeta(sm,xcoords=c(-500,500),main="DHS profile around TSS",
         line.col="blue")
```



Multiple profiles

Multiple heatmap profiles can be plotted using **multiHeatMatrix()** which takes in a **ScoreMatrixList** object. Here we used CTCF , P300 , Suz12 ,Rad21, Znf143 BAM files from genomationData package.

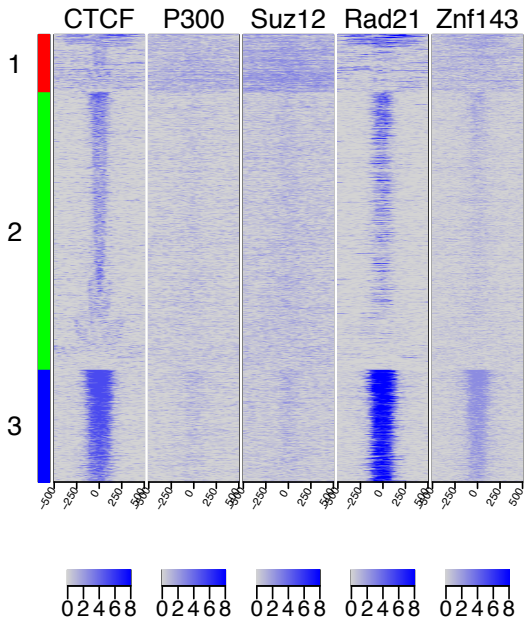
```
ctcf.peaks=readRDS("ctcf.peaks.rds")
dataPath = system.file("extdata", package = "genomationData")
bam.files = list.files(dataPath, full= T,pattern = "bam$")[c(1:4,6)]
sml = ScoreMatrixList(bam.files, ctcf.peaks, bin.num = 50,type = "bam")
names(sml)=c("CTCF","P300","Suz12","Rad21","Znf143")
multiHeatMatrix(sml, xcoords = c(-500, 500),cex.axis=0.35,common.scale = T,
                col = c("lightgray", "blue"),winsorize=c(0,95))
```



Multiple profiles

Multiple heatmap profiles can be plotted using **multiHeatMatrix()** which takes in a **ScoreMatrixList** object. Here we used CTCF , P300 , Suz12 ,Rad21, Znf143 BAM files from genomationData package.

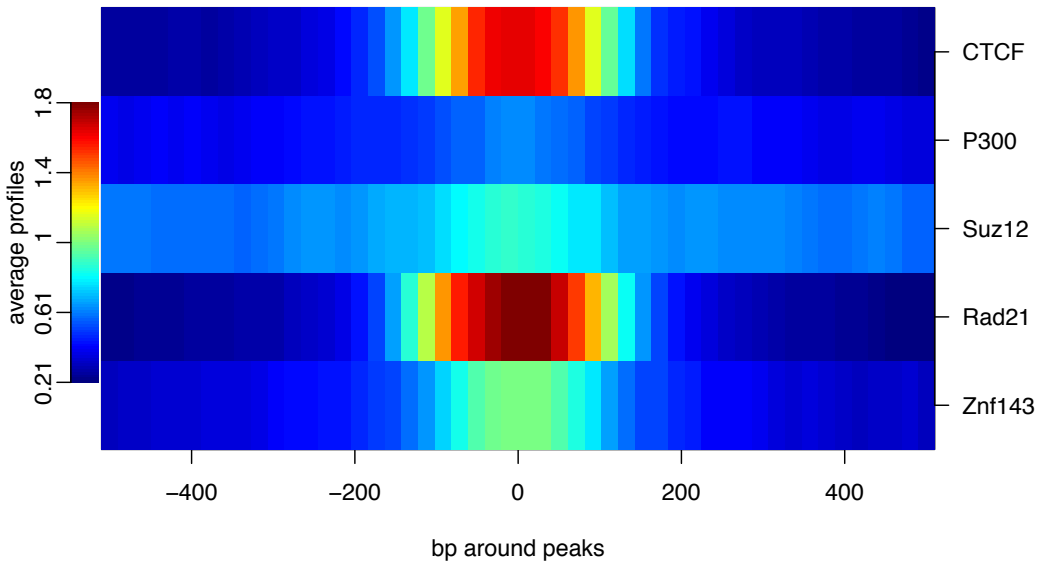
```
multiHeatMatrix(sml, xcoords = c(-500, 500),kmeans=TRUE,k=3,common.scale = T,
                cex.axis=0.4,col = c("lightgray", "blue"),winsorize=c(0,95))
```



Multiple profiles

Multiple heatmap profiles can be plotted using **multiHeatMatrix()** which takes in a **ScoreMatrixList** object. Here we used CTCF , P300 , Suz12 ,Rad21, Znf143 BAM files from genomationData package.

```
# take log2 of all matrices
sml2=scaleScoreMatrixList(sml,scalefun=function(x) log2(x+1))
heatMeta(sml2,legend.name="average profiles",xcoords=c(-500, 500),
         xlab="bp around peaks")
```



Future work...

- Explore overlap statistics between two genomic data sets: Does TF1 binding site locations overlap with TF2 sites more than expected?
- This is previously explored with GenometriCorr package. These functionality can be included in the form of a dependency.
- Performance improvement on certain functions, faster is always better...

Further information

- The genomation package is available at `http:/al2na.github.io/genomation`. You can find the link to the vignette on the webpage as well.
- Code that generated this presentation is available at `http://github.com/al2na/genomation_presentation`
- **Questions and bug reports**
  - You can view/open issues in github `https://github.com/al2na/genomation/issues?state=open`
  - You can ask questions by sending an e-mail to **genomation@googlegroups.com** or using the web interface to *google groups*
- Developed by Altuna Akalın and Vedran Franke

Session Info


```
sessionInfo()

## R version 3.0.2 (2013-09-25)
## Platform: x86_64-apple-darwin10.8.0 (64-bit)
##
## locale:
## [1] C
##
## attached base packages:
## [1] methods   grid      stats     graphics  grDevices utils     datas
## [8] base
##
## other attached packages:
## [1] genomation_0.99.0.2 knitr_1.5
##
## loaded via a namespace (and not attached):
##  [1] BSgenome_1.30.0      BiocGenerics_0.8.0   Biostrings_2.30.0
##  [4] GenomicRanges_1.14.3 IRanges_1.20.5       MASS_7.3-29
##  [7] RColorBrewer_1.0-5   RCurl_1.95-4.1       Rsamtools_1.14.1
## [10] XML_3.95-0.2         XVector_0.2.0        bitops_1.0-6
## [13] colorspace_1.2-4     data.table_1.8.10    dichromat_2.0-0
## [16] digest_0.6.3         evaluate_0.5.1       formatR_0.10
## [19] ggplot2_0.9.3.1      gridBase_0.4-6       gtable_0.1.2
```