Working with Genomic Intervals - Practical

MRC CSC Bioinformatics Core Team

15-16 December 2015

Exercises

1. `hg19Genes.txt` contains gene coordinates for Human genome hg19. Read the contents of `hg19Genes.txt` and create a GRanges object.

```
library(GenomicRanges)
hg19Gene <- read.table("hg19Genes.txt",sep="\t",header=T)

# add 'chr' prefix to chromosome name
hg19Gene$ChromosomeName <- paste("chr",hg19Gene$ChromosomeName,sep="")

hg19Gene.GR <- GRanges(seqnames=hg19Gene$ChromosomeName,
                       ranges=IRanges(start=hg19Gene$GeneStart,end=hg19Gene$GeneEnd),
                       strand=ifelse(hg19Gene$Strand==1,"+","-"),
                       EnsemblID=hg19Gene$EnsemblGeneID)
hg19Gene.GR
## GRanges object with 64162 ranges and 1 metadata column:
##                  seqnames               ranges strand   |       EnsemblID
##                     <Rle>            <IRanges>  <Rle>   |        <factor>
##       [1]          chr13 [23551994, 23552136]      -   | ENSG00000223116
##       [2]          chr13 [23708313, 23708703]      +   | ENSG00000233440
##       [3]          chr13 [23726725, 23726825]      -   | ENSG00000207157
##       [4]          chr13 [23743974, 23744736]      -   | ENSG00000229483
##       [5]          chr13 [23791571, 23791673]      -   | ENSG00000252952
##       ...            ...                  ...    ... ...             ...
##   [64158]     chrLRG_239 [  108535,   112904]      +   |         LRG_239
##   [64159]      chrLRG_24 [    5001,     9486]      +   |          LRG_24
##   [64160]     chrLRG_241 [    5001,   139211]      +   |         LRG_241
##   [64161]     chrLRG_243 [    5001,    44469]      +   |         LRG_243
##   [64162] chrHG991_PATCH [66119285, 66465398]      +   | ENSG00000261657
##   -------
##   seqinfo: 690 sequences from an unspecified genome; no seqlengths
```

2. Filter the above GRanges object for genes in chr1:1544000-2371000

```
chr1genes <- hg19Gene.GR[seqnames(hg19Gene.GR)=="chr1" &
                         start(hg19Gene.GR) > 1544000 &
                           end(hg19Gene.GR) < 2371000]
head(chr1genes)
## GRanges object with 6 ranges and 1 metadata column:
##       seqnames             ranges strand |       EnsemblID
##          <Rle>          <IRanges>  <Rle> |        <factor>
##   [1]     chr1 [1550795, 1565990]      + | ENSG00000197530
```

```
##   [2]     chr1 [1567474, 1570639]      + | ENSG00000189409
##   [3]     chr1 [1570603, 1590473]      - | ENSG00000248333
##   [4]     chr1 [1592939, 1624167]      - | ENSG00000189339
##   [5]     chr1 [1634169, 1655766]      - | ENSG00000008128
##   [6]     chr1 [1656277, 1677431]      - | ENSG00000215790
##   -------
##   seqinfo: 690 sequences from an unspecified genome; no seqlengths

# alternate
chr1genes <- subset(hg19Gene.GR,start>1544000 & end<2371000 & seqnames=="chr1")
```

3. Create a GRanges of Transcription start sites (1 bp range) for the GRanges object created in Q1.

- How to identify TSS for genes in forward/reverse strand?

```
hg19Gene$TSS <- ifelse(hg19Gene$Strand==1,hg19Gene$GeneStart,hg19Gene$GeneEnd)

hg19TSS <- GRanges(seqnames=hg19Gene$ChromosomeName,
                       ranges=IRanges(start=hg19Gene$TSS,end=hg19Gene$TSS),
                       strand=ifelse(hg19Gene$Strand==1,"+","-"),
                       EnsemblID=hg19Gene$ensembl_gene_id)
hg19TSS
## GRanges object with 64162 ranges and 0 metadata columns:
##                  seqnames              ranges strand
##                     <Rle>           <IRanges>  <Rle>
##       [1]          chr13 [23552136, 23552136]      -
##       [2]          chr13 [23708313, 23708313]      +
##       [3]          chr13 [23726825, 23726825]      -
##       [4]          chr13 [23744736, 23744736]      -
##       [5]          chr13 [23791673, 23791673]      -
##       ...            ...                 ...    ...
##   [64158]     chrLRG_239 [  108535,   108535]      +
##   [64159]      chrLRG_24 [    5001,     5001]      +
##   [64160]     chrLRG_241 [    5001,     5001]      +
##   [64161]     chrLRG_243 [    5001,     5001]      +
##   [64162] chrHG991_PATCH [66119285, 66119285]      +
##   -------
##   seqinfo: 690 sequences from an unspecified genome; no seqlengths
```

4. Create a GRanges object of human promoters with TSS ± 1000bp (using the GRanges object created in Q1). Tip: Read the documentation for `promoters` function.

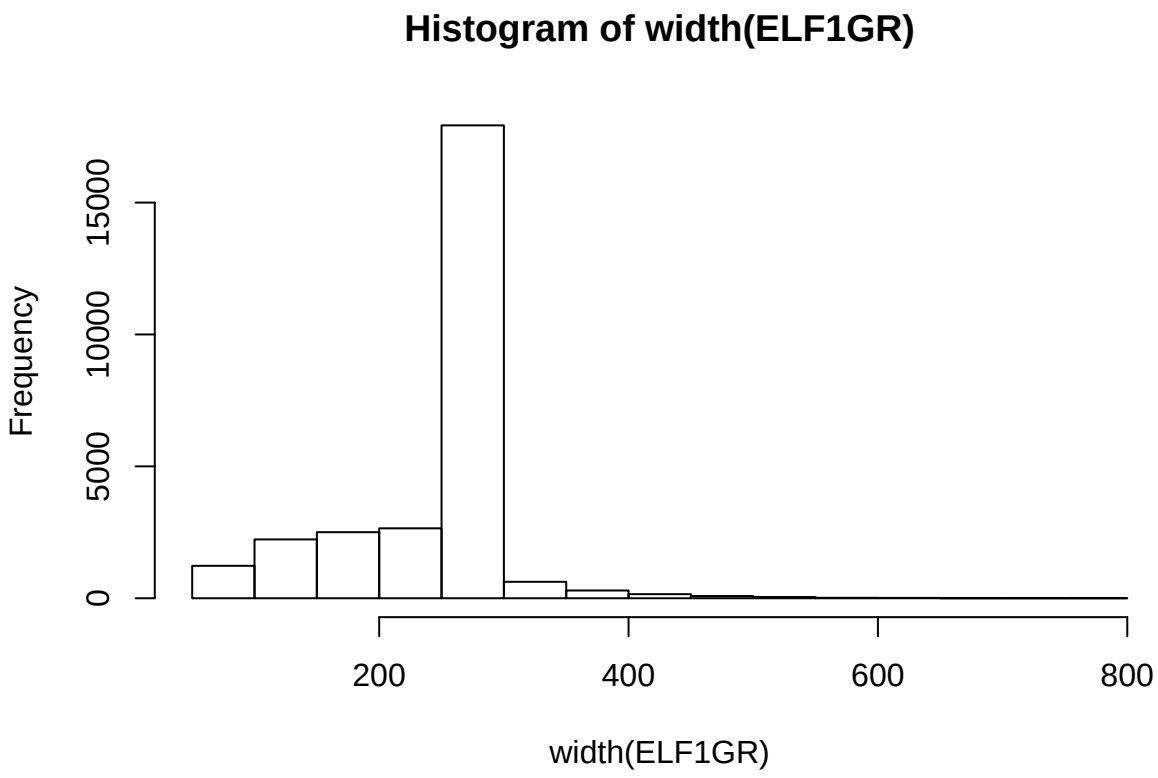
```
hg19Promoters <- promoters(hg19Gene.GR,upstream=1000,downstream=1000)
hg19Promoters
## GRanges object with 64162 ranges and 1 metadata column:
##                  seqnames              ranges strand   |       EnsemblID
##                     <Rle>           <IRanges>  <Rle>   |        <factor>
```

```
##       [1]           chr13 [23551137, 23553136]      -   | ENSG00000223116
##       [2]           chr13 [23707313, 23709312]      +   | ENSG00000233440
##       [3]           chr13 [23725826, 23727825]      -   | ENSG00000207157
##       [4]           chr13 [23743737, 23745736]      -   | ENSG00000229483
##       [5]           chr13 [23790674, 23792673]      -   | ENSG00000252952
##       ...             ...                 ...    ... ...             ...
##   [64158]     chrLRG_239 [  107535,   109534]      +   |         LRG_239
##   [64159]      chrLRG_24 [    4001,     6000]      +   |          LRG_24
##   [64160]     chrLRG_241 [    4001,     6000]      +   |         LRG_241
##   [64161]     chrLRG_243 [    4001,     6000]      +   |         LRG_243
##   [64162] chrHG991_PATCH [66118285, 66120284]      +   | ENSG00000261657
##   -------
##   seqinfo: 690 sequences from an unspecified genome; no seqlengths
```

5. Import ELF1 binding sites in K562 cell from Encode (ELF1_K562.bed) and create GRanges object.

- Import the ELF1 binding sites using `import.bed()` function from `rtracklayer` package and compare it with the above GRanges object
- Check the distribution of width of ELF1 binding sites using `hist()`
- Identify promoters (TSS ± 1kb) with ELF1 binding sites `findOverlaps()` and using `%over`
- Remember BED format uses 0-based coordinates

```
library("rtracklayer")
ELF1 <- read.table("ELF1_K562.bed",sep="\t",header=F)
ELF1GR <- GRanges(seqnames=ELF1$V1, IRanges(start=ELF1$V2+1,end=ELF1$V3))
ELF1GR_A <- import.bed("ELF1_K562.bed")

# Distribution of width of ELF1 binding sites
hist(width(ELF1GR))
```




```
# ELF1 binding sites overlap with promoters using `findOverlaps`
hg19Promoters <- promoters(hg19Gene.GR,upstream=1000,downstream=1000)
Promoter_ELF1_overlap <- findOverlaps(hg19Promoters,ELF1GR,ignore.strand=T)
Promoter_ELF1_overlap.m <- as.matrix(Promoter_ELF1_overlap)

Promoter_ELF1 <- hg19Promoters[Promoter_ELF1_overlap.m[,"queryHits"],]
Promoter_ELF1
## GRanges object with 15207 ranges and 1 metadata column:
##           seqnames              ranges strand   |       EnsemblID
##              <Rle>           <IRanges>  <Rle>   |        <factor>
##       [1]    chr13 [24039710, 24041709]      +   | ENSG00000232977
##       [2]    chr13 [24462028, 24464027]      +   | ENSG00000205861
##       [3]    chr13 [96328180, 96330179]      -   | ENSG00000247400
##       [4]    chr13 [25562064, 25564063]      -   | ENSG00000232858
##       [5]    chr13 [99228498, 99230497]      +   | ENSG00000224418
##       ...      ...                 ...    ... ...             ...
##   [15203]     chr3 [13691196, 13693195]      +   | ENSG00000224514
##   [15204]     chr3 [13973553, 13975552]      +   | ENSG00000250439
##   [15205]     chr3 [14185223, 14187222]      +   | ENSG00000228242
##   [15206]    chr10 [23002485, 23004484]      -   | ENSG00000150867
##   [15207]     chr2 [61764762, 61766761]      -   | ENSG00000082898
##   -------
##   seqinfo: 690 sequences from an unspecified genome; no seqlengths

# ELF1 binding sites overlap with promoters using `%over%`
ELF1_promoters1 <- hg19Promoters[hg19Promoters %over% ELF1GR]
ELF1_promoters1
## GRanges object with 11918 ranges and 1 metadata column:
##           seqnames              ranges strand   |       EnsemblID
```

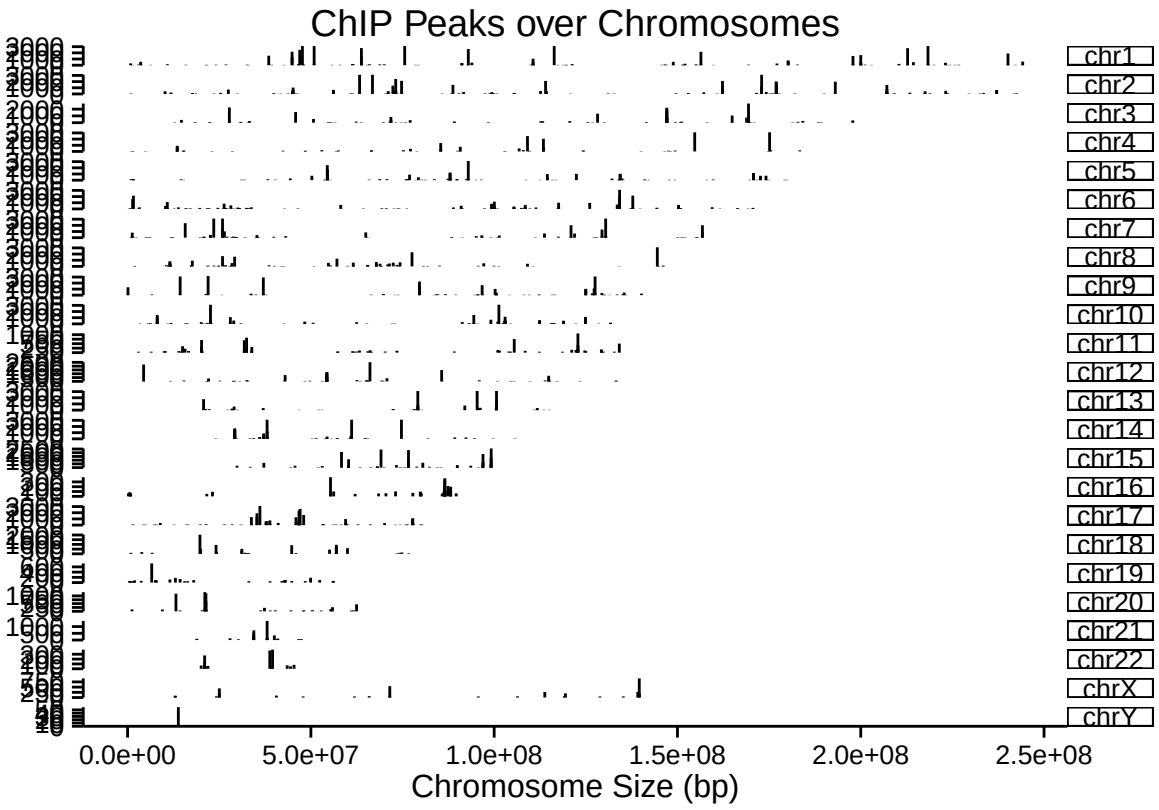
```
##              <Rle>            <IRanges>  <Rle>   |        <factor>
##       [1]    chr13 [24039710, 24041709]      +   | ENSG00000232977
##       [2]    chr13 [24462028, 24464027]      +   | ENSG00000205861
##       [3]    chr13 [96328180, 96330179]      -   | ENSG00000247400
##       [4]    chr13 [25562064, 25564063]      -   | ENSG00000232858
##       [5]    chr13 [99228498, 99230497]      +   | ENSG00000224418
##       ...      ...                 ...    ... ...             ...
##   [11914]     chr3 [13691196, 13693195]      +   | ENSG00000224514
##   [11915]     chr3 [13973553, 13975552]      +   | ENSG00000250439
##   [11916]     chr3 [14185223, 14187222]      +   | ENSG00000228242
##   [11917]    chr10 [23002485, 23004484]      -   | ENSG00000150867
##   [11918]     chr2 [61764762, 61766761]      -   | ENSG00000082898
##   -------
##   seqinfo: 690 sequences from an unspecified genome; no seqlengths
```

Note the differences in the outputs!

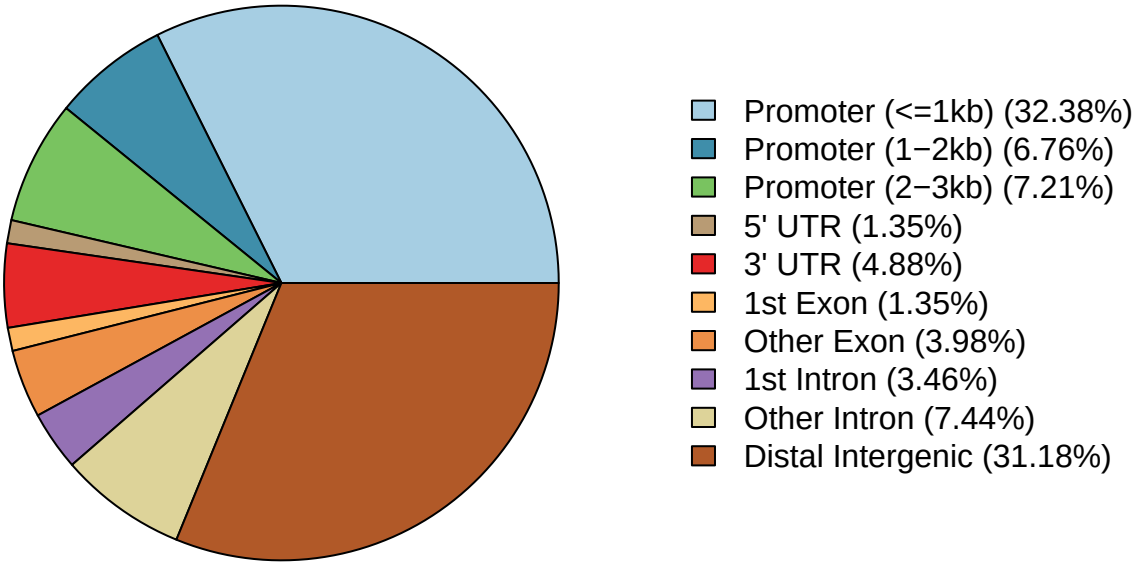
6. Import CBX6_BF peaks from `ChIPseeker` package and visualise distribution of peaks along the chromosomes. Annotate the peaks with respect to genomic regions and visualise the distribution in pie chart and bar chart.

- Get locations of files using `files <- getSampleFiles()`

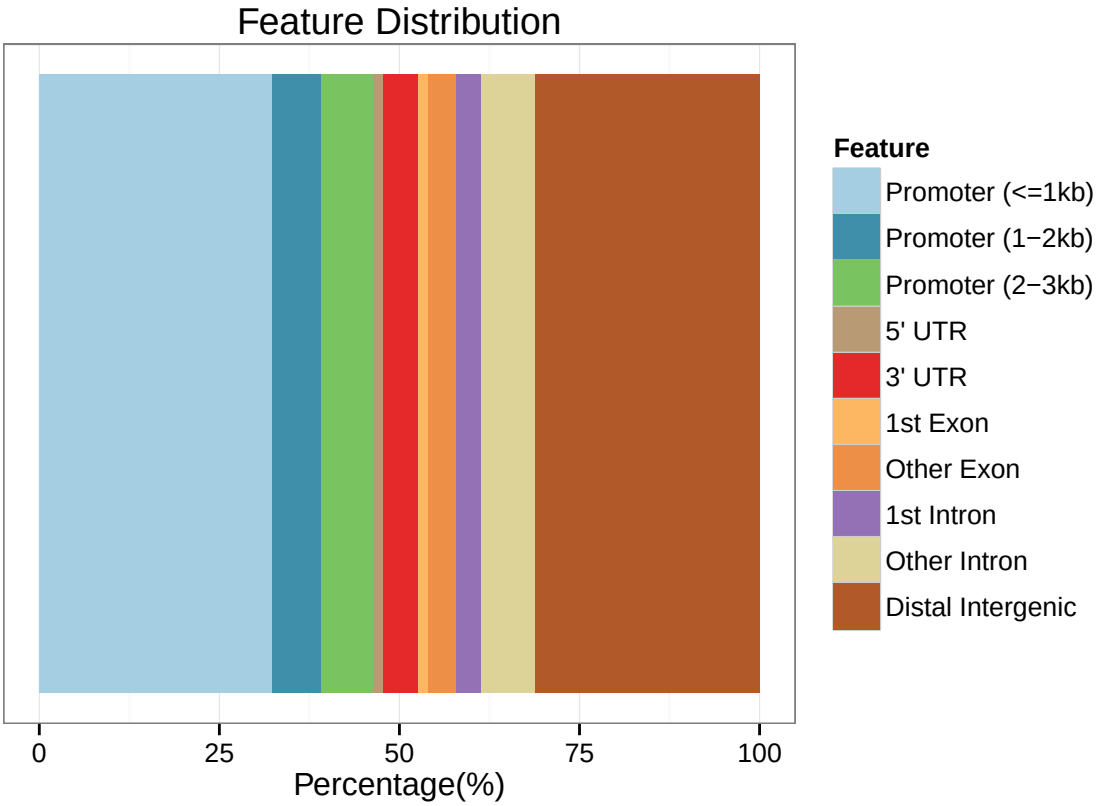
```
library(ChIPseeker)
library(TxDb.Hsapiens.UCSC.hg19.knownGene)
txdb <- TxDb.Hsapiens.UCSC.hg19.knownGene

## Sample files
files <- getSampleFiles()
peak <- readPeakFile(files[[4]])
covplot(peak, weightCol="V5")
```



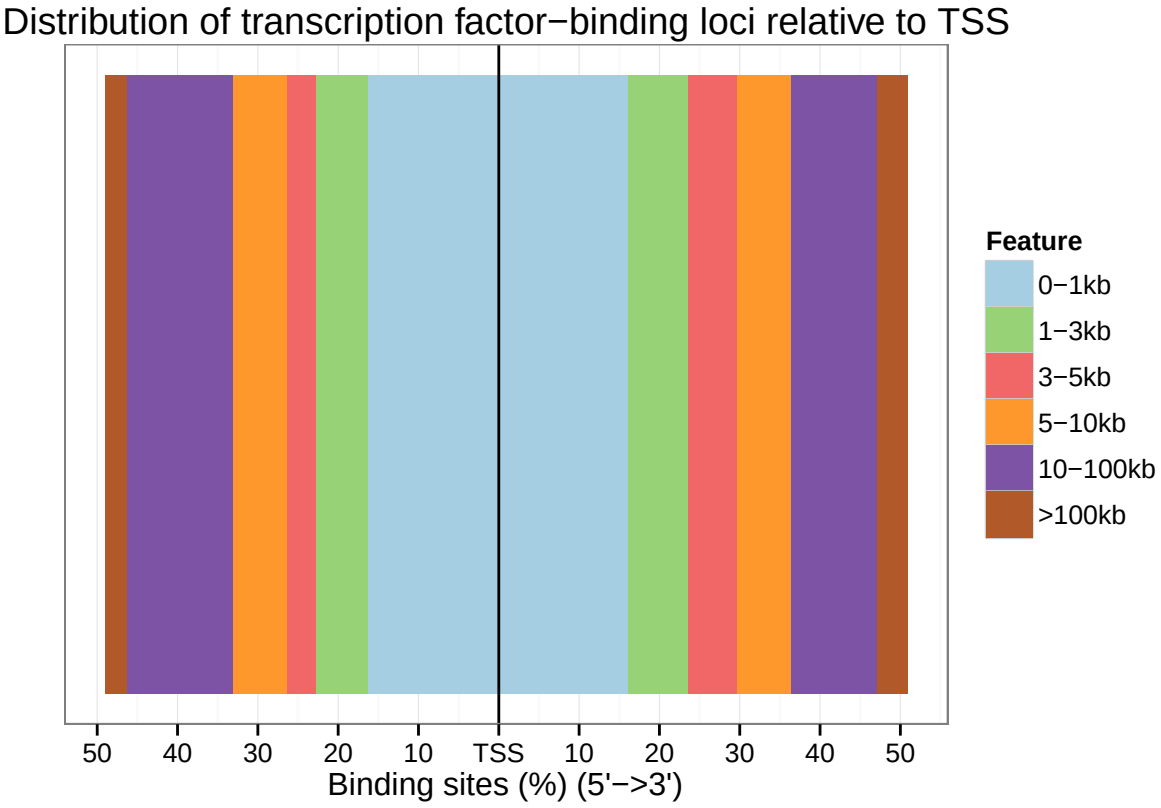

```
# Reading the ARmo_OM peak file
peakAnno <- annotatePeak(files[[4]], tssRegion=c(-3000, 3000),TxDb=txdb, annoDb="org.Hs.eg.db")
## >> loading peak file...                2015-12-14 12:59:30
## >> preparing features information...      2015-12-14 12:59:30
## >> identifying nearest features...        2015-12-14 12:59:31
## >> calculating distance from peak to TSS...   2015-12-14 12:59:32
## >> assigning genomic annotation...        2015-12-14 12:59:32
## >> adding gene annotation...          2015-12-14 13:00:03
## >> assigning chromosome lengths           2015-12-14 13:00:03
## >> done...                     2015-12-14 13:00:03

## Pie chart
plotAnnoPie(peakAnno)
```




```
## Bar chart
plotAnnoBar(peakAnno)
```




```
## Plot distance between Peaks and TSS
plotDistToTSS(peakAnno)
```


Distribution of transcription factor−binding loci relative to TSS
50
40
30
20
10
TSS
10
20
30
40
50
Binding sites (%) (5'−>3')
Feature
0−1kb
1−3kb
3−5kb
5−10kb
10−100kb
>100kb