The chromstaR user's guide

Aaron Taudt*

July 17, 2017

Contents

*aaron.taudt@gmail.com

1 Introduction

ChIP-seq has become the standard technique for assessing the genome-wide chromatin state of DNA. *chromstaR* provides functions for the joint analysis of multiple ChIP-seq samples. It allows peak calling for transcription factor binding and histone modifications with a narrow (e.g. H3K4me3, H3K27ac, ...) or broad (e.g. H3K36me3, H3K27me3, ...) profile. All analysis can be performed on each sample individually (=univariate), or in a joint analysis considering all samples simultaneously (=multivariate).

2 Outline of workflow

Every analysis with the *chromstaR* package starts from aligned reads in either BAM or BED format. In the first step, the genome is partitioned into non-overlapping, equally sized bins and the reads that fall into each bin are counted. These read counts serve as the basis for both the univariate and the multivariate peak- and broad-region calling. Univariate peak calling is done by fitting a three-state Hidden Markov Model to the binned read counts. Multivariate peak calling for $\mathcal{S}$ samples is done by fitting a $2^{\mathcal{S}}$-state Hidden Markov Model to all binned read counts.

3 Univariate analysis

3.1 Task 1: Peak calling for a narrow histone modification

Examples of histone modifications with a narrow profile are H3K4me3, H3K9ac and H3K27ac in most human tissues. For such peak-like modifications, the bin size should be set to a value between 200bp and 1000bp.

```
library(chromstaR)
```

```
## === Step 1: Binning ===
# Get an example BAM file
file <- system.file("extdata","euratrans","lv-H3K4me3-BN-male-bio2-tech1.bam",
                    package="chromstaRData")
# Get the chromosome lengths (see ?GenomeInfoDb::fetchExtendedChromInfoFromUCSC)
# This is only necessary for BED files. BAM files are handled automatically.
data(rn4_chrominfo)
head(rn4_chrominfo)
```

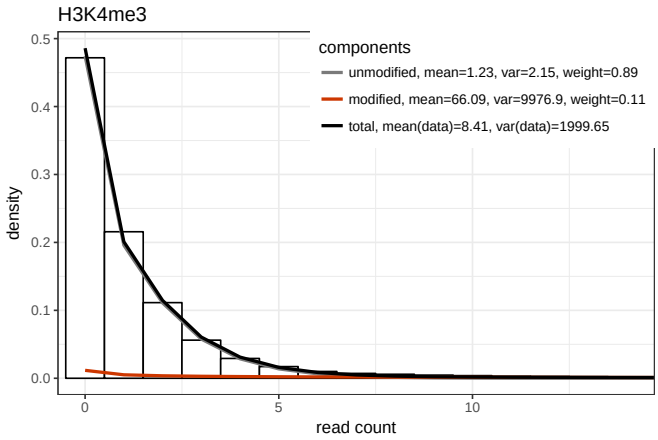
```
##   chromosome   length
## 1       chrM    16300
## 2      chr12 46782294
## 3      chr20 55268282
## 4      chr19 59218465
## 5      chr18 87265094
## 6      chr11 87759784
```

```
# We use bin size 1000bp and chromosome 12 to keep the example quick
binned.data <- binReads(file, assembly=rn4_chrominfo, binsizes=1000,
                        stepsizes=500, chromosomes='chr12')
print(binned.data)
```

```
## GRanges object with 46782 ranges and 1 metadata column:
##           seqnames               ranges strand |   counts
##              <Rle>            <IRanges>  <Rle> | <matrix>
##       [1]    chr12         [   1, 1000]      * |      0:0
##       [2]    chr12         [1001, 2000]      * |      0:0
##       [3]    chr12         [2001, 3000]      * |      0:0
##       [4]    chr12         [3001, 4000]      * |      0:0
##       [5]    chr12         [4001, 5000]      * |      0:0
##       ...      ...                  ...    ... .      ...
##   [46778]    chr12 [46777001, 46778000]      * |      2:3
##   [46779]    chr12 [46778001, 46779000]      * |      1:0
##   [46780]    chr12 [46779001, 46780000]      * |      0:0
##   [46781]    chr12 [46780001, 46781000]      * |      2:3
##   [46782]    chr12 [46781001, 46782000]      * |      1:0
##   -------
##   seqinfo: 22 sequences from an unspecified genome
```

```
## === Step 2: Peak calling ===
model <- callPeaksUnivariate(binned.data, verbosity=0)
```

```
## === Step 3: Checking the fit ===
# For a narrow modification, the fit should look something like this,
# with the 'modified'-component near the bottom of the figure
plotHistogram(model) + ggtitle('H3K4me3')
```




```
## === Step 4: Working with peaks ===
# Get the number and average size of peaks
length(model$peaks); mean(width(model$peaks))
```

```
## [1] 1245
## [1] 4008.434
```

```
# Change the false discovery rate and get number of peaks
model <- changeFDR(model, fdr=1e-4)
length(model$peaks); mean(width(model$peaks))
```

```
## [1] 913
## [1] 4861.993
```

```
## === Step 5: Export to genome browser ===
# We can export peak calls and binned read counts with
exportPeaks(model, filename=tempfile())
exportCounts(model, filename=tempfile())
```

!! It is important that the distributions are fitted correctly !! Please check section 6.4 for examples of how this plot should *not* look like and what can be done to get a correct fit.

3.2 Task 2: Peak calling for a broad histone modification

Examples of histone modifications with a broad profile are H3K9me3, H3K27me3, H3K36me3, H4K20me1 in most human tissues. These modifications usually cover broad domains of the genome, and the enrichment is best captured with a bin size between 500bp and 2000bp.

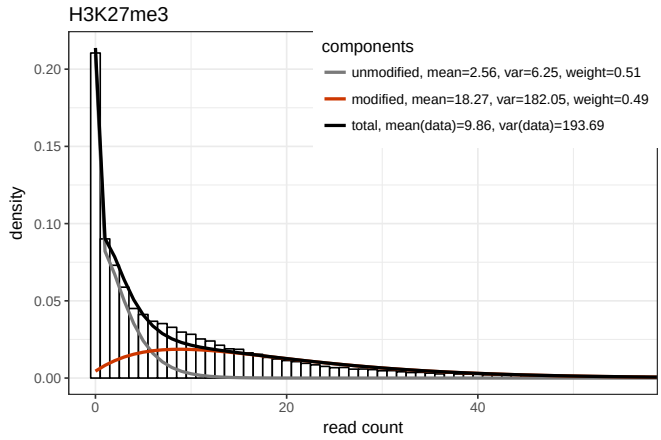
```
library(chromstaR)
```

```
## === Step 1: Binning ===
# Get an example BAM file
file <- system.file("extdata","euratrans","lv-H3K27me3-BN-male-bio2-tech1.bam",
                    package="chromstaRData")
# Get the chromosome lengths (see ?GenomeInfoDb::fetchExtendedChromInfoFromUCSC)
# This is only necessary for BED files. BAM files are handled automatically.
data(rn4_chrominfo)
head(rn4_chrominfo)
# We use bin size 1000bp and chromosome 12 to keep the example quick
binned.data <- binReads(file, assembly=rn4_chrominfo, binsizes=1000,
                        stepsizes=500, chromosomes='chr12')
```

```
## === Step 2: Peak calling ===
model <- callPeaksUnivariate(binned.data, verbosity=0)
```

```
## === Step 3: Checking the fit ===
# For a broad modification, the fit should look something like this,
# with a 'modified'-component that fits the thick tail of the distribution.
```

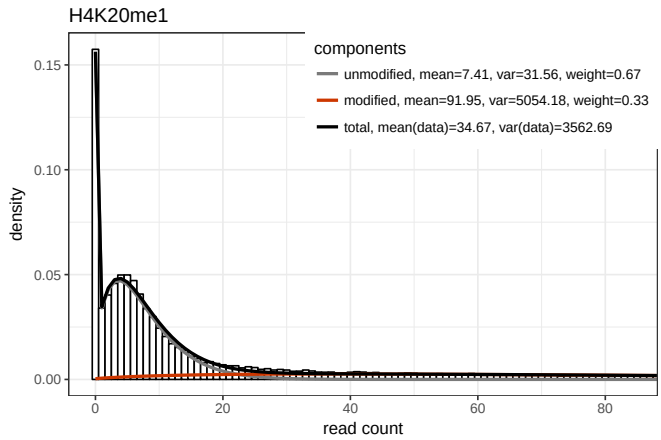
```
plotHistogram(model) + ggtitle('H3K27me3')
```




```
## === Step 4: Working with peaks ===
peaks <- model$peaks
length(peaks); mean(width(peaks))

## [1] 523
## [1] 43522.94

# Change the false discovery rate and get number of peaks
model <- changeFDR(model, fdr=1e-4)
peaks <- model$peaks
length(peaks); mean(width(peaks))

## [1] 416
## [1] 52582.93
```

```
## === Step 5: Export to genome browser ===
# We can export peak calls and binned read counts with
exportPeaks(model, filename=tempfile())
exportCounts(model, filename=tempfile())
```

```
## === Step 1-3: Another example for mark H4K20me1 ===
file <- system.file("extdata","euratrans","lv-H4K20me1-BN-male-bio1-tech1.bam",
                    package="chromstaRData")
data(rn4_chrominfo)
binned.data <- binReads(file, assembly=rn4_chrominfo, binsizes=1000,
                        stepsizes=500, chromosomes='chr12')
model <- callPeaksUnivariate(binned.data, max.time=60, verbosity=0)
plotHistogram(model) + ggtitle('H4K20me1')
```



!! It is important that the distributions are fitted correctly !! Please check section 6.4 for examples of how this plot should *not* look like and what can be done to get a correct fit.

3.3 Task 3: Peak calling for ATAC-seq, DNase-seq, FAIRE-seq, ...

Peak calling for ATAC-seq and DNase-seq is similar to the peak calling of a narrow histone modification (section 3.1). FAIRE-seq experiments seem to exhibit a broad profile with our model, so the procedure is similar to the domain calling of a broad histone modification (section 3.2).

4 Multivariate analysis

4.1 Task 1: Integrating multiple replicates

chromstaR can be used to call peaks with multiple replicates, without the need of prior merging. The underlying statistical model integrates information from all replicates to identify common peaks. It is, however, important to note that replicates with poor quality can affect the joint peak calling negatively. It is therefore recommended to first check the replicate quality and discard poor-quality replicates. The necessary steps for peak calling for an example ChIP-seq experiment with 4 replicates are detailed below.

Please note that also the other tasks in this section (Task 4.2, 4.3 and 4.4) can handle multiple replicates via specification of the `experiment.table` parameter. The following example demonstrates how to explicitly use multiple replicates for peak calling and their correlation as a basic quality control.

```
library(chromstaR)
```

```
#=== Step 1: Preparation ===
# Let's get some example data with 3 replicates in spontaneous hypertensive rat (SHR)
file.path <- system.file("extdata","euratrans", package='chromstaRData')
files.good <- list.files(file.path, pattern="H3K27me3.*SHR.*bam$", full.names=TRUE)[1:3]
# We fake a replicate with poor quality by taking a different mark entirely
files.poor <- list.files(file.path, pattern="H4K20me1.*SHR.*bam$", full.names=TRUE)[1]
files <- c(files.good, files.poor)
# Obtain chromosome lengths. This is only necessary for BED files. BAM files are
# handled automatically.
data(rn4_chrominfo)
head(rn4_chrominfo)
```

```
##   chromosome   length
## 1       chrM    16300
## 2      chr12 46782294
## 3      chr20 55268282
## 4      chr19 59218465
## 5      chr18 87265094
## 6      chr11 87759784
```

```
# Define experiment structure
exp <- data.frame(file=files, mark='H3K27me3', condition='SHR', replicate=1:4,
                  pairedEndReads=FALSE, controlFiles=NA)

# Peaks could now be called with
# Chromstar(inputfolder=file.path, experiment.table=exp, outputfolder=tempdir(),
#           mode = 'separate')
# However, to get more information on the replicates we will choose
# a more detailed workflow.
```

```
## === Step 2: Binning ===
# We use bin size 1000bp and chromosome 12 to keep the example quick
binned.data <- list()
for (file in files) {
  binned.data[[basename(file)]] <- binReads(file, binsize=1000, stepsizes=500,
                                            assembly=rn4_chrominfo, chromosomes='chr12',
                                            experiment.table=exp)
}
```

```
## === Step 3: Univariate peak calling ===
# The univariate fit is obtained for each replicate
models <- list()
for (i1 in 1:length(binned.data)) {
  models[[i1]] <- callPeaksUnivariate(binned.data[[i1]], max.time=60)
  plotHistogram(models[[i1]])
}
```

!! It is important that the distributions are fitted correctly !! Please check section 6.4 for examples of how this plot should *not* look like and what can be done to get a correct fit.

```
## === Step 4: Check replicate correlation ===
# We run a multivariate peak calling on all 4 replicates
# A warning is issued because replicate 4 is very different from the others
multi.model <- callPeaksReplicates(models, max.time=60, eps=1)
```

```
## HMM: number of states = 16
## HMM: number of bins = 46782
## HMM: maximum number of iterations = none
## HMM: maximum running time = 60 sec
## HMM: epsilon = 1
## HMM: number of experiments = 4
##  Iteration              log(P)             dlog(P)    Time in sec
##          0                -inf                   -              0
## HMM: Precomputing densities ...
##  Iteration              log(P)             dlog(P)    Time in sec
##          0                -inf                   -              1
##          1      -542989.146422                inf              1
##          2      -538374.740061         4614.406361              1
##          3      -538271.399633          103.340428              1
##          4      -538250.213731           21.185902              1
##          5      -538244.134929            6.078802              1
##          6      -538241.935145            2.199784              1
##          7      -538240.963921            0.971224              2
## HMM: Convergence reached!
## HMM: Recoding posteriors ...
```

```
## Warning in callPeaksReplicates(models, max.time = 60, eps = 1):  Your replicates cluster in 2 groups.  Consider redoing your analysis
with only the group with the highest average coverage:
## H3K27me3-SHR-rep1
## H3K27me3-SHR-rep2
## H3K27me3-SHR-rep3
## Replicates from groups with lower coverage are:
## H3K27me3-SHR-rep4
```

```
# Checking the correlation confirms that Rep4 is very different from the others
print(multi.model$replicateInfo$correlation)
```

```
##                   H3K27me3-SHR-rep1 H3K27me3-SHR-rep2 H3K27me3-SHR-rep3 H3K27me3-SHR-rep4
## H3K27me3-SHR-rep1         1.0000000         0.9999358         0.9997432        -0.3718157
## H3K27me3-SHR-rep2         0.9999358         1.0000000         0.9997217        -0.3717750
## H3K27me3-SHR-rep3         0.9997432         0.9997217         1.0000000        -0.3716530
## H3K27me3-SHR-rep4        -0.3718157        -0.3717750        -0.3716530         1.0000000
```

```
## === Step 5: Peak calling with replicates ===
# We redo the previous step without the "bad" replicate
# Also, we force all replicates to agree in their peak calls
multi.model <- callPeaksReplicates(models[1:3], force.equal=TRUE, max.time=60)
```

```
## === Step 6: Export to genome browser ===
# Finally, we can export the results as BED file
exportPeaks(multi.model, filename=tempfile())
exportCounts(multi.model, filename=tempfile())
```

4.2 Task 2: Detecting differentially modified regions

chromstaR is extremely powerful in detecting differentially modified regions in two or more samples. The following example illustrates this on ChIP-seq data for H4K20me1 in brown norway (BN) and spontaneous hypertensive rat (SHR) in left-ventricle (lv) heart tissue. The mode of analysis is called *differential*.

```
library(chromstaR)
```

```
#=== Step 1: Preparation ===
## Prepare the file paths. Exchange this with your input and output directories.
inputfolder <- system.file("extdata","euratrans", package="chromstaRData")
outputfolder <- file.path(tempdir(), 'H4K20me1-example')

## Define experiment structure, put NA if you don't have controls
data(experiment_table_H4K20me1)
print(experiment_table_H4K20me1)
```

```
##                                file     mark condition replicate pairedEndReads
## 1  lv-H4K20me1-BN-male-bio1-tech1.bam H4K20me1        BN         1          FALSE
## 2  lv-H4K20me1-BN-male-bio2-tech1.bam H4K20me1        BN         2          FALSE
## 3 lv-H4K20me1-SHR-male-bio1-tech1.bam H4K20me1       SHR         1          FALSE
##                                            controlFiles
## 1 lv-input-BN-male-bio1-tech1.bam|lv-input-BN-male-bio1-tech2.bam
## 2 lv-input-BN-male-bio1-tech1.bam|lv-input-BN-male-bio1-tech2.bam
```

```
## 3                                lv-input-SHR-male-bio1-tech1.bam
```

```
## Define assembly
# This is only necessary if you have BED files, BAM files are handled automatically.
# For common assemblies you can also specify them as 'hg19' for example.
data(rn4_chrominfo)
head(rn4_chrominfo)
```

```
##   chromosome   length
## 1       chrM    16300
## 2      chr12 46782294
## 3      chr20 55268282
## 4      chr19 59218465
## 5      chr18 87265094
## 6      chr11 87759784
```

```
#=== Step 2: Run Chromstar ===
## Run ChromstaR
Chromstar(inputfolder, experiment.table=experiment_table_H4K20me1,
          outputfolder=outputfolder, numCPU=4, binsize=1000, stepsize=500,
          assembly=rn4_chrominfo, prefit.on.chr='chr12', chromosomes='chr12',
          mode='differential')
```

```
## Results are stored in 'outputfolder' and can be loaded for further processing
list.files(outputfolder)
```

```
##  [1] "binned"               "BROWSERFILES"         "chrominfo.tsv"        "chromstaR.config"
##  [5] "combined"             "experiment_table.tsv" "multivariate"         "PLOTS"
##  [9] "README.txt"           "univariate"
```

```
model <- get(load(file.path(outputfolder,'multivariate',
                            'multivariate_mode-differential_mark-H4K20me1.RData')))
```

!! It is important that the distributions in folder outputfolder/PLOTS/univariate-distributions are fitted correctly !! Please check section 6.4 for examples of how this plot should *not* look like and what can be done to get a correct fit.

```
## === Step 3: Construct differential and common states ===
diff.states <- stateBrewer(experiment_table_H4K20me1, mode='differential',
                           differential.states=TRUE)
print(diff.states)
```

```
##   combination state
## 1       [SHR]     1
## 2        [BN]     6
```

```
common.states <- stateBrewer(experiment_table_H4K20me1, mode='differential',
                             common.states=TRUE)
print(common.states)
```

```
##   combination state
## 1          []     0
## 2    [BN+SHR]     7
```

```
## === Step 5: Export to genome browser ===
# Export only differential states
exportPeaks(model, filename=tempfile())
exportCounts(model, filename=tempfile())
exportCombinations(model, filename=tempfile(), include.states=diff.states)
```

4.3 Task 3: Finding combinatorial chromatin states

Most experimental studies that probe several histone modifications are interested in combinatorial chromatin states. An example of a simple combinatorial state would be [H3K4me3+H3K27me3], which is also frequently called "bivalent promoter", due to the simultaneous occurrence of the promoter marking H3K4me3 and the repressive H3K27me3. Finding combinatorial states with *chromstaR* is equivalent to a multivariate peak calling. The following code chunks demonstrate how to find bivalent promoters and do some simple analysis. The mode of analysis is called *combinatorial*.

```
library(chromstaR)
```

```
#=== Step 1: Preparation ===
## Prepare the file paths. Exchange this with your input and output directories.
inputfolder <- system.file("extdata","euratrans", package="chromstaRData")
outputfolder <- file.path(tempdir(), 'SHR-example')

## Define experiment structure, put NA if you don't have controls
# (SHR = spontaneous hypertensive rat)
data(experiment_table_SHR)
print(experiment_table_SHR)
```

```
##                                file     mark condition replicate pairedEndReads
## 1 lv-H3K27me3-SHR-male-bio2-tech1.bam H3K27me3       SHR         1          FALSE
## 2 lv-H3K27me3-SHR-male-bio2-tech2.bam H3K27me3       SHR         2          FALSE
## 3  lv-H3K4me3-SHR-male-bio2-tech1.bam  H3K4me3       SHR         1          FALSE
## 4  lv-H3K4me3-SHR-male-bio3-tech1.bam  H3K4me3       SHR         2          FALSE
##                        controlFiles
## 1 lv-input-SHR-male-bio1-tech1.bam
## 2 lv-input-SHR-male-bio1-tech1.bam
## 3 lv-input-SHR-male-bio1-tech1.bam
## 4 lv-input-SHR-male-bio1-tech1.bam
```

```
## Define assembly
# This is only necessary if you have BED files, BAM files are handled automatically.
# For common assemblies you can also specify them as 'hg19' for example.
data(rn4_chrominfo)
head(rn4_chrominfo)
```

```
##   chromosome   length
## 1       chrM    16300
## 2      chr12 46782294
## 3      chr20 55268282
## 4      chr19 59218465
## 5      chr18 87265094
## 6      chr11 87759784
```

```
#=== Step 2: Run Chromstar ===
## Run ChromstaR
Chromstar(inputfolder, experiment.table=experiment_table_SHR,
          outputfolder=outputfolder, numCPU=4, binsize=1000, stepsize=500,
          assembly=rn4_chrominfo, prefit.on.chr='chr12', chromosomes='chr12',
          mode='combinatorial')
```

!! It is important that the distributions in folder outputfolder/PLOTS/univariate-distributions are fitted correctly !! Please check section 6.4 for examples of how this plot should *not* look like and what can be done to get a correct fit.

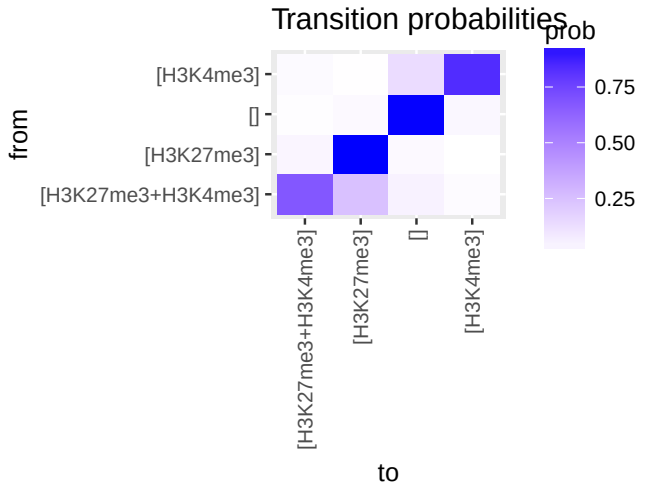
```
## Results are stored in 'outputfolder' and can be loaded for further processing
list.files(outputfolder)
```

```
##  [1] "binned"               "BROWSERFILES"         "chrominfo.tsv"        "chromstaR.config"
##  [5] "combined"             "experiment_table.tsv" "multivariate"         "PLOTS"
##  [9] "README.txt"           "univariate"
```

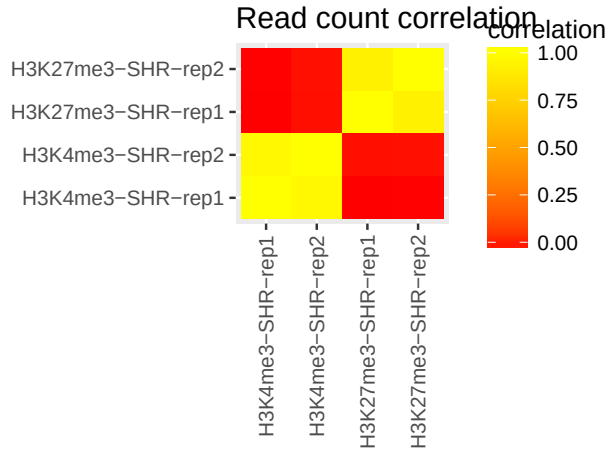
```
model <- get(load(file.path(outputfolder,'multivariate',
                            'multivariate_mode-combinatorial_condition-SHR.RData')))
# Obtain genomic frequencies for combinatorial states
genomicFrequencies(model)
```

```
## $frequency
##
##                 []          [H3K4me3]         [H3K27me3] [H3K27me3+H3K4me3]
##         0.41441153         0.09365782         0.42928904         0.06264161
##
## $domains
##
##                 []          [H3K4me3]         [H3K27me3] [H3K27me3+H3K4me3]
##               1235                678               1205                900
```

```
# Plot transition probabilities and read count correlation
heatmapTransitionProbs(model) + ggtitle('Transition probabilities')
```



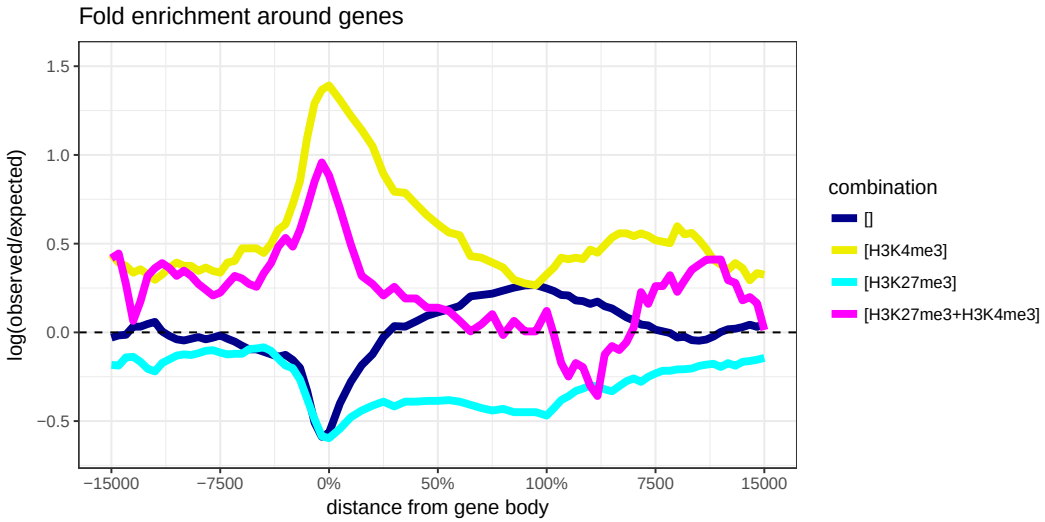

```
heatmapCountCorrelation(model) + ggtitle('Read count correlation')
```



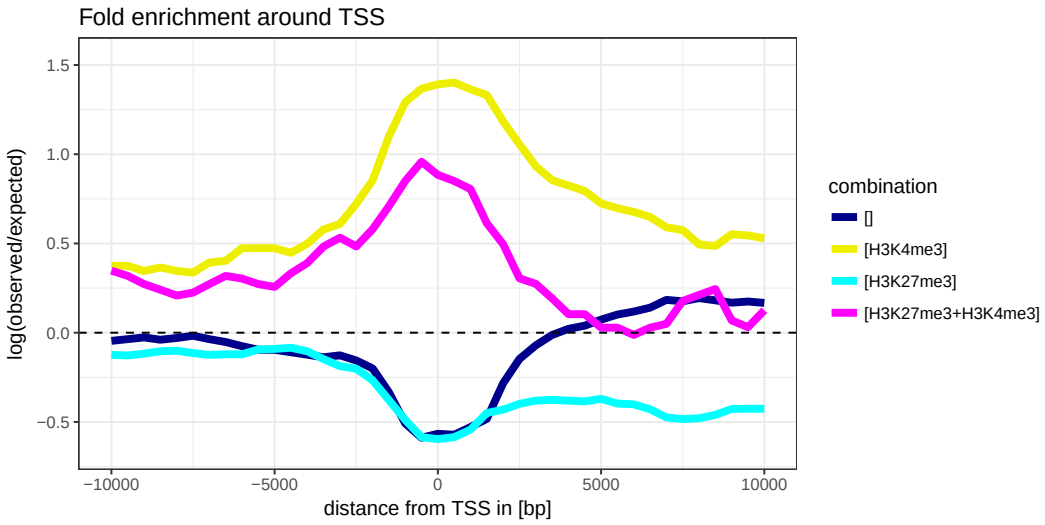

```
## === Step 3: Enrichment analysis ===
# Annotations can easily be obtained with the biomaRt package. Of course, you can
# also load them from file if you already have annotation files available.
library(biomaRt)
ensembl <- useMart('ENSEMBL_MART_ENSEMBL', host='may2012.archive.ensembl.org',
                   dataset='rnorvegicus_gene_ensembl')
genes <- getBM(attributes=c('ensembl_gene_id', 'chromosome_name', 'start_position',
                            'end_position', 'strand', 'external_gene_id',
                            'gene_biotype'),
               mart=ensembl)
# Transform to GRanges for easier handling
genes <- GRanges(seqnames=paste0('chr',genes$chromosome_name),
                 ranges=IRanges(start=genes$start, end=genes$end),
                 strand=genes$strand,
                 name=genes$external_gene_id, biotype=genes$gene_biotype)
seqlevels(genes)[seqlevels(genes)=='chrMT'] <- 'chrM'
print(genes)

## GRanges object with 29516 ranges and 2 metadata columns:
##           seqnames                ranges strand |        name        biotype
##              <Rle>             <IRanges>  <Rle> | <character>    <character>
##       [1]    chr13     [1120899, 1121213]      - |   LOC682397 protein_coding
##       [2]    chr13     [1192186, 2293551]      - |   LOC304725 protein_coding
##       [3]    chr13     [3174383, 3175216]      + |                pseudogene
##       [4]    chr13     [4377731, 4379174]      - |  D3ZPH4_RAT protein_coding
##       [5]    chr13     [4866302, 4866586]      - |  F1LZC7_RAT protein_coding
##       ...      ...                    ...    ... .         ...            ...
##   [29512]     chr6 [134310258, 134310338]      + |    SNORD113         snoRNA
##   [29513]     chr9 [  6920889,   6921049]      - |         U1          snRNA
##   [29514]    chr11 [ 40073746,  40073816]      - |    SNORD19B         snoRNA
##   [29515]     chr2 [233090372, 233090478]      - |          U6          snRNA
```

```
##   [29516]     chr6 [ 92917449,  92917541]      + |                    miRNA
##   -------
##   seqinfo: 22 sequences from an unspecified genome; no seqlengths
```

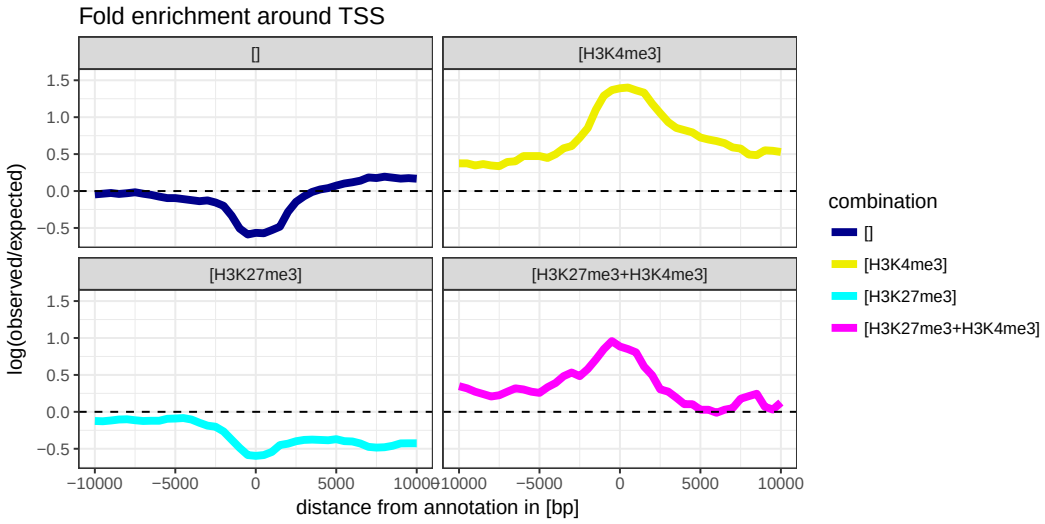
```
# We expect promoter [H3K4me3] and bivalent-promoter signatures [H3K4me3+H3K27me3]
# to be enriched at transcription start sites.
plotEnrichment(hmm = model, annotation = genes, bp.around.annotation = 15000) +
  ggtitle('Fold enrichment around genes') +
  xlab('distance from gene body')
```



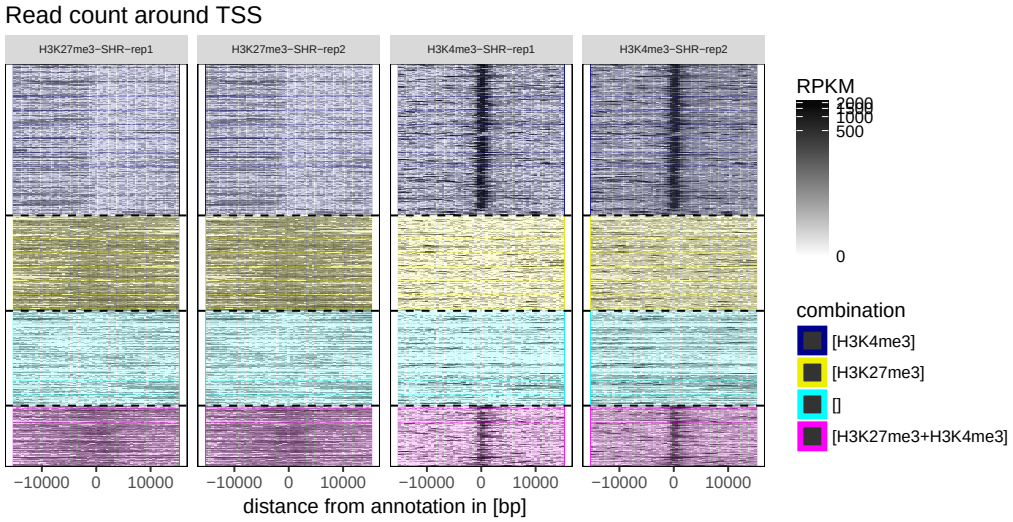

```
# Plot enrichment only at TSS. We make use of the fact that TSS is the start of a gene.
plotEnrichment(model, genes, region = 'start') +
  ggtitle('Fold enrichment around TSS') +
  xlab('distance from TSS in [bp]')
```



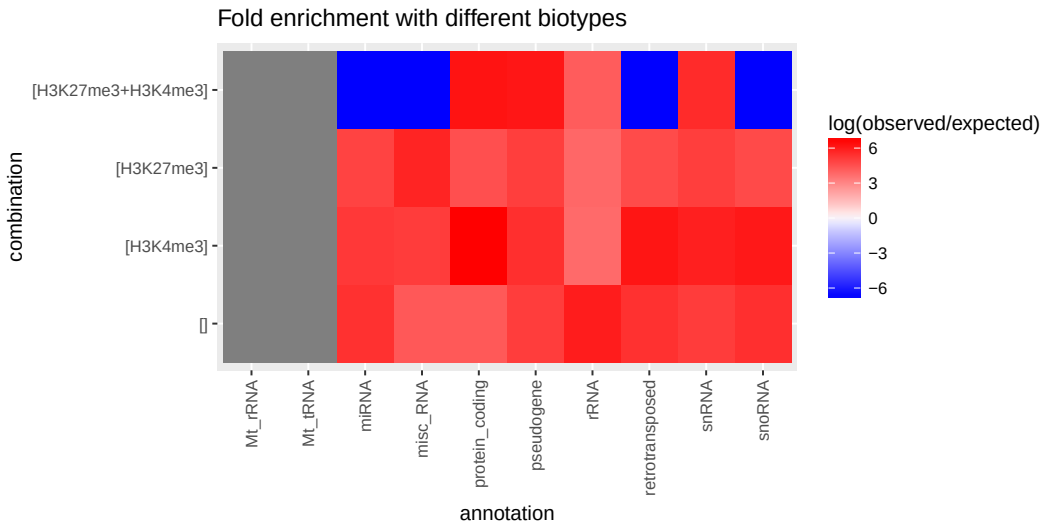

```
# Note: If you want to facet the plot because you have many combinatorial states you
# can do that with
plotEnrichment(model, genes, region = 'start') +
  facet_wrap(~ combination) + ggtitle('Fold enrichment around TSS')
```




```
# Another form of visualization that shows every TSS in a heatmap
tss <- resize(genes, width = 3, fix = 'start')
plotEnrichCountHeatmap(model, tss, bp.around.annotation = 15000) +
  theme(strip.text.x = element_text(size=6)) +
  scale_x_continuous(breaks=c(-10000,0,10000)) +
  ggtitle('Read count around TSS')
```



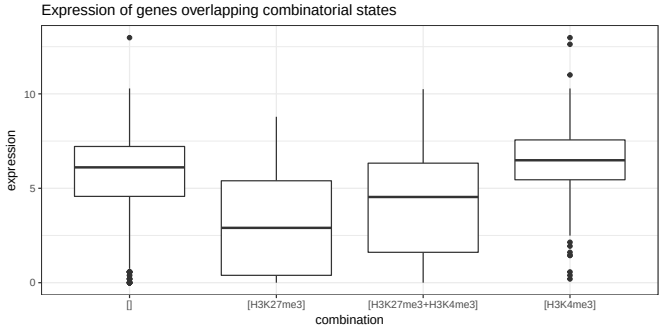

```
# Fold enrichment with different biotypes, showing that protein coding genes are
# enriched with (bivalent) promoter combinations [H3K4me3] and [H3K4me3+H3K27me3],
# while rRNA is enriched with the empty [] combination.
biotypes <- split(tss, tss$biotype)
plotFoldEnrichHeatmap(model, annotations=biotypes) + coord_flip() +
  ggtitle('Fold enrichment with different biotypes')
```



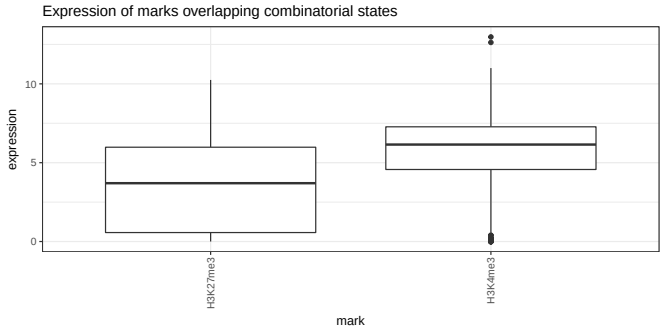

```
# === Step 4: Expression analysis ===
# Suppose you have expression data as well for your experiment. The following code
# shows you how to get the expression values for each combinatorial state.
data(expression_lv)
head(expression_lv)
```

```
##      ensembl_gene_id expression_BN expression_SHR
## 1 ENSRNOG00000000001          8.8           7.4
## 2 ENSRNOG00000000007         20.0          13.0
## 3 ENSRNOG00000000008          1.8           3.4
## 4 ENSRNOG00000000010          6.2         506.8
## 5 ENSRNOG00000000012         48.0          36.4
## 6 ENSRNOG00000000014         18.2          15.2
```

```
# We need to get coordinates for each of the genes
library(biomaRt)
ensembl <- useMart('ENSEMBL_MART_ENSEMBL', host='may2012.archive.ensembl.org',
                   dataset='rnorvegicus_gene_ensembl')
genes <- getBM(attributes=c('ensembl_gene_id', 'chromosome_name', 'start_position',
                            'end_position', 'strand', 'external_gene_id',
                            'gene_biotype'),
               mart=ensembl)
expr <- merge(genes, expression_lv, by='ensembl_gene_id')
# Transform to GRanges
expression.SHR <- GRanges(seqnames=paste0('chr',expr$chromosome_name),
                          ranges=IRanges(start=expr$start, end=expr$end),
                          strand=expr$strand, name=expr$external_gene_id,
                          biotype=expr$gene_biotype,
                          expression=expr$expression_SHR)
seqlevels(expression.SHR)[seqlevels(expression.SHR)=='chrMT'] <- 'chrM'
# We apply an asinh transformation to reduce the effect of outliers
expression.SHR$expression <- asinh(expression.SHR$expression)

## Plot
plotExpression(model, expression.SHR) +
  theme(axis.text.x=element_text(angle=0, hjust=0.5)) +
  ggtitle('Expression of genes overlapping combinatorial states')
```




```
plotExpression(model, expression.SHR, return.marks=TRUE) +
  ggtitle('Expression of marks overlapping combinatorial states')
```



4.4 Task 4: Finding differences between combinatorial chromatin states

Consider bivalent promoters defined by [H3K4me3+H3K27me3] at two different developmental stages, or in two different strains or tissues. This is an example where one is interested in *differences* between *combinatorial states*. The following example demonstrates how such an analysis can be done with *chromstaR*. We use a data set from the Euratrans project (downsampled to chr12) to find differences in bivalent promoters between brown norway (BN) and spontaneous hypertensive rat (SHR) in left-ventricle (lv) heart tissue.

`Chromstar` can be run in 4 different modes:

- *full*: Recommended mode if your (number of marks) * (number of conditions) is less or equal to 8. With 8 ChIP-seq experiments there are already $2^8 = 256$ combinatorial states which is the maximum that most computers can handle computationally for a human-sized genome at bin size 1000bp.
- **DEFAULT** *differential*: Choose this mode if you are interested in highly significant differences between conditions. The computational limit for the number of conditions is ~ 8 for a human-sized genome. Combinatorial states are not as accurate as in mode *combinatorial* or *full*.
- *combinatorial*: This mode will yield good combinatorial chromatin state calls for any number of marks and conditions. However, differences between conditions have more false positives than in mode *differential* or *full*.
- *separate*: Only replicates are processed in a multivariate manner. Combinatorial states are constructed by a simple post-hoc combination of peak calls.

```
library(chromstaR)
```

```
#=== Step 1: Preparation ===
## Prepare the file paths. Exchange this with your input and output directories.
inputfolder <- system.file("extdata","euratrans", package="chromstaRData")
outputfolder <- file.path(tempdir(), 'SHR-BN-example')

## Define experiment structure, put NA if you don't have controls
data(experiment_table)
print(experiment_table)
```

```
##                                file     mark condition replicate pairedEndReads
## 1  lv-H3K27me3-BN-male-bio2-tech1.bam H3K27me3        BN         1          FALSE
## 2  lv-H3K27me3-BN-male-bio2-tech2.bam H3K27me3        BN         2          FALSE
## 3 lv-H3K27me3-SHR-male-bio2-tech1.bam H3K27me3       SHR         1          FALSE
## 4 lv-H3K27me3-SHR-male-bio2-tech2.bam H3K27me3       SHR         2          FALSE
## 5 lv-H3K4me3-BN-female-bio1-tech1.bam  H3K4me3        BN         1          FALSE
## 6   lv-H3K4me3-BN-male-bio2-tech1.bam  H3K4me3        BN         2          FALSE
## 7  lv-H3K4me3-SHR-male-bio2-tech1.bam  H3K4me3       SHR         1          FALSE
## 8  lv-H3K4me3-SHR-male-bio3-tech1.bam  H3K4me3       SHR         2          FALSE
##                                                                 controlFiles
## 1 lv-input-BN-male-bio1-tech1.bam|lv-input-BN-male-bio1-tech2.bam
## 2 lv-input-BN-male-bio1-tech1.bam|lv-input-BN-male-bio1-tech2.bam
## 3                              lv-input-SHR-male-bio1-tech1.bam
## 4                              lv-input-SHR-male-bio1-tech1.bam
## 5                                                          <NA>
## 6                                                          <NA>
## 7                                                          <NA>
## 8                                                          <NA>
```

```
## Define assembly
# This is only necessary if you have BED files, BAM files are handled automatically.
# For common assemblies you can also specify them as 'hg19' for example.
data(rn4_chrominfo)
```

```
head(rn4_chrominfo)
```

```
##   chromosome   length
## 1       chrM    16300
## 2      chr12 46782294
## 3      chr20 55268282
## 4      chr19 59218465
## 5      chr18 87265094
## 6      chr11 87759784
```

```
#=== Step 2: Run Chromstar ===
## Run ChromstaR
Chromstar(inputfolder, experiment.table=experiment_table,
          outputfolder=outputfolder, numCPU=4, binsize=1000, stepsize=500,
          assembly=rn4_chrominfo, prefit.on.chr='chr12', chromosomes='chr12',
          mode='differential')
```

```
## Results are stored in 'outputfolder' and can be loaded for further processing
list.files(outputfolder)
```

```
##  [1] "binned"               "BROWSERFILES"         "chrominfo.tsv"        "chromstaR.config"
##  [5] "combined"             "experiment_table.tsv" "multivariate"         "PLOTS"
##  [9] "README.txt"           "univariate"
```

```
model <- get(load(file.path(outputfolder,'combined',
                            'combined_mode-differential.RData')))
```

!! It is important that the distributions in folder outputfolder/PLOTS/univariate-distributions are fitted correctly !! Please check section 6.4 for examples of how this plot should *not* look like and what can be done to get a correct fit.

```
#=== Step 3: Analysis and export ===
## Obtain all genomic regions where the two tissues have different states
segments <- model$segments
diff.segments <- segments[segments$combination.SHR != segments$combination.BN]
# Let's be strict with the differences and get only those where both marks are different
diff.segments <- diff.segments[diff.segments$differential.score >= 1.9]
exportGRangesAsBedFile(diff.segments, trackname='differential_chromatin_states',
              filename=tempfile(), scorecol='differential.score')
```

```
## Warning in exportGRangesAsBedFile(diff.segments, trackname = "differential_chromatin_states", :  Column 'differential.score' should contain
integer values between 0 and 1000 for compatibility with the UCSC convention.
```

```
## Obtain all genomic regions where we find a bivalent promoter in BN but not in SHR
bivalent.BN <- segments[segments$combination.BN == '[H3K27me3+H3K4me3]' &
                        segments$combination.SHR != '[H3K27me3+H3K4me3]']
## Obtain all genomic regions where BN and SHR have promoter signatures
promoter.BN <- segments[segments$transition.group == 'constant [H3K4me3]']

## Get transition frequencies
print(model$frequencies)
```

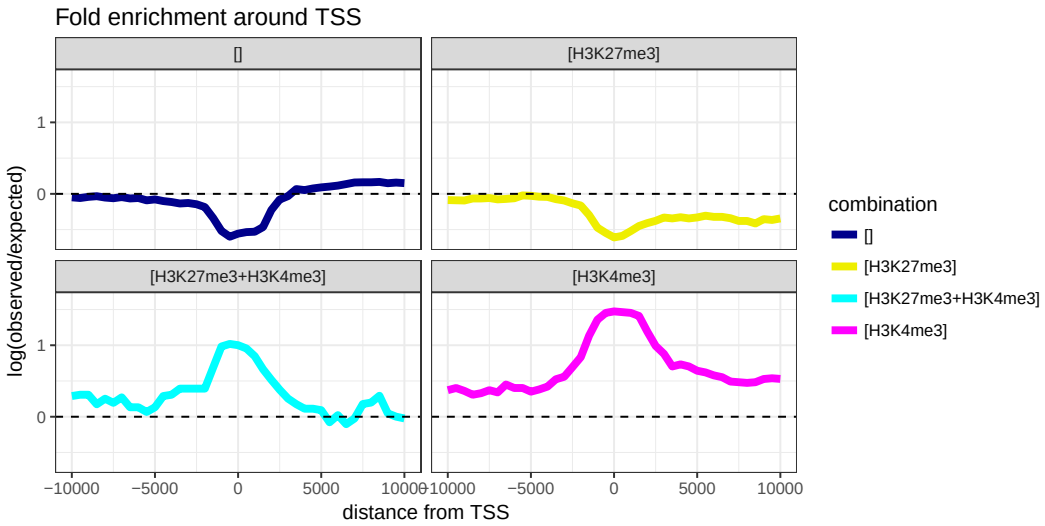
```
##         combination.BN     combination.SHR domains    frequency cumulative.frequency
## 1           [H3K27me3]          [H3K27me3]    1392 4.338528e-01            0.4338528
## 2                   []                  []    1348 4.204929e-01            0.8543457
## 3            [H3K4me3]           [H3K4me3]     854 8.563123e-02            0.9399769
## 4  [H3K27me3+H3K4me3] [H3K27me3+H3K4me3]     846 5.246676e-02            0.9924437
## 5           [H3K27me3]                  []      55 3.046043e-03            0.9954897
## 6                   []          [H3K27me3]      49 2.842974e-03            0.9983327
## 7           [H3K27me3] [H3K27me3+H3K4me3]      16 4.702663e-04            0.9988030
## 8                   []           [H3K4me3]      12 2.992604e-04            0.9991022
## 9            [H3K4me3] [H3K27me3+H3K4me3]      10 2.778847e-04            0.9993801
## 10 [H3K27me3+H3K4me3]          [H3K27me3]       5 2.030696e-04            0.9995832
## 11 [H3K27me3+H3K4me3]           [H3K4me3]       6 1.816938e-04            0.9997649
## 12           [H3K4me3]                  []       3 7.481510e-05            0.9998397
## 13                  [] [H3K27me3+H3K4me3]       1 6.412723e-05            0.9999038
## 14 [H3K27me3+H3K4me3]                  []       1 6.412723e-05            0.9999679
## 15          [H3K27me3]           [H3K4me3]       1 3.206361e-05            1.0000000
##                              group
## 1                constant [H3K27me3]
## 2                    zero transition
## 3                 constant [H3K4me3]
## 4         constant [H3K27me3+H3K4me3]
## 5           stage-specific [H3K27me3]
## 6           stage-specific [H3K27me3]
## 7                             other
## 8            stage-specific [H3K4me3]
## 9                             other
## 10                            other
## 11                            other
## 12           stage-specific [H3K4me3]
## 13 stage-specific [H3K27me3+H3K4me3]
## 14 stage-specific [H3K27me3+H3K4me3]
```

```
## 15                              other
```

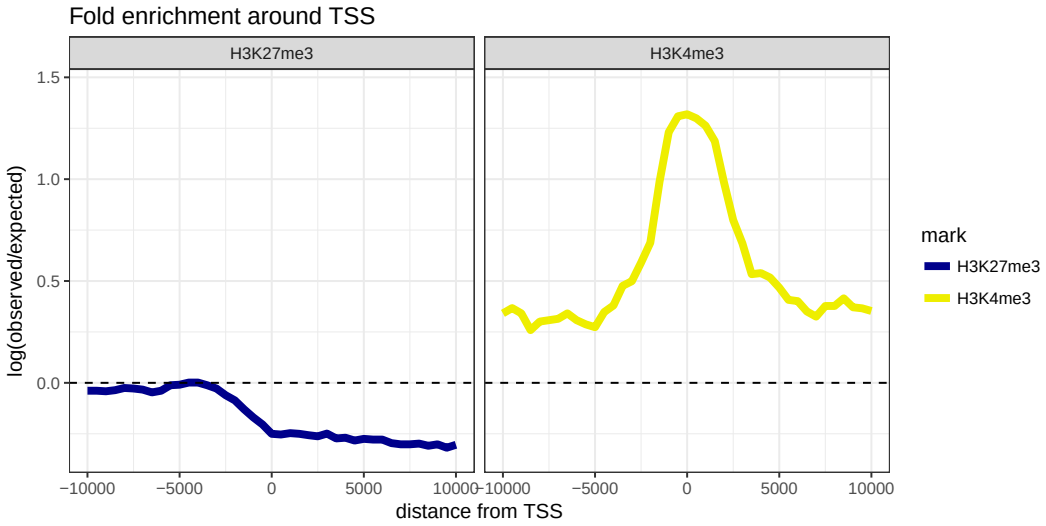
```
## === Step 4: Enrichment analysis ===
# Annotations can easily be obtained with the biomaRt package. Of course, you can
# also load them from file if you already have annotation files available.
library(biomaRt)
ensembl <- useMart('ENSEMBL_MART_ENSEMBL', host='may2012.archive.ensembl.org',
                   dataset='rnorvegicus_gene_ensembl')
genes <- getBM(attributes=c('ensembl_gene_id', 'chromosome_name', 'start_position',
                            'end_position', 'strand', 'external_gene_id',
                            'gene_biotype'),
               mart=ensembl)
# Transform to GRanges for easier handling
genes <- GRanges(seqnames=paste0('chr',genes$chromosome_name),
                 ranges=IRanges(start=genes$start, end=genes$end),
                 strand=genes$strand,
                 name=genes$external_gene_id, biotype=genes$gene_biotype)
seqlevels(genes)[seqlevels(genes)=='chrMT'] <- 'chrM'
print(genes)
```

```
## GRanges object with 29516 ranges and 2 metadata columns:
##           seqnames               ranges strand |        name        biotype
##              <Rle>            <IRanges>  <Rle> | <character>    <character>
##       [1]    chr13     [1120899, 1121213]      - |   LOC682397 protein_coding
##       [2]    chr13     [1192186, 2293551]      - |   LOC304725 protein_coding
##       [3]    chr13     [3174383, 3175216]      + |                pseudogene
##       [4]    chr13     [4377731, 4379174]      - |  D3ZPH4_RAT protein_coding
##       [5]    chr13     [4866302, 4866586]      - |  F1LZC7_RAT protein_coding
##       ...      ...                    ...    ... .          ...          ...
##   [29512]     chr6 [134310258, 134310338]      + |    SNORD113        snoRNA
##   [29513]     chr9 [  6920889,   6921049]      - |          U1         snRNA
##   [29514]    chr11 [ 40073746,  40073816]      - |    SNORD19B        snoRNA
##   [29515]     chr2 [233090372, 233090478]      - |          U6         snRNA
##   [29516]     chr6 [ 92917449,  92917541]      + |                     miRNA
##   -------
##   seqinfo: 22 sequences from an unspecified genome; no seqlengths
```

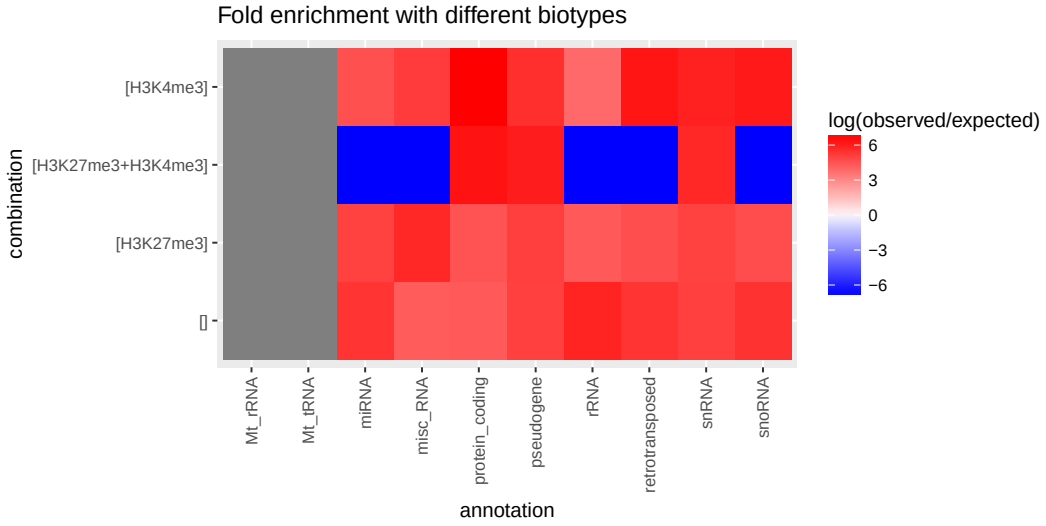
```
# We expect promoter [H3K4me3] and bivalent-promoter signatures [H3K4me3+H3K27me3]
# to be enriched at transcription start sites.
plots <- plotEnrichment(hmm=model, annotation=genes, region='start')
plots[['BN']] + facet_wrap(~ combination) +
  ggtitle('Fold enrichment around TSS') +
  xlab('distance from TSS')
```




```
plots <- plotEnrichment(hmm=model, annotation=genes, region='start', what='peaks')
plots[['BN']] + facet_wrap(~ mark) +
  ggtitle('Fold enrichment around TSS') +
  xlab('distance from TSS')
```




```
# Fold enrichment with different biotypes, showing that protein coding genes are
# enriched with (bivalent) promoter combinations [H3K4me3] and [H3K4me3+H3K27me3],
# while rRNA is enriched with the empty [] combination.
tss <- resize(genes, width = 3, fix = 'start')
biotypes <- split(tss, tss$biotype)
plots <- plotFoldEnrichHeatmap(model, annotations=biotypes)
plots[['BN']] + coord_flip() +
  ggtitle('Fold enrichment with different biotypes')
```



5 Output of function `Chromstar()`

`Chromstar()` is the workhorse of the *chromstaR* package and produces all the files that are necessary for downstream analyses. Here is an explanation of the *files* and **folders** you will find in your **outputfolder**:

- *chrominfo.tsv*:
  A tab-separated file with chromosome lengths.
- *chromstaR.config*:
  A text file with all the parameters that were used to run function `Chromstar()`.
- *experiment_table.tsv*:
  A tab-separated file of your experiment setup.

- **binned**:
  RData files with the results of the binnig step. Contains *GRanges* objects with binned genomic coordinates and read counts.
- **BROWSERFILES**:
  Bed files for upload to the UCSC genome browser. It contains files with combinatorial states ("_combinations.bed.gz") and underlying peak calls ("_peaks.bed.gz"). !!Always check the "_peaks.bed.gz" files if you are satisfied with the peak calls. If not, there are ways to make the calls stricter (see section 6.1).
- →**combined**←:
  RData files with the combined results of the uni- and multivariate peak calling steps. This is what you want to use for downstream analyses. Contains *combinedMultiHMM* objects.
  - *combined_mode-separate.RData* Simple combination of peak calls (replicates considered) without multivariate analysis.
  - *combined_mode-combinatorial.RData* Combination of multivariate results for mode='combinatorial'.
  - *combined_mode-differential.RData* Combination of multivariate results for mode='differential'.
  - *combined_mode-full.RData* Combination of multivariate results for mode='full'.
- **multivariate**:
  RData files with the results of the multivariate peak calling step. Contains *multiHMM* objects.
- **PLOTS**:
  Several plots that are produced by default. Please check the plots in subfolder **univariate-distributions** for irregularities (see section 3).
- **replicates**:
  RData files with the result of the replicate peak calling step. Contains *multiHMM* objects.
- **univariate**:
  RData files with the result of the univariate peak calling step. Contains *uniHMM* objects.

6 FAQ

6.1 The peak calls are too lenient. Can I adjust the strictness of the peak calling?

The strictness of the peak calling can be controlled with a false discovery rate. The Hidden Markov Model gives posterior probabilities for each peak, and based on these probabilites the model decides if a peak is present or not by picking the state with the highest probability. This way of peak calling leads to very lenient peak calls, and for some applications it may be desirable to obtain only very clear peaks. This can be achieved by setting a false discovery rate (which is a cutoff on the maximum posterior probability within each peak). To follow the below example, please first run step 1 and 2 from section 4.4.

```
model <- get(load(file.path(outputfolder,'combined',
                            'combined_mode-differential.RData')))
# Set a strict cutoff close to 1
model2 <- changeFDR(model, fdr=1e-4)
## Compare the number and width of peaks before and after cutoff adjustment
length(model$segments); mean(width(model$segments))
```

```
## [1] 4599
## [1] 10172.21
```

```
length(model2$segments); mean(width(model2$segments))
```

```
## [1] 3580
## [1] 13067.6
```

It is even possible to adjust the false discovery rate differently for the different marks or conditions.

```
# Set a stricter cutoff for H3K4me3 than for H3K27me3
fdrs <- c(0.1, 0.1, 0.1, 0.1, 0.01, 0.01, 0.01, 0.01)
names(fdrs) <- model$info$ID
print(fdrs)
```

```
##  H3K27me3-BN-rep1  H3K27me3-BN-rep2   H3K4me3-BN-rep1   H3K4me3-BN-rep2 H3K27me3-SHR-rep1 H3K27me3-SHR-rep2
##              0.10              0.10              0.10              0.10              0.01              0.01
##  H3K4me3-SHR-rep1  H3K4me3-SHR-rep2
##              0.01              0.01
```

```
model2 <- changeFDR(model, fdr=fdrs)
## Compare the number and width of peaks before and after cutoff adjustment
length(model$segments); mean(width(model$segments))

## [1] 4599
## [1] 10172.21

length(model2$segments); mean(width(model2$segments))

## [1] 4285
## [1] 10917.62
```

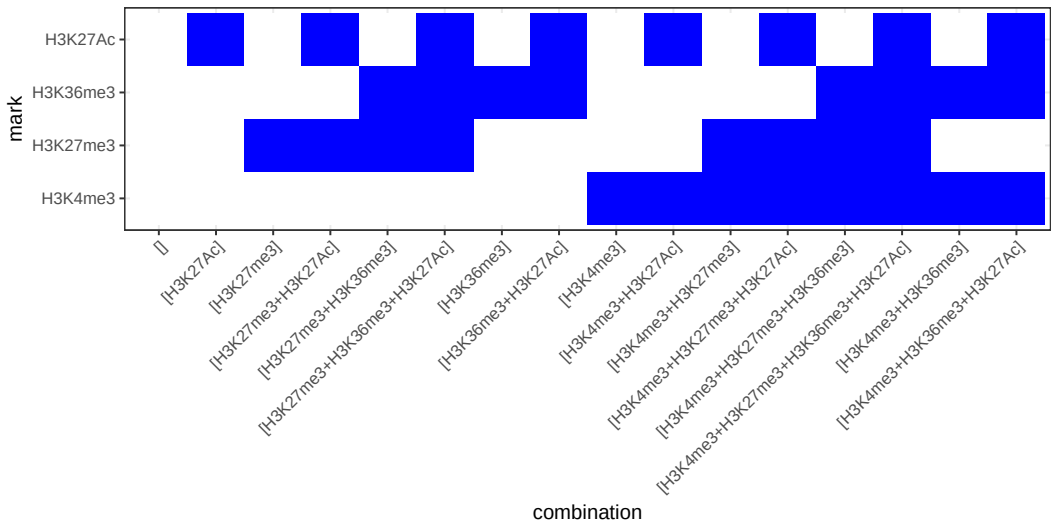
6.2 The combinatorial differences that chromstaR gives me are not convincing. Is there a way to restrict the results to a more convincing set?

You were interested in combinatorial state differences as in section 4.4 and checked the results in a genome browser. You found that some differences are convincing by eye and some are not. There are several possibilities to explore:

1. Run `Chromstar` in mode='differential' (instead of mode='combinatorial') and see if the results improve.
2. You can play with the "differential.score" (see section 4.4, step 3) and export only differences with a high score. A differential score around 1 means that one modification is different, a score close to 2 means that two modifications are different etc. The score is calculated as the sum of differences in posterior probabilities between marks.
3. Set a strict false discovery rate (previous example) to obtain only high confidence peaks.
4. Check for bad replicates that are very different from the rest and exclude them prior to the analysis.

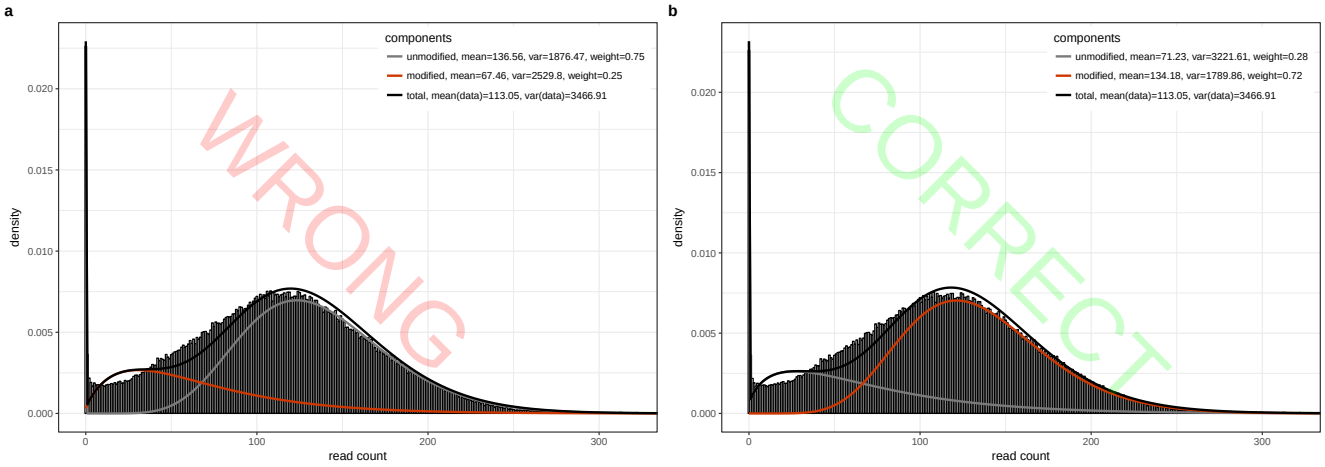
6.3 How do I plot a simple heatmap with the combinations?

```
heatmapCombinations(marks=c('H3K4me3', 'H3K27me3', 'H3K36me3', 'H3K27Ac'))
```

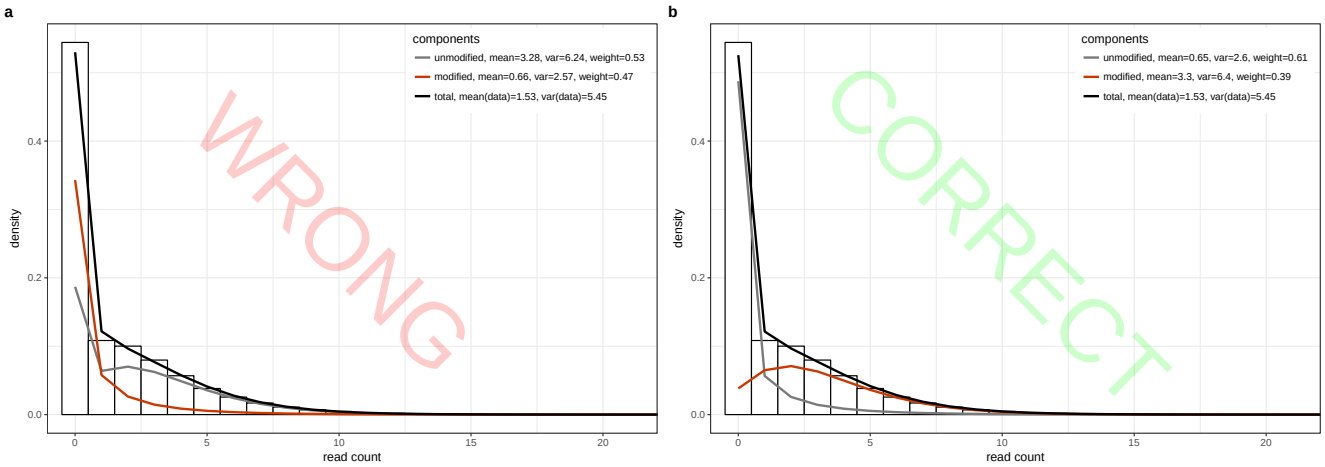


6.4 Examples of problematic distributions.

For the chromstaR peak calling to work correctly it is essential that the Baum-Welch algorithm correctly identifies unmodified (background) and modified (signal/peak) components in the data. Therefore, you should always check the plots in folder **PLOTS/univariate-distributions** for correct convergence. Here are some plots that indicate failed and succesful fitting procedures:



The plot shows data for H3K27me3 at binsize 1000bp. (a) Incorrectly converged fit, where the **modified** component (red) has lower read counts than the **unmodified** component (gray). (b) Correctly converged fit. Even here, the fit could be improved by reducing the average number of reads per bin, either by selecting a smaller binsize or by downsampling the data before using chromstaR.



The plot shows data for H3K27me3 at binsize 150bp. (a) Incorrectly converged fit, where the **modified** component (red) has a higher density at zero reads than the **unmodified** component (gray). (b) Correctly converged fit.

7 Session Info

```
toLatex(sessionInfo())
```

- R version 3.3.0 (2016-05-03), x86_64-pc-linux-gnu
- Locale: LC_CTYPE=en_US.UTF-8, LC_NUMERIC=C, LC_TIME=de_DE.UTF-8, LC_COLLATE=en_US.UTF-8, LC_MONETARY=de_DE.UTF-8, LC_MESSAGES=en_US.UTF-8, LC_PAPER=de_DE.UTF-8, LC_NAME=C, LC_ADDRESS=C, LC_TELEPHONE=C, LC_MEASUREMENT=de_DE.UTF-8, LC_IDENTIFICATION=C
- Base packages: base, datasets, graphics, grDevices, methods, parallel, stats, stats4, utils
- Other packages: BiocGenerics 0.20.0, biomaRt 2.29.2, chromstaR 1.3.1, chromstaRData 1.0.0, devtools 1.12.0, GenomeInfoDb 1.10.3, GenomicRanges 1.26.4, ggplot2 2.2.1, IRanges 2.8.2, knitr 1.15.1, Rcpp 0.12.11, S4Vectors 0.12.2
- Loaded via a namespace (and not attached): AnnotationDbi 1.36.2, assertthat 0.1, bamsignals 1.6.0, Biobase 2.34.0, BiocParallel 1.8.2, BiocStyle 2.1.14, Biostrings 2.42.1, bitops 1.0-6, codetools 0.2-14, colorspace 1.2-6, compiler 3.3.0, DBI 0.4-1, digest 0.6.10, doParallel 1.0.10, evaluate 0.10, foreach 1.4.3, GenomicAlignments 1.10.1, grid 3.3.0, gtable 0.2.0, highr 0.6, iterators 1.0.8, labeling 0.3, lattice 0.20-33, lazyeval 0.2.0, magrittr 1.5, Matrix 1.2-6, memoise 1.0.0, munsell 0.4.3, mvtnorm 1.0-6, plyr 1.8.4, RCurl 1.95-4.8, reshape2 1.4.2, Rsamtools 1.26.2, RSQLite 1.0.0, scales 0.4.1, stringi 1.1.2, stringr 1.1.0, SummarizedExperiment 1.4.0, tibble 1.2, tools 3.3.0, withr 1.0.2, XML 3.98-1.4, XVector 0.14.1, zlibbioc 1.20.0