The genomic STate ANnotation package

Benedikt Zacher[1,2,*]***, Julia Ertl***[1]***, Julien Gagneur***[1]***, Achim Tresch***[1,2,3]

[1em] [1] Gene Center and Department of Biochemistry, LMU, Munich, Germany
[2] Institute for Genetics, University of Cologne, Cologne, Germany
[3] Max Planck Institute for Plant Breeding Research, Cologne, Germany

[*]zacher (at) genzentrum.lmu.de

October 29, 2019

If you use *STAN* in published research, please cite:

Zacher, B. and Lidschreiber, M. and Cramer, P. and Gagneur, J. and Tresch, A. (2014): **Annotation of genomics data using bidirectional hidden Markov models unveils variations in Pol II transcription cycle** *Mol. Syst. Biol.* **10**:768

Contents

1 Quick start

```
## Loading library and data
library(STAN)
data(trainRegions)
data(pilot.hg19)

## Model initialization
hmm_nb = initHMM(trainRegions[1:3], nStates=10, "NegativeBinomial")

## Model fitting
hmm_fitted_nb = fitHMM(trainRegions[1:3], hmm_nb, maxIters=10)

## Calculate state path
viterbi_nb = getViterbi(hmm_fitted_nb, trainRegions[1:3])

## Convert state path to GRanges object
viterbi_nb_gm12878 = viterbi2GRanges(viterbi_nb, pilot.hg19, 200)
```

2 Introduction

Genome segmentation with hidden Markov models has become a useful tool to annotate genomic elements, such as promoters and enhancers by data integration. *STAN* (genomic **ST**ate **AN**notation) implements (bidirectional) Hidden Markov Models (HMMs) using a variety of different probability distributions, which can be used to model a wide range of current genomic data:

- Multivariate gaussian: e.g. conutinuous microarray and transformed sequencing data.
- Poisson: e.g. discrete count data from sequencing experiments.
- (Zero-Inflated) Negative Binomial: e.g. discrete count data from sequencing experiments.
- Poisson Log-Normal: e.g. discrete count data from sequencing experiments.
- Negative Multinomial: e.g. discrete count data from sequencing experiments.
- Multinomial: e.g. methylation rates from bisulfite sequencing (in this case it reduces to a Binomial) or binned nuceotide frequencies.
- Bernoulli: Initially proposed by [1] to model presence/absence of chromatin marks (another example: transcription factor binding).
- Nucleotide distribution: e.g. nucleotide frequencies in the DNA sequence.

The use of these distributions enables integrating a wide range of genomic data types (e.g. continuous, discrete, binary) to *de novo* learn and annotate the genome into a given number of 'genomic states'. The 'genomic states' may for instance reflect distinct genome-associated protein complexes or describe recurring patterns of chromatin features, i.e. the 'chromatin state'. Unlike other tools, *STAN* also allows for the integration of strand-specific (e.g. RNA) and non-strand-specific data (e.g. ChIP) [2]. In this vignette, we illustrate the use of *STAN*

by inferring 'chromatin states' from a small example data set of two Roadmap Epigenomics cell lines. Moreover, we show how to use strand-specific RNA expression with non-strand-specific ChIP data to infer 'transcription states' in yeast. Before getting started the package needs to be loaded:

```
library(STAN)
```

3 Genomic state annotation of Roadmap Epigenomics Sequencing data

The data (or observations) provided to *STAN* may consist of one or more observation sequences (e.g. chromosomes), which are contained in a list of (position x experiment) matrices. `trainRegions` is a list containing one three ENCODE pilot regions (stored in `pilot.hg19` as *GRanges* object) with data for two cell types (K562: E123, Gm12878: E116) from the Roadmap Epigenomics project. The data set contains ChIP-Seq experiments of seven histone modifications (H3K4me1, H3K4me3, H3K36me3, H3K27me3, H3K27ac and H3K9ac), as well as DNase-Seq and genomic input.

```
data(trainRegions)
names(trainRegions)
## [1] "E116.chr1.ENr231"  "E116.chr10.ENr114" "E116.chr11.ENm011"
## [4] "E123.chr1.ENr231"  "E123.chr10.ENr114" "E123.chr11.ENm011"
str(trainRegions[c(1,4)])
## List of 2
##  $ E116.chr1.ENr231: num [1:2500, 1:9] 2 4 2 0 1 1 4 4 2 14 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : NULL
##   .. ..$ : chr [1:9] "H3K4me1" "H3K4me3" "H3K36me3" "H3K27me3" ...
##  $ E123.chr1.ENr231: num [1:2500, 1:9] 2 1 5 0 1 3 8 8 7 12 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : NULL
##   .. ..$ : chr [1:9] "H3K4me1" "H3K4me3" "H3K36me3" "H3K27me3" ...
```

The genomic regions for each cell type in `trainRegions` are stored as a GRanges object in `pilot.hg19`:

```
data(pilot.hg19)
pilot.hg19
## GRanges object with 3 ranges and 1 metadata column:
##       seqnames              ranges strand |        name
##          <Rle>           <IRanges>  <Rle> | <character>
##   [1]     chr1 151158001-151658000      * |      ENr231
##   [2]    chr10   55483801-55983800      * |      ENr114
##   [3]    chr11     1743401-2349400      * |      ENm011
##   -------
##   seqinfo: 21 sequences from an unspecified genome; no seqlengths
```

Before model fitting, we calculate size factors to correct for the different different sequencing depths between cell lines.

```
celltypes = list("E123"=grep("E123", names(trainRegions)),
        "E116"=grep("E116", names(trainRegions)))
sizeFactors = getSizeFactors(trainRegions, celltypes)
sizeFactors
##       H3K4me1   H3K4me3  H3K36me3 H3K27me3   H3K9me3  H3K27ac    H3K9ac
## E123 1.055900 0.9104679 0.9397551 0.946867 1.2351436 1.112482 1.2326048
## E123 1.055900 0.9104679 0.9397551 0.946867 1.2351436 1.112482 1.2326048
## E123 1.055900 0.9104679 0.9397551 0.946867 1.2351436 1.112482 1.2326048
## E116 0.949721 1.1090610 1.0684983 1.059451 0.8400696 0.908175 0.8412481
## E116 0.949721 1.1090610 1.0684983 1.059451 0.8400696 0.908175 0.8412481
## E116 0.949721 1.1090610 1.0684983 1.059451 0.8400696 0.908175 0.8412481
##          DNase     Input
## E123 0.9252687 1.2844840
## E123 0.9252687 1.2844840
## E123 0.9252687 1.2844840
## E116 1.0878636 0.8186808
## E116 1.0878636 0.8186808
## E116 1.0878636 0.8186808
```

Genome segmentation is carried out in *STAN* using three functions: `initHMM`, `fitHMM` and `getViterbi`. `initHMM` initializes a model with `nStates` states for a given probability/emission distribution, which we set to 'PoissonLogNormal' in this example. `fitHMM` then optimizes model parameters using the Expectation-Maximization algorithm. Model parameters can be accessed with the `EmissionParams` function. Note that in this example, we set the maximal number of iteration to 10 in this case for speed reason. To ensure convergence this number should be higher in real world applications. After HMM fitting, the state annotation is calculated using the `getViterbi` function.

```
nStates = 10
hmm_poilog = initHMM(trainRegions, nStates, "PoissonLogNormal", sizeFactors)
hmm_fitted_poilog = fitHMM(trainRegions, hmm_poilog, sizeFactors=sizeFactors, maxIters=10)
viterbi_poilog = getViterbi(hmm_fitted_poilog, trainRegions, sizeFactors)
str(viterbi_poilog)
## List of 6
##  $ E116.chr1.ENr231 : Factor w/ 10 levels "1","2","3","4",..: 5 5 5 5 5 5 7 7 7 7 ...
##  $ E116.chr10.ENr114: Factor w/ 10 levels "1","2","3","4",..: 10 10 10 10 10 10 10 10 10 10 ...
##  $ E116.chr11.ENm011: Factor w/ 10 levels "1","2","3","4",..: 10 10 10 10 10 10 10 10 10 10 ...
##  $ E123.chr1.ENr231 : Factor w/ 10 levels "1","2","3","4",..: 5 5 5 5 5 5 7 7 6 6 ...
##  $ E123.chr10.ENr114: Factor w/ 10 levels "1","2","3","4",..: 10 10 10 10 10 10 10 10 10 10 ...
##  $ E123.chr11.ENm011: Factor w/ 10 levels "1","2","3","4",..: 10 10 10 10 10 10 10 10 10 10 ...
```

In order to ease the use of other genomic applications and Bioconductor packages, the viterbi path can be converted into a *GRanges* object.

```
viterbi_poilog_gm12878 = viterbi2GRanges(viterbi_poilog[1:3], regions=pilot.hg19, binSize=200)
viterbi_poilog_gm12878
## GRanges object with 387 ranges and 1 metadata column:
```

```
##         seqnames              ranges strand |        name
##            <Rle>           <IRanges>  <Rle> | <character>
##     [1]     chr1 151158001-151159201      * |           5
##     [2]     chr1 151159201-151160801      * |           7
##     [3]     chr1 151160801-151161001      * |           6
##     [4]     chr1 151161001-151161601      * |           1
##     [5]     chr1 151161601-151163001      * |           4
##     ...      ...                 ...    ... .         ...
##   [383]    chr11     2324601-2326601      * |          10
##   [384]    chr11     2326601-2327601      * |           7
##   [385]    chr11     2327601-2328001      * |           6
##   [386]    chr11     2328001-2328201      * |           7
##   [387]    chr11     2328201-2349401      * |          10
##   -------
##   seqinfo: 3 sequences from an unspecified genome; no seqlengths
```

Before giving some more details about further analysis and visualization of the models we repeat above segementations using the 'NegativeBinomial' emission functions.

```
hmm_nb = initHMM(trainRegions, nStates, "NegativeBinomial", sizeFactors)
hmm_fitted_nb = fitHMM(trainRegions, hmm_nb, sizeFactors=sizeFactors, maxIters=10)
viterbi_nb = getViterbi(hmm_fitted_nb, trainRegions, sizeFactors=sizeFactors)
viterbi_nb_gm12878 = viterbi2GRanges(viterbi_nb[1:3], pilot.hg19, 200)
```

In order to assign biologically meaningful roles to the inferred states we calculate the mean number of reads per 200 base pair bin for both segmentations.

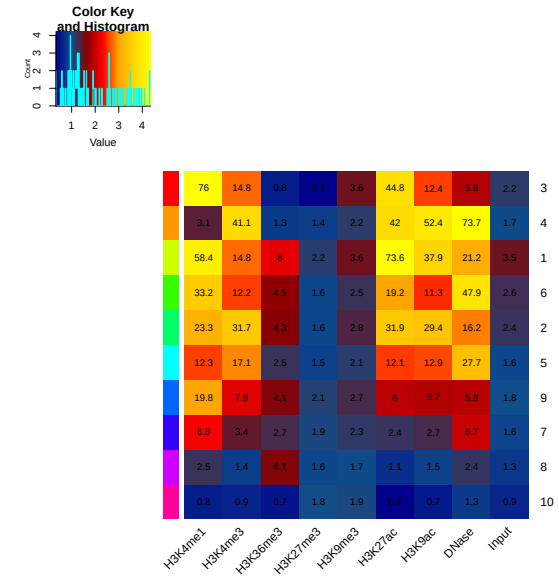
```
avg_cov_nb = getAvgSignal(viterbi_nb, trainRegions)
avg_cov_poilog = getAvgSignal(viterbi_poilog, trainRegions)
```

These are then plotted using the `heatmap.2` function (see Figure 1).

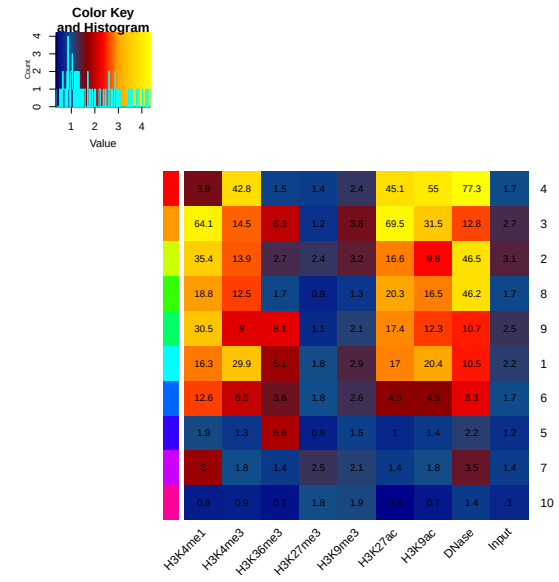
```
## specify color palette
library(gplots)
heat = c("dark blue", "dodgerblue4", "darkred", "red", "orange", "gold", "yellow")
colfct = colorRampPalette(heat)
colpal_statemeans = colfct(200)

## define state order and colors
ord_nb = order(apply(avg_cov_nb,1,max), decreasing=TRUE)
statecols_nb = rainbow(nStates)
names(statecols_nb) = ord_nb
heatmap.2(log(avg_cov_nb+1)[as.character(ord_nb),], margins=c(8,7), srtCol=45,
        RowSideColors=statecols_nb[as.character(ord_nb)], dendrogram="none",
        Rowv=FALSE, Colv=FALSE, col=colpal_statemeans, trace="none",
        cellnote=round(avg_cov_nb,1)[as.character(ord_nb),], notecol="black")
## define state order and colors
ord_poilog = order(apply(avg_cov_poilog,1,max), decreasing=TRUE)
```

```
        statecols_poilog = rainbow(nStates)
        names(statecols_poilog) = ord_poilog
        heatmap.2(log(avg_cov_poilog+1)[as.character(ord_poilog),], margins=c(8,7), srtCol=45,
                RowSideColors=statecols_poilog[as.character(ord_poilog)], dendrogram="none",
                Rowv=FALSE, Colv=FALSE, col=colpal_statemeans, trace="none",
                cellnote=round(avg_cov_poilog,1)[as.character(ord_poilog),], notecol="black")
```



(a)



(b)

Figure 1: Mean read counts of the (a) 'NegativeBinomial' and (b) 'PoissonLogNormal' state annotation

In order to visualize both *STAN* segmentations, we convert the viterbi paths and the data to *Gviz* objects.

```
library(Gviz)
from = start(pilot.hg19)[3]
to = from+300000
gvizViterbi_nb = viterbi2Gviz(viterbi_nb_gm12878, "chr11", "hg19", from, to, statecols_nb)
gvizViterbi_poilog = viterbi2Gviz(viterbi_poilog_gm12878, "chr11", "hg19", from, to,
        statecols_poilog)
gvizData = data2Gviz(trainRegions[[3]], pilot.hg19[3], binSize = 200, gen = "hg19", col="black", chrom = "ch
```

Then, we use the `plotTracks` function to plot everything (see Figure 2).

```
gaxis = GenomeAxisTrack()
data(ucscGenes)
mySize = c(1,rep(1.2,9), 0.5,0.5,3)
plotTracks(c(list(gaxis), gvizData,gvizViterbi_nb,gvizViterbi_poilog,ucscGenes["chr11"]),
        from=from, to=to, showFeatureId=FALSE, featureAnnotation="id", fontcolor.feature="black",
        cex.feature=0.7, background.title="darkgrey", lwd=2, sizes=mySize)
```

Modeling Sequencing data using other emission functions

In this section we illustrate the use of other distributions to annotate the the Roadmap Epigenomics example data set, namely the 'Poisson', 'NegativeMultinomial', 'Gaussian' and 'Bernoulli' models. The 'Poisson' model is an obvious choice when dealing with count data. However since the variance of the Poisson is equal to its mean it might not be an ideal choice for modeling Sequencing experiments, which have been shown to be overdispersed [3].

```
hmm_pois = initHMM(trainRegions, nStates, "Poisson")
hmm_fitted_pois = fitHMM(trainRegions, hmm_pois, maxIters=10)
viterbi_pois = getViterbi(hmm_fitted_pois, trainRegions)
```

The 'NegativeMultinomial' distribution for genome segmentation with HMMs was first proposed in the EpicSeg model [4]. The Negative Multinomial can be understood as a Multinomial distribution, where its overdispersion of is modeled by a Negative Binomial distribution. However, this assumes a shared overdispersion across data tracks within a state as opposed to the 'NegativeBinomial' and 'PoissonLogNormal' models which model the variance for each state and data track separately. In order to use the 'NegativeMultinomial' in *STAN* an additional data track - the sum of counts - for each bin needs to be added to the data. Internally the 'NegativeMultinomial' is modeled as a product of a 'NegativeBinomial' and a 'Multinomial' emission (see section 'Combining different emission functions' for further details):

```
simData_nmn = lapply(trainRegions, function(x) cbind(apply(x,1,sum), x))
hmm_nmn = initHMM(simData_nmn, nStates, "NegativeMultinomial")
hmm_fitted_nmn = fitHMM(simData_nmn, hmm_nmn, maxIters=10)
viterbi_nmn = getViterbi(hmm_fitted_nmn, simData_nmn)
```

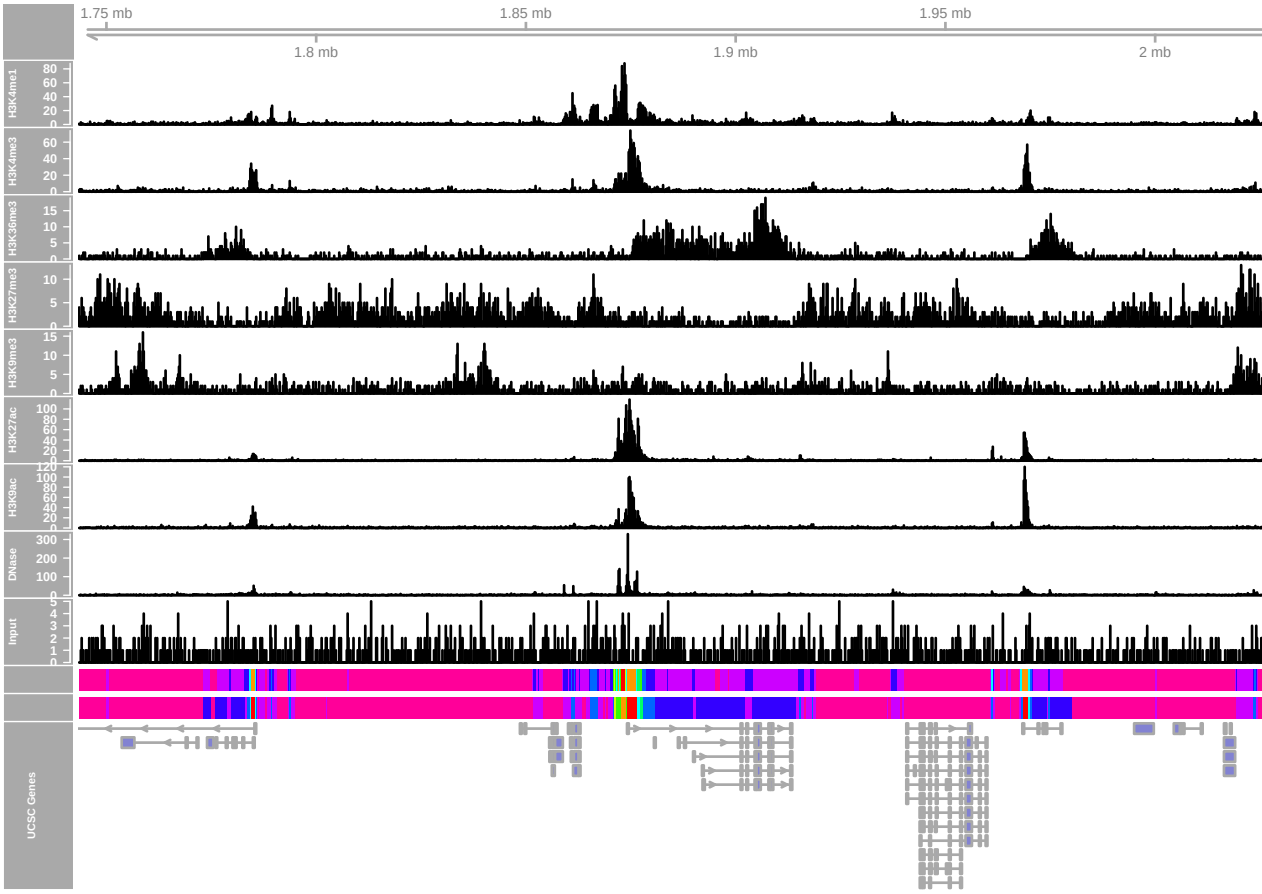


Figure 2: **Genome Browser showing the 10 data tracks used for model learning together with the 'Negativebinomial' (top) and 'PoissonLogNormal' (bottom) segmentations and known UCSC gene annotations**

In order to model the data using Gaussian distributions, it needs to be log-transformed and smoothed. This approach is implementd in Segway, a method used by the ENCODE Consortium for chromatin state annotation [5]. However, to overcome singularity of the (diagonal) covariance matrix due to the zero-inflated distribution of the transformed read counts, it uses a shared variance over states for each data track. To use gaussian distributions with Sequencing data in *STAN*, we transform the data (with the hyperbole sine function [5]) and model it using the emission 'IndependentGaussian' with a shared covariance, i.e. `sharedCov=TRUE`.

```
trainRegions_smooth = lapply(trainRegions, function(x)
            apply(log(x+sqrt(x^2+1)), 2, runningMean, 2))
hmm_gauss = initHMM(trainRegions_smooth, nStates, "IndependentGaussian", sharedCov=TRUE)
hmm_fitted_gauss = fitHMM(trainRegions_smooth, hmm_gauss, maxIters=10)
viterbi_gauss = getViterbi(hmm_fitted_gauss, trainRegions_smooth)
```

Another approach was proposed in ChromHMM, which models binarized data using an independent Bernoulli model [1]. Note, that the performance of the model highly depends on the non-trivial choice of a proper cutoff and quantitative information is lost. The latter

is especially important when predicting promoters and enhancers since these elements are both marked H3K4me1 and H3K4me3, but at different ratios. The function `binarizeData` binarizes the data using the default approach by ChromHMM [1]. The model can then be fit by specifying the 'Bernoulli' model. Note however, that initialization and model fitting are carried out differently than in the ChromHMM implementation. In particular *STAN* uses the EM algorithm while ChromHMM uses online EM. For details on the initialization, please see the `initHMM` manual.

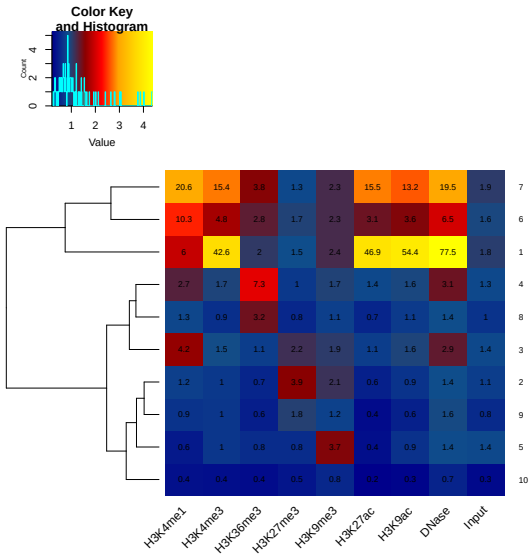
```
trainRegions_binary = binarizeData(trainRegions)
hmm_ber = initHMM(trainRegions_binary, nStates, "Bernoulli")
hmm_fitted_ber = fitHMM(trainRegions_binary, hmm_ber, maxIters=10)
viterbi_ber = getViterbi(hmm_fitted_ber, trainRegions_binary)
```

We calculate the mean read coverage for each method and segmentation:

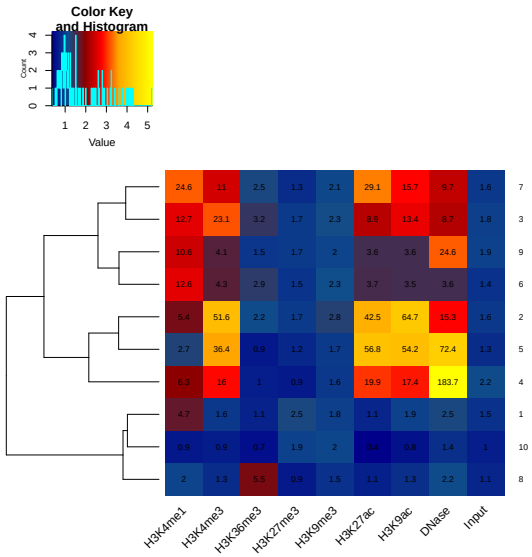
```
avg_cov_gauss = getAvgSignal(viterbi_gauss, trainRegions)
avg_cov_nmn = getAvgSignal(viterbi_nmn, trainRegions)
avg_cov_ber = getAvgSignal(viterbi_ber, trainRegions)
avg_cov_pois = getAvgSignal(viterbi_pois, trainRegions)
```

These are again plotted using the `heatmap.2` function (see Figure 3).

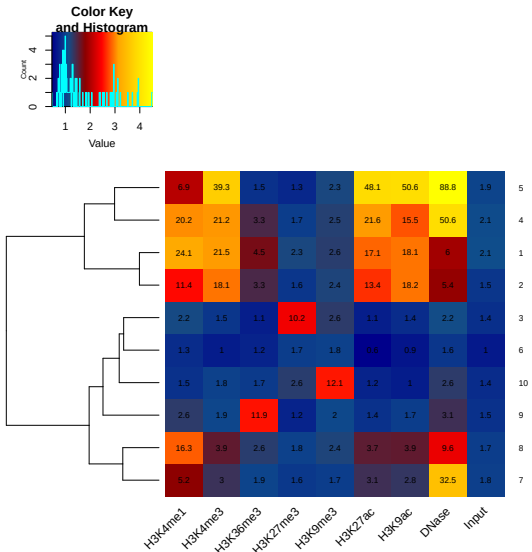
```
heatmap.2(log(avg_cov_gauss+1), margins=c(8,7),srtCol=45, dendrogram="row", Rowv=TRUE,
        Colv=FALSE, col=colpal_statemeans, trace="none", notecex=0.7, cexRow=0.75, cexCol=1,
        cellnote=round(avg_cov_gauss,1), notecol="black")
heatmap.2(log(avg_cov_nmn+1), margins=c(8,7),srtCol=45, dendrogram="row", Rowv=TRUE,
        Colv=FALSE, col=colpal_statemeans, trace="none", notecex=0.7, cexRow=0.75, cexCol=1,
        cellnote=round(avg_cov_nmn,1), notecol="black")
heatmap.2(log(avg_cov_ber+1), margins=c(8,7),srtCol=45, dendrogram="row", Rowv=TRUE,
        Colv=FALSE, col=colpal_statemeans, trace="none", notecex=0.7, cexRow=0.75, cexCol=1,
        cellnote=round(avg_cov_ber,1), notecol="black")
heatmap.2(log(avg_cov_pois+1), margins=c(8,7),srtCol=45, dendrogram="row", Rowv=TRUE,
        Colv=FALSE, col=colpal_statemeans, trace="none", notecex=0.7, cexRow=0.75, cexCol=1,
        cellnote=round(avg_cov_pois,1), notecol="black")
```


Color Key and Histogram
Count
Value
H3K4me1
H3K4me3
H3K36me3
H3K27me3
H3K9me3
H3K27ac
H3K9ac
DNase
Input

(a)


Color Key and Histogram
Count
Value
H3K4me1
H3K4me3
H3K36me3
H3K27me3
H3K9me3
H3K27ac
H3K9ac
DNase
Input

(b)


Color Key and Histogram
Count
Value
H3K4me1
H3K4me3
H3K36me3
H3K27me3
H3K9me3
H3K27ac
H3K9ac
DNase
Input

4 Integrating strand-specific and non-strand-specific data with STAN

STAN also allows for the integration of strand-specific (e.g. RNA) and non-strand-specific data (e.g. ChIP). This is donw using bidirectional hidden Markov models (bdHMMs) which were proposed in [2]. A bdHMM models a directed process using the concept of twin states, where each genomic state is split up into a pair of twin states, one for each direction (e.g. sense and antisense in context of transcription). Those twin state pairs are identical in terms of their emissions (i.e. they model the same genomic state). Currently the following models are available for bdHMMs: 'IndependentGaussian', 'Gaussian', 'NegativeBinomial', 'ZINegativeBinomial' and 'PoissonLogNormal'. We now illustrate the use of bdHMMs in *STAN* at an example data set of yeast transcription factors measured by ChIP-chip and RNA expression measured with a tiling array which was used to model the transcription cycle as a sequence of 'transcription states' in [2].
The `initBdHMM` function is used to initialize a bdHMM with 6 twin states. Note that the overall number of states in the bdHMM is 12 (6 identical twin state pairs). `dirobs` defines the directionality (or strand-specificity) of the data tracks. In `dirobs`, the first 10 data tracks are non-strand-specific ChIP-chip measurments, indicated by '0' and data track 11 and 12 are strand-specific RNA expression measurements, indicated by '1'. Note that strand-specific data tracks must be labeled as increasing pairs of integers. Thus and additional strand-specific data track pair would be labeled as a pair of '2'. Model fitting and calculation of the state annotation are carried out as for standard HMMs:

```
data(yeastTF_databychrom_ex)
dStates = 6
dirobs = as.integer(c(rep(0,10), 1, 1))
bdhmm_gauss = initBdHMM(yeastTF_databychrom_ex, dStates = dStates, method = "Gaussian", directedObs=dirobs)
bdhmm_fitted_gauss = fitHMM(yeastTF_databychrom_ex, bdhmm_gauss)
viterbi_bdhmm_gauss = getViterbi(bdhmm_fitted_gauss, yeastTF_databychrom_ex)
```

We plot the means of the multivariate gaussian distrbutions for each state (see Figure 4):

```
statecols_yeast = rep(rainbow(nStates), 2)
names(statecols_yeast) = StateNames(bdhmm_fitted_gauss)
means_fitted = EmissionParams(bdhmm_fitted_gauss)$mu
heatmap.2(means_fitted, col=colpal_statemeans,
        RowSideColors=statecols_yeast[rownames(means_fitted)],
        trace="none", cexCol=0.9, cexRow=0.9,
        cellnote=round(means_fitted,1), notecol="black", dendrogram="row",
        Rowv=TRUE, Colv=FALSE, notecex=0.9)
```

We convert the viterbi path into a `GRanges` object. Note that the directionaliy of bdHMM states is indicated by 'F' (forward) and 'R' (reverse).

```
yeastGRanges = GRanges(IRanges(start=1214616, end=1225008), seqnames="chrIV")
names(viterbi_bdhmm_gauss) = "chrIV"
viterbi_bdhmm_gauss_gr = viterbi2GRanges(viterbi_bdhmm_gauss, yeastGRanges, 8)
```

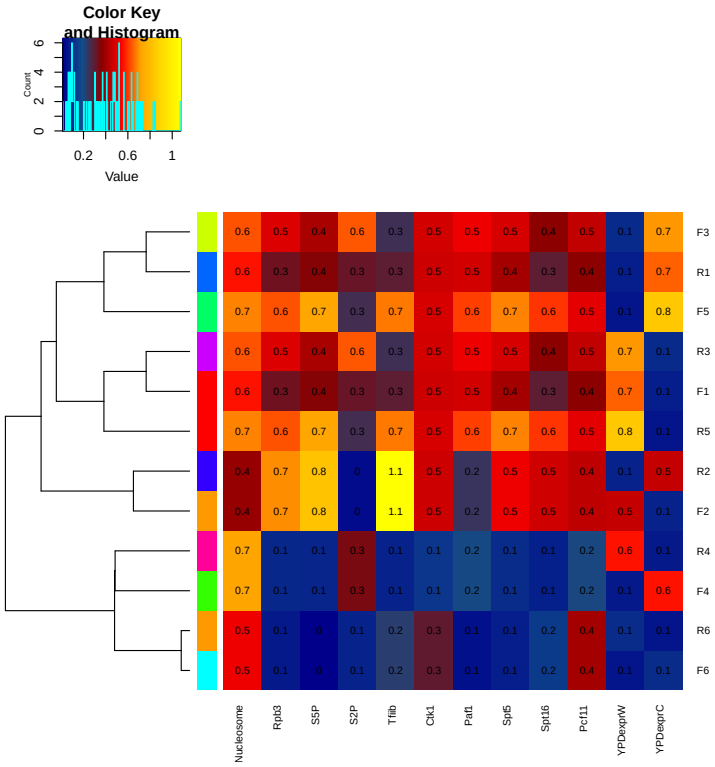


Figure 4: Mean signal 6 bdHMM twin state pairs
'F' and 'R' indicate forward and reverse direction of state pairs.

```
viterbi_bdhmm_gauss_gr
## GRanges object with 48 ranges and 1 metadata column:
##        seqnames          ranges strand |        name
##           <Rle>       <IRanges>  <Rle> | <character>
##    [1]    chrIV 1214616-1215016      * |          R3
##    [2]    chrIV 1215016-1215088      * |          F3
##    [3]    chrIV 1215088-1215224      * |          F4
##    [4]    chrIV 1215224-1215264      * |          R1
##    [5]    chrIV 1215264-1216808      * |          F4
##    ...      ...             ...    ... .         ...
##   [44]    chrIV 1224680-1224696      * |          F6
##   [45]    chrIV 1224696-1224752      * |          R6
##   [46]    chrIV 1224752-1224880      * |          R5
##   [47]    chrIV 1224880-1224928      * |          F1
##   [48]    chrIV 1224928-1225008      * |          R5
##   -------
##   seqinfo: 1 sequence from an unspecified genome; no seqlengths
```

Next, we visualize the data, state annotation and together with SGD genes using *Gviz* (see Figure 5):

```
        chr = "chrIV"
        gen = "sacCer3"
        gtrack <- GenomeAxisTrack()

        from=1217060
        to=1225000
        forward_segments = grep("F", viterbi_bdhmm_gauss_gr$name)
        reverse_segments = grep("R", viterbi_bdhmm_gauss_gr$name)
        gvizViterbi_yeast = viterbi2Gviz(viterbi_bdhmm_gauss_gr[forward_segments],
                "chrIV", "sacCer3", from, to, statecols_yeast)
        gvizViterbi_yeast2 = viterbi2Gviz(viterbi_bdhmm_gauss_gr[reverse_segments],
                "chrIV", "sacCer3", from, to, statecols_yeast)

        gvizData_yeast = data2Gviz(obs = yeastTF_databychrom_ex[[1]], regions = yeastGRanges, binSize = 8, gen = "sac
        gaxis = GenomeAxisTrack()
        data(yeastTF_SGDGenes)
        mySize = c(1,rep(1,12), 0.5,0.5,3)

        plotTracks(c(list(gaxis), gvizData_yeast,gvizViterbi_yeast,gvizViterbi_yeast2,
                list(yeastTF_SGDGenes)), cex.feature=0.7, background.title="darkgrey", lwd=2,
                sizes=mySize, from=from, to=to, showFeatureId=FALSE, featureAnnotation="id",
                fontcolor.feature="black", cex.feature=0.7, background.title="darkgrey",
                showId=TRUE)
```

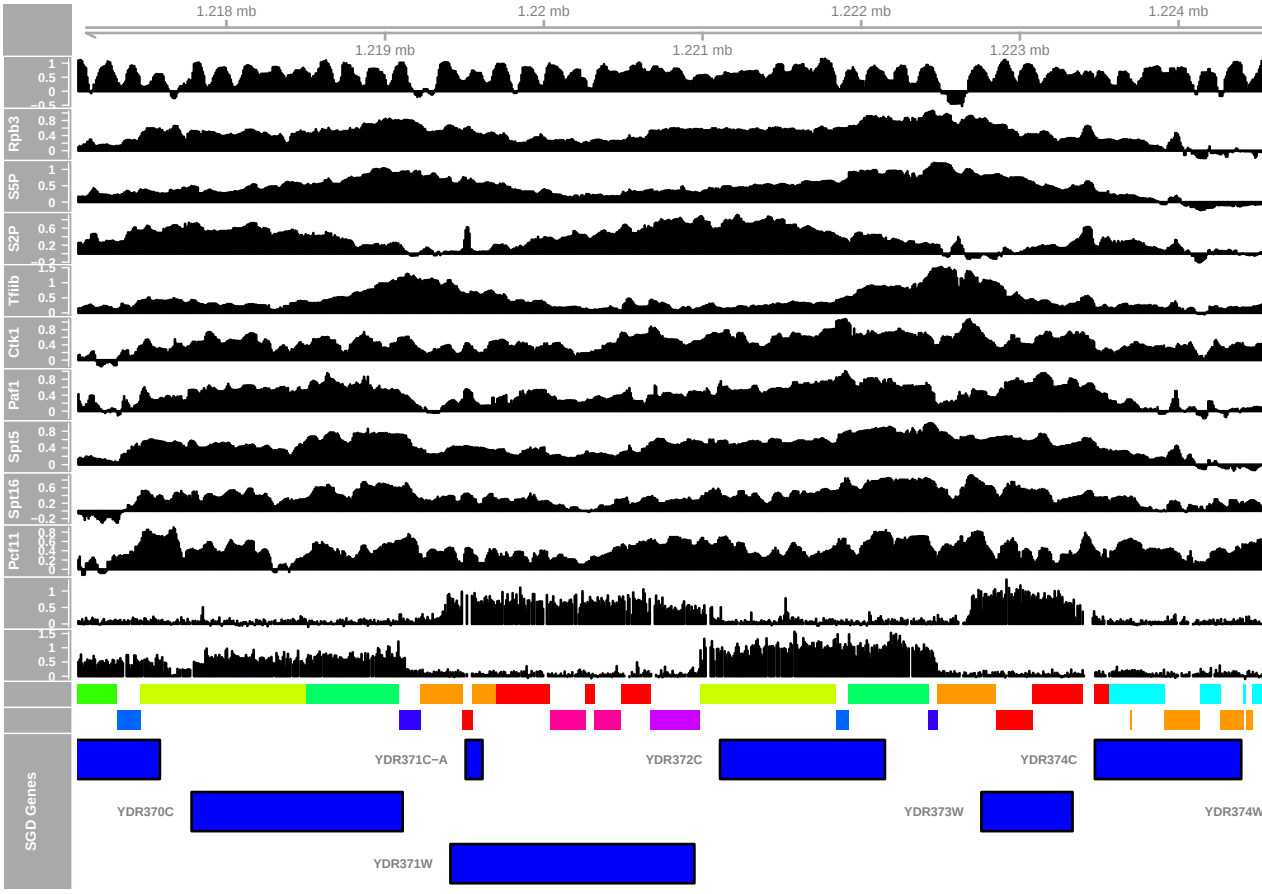


Figure 5: Genome Browser showing the 12 data tracks used for model learning together with the segmentations and known SGD gene annotations

5 Concluding Remarks

This vignette was generated using the following package versions:

- R version 3.6.1 (2019-07-05), `x86_64-w64-mingw32`
- Locale: `LC_COLLATE=C`, `LC_CTYPE=English_United States.1252`, `LC_MONETARY=English_United States.1252`, `LC_NUMERIC=C`, `LC_TIME=English_United States.1252`
- Running under: `Windows Server 2012 R2 x64 (build 9600)`
- Matrix products: default
- Base packages: base, datasets, grDevices, graphics, grid, methods, parallel, stats, stats4, utils
- Other packages: BiocGenerics 0.32.0, GenomeInfoDb 1.22.0, GenomicRanges 1.38.0, Gviz 1.30.0, IRanges 2.20.0, S4Vectors 0.24.0, STAN 2.14.0, gplots 3.0.1.1, knitr 1.25, poilog 0.4

- Loaded via a namespace (and not attached): AnnotationDbi 1.48.0, AnnotationFilter 1.10.0, BSgenome 1.54.0, Biobase 2.46.0, BiocFileCache 1.10.0, BiocManager 1.30.9, BiocParallel 1.20.0, BiocStyle 2.14.0, Biostrings 2.54.0, DBI 1.0.0, DelayedArray 0.12.0, Formula 1.2-3, GenomeInfoDbData 1.2.2, GenomicAlignments 1.22.0, GenomicFeatures 1.38.0, Hmisc 4.2-0, KernSmooth 2.23-16, Matrix 1.2-17, ProtGenerics 1.18.0, R6 2.4.0, RColorBrewer 1.1-2, RCurl 1.95-4.12, RSQLite 2.1.2, Rcpp 1.0.2, Rsamtools 2.2.0, Rsolnp 1.16, SummarizedExperiment 1.16.0, VariantAnnotation 1.32.0, XML 3.98-1.20, XVector 0.26.0, acepack 1.4.1, askpass 1.1, assertthat 0.2.1, backports 1.1.5, base64enc 0.1-3, biomaRt 2.42.0, biovizBase 1.34.0, bit 1.1-14, bit64 0.9-7, bitops 1.0-6, blob 1.2.0, caTools 1.17.1.2, checkmate 1.9.4, cluster 2.1.0, colorspace 1.4-1, compiler 3.6.1, crayon 1.3.4, curl 4.2, data.table 1.12.6, dbplyr 1.4.2, dichromat 2.0-0, digest 0.6.22, dplyr 0.8.3, ensembldb 2.10.0, evaluate 0.14, foreign 0.8-72, gdata 2.18.0, ggplot2 3.2.1, glue 1.3.1, gridExtra 2.3, gtable 0.3.0, gtools 3.8.1, hms 0.5.1, htmlTable 1.13.2, htmltools 0.4.0, htmlwidgets 1.5.1, httr 1.4.1, lattice 0.20-38, latticeExtra 0.6-28, lazyeval 0.2.2, magrittr 1.5, matrixStats 0.55.0, memoise 1.1.0, munsell 0.5.0, nnet 7.3-12, openssl 1.4.1, pillar 1.4.2, pkgconfig 2.0.3, prettyunits 1.0.2, progress 1.2.2, purrr 0.3.3, rappdirs 0.3.1, rlang 0.4.1, rmarkdown 1.16, rpart 4.1-15, rstudioapi 0.10, rtracklayer 1.46.0, scales 1.0.0, splines 3.6.1, stringi 1.4.3, stringr 1.4.0, survival 2.44-1.1, tibble 2.1.3, tidyselect 0.2.5, tools 3.6.1, truncnorm 1.0-8, vctrs 0.2.0, xfun 0.10, yaml 2.2.0, zeallot 0.1.0, zlibbioc 1.32.0

References

[1] J. Ernst and M. Kellis. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.*, 28(8):817–825, Aug 2010.

[2] B. Zacher, M. Lidschreiber, P. Cramer, J. Gagneur, and A. Tresch. Annotation of genomics data using bidirectional hidden Markov models unveils variations in Pol II transcription cycle. *Mol. Syst. Biol.*, 10:768, 2014.

[3] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biol.*, 11(10):R106, 2010.

[4] Alessandro Mammana and Ho-Ryun Chung. Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome. *Genome Biology*, 16(1):151, 2015.

[5] Michael M Hoffman, Orion J Buske, Jie Wang, Zhiping Weng, Jeff a Bilmes, and William Stafford Noble. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods*, 9(5):473–476, 2012.