biomvRCNS: Copy Number study and Segmentation for multivariate biological data.

Yang Du*

August 27, 2013

Abstract

With high throughput experiments like tiling array and NGS, researchers are looking for continuous homogeneous segments or signal peaks, which would represent chromatin states, methylation ratio, transcripts or genome regions of deletion and amplification. While in a normal experimental set-up, these profiles would be generated for multiple samples or conditions with replicates. In the package *biomvRCNS*, a Hidden Semi Markov Model and one homogeneous segmentation model are implemented and tailored to handle multiple genomic profiles, with the aim of assisting in transcripts detection using high throughput technology like RNA-seq or tiling array, and copy number analysis using aCGH or targeted sequencing.

1 Introduction

To begin with *biomvRCNS*, load the package and read the manual page.

```
> library(biomvRCNS)
```

In the package, 3 main functions are provided for the batch processing of multiple chromosome regions across samples: `biomvRhsmm`, a hidden semi Markov model (HSMM); `biomvRseg`, a maximum likelihood based homogeneous segmentation model; and a third `biomvRmgmr`, custom batch function using max-gap-min-run algorithm. In the following sections we will illustrate their functionalities using example data. Currently the package does not deal with data correction, so input should be normalized by reference or paired sample and corrected for factor of interest before passing down.

2 Example of array CGH data set of Coriell cell lines

Extracted from packge *DNACopy* [Olshen et al., 2004], the `coriell` data contains two aCGH studies (GM05296 and GM13330) of Corriel cell lines taken from Snijders et al. [2001]. In particular, with 2271 mapped features in total across 22 autosomes and chromosome X.

All three main functions accept common data matrix plus positional information as input or a *GRanges* object with data matrix stored in the meta columns. To get started, we first build a *GRanges* object from *data.frame*.

```
> data('coriell', package='biomvRCNS')
> head(coriell, n=3)
```

*email: duyang@fbn-dummerstorf.de

```
       Clone Chromosome Position Coriell.05296 Coriell.13330
1 GS1-232B23          1        1      0.000359      0.207470
2 RP11-82d16          1      469      0.008824      0.063076
3 RP11-62m23          1     2242     -0.000890      0.123881
```

```
> xgr<-GRanges(seqnames=paste('chr', coriell[,2], sep=''),
+         IRanges(start=coriell[,3], width=1, names=coriell[,1]))
> values(xgr)<-DataFrame(coriell[,4:5], row.names=NULL)
> xgr<-sort(xgr)
> head(xgr, n=3)
```

```
GRanges with 3 ranges and 2 metadata columns:
             seqnames       ranges strand | Coriell.05296 Coriell.13330
                <Rle>    <IRanges>  <Rle> |     <numeric>     <numeric>
  GS1-232B23     chr1 [   1,    1]      * |      0.000359       0.20747
  RP11-82d16     chr1 [ 469,  469]      * |      0.008824      0.063076
  RP11-62m23     chr1 [2242, 2242]      * |      -0.00089      0.123881
  ---
  seqlengths:
    chr1 chr10 chr11 chr12 chr13 chr14 chr15 chr16 ...  chr3  chr4  chr5  chr6  chr7  chr8  chr9
      NA    NA    NA    NA    NA    NA    NA    NA ...    NA    NA    NA    NA    NA    NA    NA
```

Please be sure that the data is sorted with respect to their positions before feeding to the models.

2.1 Genomic segmentation with Hidden-semi Markov model

First we use the hidden-semi Markov model with the batch function `biomvRhsmm`, which will sequentially process each chromosome identified by the *seqnames* (using dummy name when no *GRanges* supplied in `x` or `xRange`), thus for non-continuous regions on the same chromosome user should give different *seqnames* to each part of the data. Within this package, there is one argument *grp*, for all main batch functions, which is used to assign data columns to groups according to the experimental design, say technical replicates or biological replicates. Sample columns within the same group could be treated simultaneously in the modelling process as well as iteratively. [1] In this example, the two profiles are considered independent and not similar, thus been given different values in the *grp* vector. Additionally there is a built-in automatic grouping method, given a valid clustering method *cluster.m* and *grp* set to `NULL`. By default, all data columns are assumed to be from the same group.

```
> rhsmm<-biomvRhsmm(x=xgr, maxbp=4E4, J=3, soj.type='gamma',
+         com.emis=F, emis.type='norm', prior.m='quantile', grp=c(1,2))
```

```
> show(rhsmm)
```

```
Object is of class: 'biomvRCNS'
List of parameters used in the model:
J, maxk, maxbp, maxgap, soj.type, emis.type, q.alpha, r.var, iterative, cMethod, maxit, tol,
grp, cluster.m, avg.m, prior.m, trim, na.rm, soj.par, emis.par

The segmented ranges:
GRanges with 197 ranges and 3 metadata columns:
        seqnames           ranges strand   |        SAMPLE STATE        AVG
```

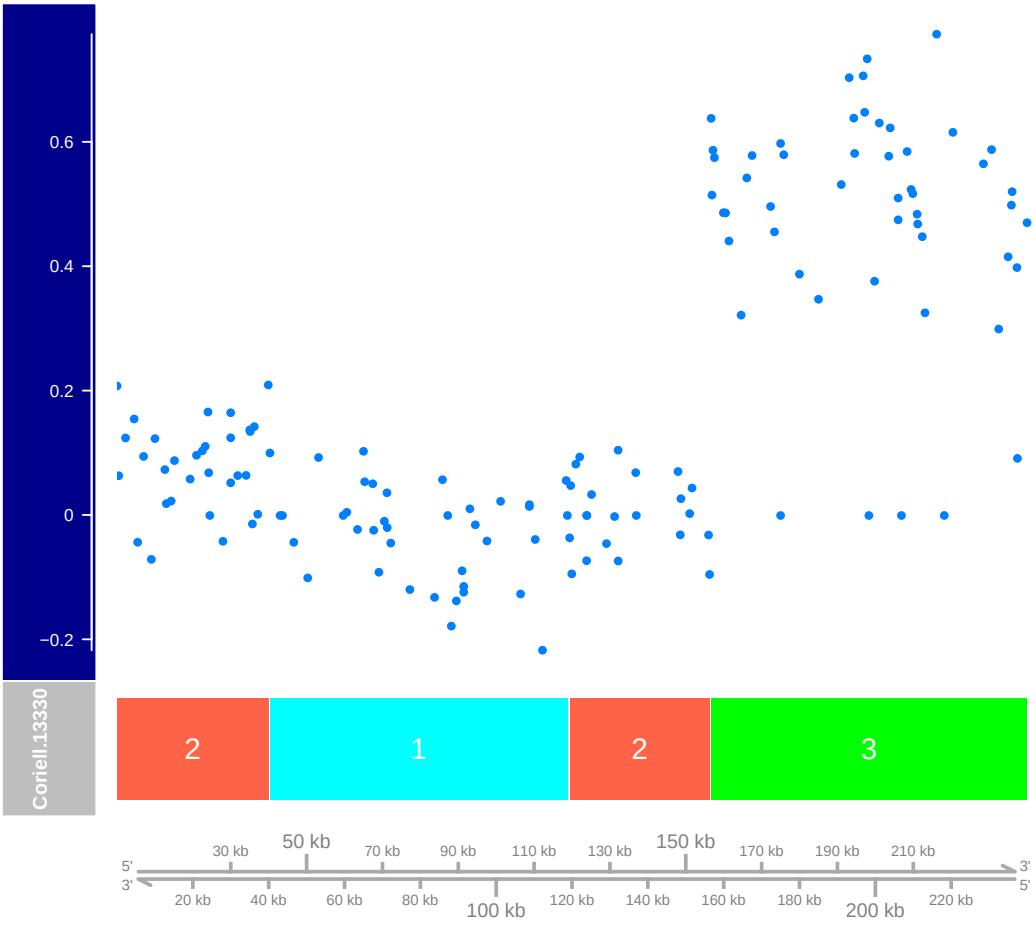
[1] Simultaneous treatment within group is currently available for *emis.type* equals 'mvnorm' or 'mvt' in `biomvRhsmm`, *poolGrp=TRUE* in `biomvRmgmr` and *twoStep=FALSE* in `biomvRseg`.

```
        <Rle>        <IRanges>  <Rle>   |         <Rle> <Rle>      <Rle>
  [1]     chr1 [     1,  35001]      *   | Coriell.05296     2  0.0045915
  [2]     chr1 [ 40327,  85833]      *   | Coriell.05296     2  0.0129660
  [3]     chr1 [ 91001, 132148]      *   | Coriell.05296     2  0.0066565
  [4]     chr1 [147954, 193091]      *   | Coriell.05296     2 -0.0056100
  [5]     chr1 [203907, 235001]      *   | Coriell.05296     2 -0.0028350
  ...      ...              ...    ... ...           ...   ...        ...
[193]     chr9  [33136,  75244]      *   | Coriell.13330     2   -0.02824
[194]     chr9  [78040, 115001]      *   | Coriell.13330     2   -0.01081
[195]     chr9  [ 8639,  14050]      *   | Coriell.13330     1   -0.19048
[196]     chr9  [24326,  33001]      *   | Coriell.13330     1   -0.16646
[197]     chr9  [75733,  77808]      *   | Coriell.13330     3    0.09102
---
seqlengths:
  chr1 chr10 chr11 chr12 chr13 chr14 chr15 chr16 ...  chr3  chr4  chr5  chr6  chr7  chr8  chr9
    NA    NA    NA    NA    NA    NA    NA    NA ...    NA    NA    NA    NA    NA    NA    NA
```

In the above run, we limit the model complexity by setting the *maxbp* to $4E4$, which will restrict the maximum sojourn time to *maxbp*. J is the number of states in the HSMM model, this argument can be given explicitly or estimated from prior information provided in *xAnno*. Argument *soj.type* defines the type of sojourn distribution; with Gamma distributed sojourn, the neighbouring position will tend to have the same state, and transit to other states if far apart. In this way the sojourn distribution fully incorporate the positional information into the probabilistic framework. Argument *emis.type* controls the distribution of emission probability, in this case the log2 ratio of aCGH data is considered to follow Normal distribution. The emission density could be estimated using all data or only data on the repective region or chromosome (identified by unique `seqnames`), controlling via *com.emis*. The prior of the emission parameters could be controlled by supplying *q.alpha* and *r.var* with *prior.m='quantile'*, or automatically determined through a clustering process with *prior.m='cluster'*. The function will then call C codes and estimate the most likely state sequence, with either *cMethod='F-B'* or *cMethod='Viterbi'*. The *F-B* method (default) uses a forward-backward algorithm described in Guédon [2003], which gives a smooth state sequence, whereas the Viterbi algorithm with *cMethod='Viterbi'* will use the state profile estimated by the forward-backward algorithm and rebuild the most likely state sequence. The function returns an object of class *biomvRCNS*, in which the `res` slot is a *GRanges* object containing the summary of each estimated segments. There are three meta columns: column `SAMPLE` gives the column name of which sample this segment belongs to; column `STATE`, the estimated state for each segment, the lower state number represents state with lower mean value, thus in this example, a state of 1 could represent region of deletion and 3 for region of duplication, whereas state 2 could be considered copy neutral; column `AVG`, gives the segment average value, which could take the form of (trimmed) mean or median controlled by *avg.m*. The original input is also kept and returned in slot `x` with the estimated most likely state assignment and associated probability.

A `plot` method has been implemented for *biomvRCNS* object using package *Gviz*, by default the `plot` method tries to output graphics to multiple EPS/PDF files for each chromosome region and sample. Here we set *tofile=FALSE* to output graphics to the current device, and only show resulting graphics for chromosome 1 from sample Coriell.13330.

```
> obj<-biomvRGviz(exprgr=xgr[seqnames(xgr)=='chr1', 'Coriell.13330'],
+         seggr=rhsmm@res[mcols(rhsmm@res)[,'SAMPLE']=='Coriell.13330'], tofile=FALSE)
```



2.2 Using other methods provided in the package

In this section, we use the other two batch functions to process the `coriell` data. First we use `biomvRseg`, in which a similar segmentation method like in the package *tillingArray* [Huber et al., 2006] is implemented and extended to handle Poisson and Negative binomial distributed data. The function shares several argument with `biomvRhsmm`, like *maxbp* and *grp*. The *maxseg* gives the maximum number of segment per chromosome region, while the optimal number of segment per chromosome region is determined internally by assessing the likelihood with optional penalty terms, by default *penalty='BIC'* is used. Another option is to use modified Bayes information criterion *penalty='mBIC'* [Zhang and Siegmund, 2007], as in the CBS algorithm used in *DNAcopy*. The function proceed in the following manner: assuming within each group sample columns exhibit similar patterns, and thus be processed simultaneously in the first step. By maximizing the likelihood the optimal number of segments is selected for each group. And in a second step if *twoStep=TRUE* or merging is necessary, the candidate segments produced in the first step are merged with respect to each sample, thus forcing sample columns in the same group to have a more unified segmentation result yet keeping it possible to have sample specific pattern.

```
> rseg<-biomvRseg(x=xgr, maxbp=4E4, maxseg=10, family='norm', grp=c(1,2))
```

```
> head(rseg@res)
```

```
GRanges with 6 ranges and 3 metadata columns:
      seqnames           ranges strand |        SAMPLE        AVG       STATE
         <Rle>        <IRanges>  <Rle> |   <character>  <numeric> <character>
  [1]     chr1 [     1,  35001]      * | Coriell.05296  0.0045915         LOW
  [2]     chr1 [ 35106,  36207]      * | Coriell.05296  0.1335200        HIGH
  [3]     chr1 [ 37117,  85833]      * | Coriell.05296  0.0129660        HIGH
  [4]     chr1 [ 87215,  89493]      * | Coriell.05296 -0.1010400         LOW
  [5]     chr1 [ 91001, 132148]      * | Coriell.05296  0.0066565         LOW
  [6]     chr1 [132171, 136942]      * | Coriell.05296  0.1421530        HIGH
  ---
  seqlengths:
    chr1 chr10 chr11 chr12 chr13 chr14 chr15 chr16 ...  chr3  chr4  chr5  chr6  chr7  chr8  chr9
      NA    NA    NA    NA    NA    NA    NA    NA ...    NA    NA    NA    NA    NA    NA    NA
```

After the example run, the function returns a *biomvRCNS* object, containing similar information as the previous `biomvRhsmm` run, except that the *STATE* column now only have a binary state value of either "HIGH" or "LOW", which is simply graded as 'HIGH' if the segment average is higher than the grand average of the whole region, and 'LOW' otherwise.

It is also possible to use the simple max-gap-min-run algorithm to segment aCGH profiles, by calling `biomvRmgmr`. But due to the binary nature of the algorithm, one have to run twice in order to get both extremely high and low segments, then combine the resulting *GRanges* manually.

```
> rmgmrh<-biomvRmgmr(xgr, q=0.9, high=T, maxgap=1000, minrun=2500, grp=c(1,2))
> rmgmrl<-biomvRmgmr(xgr, q=0.1, high=F, maxgap=1000, minrun=2500, grp=c(1,2))
> res<-c(rmgmrh@res, rmgmrl@res)
```

3 Example of transcript detection with RNA-seq data from ENCODE

The data contains gene expressions and transcript annotations in the region of the human TP53 gene (chr17:7,560,001-7,610,000 from the Human February 2009 (GRCh37/hg19) genome assembly), which is part of the long RNA-seq data generated by ENCODE [Consortium, 2004] /Cold Spring Harbor Lab, containing 2 cell types (GM12878 and K562) with 2 replicates each. The libraries were sequenced on the Illumina GAIIx platform as paired-ends for 76 or 101 cycles for each read. The average depth of sequencing was 200 million reads (100 million paired-ends). The data were mapped against hg19 using Spliced Transcript Alignment and Reconstruction (STAR).

To generate local read counts, alignment files were pulled from UCSC (`http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCshlLongRnaSeq/`) using package *Rsamtools*. And subsequently reads were counted in each non-overlapping unit sized window for the region (chr17:7,560,001-7,610,000). In the pre-compiled data `encodeTP53` , a window size of 25bp was used with the chunk of code below.

```
> winsize<-25
> cgr<-GRanges("chr17", strand='-',
+         IRanges(start=seq(7560001, 7610000, winsize), width =winsize))
> bf<-system.file("extdata", "encodeFiles.txt", package = "biomvRCNS")
> bamfiles<-read.table(bf, header=T, stringsAsFactors=F)
> library(Rsamtools)
> which<-GRanges("chr17", IRanges(7560001, 7610000))
> param<-ScanBamParam(which=which, what=scanBamWhat())
> for(i in seq_len(nrow(bamfiles))){
+         frd<-scanBam(bamfiles[i,1], param=param)
+         frdgr<-GRanges("chr17", strand=frd[[1]]$strand,
+                 IRanges(start=frd[[1]]$pos , end = frd[[1]]$pos+frd[[1]]$qwidth-1))
```

```
+         mcols(cgr)<-DataFrame(mcols(cgr), DOC=countOverlaps(cgr, frdgr))
+ }
```

Alternatively one can also operate on base pair resolution, in which case a *Rle* object should be preferred to store the count data for lower memory footprint and better efficiency. Also to speed things up, one could set *useMC=T* to enable parallel processing of multiple `seqnames`, the number of parallel process could be set by *options(mc.cores=n)*.

```
> cgr<-GRanges("chr17", strand='-',
+         IRanges(seq(7560001, 7610000), width=1))
> bf<-system.file("extdata", "encodeFiles.txt", package = "biomvRCNS")
> bamfiles<-read.table(bf, header=T, stringsAsFactors=F)
> library(Rsamtools)
> which<-GRanges("chr17", IRanges(7560001, 7610000))
> param<-ScanBamParam(which=which, flag=scanBamFlag(isMinusStrand=TRUE))
> for(i in seq_len(nrow(bamfiles))){
+         cod<-coverage(BamFile(bamfiles[i,1]), param=param)[['chr17']][7560001:7610000]
+         mcols(cgr)<-DataFrame(mcols(cgr), DOC=cod)
+ }
```

The pre-compiled data `encodeTP53` also includes the regional annotation of TP53 RNAs isoforms, `gmgr`, which were derived from the ENCODE Gene Annotations (GENCODE), `http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeGencodeV4/wgEncodeGencodeManualV4.gtf.gz)` , and subset to only isoforms of TP53 gene. The annotation object `gmgr` could be rebuilt with the following lines using the included file under `extdata`.

```
> af<-system.file("extdata", "gmodTP53.csv", package = "biomvRCNS")
> gtfsub<-read.table(af, fill=T, stringsAsFactors=F)
> idx<-gtfsub[,3]=='CDS' | gtfsub[,3]=='UTR'
> gmgr<-GRanges("chr17", IRanges(start=gtfsub[idx, 4], end=gtfsub[idx, 5],
+         names=gtfsub[idx, 13]), strand='-', TYPE=gtfsub[idx, 3])
```

We first load the `encodeTP53` data, poll the read counts for each cell type and add 1 to the base count to increase stability.

```
> data(encodeTP53)
> cgr<-encodeTP53$cgr
> gmgr<-encodeTP53$gmgr
> mcols(cgr)<-DataFrame(
+         Gm12878=1+rowSums(as.matrix(mcols(cgr)[,1:2])),
+         K562=1+rowSums(as.matrix(mcols(cgr)[,3:4])) )
```

For count data from sequencing, the *emis.type* could be set to either `'pois'` or `'nbinom'`, though `'pois'` is preferred for sharp boundary detection. For the sojourn settings, instead of using the uninformative flat prior, we here use estimates from other data source as a prior. We load the *TxDb.Hsapiens.UCSC.hg19.knownGene* known gene database, and pass the *TranscriptDb* object to *xAnno*. Then internally sojourn parameters and state number J will be estimated from *xAnno* by calling function `sojournAnno`. When given a *TranscriptDb* object to *xAnno*, state number would be set to 3 and each represents 'intergenic', 'intron' and 'exon'. One can also supply a named *list* object with initial values for parameters of distribution specified by *soj.type*. For emission, since the highly dispersed nature of count data, we set the prior for emission mean to be more extreme, with *q.alpha=0.01*.

```
> library(TxDb.Hsapiens.UCSC.hg19.knownGene)
> txdb<-TxDb.Hsapiens.UCSC.hg19.knownGene
> rhsmm<-biomvRhsmm(x=cgr, xAnno=txdb, maxbp=1E3, soj.type='gamma',
+         emis.type='pois', prior.m='quantile', q.alpha=0.01)
```

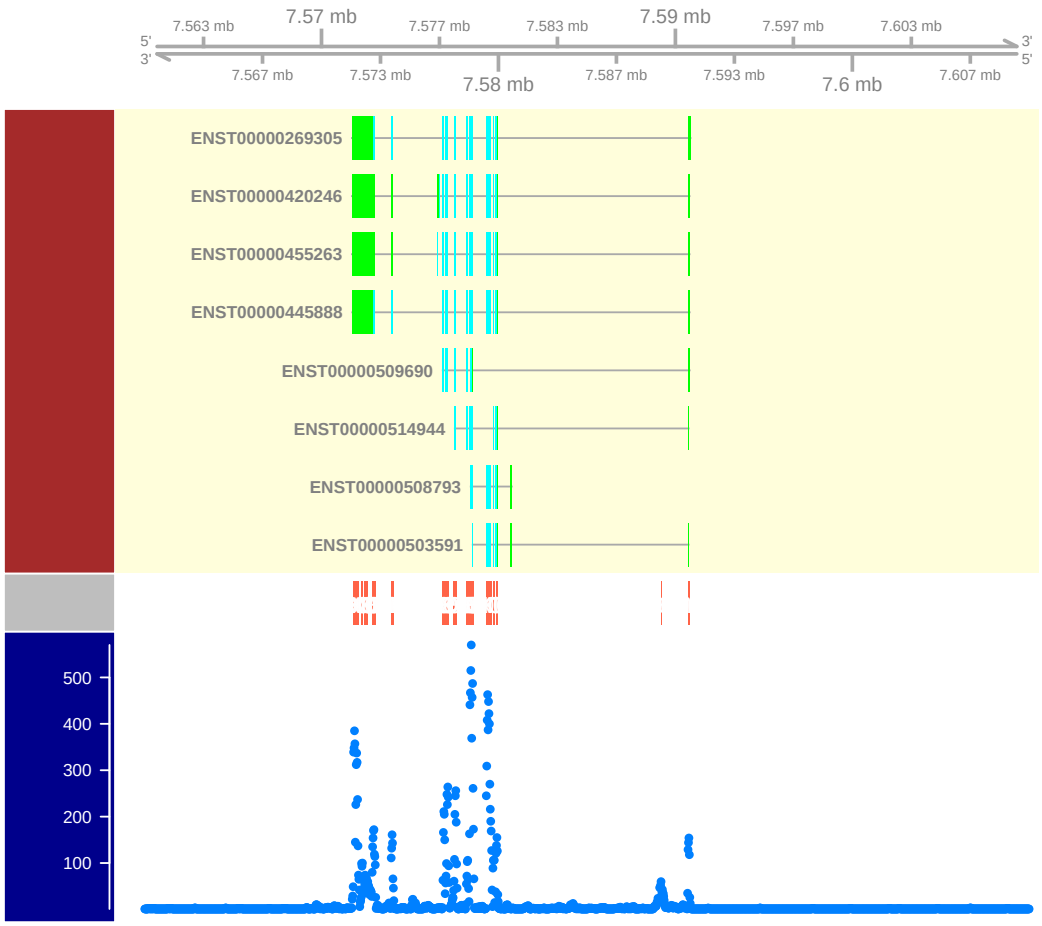
As in the ENCODE guide [Consortium, 2011], the study identified the p53 isoform observed in K562 cells has a longer 3'UTR than the isoform seen in the GM12878 cell line. So here we plot our model estimates and consider the third state, namely 'exon', to represent detected transcripts. And the HSMM model clearly picked up the extra transcripts of the K562 cell line at the 3'UTR.

```
> rhsmm@res[mcols(rhsmm@res)[,'STATE']=='exon']
```

```
GRanges with 52 ranges and 3 metadata columns:
       seqnames             ranges strand   |  SAMPLE STATE   AVG
          <Rle>          <IRanges>  <Rle>   |   <Rle> <Rle> <Rle>
   [1]    chr17 [7571801, 7572125]      -   | Gm12878  exon   312
   [2]    chr17 [7572251, 7572350]      -   | Gm12878  exon    96
   [3]    chr17 [7572426, 7572550]      -   | Gm12878  exon    61
   [4]    chr17 [7572601, 7572625]      -   | Gm12878  exon    60
   [5]    chr17 [7572851, 7573050]      -   | Gm12878  exon   127
   ...      ...                ...    ... ...     ...   ...   ...
  [48]    chr17 [7588926, 7589525]      -   |    K562  exon  15.0
  [49]    chr17 [7589676, 7589825]      -   |    K562  exon   9.0
  [50]    chr17 [7589851, 7589875]      -   |    K562  exon   5.0
  [51]    chr17 [7590701, 7590800]      -   |    K562  exon  14.5
  [52]    chr17 [7591976, 7592050]      -   |    K562  exon   6.0
  ---
  seqlengths:
   chr17
      NA
```

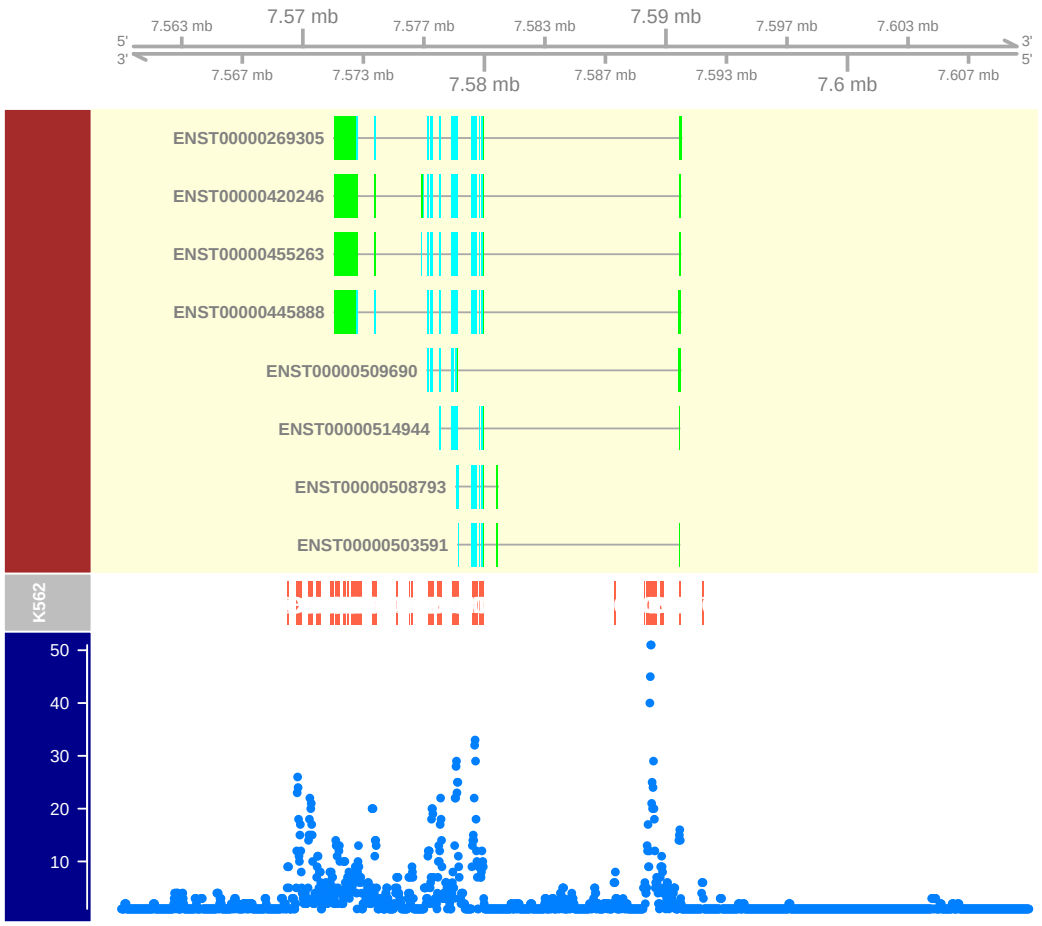
```
> g<-mcols(rhsmm@res)[,'STATE']=='exon' & mcols(rhsmm@res)[,'SAMPLE']=='Gm12878'
> obj<-biomvRGviz(exprgr=cgr[,'Gm12878'], gmgr=gmgr,
+         seggr=rhsmm@res[g], plotstrand='-', regionID='TP53',  tofile=FALSE)
```

›53@chr17.7560000−7610000@Gm128




```
> k<-mcols(rhsmm@res)[,'STATE']=='exon' & mcols(rhsmm@res)[,'SAMPLE']=='K562'
> obj<-biomvRGviz(exprgr=cgr[,'K562'], gmgr=gmgr,
+   seggr=rhsmm@res[k], plotstrand='-', regionID='TP53', tofile=FALSE)
```

TP53@chr17.7560000−7610000@K562



Now we can locate those novel detected fragments in K562 cell line comparing to the annotation and those detected in Gm12878 cell line. One can then follow up those findings either by gene structure prediction using local nucleotides composition or by experimental validation.

```
> nK2gm<-findOverlaps(rhsmm@res[k], gmgr)@queryHits
> nK2G<-findOverlaps(rhsmm@res[k], rhsmm@res[g])@queryHits
> rhsmm@res[k][setdiff(seq_len(sum(k)), unique(c(nK2G, nK2gm)))]
```

```
GRanges with 19 ranges and 3 metadata columns:
       seqnames              ranges strand   | SAMPLE STATE   AVG
          <Rle>           <IRanges>  <Rle>   |  <Rle> <Rle> <Rle>
   [1]    chr17 [7569151, 7569250]      -   |   K562  exon     9
   [2]    chr17 [7569626, 7569950]      -   |   K562  exon    12
   [3]    chr17 [7570276, 7570550]      -   |   K562  exon    15
   [4]    chr17 [7570751, 7570850]      -   |   K562  exon    10
   [5]    chr17 [7570876, 7571000]      -   |   K562  exon     8
   ...      ...                ...    ... ...    ...   ...   ...
  [15]    chr17 [7587126, 7587225]      -   |   K562  exon     6
  [16]    chr17 [7588826, 7588850]      -   |   K562  exon     6
  [17]    chr17 [7589676, 7589825]      -   |   K562  exon     9
  [18]    chr17 [7589851, 7589875]      -   |   K562  exon     5
```

```
  [19]    chr17 [7591976, 7592050]      -   |   K562  exon     6
  ---
  seqlengths:
   chr17
      NA
```

The other 2 batch functions could also be similarly applied here.

```
> rseg<-biomvRseg(x=cgr, maxbp=1E3, maxseg=20, family='pois')
> rmgmr<-biomvRmgmr(x=cgr, q=0.99, maxgap=50, minrun=100)
```

4 Example of differentially methylated region (DMR) detection

As an example, we include a toy dataset extracted from *BiSeq* [Hebestreit et al., 2013], which is a small subset of a published study [Schoofs et al., 2013], comprising intermediate differential methylation results prior to DMR detection. We first load the `variosm` data, the data contains a *GRanges* object `variosm` with two meta columns: '`meth.diff`', methylation difference between the two sample groups; '`p.val`', significance level from the Wald test. What we will show here latter could be applied on other pipelines as well, using similar data input.

```
> data(variosm)
> head(variosm, n=3)
```

```
GRanges with 3 ranges and 2 metadata columns:
      seqnames           ranges strand |          meth.diff                p.val
         <Rle>        <IRanges>  <Rle> |          <numeric>            <numeric>
  [1]     chr1 [872335, 872335]      * | 0.0889114878437597  0.00113176519852419
  [2]     chr1 [872369, 872369]      * | 0.0970307417273086 0.000767802697850362
  [3]     chr1 [872370, 872370]      * |  0.096558901255241 0.000834745131903375
  ---
  seqlengths:
   chr1 chr2
     NA   NA
```

In the *BiSeq* work-flow, they use an approach similar to the max-gap-min-run algorithm to define the DMR boundaries, by prior filtering and comparing the differential test statistics with a user specified significance level in the candidate regions. The positional information of methylation sites is taking into account by locating and testing highly correlated cluster regions in the filtering process.

We now use the `biomvRhsmm` model to detect DMR, since there are mainly two types of measurement associated with differential methylation studies like we have here, one is the difference in the methylation ratio and the other one is the significance level from differential test. The methylation difference gives information about the directionality of the change as well as the size, and the significance level gives the confidence in claiming differential events.

So here we utilize both information for the DMR detection. We implicitly ask the model to give 3 states, since J is default to 3, in which case the three states may each represent hypomethylated regions, undefined null regions, and hypermethylated regions respectively when modelling `meth.diff`; While modelling significance level these states would represent highly confident regions, lowly confident regions or / and null results. For both scenarios, we are more interested in extreme states where we have consistent differences and low P-values. However the distribution of `p.val` and `meth.diff` are both highly asymmetric, we thus enable the cluster mode for emission prior initialization by setting *prior.m*='cluster' . And due to the non-uniformly located CpG sites, one may split inter-spreading long segments with parameter *maxgap*=100.

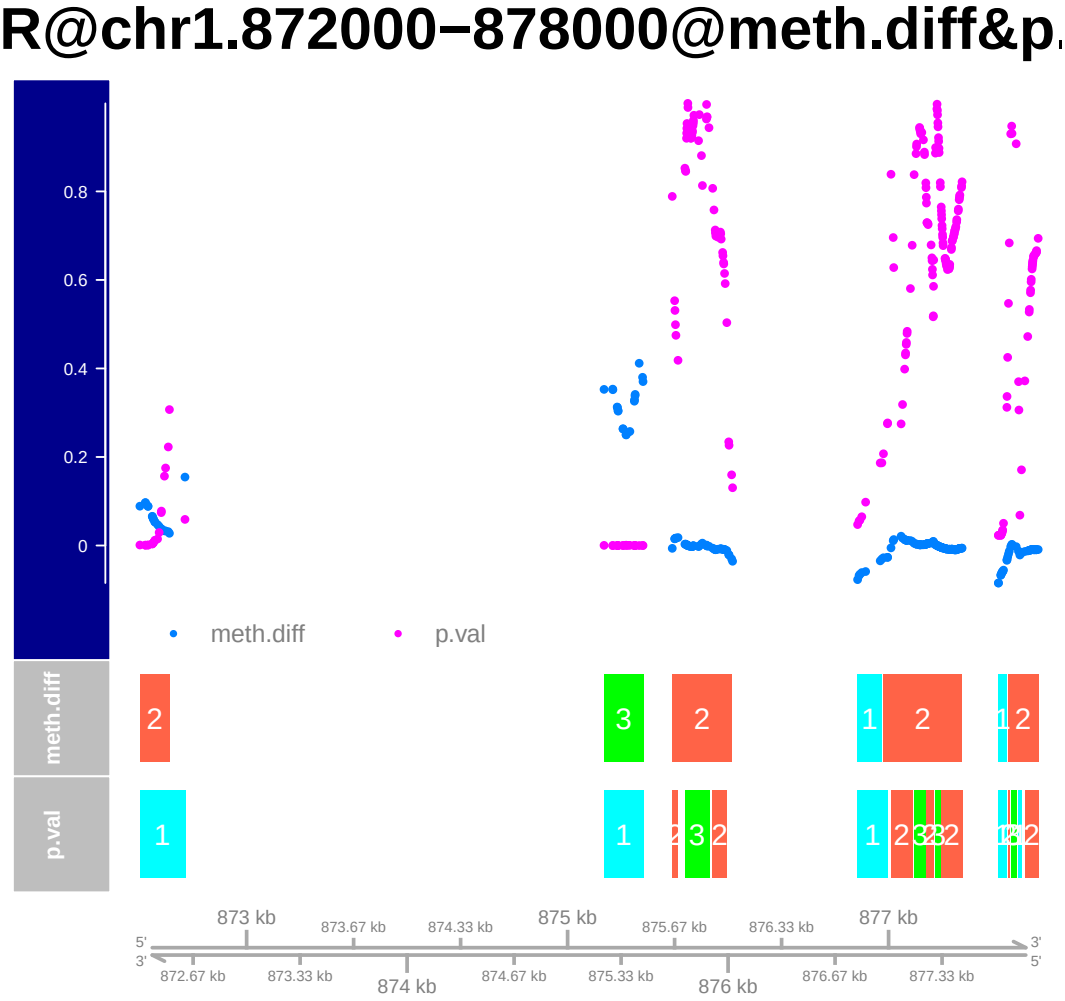
```
> rhsmm<-biomvRhsmm(x=variosm, maxbp=100, prior.m='cluster', maxgap=100)


> hiDiffgr<-rhsmm@res[mcols(rhsmm@res)[,'STATE']!=2
+         & mcols(rhsmm@res)[,'SAMPLE']=='meth.diff']
> dirNo<-mcols(hiDiffgr)[,'STATE']=='1' & mcols(hiDiffgr)[,'AVG']>0 |
+         mcols(hiDiffgr)[,'STATE']=='3' & mcols(hiDiffgr)[,'AVG']<0
> hiDiffgr<- hiDiffgr[!dirNo]
> loPgr<-rhsmm@res[mcols(rhsmm@res)[,'STATE']==1
+         & mcols(rhsmm@res)[,'SAMPLE']=='p.val']
> DMRs<-intersect(hiDiffgr, loPgr)
> idx<-findOverlaps(variosm, DMRs, type='within')
> mcols(DMRs)<-DataFrame(aggregate(as.data.frame(mcols(variosm[idx@queryHits])),
+         by=list(DMR=idx@subjectHits), FUN=median))
> DMRs


GRanges with 5 ranges and 3 metadata columns:
      seqnames           ranges strand |       DMR           meth.diff               p.val
         <Rle>        <IRanges>  <Rle> | <integer>           <numeric>             <numeric>
  [1]     chr1 [875227, 875470]      * |         1   0.319474180010203 6.67719258580069e-06
  [2]     chr1 [876807, 876958]      * |         2 -0.0610821869609038   0.0650032810337598
  [3]     chr1 [877684, 877738]      * |         3 -0.0612300842467785    0.028446386208542
  [4]     chr2 [ 46126,  46280]      * |         4   0.410085244128014 1.81853017591375e-07
  [5]     chr2 [ 46389,  46558]      * |         5   0.448231719164192 1.89081904128487e-06
  ---
  seqlengths:
   chr1 chr2
     NA   NA
```

After the model fitting, by intersecting regions with extreme `meth.diff` and regions with low `p.val`, we can locate those detected DMRs, returned with their average `meth.diff` and `p.val`. Comparing to the regions detected in the *BiSeq* vignette, the two sets of regions are largely similar except for one region (chr1:872335,872386), which in our case the `meth.diff` has not been considered high enough. However it is worth mentioning that due to the filtering applied in their work-flow, they built wider regions out of a smaller set of more significant sites; while in our case, the regions are more refined and especially we identified two hypomethylated regions.

```
> s<-seqnames(rhsmm@res) == 'chr1'
> e<-seqnames(variosm) == 'chr1'
> obj<-biomvRGviz(exprgr=variosm[e],  seggr=rhsmm@res[s], regionID='DMR', tofile=FALSE)
```



5 More

To be continued ...

6 Session information

```
> sessionInfo()
```

```
R Under development (unstable) (2013-04-12 r62558)
Platform: x86_64-unknown-linux-gnu (64-bit)

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C                   LC_TIME=en_US.UTF-8
 [4] LC_COLLATE=en_US.UTF-8     LC_MONETARY=en_US.UTF-8        LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=C                 LC_NAME=C                      LC_ADDRESS=C
[10] LC_TELEPHONE=C             LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
```

```
attached base packages:
[1] grid      parallel  stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] cluster_1.14.4                          TxDb.Hsapiens.UCSC.hg19.knownGene_2.9.2
 [3] GenomicFeatures_1.13.36                 AnnotationDbi_1.23.18
 [5] Biobase_2.21.6                          biomvRCNS_1.1.7
 [7] Gviz_1.5.10                             GenomicRanges_1.13.37
 [9] XVector_0.1.0                           IRanges_1.19.26
[11] BiocGenerics_0.7.4

loaded via a namespace (and not attached):
 [1] biomaRt_2.17.2     Biostrings_2.29.15 biovizBase_1.9.2   bitops_1.0-6
 [5] BSgenome_1.29.1    colorspace_1.2-2   DBI_0.2-7          dichromat_2.0-0
 [9] Hmisc_3.12-2       labeling_0.2       lattice_0.20-23    munsell_0.4.2
[13] mvtnorm_0.9-9995   plyr_1.8           RColorBrewer_1.0-5 RCurl_1.95-4.1
[17] rpart_4.1-2        Rsamtools_1.13.30  RSQLite_0.11.4     rtracklayer_1.21.9
[21] scales_0.2.3       stats4_3.1.0       stringr_0.6.2      tools_3.1.0
[25] XML_3.98-1.1       zlibbioc_1.7.0
```

References

The ENCODE Project Consortium. The ENCODE (ENCyclopedia of DNA elements) project. *Science*, 306(5696): 636–640, 2004.

The ENCODE Project Consortium. A User's Guide to the Encyclopedia of DNA Elements (ENCODE). *PLoS Biol*, 9(4):e1001046, 04 2011. doi: 10.1371/journal.pbio.1001046. URL http://dx.doi.org/10.1371%2Fjournal.pbio.1001046.

Yann Guédon. Estimating Hidden Semi-Markov Chains from Discrete Sequences. *Journal of Computational and Graphical Statistics*, 12(3):604–639, 2003. ISSN 10618600. doi: 10.2307/1391041. URL http://dx.doi.org/10.2307/1391041.

Katja Hebestreit, Martin Dugas, and Hans-Ulrich Klein. Detection of Significantly Differentially Methylated Regions in Targeted Bisulfite Sequencing Data. *Bioinformatics*, 2013. doi: 10.1093/bioinformatics/btt263. URL http://bioinformatics.oxfordjournals.org/content/early/2013/05/08/bioinformatics.btt263.abstract.

Wolfgang Huber, Joern Toedling, and Lars M. Steinmetz. Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics*, 22:1963–1970, 2006.

Adam B. Olshen, E. S. Venkatraman, Robert Lucito, and Michael Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572, 2004. doi: 10.1093/biostatistics/kxh008. URL http://biostatistics.oxfordjournals.org/content/5/4/557.abstract.

Till Schoofs, Christian Rohde, Katja Hebestreit, Hans-Ulrich Klein, Stefanie Göllner, Isabell Schulze, Mads Lerdrup, Nikolaj Dietrich, Shuchi Agrawal-Singh, Anika Witten, Monika Stoll, Eva Lengfelder, Wolf-Karsten Hofmann, Peter Schlenke, Thomas Büchner, Klaus Hansen, Wolfgang E. Berdel, Frank Rosenbauer, Martin Dugas, and Carsten Müller-Tidow. DNA methylation changes are a late event in acute promyelocytic leukemia and coincide with loss of transcription factor binding. *Blood*, 121(1):178–187, 2013. doi: 10.1182/blood-2012-08-448860. URL http://bloodjournal.hematologylibrary.org/content/121/1/178.abstract.

Antoine M Snijders, Norma Nowak, Richard Segraves, Stephanie Blackwood, Nils Brown, Jeffrey Conroy, Greg Hamilton, Anna Katherine Hindle, Bing Huey, Karen Kimura, et al. Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature genetics*, 29:263–264, 2001.

Nancy R. Zhang and David O. Siegmund. A Modified Bayes Information Criterion with Applications to the Analysis of Comparative Genomic Hybridization Data. *Biometrics*, 63(1):22–32, 2007. ISSN 1541-0420. doi: 10.1111/j.1541-0420.2006.00662.x. URL http://dx.doi.org/10.1111/j.1541-0420.2006.00662.x.