CRAN GMD: Data Processing (0.3.3)

Measure Similarity between Histone Modifications

Xiaobei Zhao*

Modified: 2014-08-26 Compiled: 2014-8-27

You may find the latest version of *GMD* and this documentation at,
http://CRAN.R-project.org/package=GMD

Keywords: histone modifications, histogram, distance, heatmap, alignment, GMD

Contents

1 Introduction and scope

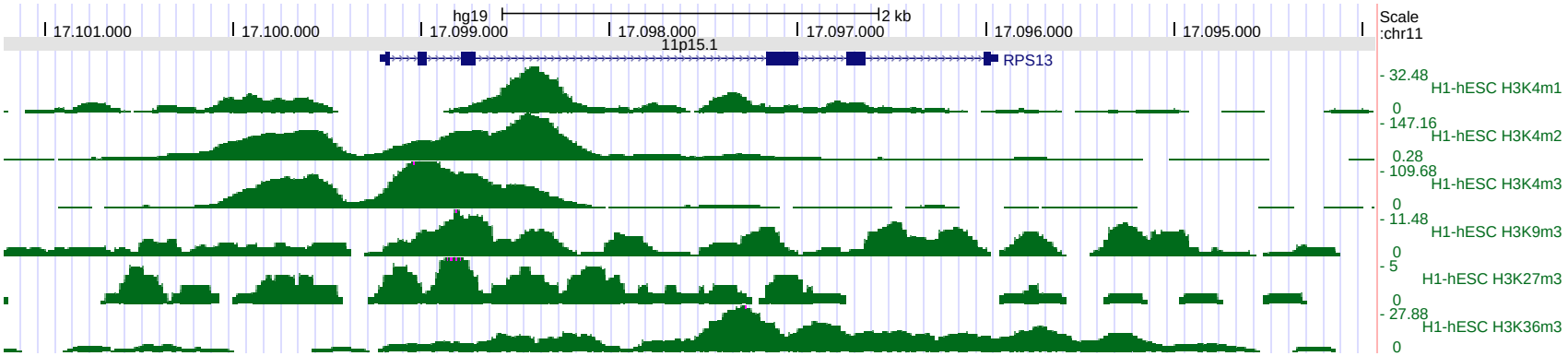
This is a tutorial to explain how biological data is processed to be used in *GMD*. For detailed information of the *GMD* functionality, please refer to the Reference Manual and the Vignette[1][1]. One case study is presented to measure the similarity of histone modifications using `BigWig` files from "Histone Modifications by ChIP-seq from ENCODE/Broad Institute". The idea and process is similar for other data formats (`bed`, `bam`, etc.).

*Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, 450 West Dr, Chapel Hill, NC 27599, USA. xiaobei@binf.ku.dk

[1] cran.r-project.org/package=GMD

2 Case study: measure similarity between histone modifications

We are going to investigate the histone modifications around the gene *RPS13 (chr11:17095939-17099220), reverse strand* generated by ChIP-seq from ENCODE/Broad Institute as shown below:



Let's start with a few `BigWig` files of histone modifications downloaded via the "UCSC Genome Browser"[2] (assembly: hg19), for instance, there is a `BigWig` file for H3K4me1 wgEncodeBroadHistoneH1hescH3k4me1StdSig.bigWig.

[2] http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHistone/

2.1 Convert to `BedGraph`

We first download the `BigWig` files and convert them into `BedGraph` using the `bigWigToBedGraph`[3] program at a UNIX-like terminal,

```
bigWigToBedGraph wgEncodeBroadHistoneH1hescH3k4me1StdSig.bigWig H3K4me1.full.BedGraph
```

For a minimal demonstration, we will extract only a subregion of the data - the regions around the gene *RPS13 (chr11:17095939-17099220)* with `+/-2000` base pairs flanking the gene body. Note: the start position is zero-based.

```
bigWigToBedGraph wgEncodeBroadHistoneH1hescH3k4me1StdSig.bigWig H3K4me1.BedGraph \
-chrom=chr11 -start=17093938 -end=17101220

head H3K4me1.BedGraph

chr11   17093950        17093975        0.92
chr11   17093975        17094000        1.84
chr11   17094000        17094025        2
chr11   17094025        17094050        2
chr11   17094050        17094075        2
chr11   17094075        17094100        2
chr11   17094100        17094125        2
chr11   17094125        17094150        2
chr11   17094150        17094175        1.08
chr11   17094175        17094200        0.16
```

This `BedGraph` has four columns: `chr`, `start`, `end` and `datavalue`[4].

2.2 Convert to a vector of depth-like signals

The above `H3K4me1.BedGraph` file is packed with the *GMD* package in the `extdata/hg19` subdirectory under the top level directory. Then we can read the file in *R* and make downstream analysis.

[3] http://hgdownload.cse.ucsc.edu/admin/exe/
[4] https://genome.ucsc.edu/goldenPath/help/bedgraph.html

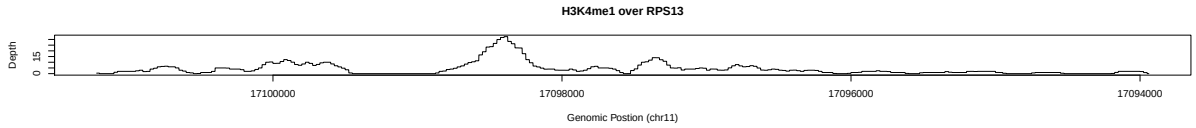
```
> require(GMD)
> ## The file path of ecternal data
> id <- "H3K4me1"
> inFpath <- system.file('extdata/hg19',paste0(id,'.BedGraph.gz'),package="GMD",mustWork=TRUE)
> print(inFpath)

[1] "/tmp/RtmpWbWXAq/Rinst82f73dcdbd5c/GMD/extdata/hg19/H3K4me1.BedGraph.gz"

> ## Convert to a vector of depth-like signals
> res <- bedgraph.to.depth(inFpath,chr="chr11",start=17093938,end=17101220,reverse=TRUE)
> str(res)

 Named num [1:7282] 0.44 0.44 0.44 0.44 0.44 0.44 0.44 0.44 0.44 0.44 ...
 - attr(*, "names")= chr [1:7282] "17101220" "17101219" "17101218" "17101217" ...

> ## Visualize the pattern of the signals
> plot(as.numeric(names(res)),res,type="l",xlim=rev(range(as.numeric(names(res)))),
+      main="H3K4me1 over RPS13", ylab="Depth", xlab="Genomic Postion (chr11)"
+      )
```



Similarly, we have `BedGraph` files of other histone modifications. We will read them in *R*.

```
> data_test <- list()
> ids <- c("H3K4me1","H3K4me2","H3K4me3","H3K9me3","H3K27me3","H3K36me3")
> for ( id in ids) {
+   inFpath <- system.file('extdata/hg19',paste0(id,'.BedGraph.gz'),package="GMD",mustWork=TRUE)
+   res <- bedgraph.to.depth(inFpath,chr="chr11",start=17093938,end=17101220,reverse=TRUE)
+   data_test[[id]] <- res
+ }
> str(data_test)

List of 6
 $ H3K4me1 : Named num [1:7282] 0.44 0.44 0.44 0.44 0.44 0.44 0.44 0.44 0.44 0.44 ...
  ..- attr(*, "names")= chr [1:7282] "17101220" "17101219" "17101218" "17101217" ...
 $ H3K4me2 : Named num [1:7282] 2 2 2 2 2 2 2 2 2 2 ...
  ..- attr(*, "names")= chr [1:7282] "17101220" "17101219" "17101218" "17101217" ...
 $ H3K4me3 : Named num [1:7282] 0 0 0 0 0 0 0 0 0 0 ...
  ..- attr(*, "names")= chr [1:7282] "17101220" "17101219" "17101218" "17101217" ...
 $ H3K9me3 : Named num [1:7282] 1.32 1.32 1.32 1.32 1.32 1.32 1.32 1.32 1.32 1.32 ...
  ..- attr(*, "names")= chr [1:7282] "17101220" "17101219" "17101218" "17101217" ...
 $ H3K27me3: Named num [1:7282] 0.72 0.72 0.72 0.72 0.72 0.72 0.72 0.72 0.72 0.72 ...
  ..- attr(*, "names")= chr [1:7282] "17101220" "17101219" "17101218" "17101217" ...
 $ H3K36me3: Named num [1:7282] 1 1 1 1 1 1 1 1 1 1 ...
  ..- attr(*, "names")= chr [1:7282] "17101220" "17101219" "17101218" "17101217" ...
```

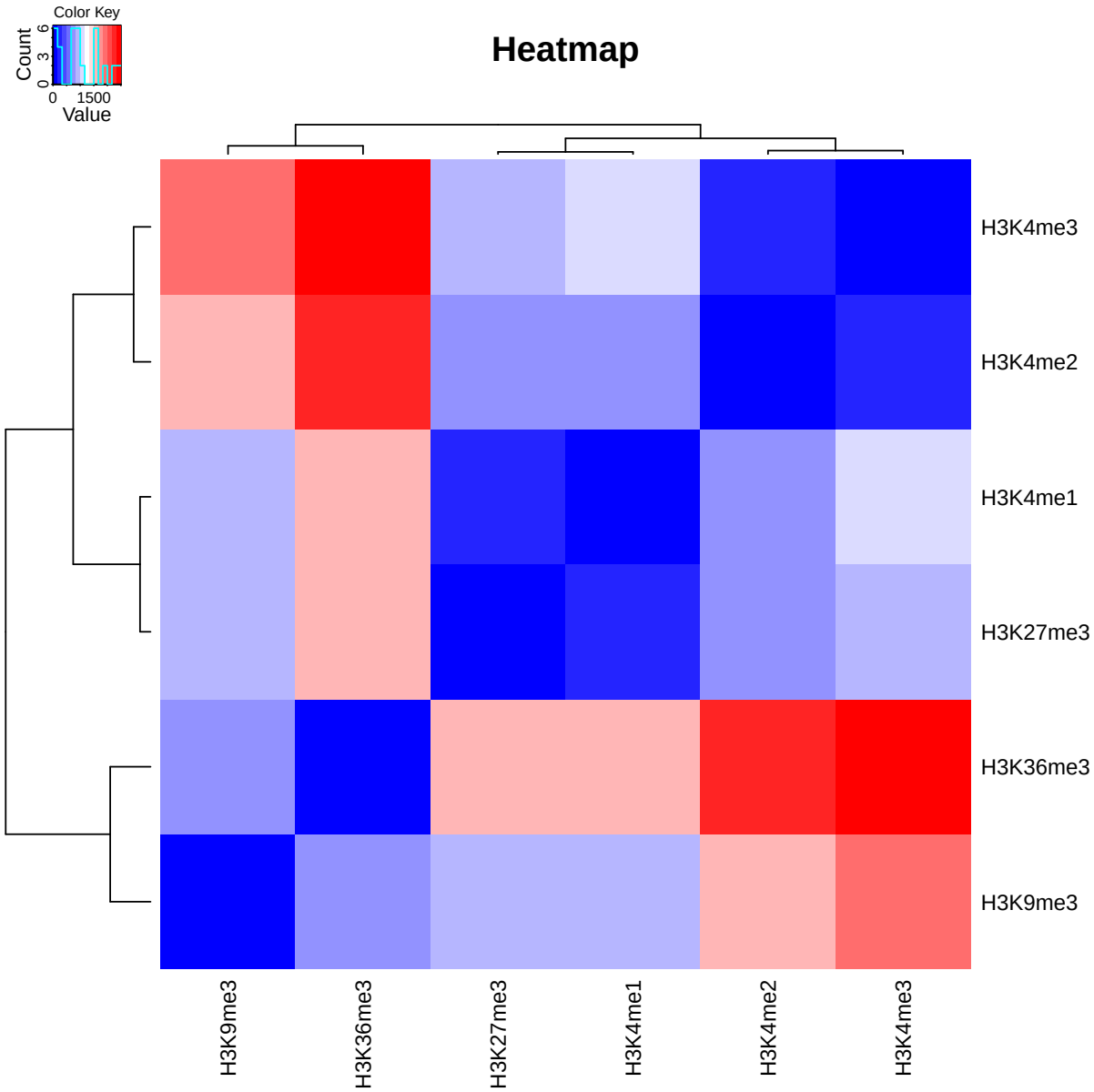
2.3 Distance measure

```
> x <- gmdm(data_test,sliding=FALSE)
> print(x)


           H3K4me1   H3K4me2   H3K4me3   H3K9me3  H3K27me3  H3K36me3
H3K4me1     0.0000  645.2440  885.6236  726.9927  226.9335 1268.7752
H3K4me2   645.2440    0.0000  253.4012 1332.1337  560.4015 1812.8188
H3K4me3   885.6236  253.4012    0.0000 1555.7017  793.3465 2009.9961
H3K9me3   726.9927 1332.1337 1555.7017    0.0000  794.9317  650.3151
H3K27me3  226.9335  560.4015  793.3465  794.9317    0.0000 1336.9398
H3K36me3 1268.7752 1812.8188 2009.9961  650.3151 1336.9398    0.0000
```

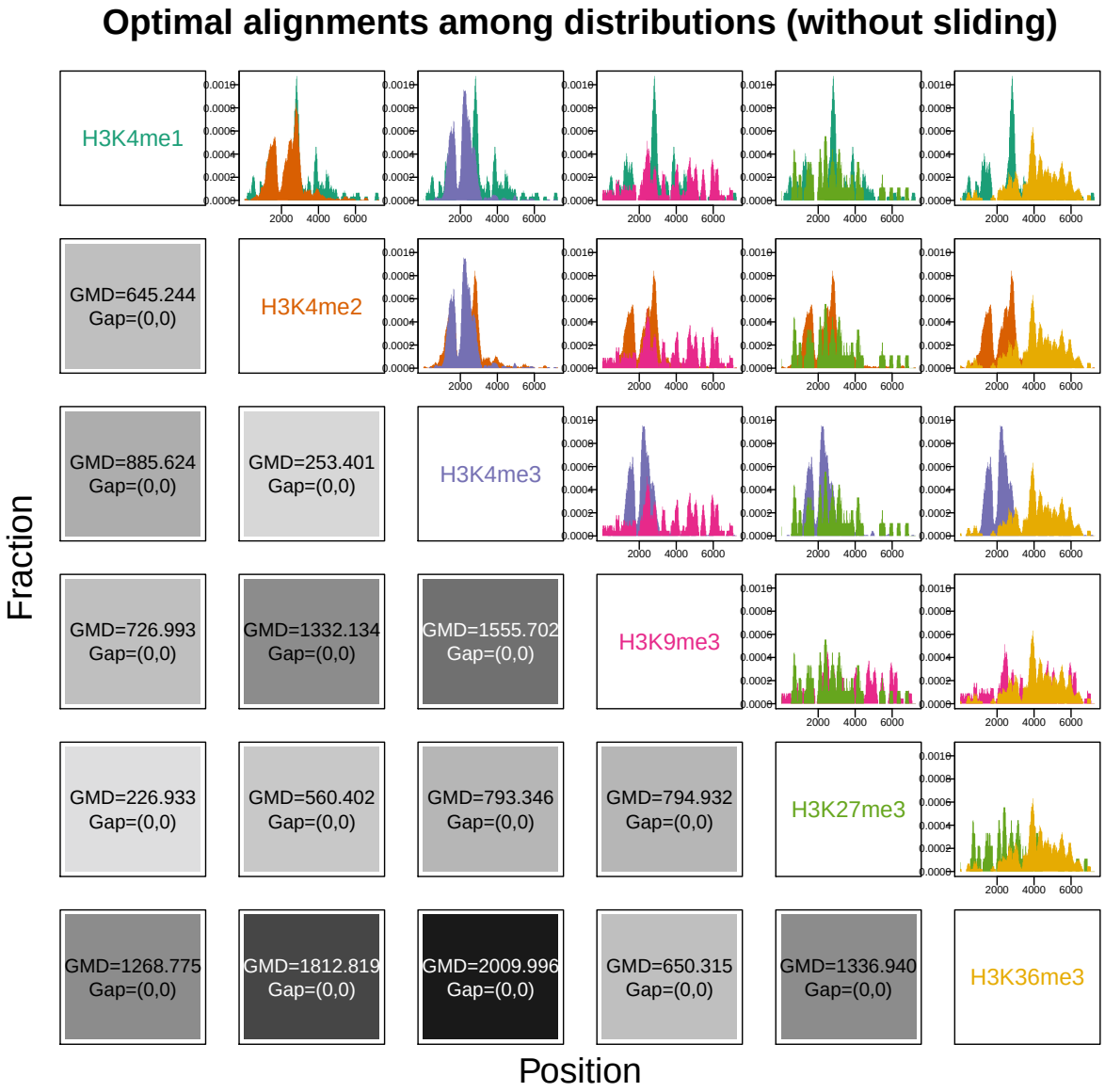
2.4 Heatmap of the distance matrix

```
> heatmap.3(x,dendrogram="both",cexCol=0.85,cexRow=0.85)
```



2.5 Alignment of the distributions (without sliding)

```
> plot(x,cex.text=2,type="polygon",if.plot.new=FALSE)
> #- setting if.plot.new to TRUE to pop-up a auto-adjust window
```



References

[1] Xiaobei Zhao and Albin Sandelin. *GMD: Generalized Minimum Distance of distributions*, 2014. R package.