Joint Analysis of Multiple ChIP-seq Datasets with the 'jmosaics' Package

Xin Zeng[1] and Sündüz Keleş[1,2]

[1]Department of Statistics, University of Wisconsin Madison, WI

[2]Department of Statistics and of Biostatistics and Medical Informatics,

University of Wisconsin, Madison, WI

August 2, 2012

1 Overview

This document provides an introduction to the joint analysis of multiple ChIP-seq datasets with the 'jmosaics' package. R package 'jmosaics' implements jMOSAiCS: Joint Analysis of Multiple ChIP-seq Datasets, proposed in [4]. It detects combinatorial enrichment patterns across multiple ChIP-seq datasets and is applicable with ChIP-seq data of both transcription factor binding and histone modifications. In this document, we use data from a ChIP- seq experiment of H3K27me3 and H3K4me1 in G1E cells [3] as described in [4]. For illustration purposes, we only utilize reads mapping to chromosome 10. The package can be loaded with the command:

```
R> library("jmosaics")
```

2 Workflow

'jmosaics' utilizes R package 'mosaics'[2] for modeling read counts of each individual dataset. We therefore start with an overview of the relevant 'mosaics' functions for reading in data and model fitting. 'mosaics' is available through both Bioconductor (`http://www.bioconductor.org/packages/release/bioc/html/mosaics.html`) and Galaxy Tool Shed (`http://toolshed.g2.bx.psu.edu/`) under "Sequence Analysis".

2.1 readBins

This function from package 'mosaics' is used to read bin-level data into the R Environment. Bin-level data is easily obtained from the aligned read files by the constructBin function of the 'mosaics' package. constructBin supports multiple alignment formats from the Eland and Bowtie aligners. 'jmosaics' currently allows two-sample analysis with ChIP and control (input) data and also two-sample analysis with mappability and GC features in addition to ChIP and input data. The following reads in ChIP and input bin-level data for each individual experiment:

```
> bin1 <- readBins(type = c("chip","input"),
 +fileName = c(system.file(file.path("extdata","h3k27me3_chip_chr10.txt"),
 +package="jmosaics"),
 +system.file(file.path("extdata","h3k27me3_input_chr10.txt"),
 +package="jmosaics")))
> bin2 <- readBins(type = c("chip","input"),
 +fileName = c(system.file(file.path("extdata","h3k4me1_chip_chr10.txt"),
 +package="jmosaics"),
 +system.file(file.path("extdata","h3k4me1_input_chr10.txt"),
 +package="jmosaics")))
```

For the two-sample analysis that adjusts for mappability and GC biases, pre-processed bin-level ChIP data, control sample data, mappability score, GC content score, and sequence ambiguity score can be read in as:

```
> bin1 <- readBins(type = c("chip", "M", "GC", "N","input"),
+ fileName = c("h3k27me3_chip_chr10.txt",
+ "./M_chr10.txt", "./GC_chr10.txt", "./N_chr10.txt",
+ "h3k27me3_input_chr10.txt"))
> bin2 <- readBins(type = c("chip", "M", "GC", "N","input"),
+ fileName = c("h3k4me1_chip_chr10.txt",
+ "./M_chr1.txt", "./GC_chr10.txt", "./N_chr10.txt",
+ "h3k4me1_input_chr10.txt"))
```

Details regarding these pre-processed bin-level data are available in the 'mosaics' package.

2.2 readBinsMultiple

This function matches the bin coordinates of multiple ChIP-seq datasets. A list of bin-level data ('origin_bin' below) is used as input.

```
> origin_bin <- list(bin1,bin2)
> bin <- readBinsMultiple(origin_bin)
> str(bin)
 $ :Formal class 'BinData' [package "mosaics"] with 7 slots
  .. ..@ chrID      : chr [1:649776] "chr10" "chr10" "chr10" "chr10" ...
  .. ..@ coord      : num [1:649776] 0 200 400 600 800 1000 1200 1400 1600 1800 ...
  .. ..@ tagCount   : num [1:649776] 0 0 0 0 0 0 0 0 0 0 ...
  .. ..@ mappability: num(0)
  .. ..@ gcContent  : num(0)
  .. ..@ input      : num [1:649776] 0 0 0 0 0 0 0 0 0 0 ...
```

```
  .. ..@ dataType   : chr "unique"
 $ :Formal class 'BinData' [package "mosaics"] with 7 slots
  .. ..@ chrID      : chr [1:649776] "chr10" "chr10" "chr10" "chr10" ...
  .. ..@ coord      : num [1:649776] 0 200 400 600 800 1000 1200 1400 1600 1800 ...
  .. ..@ tagCount   : num [1:649776] 0 0 0 0 0 0 0 0 0 0 ...
  .. ..@ mappability: num(0)
  .. ..@ gcContent  : num(0)
  .. ..@ input      : num [1:649776] 0 0 0 0 0 0 0 0 0 0 ...
  .. ..@ dataType   : chr "unique"
```

The output ('bin') is a list of all the bin-level data with matching bin coordinates.

2.3 mosaicsFit

We are now ready to fit a MOSAiCS model using the mosaicsFit function from the 'mosaics' package. Each bin-level data in the list (e.g., bin[[1]]) is used as input. A MOSAiCS model is fitted with the command:

```
> fit1 <- mosaicsFit(bin[[1]], analysisType = "IO", bgEst="automatic")
> fit2 <- mosaicsFit(bin[[2]], analysisType = "IO", bgEst="automatic")
```

'`analysisType="IO"`' indicates implementation of the two-sample analysis. '`bgEst`' argument determines background estimation approach. '`bgEst="matchLow"`' estimates background distribution using only bins with low tag/read counts. '`bgEst="rMOM"`' estimates background distribution using robust method of moments (MOM) and can be used for tuning if the goodness of fit plot from using '`bgEst="automatic"`' is not satisfactory.

After fitting each dataset separately, an R list should be generated as follows:

```
> fit <- list(fit1, fit2)
```

This list of 'mosaics' object fits constitutes the main input for detecting enrichment patterns with 'jmosaics'.

2.4 jmosaicsPattern

This is the main function for obtaining enriched regions and combinatorial enrichment patterns. It allows false discovery rate control through the '`FDR`' parameter and filtering with respect to a minimum average ChIP tag count across the bins within a region through the '`thres`' parameter for the B- and E-layer analyses. When B variable is 1, the region is enriched in at least one of the datasets. We first use the posterior probabilities of the region-specific B variables for false discovery rate control and then refine initial set of enriched regions by '`thres`'. Initial regions which satisfy the minimum average ChIP tag count requirement as implied by '`thres`' variable in at least one dataset are reported in the object 'B_peak'. E-layer analysis declares enrichment for each dataset separately based on posterior probabilities of the region- and dataset-specific E variables and the average ChIP tag counts. The combinatorial enrichment pattern is then assigned as the pattern with the maximum joint posterior probability of the E variables. For D datasets, we can observe up to 2^D enrichment patterns for a genomic region. For example, for $D = 2$, $\{(0,0),(0,1),(1,0),(1,1)\}$ denote the set of possible patterns: $(0,0)$: not enriched in either of the samples; $(1,0)$: enriched only in sample 1; $(0,1)$: enriched only in sample 2; $(1,1)$: enriched in both samples.

```
> result <- jmosaicsPattern(fit, region_length=1, FDR=0.01, thres=c(10,10),
+type=c('B', 'E', 'Pattern'), patternInfo='FALSE')
> str(result)
List of 3
 $ E_LAYER:List of 2
  ..$ :'data.frame': 43916 obs. of  8 variables:
  .. ..$ chrID           : Factor w/ 1 level "chr10": 1 1 1 1 1 1 1 1 1 1 ...
```

```
  .. ..$ PeakStart       : num [1:43916] 3008800 3009000 3021200 3030200 3030400 ...
  .. ..$ PeakStop        : num [1:43916] 3008999 3009199 3021399 3030399 3030599 ...
  .. ..$ Postprob        : num [1:43916] 0.0434 0.0495 0.0495 0.0285 0.0461 ...
  .. ..$ ChipCount       : num [1:43916] 14 14 14 16 17 18 17 13 14 21 ...
  .. ..$ InputCount      : num [1:43916] 1 1 1 3 4 3 1 0 0 5 ...
  .. ..$ InputCountScaled: num [1:43916] 2.3 2.3 2.3 6.9 9.2 ...
  .. ..$ Log2Ratio       : num [1:43916] 2.184 2.184 2.184 1.105 0.819 ...
  ..$ :'data.frame': 46562 obs. of  8 variables:
  .. ..$ chrID           : Factor w/ 1 level "chr10": 1 1 1 1 1 1 1 1 1 1 ...
  .. ..$ PeakStart       : num [1:46562] 3049200 3108800 3134600 3197600 3197800 ...
  .. ..$ PeakStop        : num [1:46562] 3049399 3108999 3134799 3197799 3197999 ...
  .. ..$ Postprob        : num [1:46562] 0.04346 0.00603 0.04445 0.00675 0.02183 ...
  .. ..$ ChipCount       : num [1:46562] 13 19 26 51 34 20 23 19 18 18 ...
  .. ..$ InputCount      : num [1:46562] 0 0 2 5 3 0 0 0 0 0 ...
  .. ..$ InputCountScaled: num [1:46562] 0 0 14.1 35.2 21.1 ...
  .. ..$ Log2Ratio       : num [1:46562] 3.807 4.322 0.839 0.521 0.661 ...
 $ B_LAYER:'data.frame': 86398 obs. of  8 variables:
  ..$ chrID         : Factor w/ 1 level "chr10": 1 1 1 1 1 1 1 1 1 1 ...
  ..$ PeaksStart    : num [1:86398] 3008800 3009000 3010400 3021200 3021400 ...
  ..$ PeakStop      : num [1:86398] 3008999 3009199 3010599 3021399 3021599 ...
  ..$ Postprob      : num [1:86398] 0.0207 0.0238 0.0435 0.0238 0.0264 ...
  ..$ ChipCount_E_1 : num [1:86398] 14 14 11 14 10 16 16 17 18 17 ...
  ..$ InputCount_E_1: num [1:86398] 1 1 1 1 0 4 3 4 3 1 ...
  ..$ ChipCount_E_2 : num [1:86398] 12 11 17 11 8 12 16 9 11 10 ...
  ..$ InputCount_E_2: num [1:86398] 1 1 1 1 0 4 3 4 3 1 ...
 $ Pattern:'data.frame': 649776 obs. of  8 variables:
  ..$ chrID             : Factor w/ 1 level "chr10": 1 1 1 1 1 1 1 1 1 1 ...
```

```
  ..$ RegionStart       : num [1:649776] 0 200 400 600 800 1000 1200 1400 1600 1800 ...
  ..$ RegionStop        : num [1:649776] 200 400 600 800 1000 1200 1400 1600 1800 2000 ...
  ..$ Enrichment Pattern: Factor w/ 4 levels "00","01","10",..: 1 1 1 1 1 1 1 1 1 1 ...
  ..$ aveChipCount_E_1  : num [1:649776] 0 0 0 0 0 0 0 0 0 0 ...
  ..$ aveInputCount_E_1 : num [1:649776] 0 0 0 0 0 0 0 0 0 0 ...
  ..$ aveChipCount_E_2  : num [1:649776] 0 0 0 0 0 0 0 0 0 0 ...
  ..$ aveInputCount_E_2 : num [1:649776] 0 0 0 0 0 0 0 0 0 0 ...
```

'jmosaicsPattern' function returns objects based on 'type'. 'B_LAYER' object is a list of regions which are enriched in at least one dataset. Peak information can be accessed by 'chrID', 'PeakStart', 'PeakStop', 'Postprob', 'ChipCount_E_*', 'InputCount_E_*', where * indicates the indices of the datasets. Each list of 'E_LAYER' object reports enriched regions for each individual dataset and includes variables 'chrID', 'PeakStart', 'PeakStop', 'Postprob', 'ChipCount', 'InputCount', ''InputCountScaled', 'Log2Ratio'. 'Pattern' object reports enrichment patterns across all the input regions. The argument of 'patternInfo' lets users decide whether to report the average ChIP and input tagCount of all the regions for each dataset.

For example, results for region chr10: 5018000-5018600, which includes three enriched bins for both H3k4m1 and H3k27m3 datasets, can be accessed through element 'Pattern' of the jmosaics list as follows:

```
> result$Pattern[25091:25093,]
 chrID RegionStart RegionStop Enrichment Pattern aveChipCount_E_1
 chr10     5018000    5018200                 11               40
 chr10     5018200    5018400                 11               45
 chr10     5018400    5018600                 11               30
 aveInputCount_E_1 aveChipCount_E_2 aveInputCount_E_2
 2                        28              2
```

```
 1                        35              1
 1                        27              1
```

These three enriched bins are also listed in the object 'E_LAYER' for each dataset.

```
> result$E_LAYER[[1]][302:304,]
chrID PeakStart PeakStop Postprob     ChipCount InputCount InputCountScaled Log2Ratio
chr10   5018000  5018199 3.114455e-08    40          2         4.601873    2.871643
chr10   5018200  5018399 1.923238e-10    45          1         2.300936    3.800687
chr10   5018400  5018599 1.716286e-06    30          1         2.300936    3.231321

> result$E_LAYER[[2]][137:139,]
chrID PeakStart PeakStop Postprob     ChipCount InputCount InputCountScaled Log2Ratio
chr10   5018000  5018199 0.014876554    28          2         14.09392    0.9420854
chr10   5018200  5018399 0.002710841    35          1          7.04696    2.1614812
chr10   5018400  5018599 0.011020220    27          1          7.04696    1.7989111
```

These bins are also reported in object 'B_LAYER'.

```
> result$B_LAYER[538:540,]
chrID PeaksStart PeakStop Postprob ChipCount_E_1 InputCount_E_1
chr10    5018000  5018199 1.055964e-08   40              2
chr10    5018200  5018399 2.667433e-11   45              1
chr10    5018400  5018599 5.229809e-07   30              1
ChipCount_E_2 InputCount_E_2
28              2
35              1
27              1
```

Plotting functionality for the 'jmosaics' package is supported by the 'dpeak' package [1]. This package can be used to extract reads corresponding to specified regions ('dpeakRead' function) and generate coverage plots as in Figure 1.

References

[1] D. Chung, K. Myers, J. Grass, D. Park, P. Kiley, R. Landick, and S. Keleş. dPeak: High Resolution Identification of Transcription Factor Binding Sites from PET and SET ChIP-Seq Data. Submitted, August 2012.

[2] P. F. Kuan, D. Chung, G. Pan, J. Thomson, R. Stewart, and S. Keleş. A statistical framework for the analysis of ChIP-Seq data. *Journal of the American Statistical Association*, 106:891–903, 2011. Software available on Galaxy `http://toolshed.g2.bx.psu.edu/` and also on Bioconductor `http://bioconductor.org/packages/2.8/bioc/html/mosaics.html`.

[3] W. Wu, Y. Cheng, C. A. Keller, J. Ernst, S. A. Kumar, T. Mishra, C. Morrissey, C. M. Dorman, K-B. Chen, D. Drautz, B. Giardine, Y. Shibata, L. Song, M. Pimkin, G. E. Crawford, T. S. Furey, M. Kellis, W. Miller, J. Taylor, S. C. Schuster, Y. Zhang, F. Chiaromonte, G. A. Blobel, M. J. Weiss, and R. C. Hardison. Dynamics of the epigenetic landscape during erythroid differentiation after GATA1 restoration. *Genome Research*, 21(10):1659–1671, 2011.

[4] X. Zeng, R. Sanalkumar, E. H. Bresnick, H. Li, Q. Chang, and S. Keleş. jMOSAiCS: Joint Analysis of Multiple ChIP-seq Datasets. Submitted, August 2012.

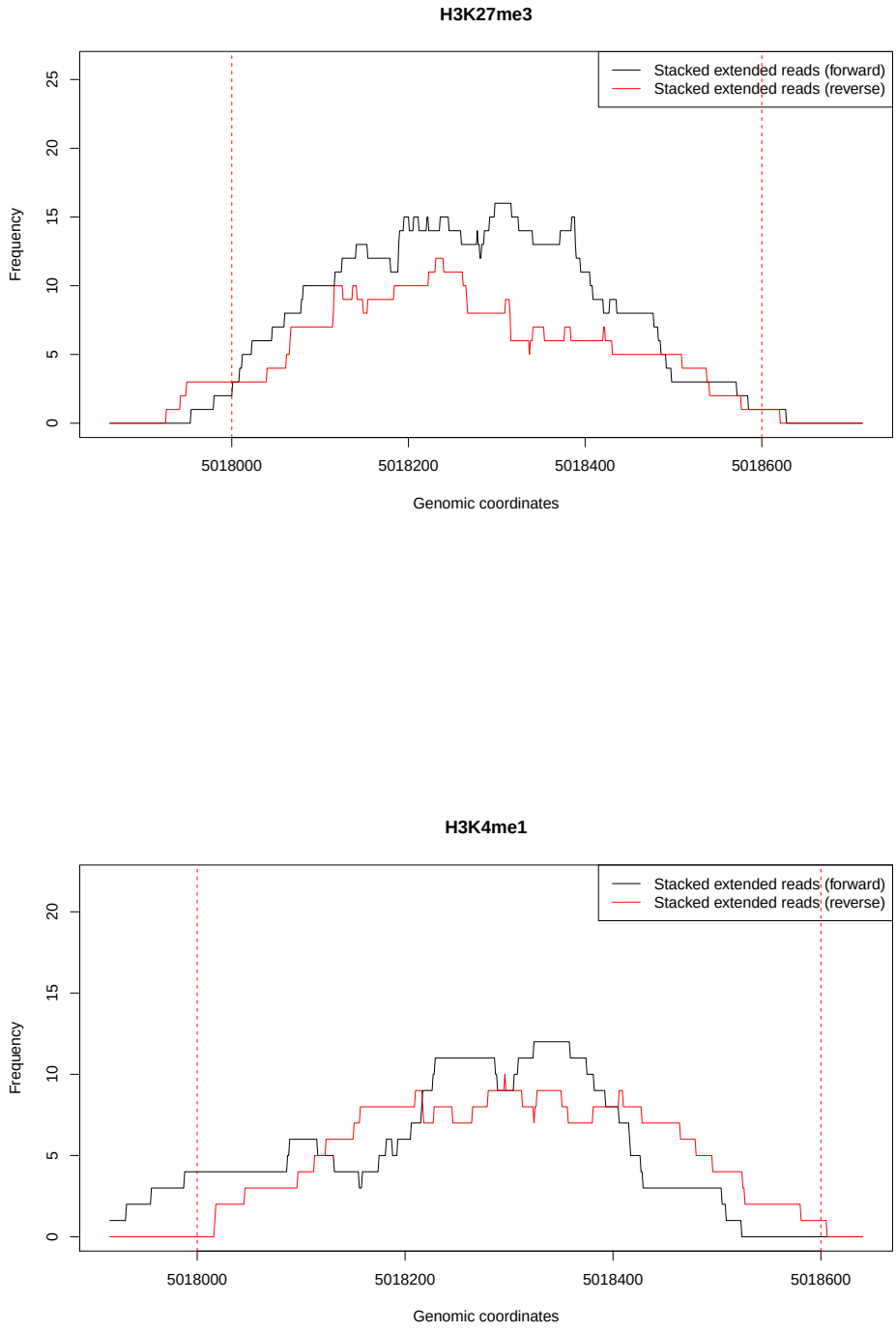

Figure 1: Coverage plot for region chr1: 5018000 – 5018600.