BIOINFORMATICS APPLICATIONS NOTE

Vol. 26 no. 5 2010, pages 678–679
doi:10.1093/bioinformatics/btq023

Genome analysis

Advance Access publication January 19, 2010

rMAT - an R/Bioconductor package for analyzing ChIP-chip experiments

Arnaud Droit[1,*], Charles Cheung[2] and Raphael Gottardo[1,*]

[1]Institut de recherches cliniques de Montreal, 110, avenue des Pins Ouest, Montreal, QC H2W 1R7, Canada and [2]Department of Biostatistics, University of Washington, Seattle, WA 98195–7232, USA

Associate Editor: Martin Bishop

ABSTRACT

Summary: Chromatin immunoprecipitation combined with DNA microarrays (ChIP-chip) has evolved as a popular technique to study DNA–protein binding or post-translational chromatin/histone modifications at the genomic level. However, the raw microarray intensities generate a massive amount of data, creating a need for efficient analysis algorithms and statistical methods to identify enriched regions.

Results: We present a fast, free and powerful, open source R package, `rMAT`, that allows the identification of regions enriched for transcription factor binding sites in ChIP-chip experiments on Affymetrix tiling arrays.

Availability: The R-package `rMAT` is available from the Bioconductor web site at http://bioconductor.org and runs on Linux, MAC OS and MS-Windows. `rMAT` is distributed under the terms of the Artistic Licence 2.0.

Contact: arnaud.droit@ircm.qc.ca; raphael.gottardo@ircm.qc.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 2, 2009; revised on January 13, 2010; accepted on January 14, 2010

1 INTRODUCTION

Chromatin immunoprecipitation (ChIP) has become an important assay for the genome-wide study of protein–DNA interactions and gene regulation. In a typical ChIP experiment, protein complexes on contrast with DNA are cross-linked to their binding sites, the chromatin is sheared into short fragments and the specific DNA fraction interacting with the protein of interest is isolated by immunoprecipitation. A genome-wide readout of the binding sites is produced by hybridization of the DNA pool to a tilling array (Chip-chip) (Lieb *et al*., 2003).

One problem in the analysis of ChIP-chip, and microarrays in general, is that the fluorescence intensity values obtained from microarray hybridization are not directly comparable because of systematic probe biases due to non-specific binding affinity. If not accounted for, such biases can severely deteriorate any subsequent results. The statistical method of normalization aims at making the probe measurements more comparable by reducing these biases. Once the data have been normalized, probe measurements can be processed and enriched regions detected. Several tools and methods have already been proposed for analyzing chip-chip data. Johnson *et al*. (2006) introduced the first normalization method for Affymetrix chip-chip arrays based on nucleotide composition. Toedling *et al*. (2007) have proposed `Ringo`, an open source R package that facilitates the analysis of Nimblegen and Agilent chip-chip experiments. TileProbe is a recent extension of MAT which uses publicly available datasets to improve upon MAT's original normalization (Judy and Ji, 2009). The authors have shown that TileProbe can perform better than MAT but it requires a large number of independent arrays to estimate the probe effect.

In this article, we present an open-source R package `rMAT`, which is based on the popular MAT software. `rMAT` has been written in C and R and provides an efficient implementation of the functionality of MAT as well novel statistical normalization techniques not available in the original MAT. We show that these model refinements can improve normalization and increase power when detecting enriched regions. In addition, `rMAT` is well integrated with other Bioconductor packages (Gentleman *et al*., 2004), which makes it easy for users to construct sophisticated analysis approaches that also leverage other R/Bioconductor functionally; see Supplementary Material Figure 1.

2 METHODS

Architecture: In `rMAT`, the source code is written in C for optimal utilization of system resources, and wrapped in more user friendly R code. In addition, `rMAT` makes use of Grand Central Dispatch on Mac OS X 10.6 and above, which greatly facilitates parallel processing when several arrays are to be processed. `rMAT` adopts a formal object-oriented programming discipline, making use of the S4 system to define classes and methods. This allows for better usability and integration with other R/Bioconductor packages.

Data input: `rMAT` is able to work directly with CEL and BPMAP files provided by Affymetrix; it provides a function that can be used to efficiently parse all input files and create a *tileSet* object containing all necessary information: probe intensities, sequences and genomic locations. On the Estrogen Receptor ChIP-chip data published in Carroll *et al*. (2005), the data can be read with the following commands,

```
> bpmap <- "CHIP_A.NCBIv35.bpmap"
> cel <- c("ER1.CEL", "ER3.CEL", "ER4.CEL", "INP1.CEL",
+      "INP3.CEL","INP4.CEL")
> ERset <- BPMAPCelParser(bpmap, cel, seqName = "chr21")
```

where the 'seqName' argument specifies that only data on chr21 should be read; by default all probes are read.

Statistical analysis: We now describe each analysis step implemented in `rMAT`.

Normalization: The purpose of normalization is to adjust for systematic biases that arise from variation in the microarray technology rather than from biological differences between the printed probes. One of the major

*To whom correspondence should be addressed.

biases in tiling arrays is the so called 'probe effect', which has been shown to be linked to probe sequence composition (Johnson *et al*., 2006). In `rMAT`, the normalization is done in two stages: (i) a prediction model for the probe intensities is derived from their sequence compositions and (ii) each probe is normalized by subtracting its predicted intensity (representing the bias) from the observed intensity. `rMAT` implements the original MAT normalization by default but we have also added a 'robust' option and a novel normalization model based on nucleotide pairs. The robust option is a variant of the original model where the Gaussian errors are replaced by t errors with 4 degrees of freedom, which provides a good robust alternative to the Gaussian distribution. The model based on nucleotide pairs is an improvement of the original model where nucleotide pairs are used to model the probe affinity. We refer the reader to the Supplementary Material for more details on the original MAT model and improvements and an illustration of the improvements on two datasets. In `R`, the normalization based on the improved model with robust errors can be done as follows:

```
> ERsetNorm <- NormalizeProbes(ERset, method = "PairBinned",
 robust = TRUE)
```

which creates a novel *tilingSet* object replacing the raw intensities with the normalized ones.

Data smoothing: Once the data have been normalized, the next step consists in smoothing probe intensities and calculating a score for each probe that will be used to detect enriched regions. Sonication during ChIP procedure shears the DNA to ∼200–1000 bp. As in the MAT version, `rMAT` smooths normalized probe intensities using a sliding window trimmed mean statistics of default size 600 bp, which corresponds to the average fragment size after sonication. The MAT scores can be calculated as follows:

```
> ERscore <- computeMATScore(ERsetNorm, cName = "INP")
```

where the 'cName' argument specifies that all filenames containing 'INP', short for input, are control samples and should be used as such during the score calculation. The calculated scores are stored as a *RangedData* object giving the start/end, chromosome and score of each probe.

Detecting enriched regions: Once scores have been calculated, the score cutoff to call enriched regions can be set arbitrarily, determined based on random-simple qPCR validation, or based on a *P*-value or a false discovery rate (FDR) cutoff. The *P*-value and FDR are calculated using an empirical null distribution as described in Johnson *et al*. (2006). In `rMAT`, this can be done as follows: where the option 'method' can be either 'score' for MAT score, 'pValue' or 'FDR' and 'threshold' specifies the desired cutoff. Here we use the default *P*-value cutoff of $1e-5$.

Data output: Export of *RangedData* instances is supported in the following formats: Browser Extended Display and UCSC web browser compatible wiggle (.wig) file for visualization of raw alignment results through the UCSC Genome Browser and other genome browsers. All of these functionalities are provided by the `rtracklayer` package and are simply done with the `export` function.

Graphical representation: Finally, the *RangedData* objects returned by `rMAT` can also be used for visualization. Here we only discuss two visualization options provided by the `GenomeGraphs` and `rtracklayer` packages. `GenomeGraphs` uses the `biomaRt` package to perform live annotation queries to Ensembl and can combine these to custom tracks to provide high quality graphics; see Figure 1 for an example. `rtracklayer` provides an extensible framework for interacting with multiple genome browsers (e.g. UCSC genome browser, argo) and can also be used to manipulate annotations and custom tracks, just as a regular genome browser; see Figure 5 in Supplementary Material.

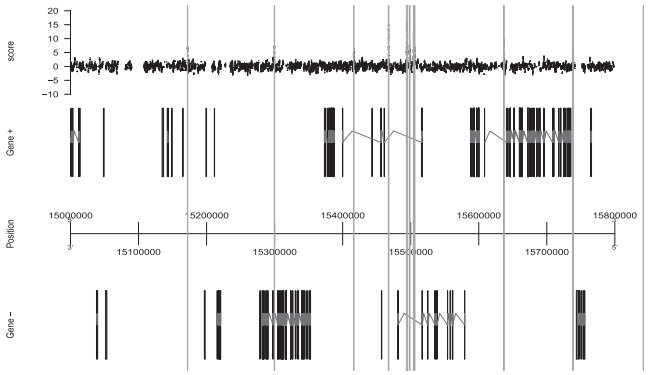

Fig. 1. Example of a graphical representation of `rMAT`-enriched regions with the `GenomeGraphs` package. The score (top track) are the MAT scores. Gene + (respectively −) are the gene annotations on the positive (respectively negative) strand. The vertical rectangles/bars show the enriched regions detected by rMAT at the default cutoff.

ACKNOWLEDGEMENT

The authors thank Wei Li for helpful discussions about the MAT software.

Funding: Natural Sciences and Engineering Research Council of Canada.

Conflict of Interest: none declared.

REFERENCES

Carroll,J.S. *et al*. (2005) Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell*, **122**, 33–43.

Gentleman,R.C. *et al*. (2004) Bioconductor: open software development for computational biology and Bioinformatics. *Genome Biol.*, **5**, R80.

Johnson,D.S. *et al*. (2006) Model-based analysis of tiling-arrays for chip-chip. *Proc. Natl Acad. Sci. USA*, **103**, 12457–12462.

Judy,J.T. and Ji,H. (2009) TileProbe: modeling tiling array probe effects using publicly available data. *Bioinformatics*, **25**, 2369–2375.

Lieb,J.D. *et al*. (2003) Genome-wide mapping of protein-DNA interactions by chromatin immunoprecipitation and DNA microarray hybridization. *Methods Mol. Biol.*, **224**, 99–109.

Toedling,J. *et al*. (2007) Ringo–an R/Bioconductor package for analyzing ChIP-chip readouts. *BMC Bioinformatics*, **8**, 443.