

Model Based Analysis of Tiling Arrays

The rMAT package.

Charles Cheunga*and Raphael Gottardo†and Arnaud Droit‡

April 29, 2009

A step-by-step guide in the analysis of tiling array data using the rMAT package in R

Contents

I	Licensing	3
II	Introduction	3
III	Loading the rMAT Package	4
IV	Loading in the data	4
V	Reading BMAP and CEL files	4
VI	Normalization	6
VII	Finding the Enriched Regions	6

*cykc@interchange.ubc.ca

†raphael.gottardo@ircm.qc.ca

‡arnaud.droit@ircm.qc.ca

1	Creating an Ensembl annotation graphic	7
2	Plotting a Gene	8
3	Overlay for enriched regions	8
VIII	Appendix: Installing rMAT	9

Part I

Licensing

Under the Artistic license 2.0, you are free to use and redistribute this software. However, we ask you to cite the following paper if you use this software for publication.

W. E. Johnson, Li, W., Meyer, C. A., Gottardo, R., Carroll, J. S., Brown, M., and Liu, X. S. (2006). Model-based analysis of tiling-arrays for ChIP-chip. PNAS 103:12457-12462.

Part II

Introduction

In our guide, we include examples of code that we hope will help you when using the rMAT package. The examples are kept at the basic level for ease of understanding. Some of the options in the functions have been set by default. To learn more about the exact parameters and usage of each function, you may type `help(FUNCTION_NAME)` of the function of interest in R after the rMAT package is loaded.

The probe sequence information of an Affymetrix tiling array is stored in the .BPMAP file, while the corresponding expression values (intensity signals) of each experiment is stored separately in each .CEL file. The BPMAP file contains different sequences that describe different contents in the array. For instance, the first sequence may contain probes from chromosome 1 while the second sequence may contain probes from chromosome 2. Each probe would include information such as its Perfect Match base pair sequence (ie. AGCTTCGAAGCTTCGAAGCTTCGAG), location on chromosome, X and Y coordinates, etc. The CEL file does not know anything about the design of the array; it is just a file with columns such as X coordinate, Y coordinate, expression value, and other auxiliary columns. For each array experiment (ie. mock, treated with reagent X, treated with reagent Y), we have one CEL file. These two type of files are stored in a binary format and require a parser (reader) to read its content meaningfully. The common goal in analyzing this type of data is to find activities (DNA-protein interaction, transcription, etc) in specific chromosomal regions. This

package focus on detecting DNA-protein interactions from ChIP-chip experiments. Though, many of the functions are more general than that. The steps in analyzing the tiling array data are as follows.

Part III

Loading the rMAT Package

To load the rMAT package in R, we type

```
> library(rMAT)
```

Part IV

Loading in the data

The next step in a typical analysis is to load in data from Affymetrix CEL files. The data used in this example are available in this package in inst/doc folder.

In this documentation, the path for the data are : /rMAT/inst/doc folder.

Part V

Reading BPMAp and CEL files

Reading the design of tiling array

The first step in analyzing tiling array data is to understand the content of the data. To understand the design, we would explore the header section of the BPMAp file using the function `ReadBPMApAllSeqHeader`. `ReadBPMApAllSequence` takes in the filename of the BPMAp file as an argument. The filename is formatted as a string literal (characters) in unix path format and stored in the variable `BPMApFile`, which is then used by **`ReadBPMApAllSeqHeader`** to specify which BPMAp file to read.

```
> pwd <- ""
> path <- system.file("doc/Sc03b_MR_v04_10000.bpmap", package = "rMAT")
> bpmapFile = paste(pwd, path, sep = "")
> seqHeader <- ReadBPMApAllSeqHeader(bpmapFile)
```

Specifying the filenames

From the above header content, the information we want to obtain is the direct mapping from sequence number to chromosome number. Sequence number is stored in the seqNum column while chromosome number can be read from the Name column, which describes the content of the **sequence**. We would like to read the BMAP and CEL files and merge them by X and Y coordinate so information such as probe sequence and location along the chromosome would pair up with the corresponding expression value. We have already specified the location of the BMAP file in BMAP-File variable, so now let's specify the location of the CEL files. Because BMAPCelParser allows us to parse multiple CEL files simultaneously, we can store the location of multiple files in a vector using `c()` each separate by `", "`.

```
> pathCEL <- system.file("doc/Swr1WTIP_Short.CEL", package = "rMAT")
> arrayFile <- paste(pwd, c(pathCEL), sep = "")
```

Calling BMAPCelParser

We are now ready to call the BMAPCelParser.

```
> ScSet <- BMAPCelParser(bmapFile, arrayFile, verbose = FALSE,
+   groupName = "Sc")
```

groupName corresponding to the genome name used. In this example, we specified *saccharomyces cerevisiae* genome (Sc)

This function returns an object of class `tilingSet` containing all necessary information: probe sequences, genomic positions, chromosomes as well as the probe intensities.

The list of vectors of the merged data are now stored in `ScSet`. Let's explore the (partial) content of `ScSet`.

```
> summary(ScSet)

Genome interrogated:  Sc03b_MR_v04_10000
Chromosome(s) interrogated: 1
Sample name(s):  Swr1WTIP_Short
The total number of probes is: 10000
Preprocessing Information
- Transformation: log
- Normalization: none
```

We are now ready to normalize the raw data. Normalization is a procedure to transform raw data into the so-called normalized expression data so expression values from different tiling arrays can be compared.

Part VI

Normalization

The `NormalizeProbes` function allows users to normalize expression values of different experiments with one command, as long as all those experiments use the same BMAP tiling design file. We can load these raw expression values in batch using `cbind()`. `NormalizeProbes` also requires users to specify the sequence vector. In this case, it is a vector of characters containing the 25 base pair sequence of each probe. (Right now, Normalization works for reading 25mer only.)

For a complete list of parameters for `NormalizeProbes`, please refer to `help(NormalizeProbes)`. We are now ready to run the command.

```
> ScSetNorm <- NormalizeProbes(ScSet, method = "MAT", robust = FALSE,  
+   all = FALSE, standard = TRUE, verbose = FALSE)
```

The user can choose from "MAT", or "PairBinned" normalization method. The Pair option also takes into account of the interaction between adjacent pairs along the probe as covariates for linear regression.

The output in this example is saved in `ScSetNorm`.

Let's explore the (partial) content of `ScSetNorm`.

```
> summary(ScSetNorm)
```

```
Genome interrogated:  Sc03b_MR_v04_10000  
Chromosome(s) interrogated: 1  
Sample name(s):  Swr1WTIP_Short  
The total number of probes is: 10000  
Preprocessing Information  
- Transformation: log  
- Normalization: MAT standardized
```

Part VII

Finding the Enriched Regions

After normalization, we are ready to find enriched regions. We can adjust the threshold for detection of enriched regions. A higher threshold provides a stricter criterion and thus less regions are expected. A higher threshold value also means that the enriched regions would be shorter on average. Please note that the threshold for MATScore with control and without control can be very different because with control the expression values from the Immunoprecipitated data is subtracted from the that of the Control. The generated list of MAT enriched regions will be written to a file defined by parameter output if it is specified. For more details about the overall procedure please refer to the paper cited above.

For a comprehensive list of parameters you can adjust in MATScore, please refer to `help(MATScore)`. Another note is that if FDR is used, threshold should be set in the range between 0 and 1.

```
> ScScore <- MATScore(ScSetNorm, cName = NULL, dMax = 600, nProbesMin = 8,
+   dMerge = 300, method = "score", threshold = 5, verbos = TRUE,
+   bedName = "MyBedFile")

** Calculating MAT Scores **
** Finished processing 10000 probes on 1 arrays **
** Number of Enriched regions is 1 **
** Creating bed file **
```

1 Creating an Ensembl annotation graphic

rMAT results can benefit from integrated visualisation of the genomic information. We have decided to use the GenomeGraphs package. This package uses the biomaRt package to deliver queries to Ensembl e.g. gene/transcript structures to viewports of the grid package, resulting in genomic information plotted together with your data.

To load the GenomeGraphs package in R, we type

```
> library(GenomeGraphs)
```

2 Plotting a Gene

If one wants to plot annotation information from Ensembl then you need to connect to Ensembl BioMart using the `useMart` function of the `biomaRt` package.

```
> mart <- useMart("ensembl", dataset = "scerevisiae_gene_ensembl")
```

If you are interested in plotting a whole gene region, you should create a `GeneRegion` object. In the example below we will retrieve the genes of the chromosome (I) between 1 and 200000. We added a genomic axis as well to give us the base positions.

```
> genomeAxis <- makeGenomeAxis(add53 = TRUE, add35 = TRUE)
> minbase <- 1
> maxbase <- 1e+05

> genesplus <- makeGeneRegion(start = minbase, end = maxbase, strand = "+",
+   chromosome = "I", biomaRt = mart)
> genesmin <- makeGeneRegion(start = minbase, end = maxbase, strand = "-",
+   chromosome = "I", biomaRt = mart)
```

We create a Generic Array for chromosome I only

```
> MatScore <- makeGenericArray(intensity = as.matrix(ScScore@score[ScScore@featureChr
+   "chr1"]), probeStart = ScScore@featurePosition[ScScore@featureChromosome ==
+   "chr1"], dp = DisplayPars(size = 1, color = "black", type = "l"))
```

3 Overlay for enriched regions

Overlays can be used to look enriched regions of the plot.

```
> featurePositionForRegion <- ScScore@featurePosition[ScScore@featureChromosome ==
+   "chr1" & ScScore@featurePosition < maxbase & ScScore@featurePosition >
+   minbase]
> regIndexForRegion <- ScScore@regIndex[ScScore@featureChromosome ==
+   "chr1" & ScScore@featurePosition < maxbase & ScScore@featurePosition >
+   minbase]
> RegionUnique <- unique(regIndexForRegion[regIndexForRegion >
+   0])
> rectList <- vector("list", length(RegionUnique))
```



```

> for (i in 1:length(RegionUnique)) {
+   m <- min(featurePositionForRegion[regIndexForRegion == RegionUnique[i]])
+   M <- max(featurePositionForRegion[regIndexForRegion == RegionUnique[i]])
+   rectList[i] <- makeRectangleOverlay(start = m, end = M, region = c(1,
+     4), dp = DisplayPars(color = "green", alpha = 0.1))
+ }

> gdPlot(list(score = MatScore, `Gene +` = genesplus, Position = genomeAxis,
+   `Gene -` = genesmin), minBase = minbase, maxBase = maxbase,
+   labelCex = 1, overlays = rectList)

```

Part VIII

Appendix: Installing rMAT

To build the `rMAT` package from source, make sure that the following is present in your system:

- GNU Scientific Library (GSL)
- Basic Linear Algebra Subprograms (BLAS)
- a C compiler

GSL can be downloaded at <http://www.gnu.org/software/gsl/>. In addition, the package uses BLAS to perform basic vector and matrix operations. Please go to <http://www.netlib.org/blas/faq.html#5> for a list of optimized BLAS libraries for a variety of computer architectures. For instance, Mac users may use the built-in `vecLib` framework, while users of Intel machines may use the Math Kernel Library (MKL). A C compiler is needed to build the package as the core of the `rMAT` function is coded in C.

For the package to be installed properly you might have to type the following command before installation:

```
export LD_LIBRARY_PATH="/path/to/GSL/:/path/to/BLAS/:$LD_LIBRARY_PATH"
```

which will tell **R** where your GSL and BLAS libraries (see below for more details about BLAS libraries) are. Note that this might have already been configured on your system, so you might not have to do so. In case you need to do it, you might consider copying and pasting the line in your `.bashrc` so that you do not have to do it every time.

Now you are ready to install the package:

```
R CMD INSTALL rMAT_x.y.z.tar.gz
```

The package will look for a BLAS library on your system, and by default it will choose `gslcblas`, which is not optimized for your system. To use an optimized BLAS library, you can use the `--with-blas` argument which will be passed to the `configure.ac` file. For example, on a Mac with `vecLib` pre-installed the package may be installed via:

```
R CMD INSTALL rMAT_x.y.z.tar.gz --configure-args="--with-blas='-  
framework vecLib'"
```

On a 64-bit Intel machine which has MKL as the optimized BLAS library, the command may look like:

```
R CMD INSTALL rMAT_x.y.z.tar.gz --configure-args="--with-blas='-  
L/usr/local/mkl/lib/em64t/ -lmkl -lguide -lpthread'"
```

where `/usr/local/mkl/lib/em64t/` is the path to MKL.

If you prefer to install a prebuilt binary, you need GSL for successful installation.