MethylMix
An R package for identifying DNA methylation driven genes

Olivier Gevaert

October 13, 2014

Stanford Center for Biomedical Informatics
Department of Medicine
1265 Welch Road
Stanford CA, 94305-5479

1 Getting started

Installing the package. To install the *MethylMix* package, the easiest way is through bioconductor:

```
> source("http://bioconductor.org/biocLite.R")
> biocLite(MethylMix)
```

Other ways to install MethylMix is to first download the appropriate file for your platform from the Bioconductor website http://www.bioconductor.org/. For Windows, start R and select the `Packages` menu, then `Install package from local zip file`. Find and highlight the location of the zip file and click on open. For Linux/Unix, use the usual command `R CMD INSTALL` or install from bioconductor.

Loading the package. To load the *MethylMix* package in your R session, type `library(MethylMix)`.

Help files. Detailed information on *MethylMix* package functions can be obtained in the help files. For example, to view the help file for the function `MethylMix` in a Rsession, use `?MethylMix`.

2 Introduction

DNA methylation is a mechanism whereby a methyl-group is added onto a CpG site. Methylation of these CpG sites is associated with gene silencing and is an important mechanism for normal tissue development and is often involved in diseases such as cancer. Recently, many high throughput data

has been generated profiling CpG site methylation on a genome wide bases. This has created large amounts of data on DNA methylation for many disease. Computational analysis of DNA methylation data is required to identify potentiall aberrant DNA methylation compared to normal tissue. MethylMix was developed to tackle this question using a computational approach. MethylMix identifies differential and functional DNA methylation by using a beta mixture model to identify subpopulations of samples with different DNA methylation compared to normal tissue. Functional DNA methylation refers to a significant negative correlation based on matched gene expression data. Together MethylMix outputs hyper and hypomethylated genes which can be used for downstream analysis, and are called MethylMix drivers. MethylMix was designed for cis-regulated promoter differential methylation and works best when specific CpG sites are profiled associated with a gene. For example using data from the 27k and 450k Infinium platforms.

3 Data access and preprocessing

The data in this vignette is accessible at The Cancer Genome Atlas (TCGA) portal. A programmatic way of donwloading data is through the firehose_get tool developed by the broad institute ("https://confluence.broadinstitute.org/display/GDAC/Download"). Firehose_get provides a unified way to download data for all cancer sites and all platforms. We suggest to only use Level3 data for users unfamiliar with the TCGA data. For MethylMix two data types are relevant, DNA methylation data and gene expression data. The methylation data is provided using two platforms the 27k and 450k Illumina platform. We suggest to start from the methylation files marked as *Merge_methylation__humanmethylation*. The methylation archives for both platforms contain the beta-values for each sample and gene. For gene expression there are also two options, either microarray data *Merge_transcriptome__agilentg4502a* or RNA sequencing data *Merge_rnaseqv2__illuminahiseq_rnaseqv2*.
The only preprocessing we recommend for both data sets is to correct for batch effects. We use Combat to adjust for batch effects. One can either donwload the Combat function from the authors website, or use the *sva* bioconductor package which contains the Combat script. The batch information for each sample can be found in the Biospecimen Metadata Browser at the TCGA data portal "https://tcga-data.nci.nih.gov/uuid/uuidBrowser.htm". Next to take care of the multiple probes that are available for each gene for the methylation data, we use a clustering approach that is explained in section 5.

4 Data input for MethylMix

To run MethylMix at least a methylation data set of a particular disease is required. This will allow to identify methylation states associated with a disease for each gene of interest.

```
> library(MethylMix)
> data(METcancer)
> head(METcancer)
```

Adding an appropriate normal or baseline methylation data set increases MethylMix functionality significantly by also being able to distinguish between hyper or increased methylation vs. hypo or decreased methylation with respect to the normal or baseline methylation data set.

```
> library(MethylMix)
> data(METnormal)
> head(METnormal)
```

Finally, if matched gene expression data is available for the same samples that methylation data was available studying this disease, MethylMix will also identify functional differnential methylation by focusing only on differentialy methylation that has a significant inversely correlated effect with gene expression. Each of these three data sets are matrix objects with genes in the rows with unique rownames (e.g. gene symbols) and samples or patients in the columns with unique patient names.

```
> library(MethylMix)
> data(MAcancer)
> head(MAcancer)
```

5 Annotation of CpG probes

When only probe level Illumina data is available, mapping probes to genes is recommended before building mixture models. This allows to focus on cis- regulated differential methylation by only focusing on differential methylation of CpG sites to their closest gene transcripts. Both the 27k and 450k Illumina platforms have database R packages that provide the necessary mapping information. The next example maps the probes of the 27k Illumina platform to Gene Symbol ids.

```
> library(IlluminaHumanMethylation27k.db)
> ProbeToSymbol <- IlluminaHumanMethylation27kSYMBOL
> mapped_probes <- mappedkeys(ProbeToSymbol)
> mapped_probes_List <- as.list(ProbeToSymbol[mapped_probes])
> mapped_probes_List[3:4]
```

Other packages can be used as well such as the FDb.InfiniumMethylation.hg19 package, which provides annotation for both 27k and 450k but requires a 1GB download.

6 MethylMix on glioblastoma TCGA data from the 27k Infinium platform

An example data set from the glioblastoma TCGA project is available in MethylMix. The DNA methylation data for cancer and normal samples from the 27k Infinium platform, and matched gene expression data can be loaded as follows:

```
> library(MethylMix)
> data(METcancer)
> data(METnormal)
> data(MAcancer)
```

Next, we can run MethylMix to identify hypo and hypermethylated genes as follows:

```
> MethylMixResults = MethylMix(METcancer,METnormal,MAcancer)
```

7 MethylMix on breast cancer data for CDH1 TCGA data from the 450k Infinium platform

MethylMix can also be applied on Infinium 450k data. To illustrate this we extracted DNA cancer and normal methylation data for a single gene for the breast cancer TCGA project. We focused on CDH1 differential methylation. First the cancer and normal DNA methylation for CDH1 is loaded and the matched gene expression data:

```
> data(METcancer_CDH1)
> data(METnormal_CDH1)
> data(MAcancer_CDH1)
```

On the 450k platform many more probes are available for every gene. MethylMix can be run on each probe independently however, this will require significant computational resources and many probes show correlated methylation profiles. Therefore, a clustering algorithm can be used to reduce the number of probes and create CpG clusters with similar methylation profiles. For example, the gene CDH1 has 63 probes in the breast cancer TCGA data and using hierarchical clustering we can cluster these probes into CpG clusters as follows:

```
> ProbeCorrelation=cor(t(METcancer_CDH1),method='pearson')
> ClusterResults=hclust(as.dist(1-ProbeCorrelation), method = "complete",
+                       members = NULL)
> plot(ClusterResults)
```

We used hierarchical culstering with complete linkage and the 1- pearson correlation coefficient as a distance measure. Next, we can use a cutoff of corresponding to a minimum correlation of 0.3 in each cluster to define CpG clusters. Then we summarize the clusters by taking the average of the tightly correlated probes in each cluster:

```
> METcancer_CDH1_Clustered=matrix(0,0,length(colnames(METcancer_CDH1)))
> METnormal_CDH1_Clustered=matrix(0,0,length(colnames(METnormal_CDH1)))
> Clusternames=c()
> Clusters=cutree(ClusterResults,h=0.7)
> GeneProbes=rownames(METcancer_CDH1)
> for (i in 1:length(unique(Clusters))) {
+      tmpGeneProbes=GeneProbes[Clusters==i]
+      if (length(tmpGeneProbes)>1) {
+           tmpAveragedProfile=colMeans(METcancer_CDH1[tmpGeneProbes,])
+           METcancer_CDH1_Clustered=rbind(METcancer_CDH1_Clustered,
+                                          tmpAveragedProfile)
+           # Same for normal
+           tmpAveragedProfile=colMeans(METnormal_CDH1[tmpGeneProbes,])
+           METnormal_CDH1_Clustered=rbind(METnormal_CDH1_Clustered,
+                                          tmpAveragedProfile)
```

```
+      } else {
+           METcancer_CDH1_Clustered=rbind(METcancer_CDH1_Clustered,
+                                          METcancer_CDH1[tmpGeneProbes,])
+           METnormal_CDH1_Clustered=rbind(METnormal_CDH1_Clustered,
+                                          METnormal_CDH1[tmpGeneProbes,])
+      }
+      Clusternames=c(Clusternames,paste('CDH1---Cluster',i,sep=""))
+ }
> rownames(METcancer_CDH1_Clustered)=Clusternames
> rownames(METnormal_CDH1_Clustered)=Clusternames
```

For CDH1 in the TCGA breast cancer data, this results in seven CpG clusters, reducing the dimensionality nine fold from the original 63 probes. Now we can build a MethylMix model for each of these seven clusters and investigate which CDH1 CpG cluster is functionally and differentially methylated in breast cancer.

```
> MethylMixResults=MethylMix(METcancer_CDH1_Clustered,METnormal_CDH1_Clustered,
+                            MAtumor_CDH1,OutputRoot='',Parallel=FALSE)
> MethylMix_PlotModel('CDH1---Cluster3',METcancer_CDH1_Clustered,MethylMixResults,
+                     MAtumor_CDH1,METnormal_CDH1_Clustered)
```

It turns out that one CDH1 cluster, cluster 3, is hypomethylated in 40% of breast cancer samples.

8 Session Information

```
> toLatex(sessionInfo())
```

- R version 3.1.1 Patched (2014-09-25 r66681), x86_64-unknown-linux-gnu
- Locale: LC_CTYPE=en_US.UTF-8, LC_NUMERIC=C, LC_TIME=en_US.UTF-8, LC_COLLATE=C, LC_MONETARY=en_US.UTF-8, LC_MESSAGES=en_US.UTF-8, LC_PAPER=en_US.UTF-8, LC_NAME=C, LC_ADDRESS=C, LC_TELEPHONE=C, LC_MEASUREMENT=en_US.UTF-8, LC_IDENTIFICATION=C
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Other packages: MethylMix 1.0.0
- Loaded via a namespace (and not attached): BiocStyle 1.4.0, RColorBrewer 1.0-5, RPMM 1.20, cluster 1.15.3, codetools 0.2-9, doParallel 1.0.8, foreach 1.4.2, iterators 1.0.7, optimx 2013.8.6, parallel 3.1.1, tools 3.1.1