The ENmix User's Guide
Analyzing Illumina HumanMethylation450 BeadChip

Zongli Xu, Liang Niu, Leping Li and Jack A. Taylor

Modified: June 26, 2015. Compiled: July 17, 2015

1 Introduction

The ENmix package provides tools to preprocess data from Illumina HumanMethylation450 array. It includes functions to improve data quality and to prepare clean dataset for EWAS and other DNA methylation analyses. The ENmix uses the same data structure as R package minfi, and is compatible with several R packages, such as minfi and waterMelon, and provides complementary functions for data quality control, background correction, inter-array normalization and variance exploretion.

The Illumina Infinium HumanMethylation450 BeadChip has a complicated design. The array uses bisulfite converted DNA to estimate methylated (M) and unmethylated (U) allele intensity at individual CpG site. Two different assay chemistries are employed. The Infinium I assay is used for 28% (135, 476) of the CpGs on array and has 2 bead types for each CpG locus: one for the methylated and one for the unmethylated alleles. Signal intensities for both alleles at a locus are scanned on the same color channel (Cy3 green for some loci and Cy5 red for others). For a given type I bead, the intensity of the unused color channel has been proposed as a means to estimate background, and termed the out-of-band (oob) intensity. The Infinium II assay is used for 72% (350, 036) of the CpGs on the array and uses a single bead type per CpG. It utilizes two different colors to represent the two different alleles. These are assessed via single base extension with guanine (labeled with Cy3) for methylated, or adenine (labeled with Cy5) for unmethylated alleles. The HumanMethylation450K Beadchip has 850 internal control probes to monitor experimental procedures at different steps, including 613 negative control probes to measure background intensity and 186 non-polymorphic control probes that can be used to monitor color channel difference.

2 Citation

If you are using ENmix package, please cite this publication:

Zongli Xu, Liang Niu, Leping Li and Jack A. Taylor ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip, Under review.

Thanks for your help!

3 Setting up the data

The first step is to import array raw data files (*.idat) using functions provided in R package minfi to create an object of `RGChannelSetExtended`.

```
> library(ENmix)
> require(minfi)
> #see minfi user's guide for the format of sample_sheet.txt file
> targets <- read.table("./sample_sheet.txt", header=T)
> rgSet <- read.450k.exp( targets = targets, extended = TRUE)
> # or read in all idat files under a directory
> rgSet <- read.450k.exp(base = "path_to_directory_idat_files",
+ targets = NULL, extended = TRUE, recursive=TRUE)
```

When methylation IDAT raw data files are not available, such as in most publically available datasets, users can use methylated (M) and unmethylated (U) intensity data to create an object of `MethylSet`.

```
> M<-matrix_for_methylated_intensity
> U<-matrix_for_unmethylated_intensity
> pheno<-as.data.frame(cbind(colnames(M), colnames(M)))
> names(pheno)<-c("Basename","filenames")
> rownames(pheno)<-pheno$Basename
> pheno<-AnnotatedDataFrame(data=pheno)
> anno<-c("IlluminaHumanMethylation450k", "ilmn12.hg19")
> names(anno)<-c("array", "annotation")
> mdat<-MethylSet(Meth = M, Unmeth = U, annotation=anno,
+ phenoData=pheno)
```

As an example for testing, users can use IDAT files installed in R data package minfiData to create an object of `RGChannelSetExtended`.

```
> library(ENmix)
> require(minfi)
```

```
> require(minfiData)
> sheet <- read.450k.sheet(file.path(find.package("minfiData"),
+ "extdata"), pattern = "csv$")
> rgSet <- read.450k.exp(targets = sheet, extended = TRUE)
```

4 Quality Control

4.1 Internal control probes

Illumina 450k chip incorporated 15 different types of internal control probes (total of 848 probes). The function `plotCtrl` can plot intensity values for each type of the controls to evaluate data quality or the performance of specific steps in the process flow. See Illumina Infinium HD Methylation Assay for detailed description on how to interpret these control figures. Here is a list of the control types:

Control types	**Number of probes**
Sample-Independent Controlsi	
STAINING	4
EXTENSION	4
HYBRIDIZATION	3
TARGET REMOVAL	2
RESTORATION	1
Sample-Dependent Controls	
BISULFITE CONVERSION I	12
BISULFITE CONVERSION II	4
SPECIFICITY I	12
SPECIFICITY II	3
NON-POLYMORPHIC	4
NORM_A	32
NORM_C	61
NORM_G	32
NORM_T	61
NEGATIVE	613

```
> plotCtrl(rgSet)
```

These controls can also be plotted in user specified order to check how experimental factors affect methylation measures, such as batch, plate, array or array location.

```
> pinfo=pData(rgSet)
> IDorder=rownames(pinfo)[order(pinfo$Slide,pinfo$Array)]
```

```
> plotCtrl(rgSet,IDorder)
```

4.2 Data distribution

The distribution plot for methylation intensities or beta values is very useful for data summary, visual inspection and identification of outlier samples. Density plot was routinely generated to display data distribution in methylation studies. However density plot is difficult to understand for many investigators. As noted in the man page of `multidensity`, density plot may not able to display data distribution accurately for some data because of the smooth function. Furthermore, the function is computation intensive, and can take several hours to produce density plot for large samples.

Frequency polygon plot is a better alternative for inspection of data distribution. It can accurately reflect data distribution and easy to understand, just like histogram. It is much faster, and only take a few minutes to produce a distribution plot for >1000 samples.

```
> mraw <- preprocessRaw(rgSet)
> #total intensity plot is userful for data quality inspection
> multidensity(assayData(mraw)$Meth+assayData(mraw)$Unmeth)
> #the following code is to compare frequency polygon plot and
> # density plot
> beta<-getBeta(mraw, "Illumina")
> anno=getAnnotation(rgSet)
> beta1=beta[anno$Type=="I",]
> beta2=beta[anno$Type=="II",]
> library(geneplotter)
> jpeg("dist.jpg",height=1200,width=800)
> par(mfrow=c(3,2))
> multidensity(beta,main="Multidensity")
> multifreqpoly(beta,main="Multifreqpoly")
> multidensity(beta1,main="Multidensity: Infinium I")
> multidensity(beta2,main="Multidensity: Infinium II")
> multifreqpoly(beta1,main="Multifreqpoly: Infinium I")
> multifreqpoly(beta2,main="Multifreqpoly: Infinium II")
> dev.off()
```

4.3 Filtering out low quality samples and probes

Data quality measures, including detection P values, number of bead for each methylation reads and average intensities for bisulfite conversion probes can be extracted using function `QCinfo` from an object of `RGChannelSetExtended`. The information can be used for identifying low quality data points. Samples or probes with large percentage of low quality data can be excluded

using function `QCfilter`. Figures and data outputted by function `QCinfo` can be used to select appropriate thresholds (arguments samplethre, CpGthre and bisulthre). Users can also specify a list of samples or CpGs to be filtered out by using options outid and outCpG. Frequency polygon plot for total intensity (methylated intensity + unmethylated intensity) and methylation beta values before and after filtering can be generated using argument plot=TRUE.

```
> qc<-QCinfo(rgSet)
> mraw <- preprocessRaw(rgSet)
> mraw<-QCfilter(mraw, qcinfo=qc, samplethre = 0.01, CpGthre = 0.05
+     ,bisulthre=5000,plot=TRUE, outid=NULL, outCpG=NULL)
```

4.4 Filtering out outliers and low quality data values

Outlier and low quality data values can affect association analysis, function `rm.outlier` can be used to replace these data points with assign missing values. Outlier was defined as value smaller than 3 times IQR from the lower quartile or larger than 3 times IQR from the upper quartile. Some analysis method do not allow missing value, argument impute=TRUE in the function can be specified to impute missing data using k-nearest neighbors method.

```
> #filter out outliers
> b1=rm.outlier(beta)
> #filter out low quality and outlier values
> b2=rm.outlier(beta,qcscore=qcscore)
> #filter out low quality and outlier values, remove rows and columns
> # with too many missing values
> b3=rm.outlier(beta,qcscore=qcscore,rmcr=TRUE)
> #filter out low quality and outlier values, remove rows and columns
> # with too many missing values, and then do imputation
> b3=rm.outlier(beta,qcscore=qcscore,rmcr=TRUE,impute=TRUE)
```

5 Background correction

Function `preprocessENmix` incorporated a model based background correction method ENmix, which models methylation signal intensities with a flexible exponential-normal mixture distribution, together with a truncated normal distribution to model background noise. User can also specify a list of poor performance CpGs to be excluded before background correction using argument exCpG.

```
> mdat<-preprocessENmix(rgSet, bgParaEst="oob", dyeCorr=TRUE,
+ exCpG=NULL, nCores=6)
```

```
> mdat<-QCfilter(mdat, qcinfo=qc, samplethre = 0.01, CpGthre = 0.05
+     ,bisulthre=5000,plot=TRUE, outid=NULL, outCpG=NULL)
```

6 Inter-array normalization

Function `normalize.quantile.450k` can be used to perform quantile normalization on methylation intensity value

```
> mdat<-normalize.quantile.450k(mdat, method="quantile1")
```

7 Parallele computing wrappers

DNA methylation array analysis is computation intensive. To take advandage of the widely available multi-core computation resources, we provided a few parallele computing wrappers to speed up the analysis.

7.1 Probe type bias correction

Function `bmiq.mc` is a multi-core parallel computing wrapper for the `BMIQ` function in R package watermelon.

```
> beta<-bmiq.mc(mdat, nCores=6)
```

7.2 Batch effect correction

Function `ComBat.mc` is a multi-core parallel computing wrapper for the `ComBat` function in R package sva.

```
> batch<-factor(pData(mdat)$Slide)
> betaC<-ComBat.mc(beta, batch, nCores=6, mod=NULL)
```

8 Principal component regression analysis plot

Principal component regression can be used to explore methylation data variance structure (or source of variance) and identifying possible confounding variables for association analysis. Principal components are derived using singular value decomposition method in standardized (for each probe) beta value matrix.

```
> cov<-data.frame(group=pData(mdat)$Sample_Group,
+     slide=factor(pData(mdat)$Slide))
> pcrplot(beta, cov, npc=6)
```

9 Multimodal CpGs

Function `nmode.mc` uses an empirical approach to identify multimodal distributed CpGs. When measured in a population of people the majority of CpGs on the Illumina HumanMethylation450 BeadChip have unimodal distributions of DNA methylation values with relatively small between-person variation. However, some CpGs (typically 10,000 often seemingly the result of SNPs in the probe region) may have multimodal distributions of methylation values with sizeable differences between modes and greater between-person variation. These multimodal distributed data are usually caused by SNP effect, problematic probe design or some other unknown artifacts instead of actual methylation level and thus should be excluded from DNA methylation analysis. Researchers have often excluded CpGs based on SNP annotation information. However, because SNP annotation always depends on population origin, we found that this approach alone may exclude many well-distributed (unimodal) CpGs, while still failing to identify other multi-modal CpGs. We developed an empirical approach to identify CpGs that are not uni-modally distributed, so that researchers can make more informed decisions about whether to exclude them in their particular study populations and analyses.

```
> nmode<- nmode.mc(beta, minN = 3, modedist=0.2, nCores = 5)
```

To illustrate the function, we have applied the approach in two different datasets: Sister study data (SS, n=200, Harlid et al, Plos one, 2015) and Norway Facial Clefts Study (NCL, n=891, Markunas et al. Environ Health Perspect. 2014) datasets. To simplify our illustration, we excluded 52,238 CpGs on X or Y chromosome, and non-specific probes annotated by Price et al (Price et al, Epigenetics Chromatin, 2013).

From the following summary table we can see this approach can correctly identifies all 65 known SNP probes as not unimodal in both datasets. More than 90% of CpGs with an annotated SNP in probe region are uni-modal in both populations. For CpGs with any SNP at CpG target site, more than 80% are unimodal, and even 50% of CpGs with an annotated common (minor allele frequency > 0.05) SNP at target site have unimodal distributions.

Table 1. Number of unimodal and multimodal CpGs in two independent datasets.

	Unimodal in NCL				Concordance	% unimodal	
	Yes		No				
	Unimodal in SS		Unimodal in SS			NCL	SS
	Yes	No	Yes	No			
SNP probe	0	0	0	65	1.00	0	0
SNP in target site with MAF 0.05	2785	493	315	2584	0.87	0.531	0.502
SNP in target site#	14131	810	409	2844	0.93	0.821	0.799
SNP in probe#	109480	2140	676	3227	0.98	0.966	0.954
No SNP in probe#	314502	2264	443	542	0.99	0.997	0.991

*Based on 1000 genome project data for European population
based on annotation by Price et al, Epigenetics Chromatin, 2013

10 Example Analysis

Working with IDAT files

```
> library(ENmix)
> #read in data
> sheet <- read.450k.sheet(file.path(find.package("minfiData"),
+     "extdata"), pattern = "csv$")
> rgSet <- read.450k.exp(targets = sheet, extended = TRUE)
> #control plots
> plotCtrl(rgSet)
> #QC info
> qc<-QCinfo(rgSet)
> mraw <- preprocessRaw(rgSet)
> beta<-getBeta(mraw, "Illumina")
> #distribution plot
> multidensity(beta,main="Methylation Beta value distribution")
> #Search for multimodal CpGs
> #sample size in this example data is too small for this purpose!
> bb=beta; bb[qc$detP>0.05 | qc$nbead<3]=NA #exclude low quality data first
> nmode<-nmode.mc(bb, minN = 3, modedist=0.2, nCores = 6)
> #background correction and dye bias correction
> mdat<-preprocessENmix(rgSet, bgParaEst="oob", dyeCorr=TRUE, nCores=6)
> #exclude samples or CpGs with poor data quality
> #user specified CpG list to be excluded
> outCpG = names(nmode)[nmode>1]
```

```
> mdat<-QCfilter(mdat, qcinfo=qc, samplethre = 0.01, CpGthre = 0.05
+     ,plot=TRUE, outid=NULL, outCpG=outCpG)
> #inter-array normalization
> mdat<-normalize.quantile.450k(mdat, method="quantile1")
> #probe-type bias adjustment
> beta<-bmiq.mc(mdat, nCores=6)
> # Principal component regression analysis plot
> cov<-data.frame(group=pData(mdat)$Sample_Group,
+     slide=factor(pData(mdat)$Slide))
> pcrplot(beta, cov, npc=6)
> #filter out low quality and outlier values, remove rows and columns
> #with too many missing value, and then do imputation
> beta <- rm.outlier(beta,qcscore=qc,rmcr=TRUE,impute=TRUE)
> #batch correction
> batch<-factor(pData(mdat)[colnames(beta),]$Slide)
> betaC<-ComBat.mc(beta, batch, nCores=6, mod=NULL)
```

11 SessionInfo

- R version 2.14.1 (2011-12-22), `x86_64-unknown-linux-gnu`
- Locale: `LC_CTYPE=en_US.UTF-8`, `LC_NUMERIC=C`, `LC_TIME=en_US.UTF-8`, `LC_COLLATE=en_US.UTF-8`, `LC_MONETARY=en_US.UTF-8`, `LC_MESSAGES=en_US.UTF-8`, `LC_PAPER=C`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=en_US.UTF-8`, `LC_IDENTIFICATION=C`
- Base packages: base, datasets, graphics, grDevices, methods, stats, utils
- Loaded via a namespace (and not attached): tools 2.14.1

12 References

Zongli Xu, Liang Niu, Leping Li and Jack A. Taylor, ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip, Under review

Illumina Inc., Infinium HD Assay Methylation Protocol Guide, Illumina, Inc. San Diego, CA.

Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD and Irizarry RA (2014). Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA Methylation microarrays. Bioinformatics, 30(10), pp. 13631369.

Pidsley, R., CC, Y.W., Volta, M., Lunnon, K., Mill, J. and Schalkwyk, L.C. (2013) A data-driven approach to preprocessing Illumina 450K methylation array data. BMC genomics, 14, 293.

Teschendorff AE et. Al (2013). A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics.

Johnson, WE, Rabinovic, A, and Li, C (2007). Adjusting batch effects in microarray expression data using Empirical Bayes methods. Biostatistics 2007 8(1):118-127.

Harlid S, Xu Z, Panduri V, D'Aloisio AA, DeRoo LA, Sandler DP, Taylor JA. In utero exposure to diethylstilbestrol and blood DNA methylation in women ages 40-59 years from the sister study. PLoS One. 2015 Mar 9;10(3):e0118757. doi: 10.1371/journal.pone.0118757. PMID: 25751399

Markunas CA, Xu Z, Harlid S, Wade PA, Lie RT, Taylor JA, Wilcox AJ. Identification of DNA methylation changes in newborns related to maternal smoking during pregnancy.Environ Health Perspect. 2014 Oct;122(10):1147-53. doi: 10.1289/ehp.1307892. PMID: 24906187

Price ME1, Cotton AM, Lam LL, Farr P, Emberly E, Brown CJ, Robinson WP, Kobor MS.Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. Epigenetics Chromatin. 2013 Mar 3;6(1):4. doi: 10.1186/1756-8935-6-4. PMID: 23452981