The ENmix User's Guide

Zongli Xu, Liang Niu and Jack A Taylor

Modified: October 9 2020. Compiled: May 2, 2021

Contents

1 Introduction

The ENmix package provides a set of quality control, preprocessing/correction and data analysis tools for Illumina Methylation Beadchips. It includes functions to read in raw idat data, background correction, dye bias correction, probe-type bias adjustment, along with a number of additional tools. These functions can be used to remove unwanted experimental noise and thus to improve accuracy and reproducibility of methylation measures. ENmix functions are flexible and transparent. Users have option to choose a single pipeline command to finish all data preprocessing steps (including quality control, background correction, dye-bias adjustment, between-array normalization and probe-type bias correction) or to use individual functions sequentially to perform data pre-processing in a more customized manner. In addition the ENmix package has selectable complementary functions for efficient data visualization (such as QC plots, data distribution plot, manhattan plot and Q-Q plot), quality control (identifing and filtering low quality data points, samples, probes, and outliers, along with imputation of missing values), identification of probes with multimodal distributions due to SNPs or other factors, exploration of data variance structure using principal component regression analysis plot, preparation of experimental factors related surrogate control variables to be adjusted in downstream statistical analysis, an efficient algorithm oxBS-MLE to estimate 5-methylcytosine and 5-hydroxymethylcytosine level; estimation of celltype proporitons; methlation age calculation and differentially methylated region (DMR) analysis.

Most ENmix package can also support the data structure used by several other related R packages, such as minfi, wateRmelon and ChAMP, providing straightforward integration of ENmix-corrected datasets for subsequent data analysis.

The software is designed to support large scale data analysis, and provides multi-processor parallel computing options for most functions.

2 Citation

The following publications can be refered to learn more about the methods implemented in this package.

Xu Z, Niu L, Li L, Taylor JA. ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip, Nucleic Acids Research, 2015

Niu L, Xu Z, Taylor JA. RCP: a novel probe design bias correction method for Illumina Methylation BeadChip. Bioinformatics, 2016

Xu Z, Langie SA, De Boever P, Taylor JA, Niu L. RELIC: a novel dye-bias correction method for Illumina Methylation BeadChip. BMC Genomics. 2017

Xu Z, Taylor JA, Leung YK, Ho SM, Niu L. oxBS-MLE: an efficient method to estimate 5-methylcytosine and 5-hydroxymethylcytosine in paired bisulfite and oxidative bisulfite treated

DNA. Bioinformatics. 2016

Xu Z, Taylor JA. Reliability of DNA methylation measures using Illumina methylation BeadChip. Epigenetics 2020

Xu Z, Xie C, Taylor JA, Niu L. ipDMR: Identification of differentially methylated regions with interval P values. Bioinformatics, 2020

Xu Z, Niu L, Taylor JA. ENmix: a comprehensive R package for the analysis of Illumina DNA methylation arrays. in review

3 List of functions

Function	Description
Data acquisition	
readidat	Read idat files into R
readmanifest	Read array manifest file into R
Quality control	
QCinfo	Extract and visualize QC information
plotCtrl	Generate internal control plots
getCGinfo	Extract CpG probe annotation information
calcdetP	Compute detection P values
qcfilter	Remove low quality values, samples or CpGs; remove outlier samples and perform imputation
QCfilter	Sample, CpG, low quality data filter
nmode	Identify gap probes
dupicc	Calculate ICC with data for duplicates
freqpoly	Frequency polygon plot for single variable
multifreqpoly	Frequency polygon plot for multiple variables
Preprocessing	
mpreprocess	Preprocessing pipeline
preprocessENmix	ENmix background correction and dye bias correction
relic	RELIC dye bias correction
norm.quantile	Quantile normalization
rcp	RCP probe design type bias correction
Differential analysis	
ipdmr	ipDMR differentially methylated region analysis
combp	Combp differentially methylated region analysis
Other functions	
oxBS.MLE	MLE estimates of 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC)
estimateCellProp	Estimate cell type proportions
methyAge	Calculate methylation age
predSex	Estimate sample sex
ctrlsva	Derive surrogate variables to control for experimental confoundings using non-negative control probes
pcrplot	Principal component regression plot
mhtplot	P value manhattan plot
p.qqplot	P value Q-Q plot
B2M	Convert Beta value to M value
M2B	Convert M value to Beta value

4 Example Analysis

ENmix functions are easy to run. The following examples demonstrate how to perform a typical QC and data-preprocessing, as well as in depth methylation analyses.

4.1 Example 1: using pipeline

The pipeline function `mpreprocess` can be used to perform quality control and data preprocessing in just two sentences.

```
> rgSet <- readidat(datapath)
> beta=mpreprocess(rgSet)
```

These two sentences can achieve the following tasks:

- Read in raw idat files
- Data preprocessing
  - background correction
  - dye-bias correction
  - between-array normalization
  - probe-type bias adjustment
- Quality control
  - Identify and exclude low quality probes
  - Identify and exclude low quality samples
  - Identify and exclude outlier samples
  - Remove low quality and outlier data points
  - Imputation of missing values
- Visualization
  - Data distribution plots before and after low quality sample removal
  - Plots of quality controls measures (detp, bisulfite) to guide selection of thresholds

The following is an executable example.

```
> library(ENmix)
> #read in data
> require(minfiData)
> #read in IDAT files
```

```
> path <- file.path(find.package("minfiData"),"extdata")
> rgSet <- readidat(path = path,recursive = TRUE)
> #data pre-processing
> beta=mpreprocess(rgSet,nCores=6)
> #quality control, data pre-processing and imputation
> beta=mpreprocess(rgSet,nCores=6,qc=TRUE,fqcfilter=TRUE,
+       rmcr=TRUE,impute=TRUE)
```

4.2 Example 2: using individual function

The example code below is basically doing the same thing as in example 1, but has more customized options.

```
> library(ENmix)
> #read in data
> path <- file.path(find.package("minfiData"),"extdata")
> rgSet <- readidat(path = path,recursive = TRUE)
> #QC info
> qc<-QCinfo(rgSet)
> #background correction and dye bias correction
> #if qc info object is provided, the low quality or outlier samples
> # and probes will be excluded before background correction
> mdat<-preprocessENmix(rgSet, bgParaEst="oob", dyeCorr="RELIC",
+                       QCinfo=qc, nCores=6)
> #between-array normalization
> mdat<-norm.quantile(mdat, method="quantile1")
> #probe-type bias adjustment
> beta<-rcp(mdat,qcscore=qc)
> #assign missing to low quality and outlier data points, remove samples
> # and probes with too many missing data, and do imputation
> beta <- qcfilter(beta,qcscore=qc,rmcr=TRUE,impute=TRUE)
```

4.3 Example 3: A more comprehensive example

This example is to demonstrate the flexibility of the package functions. Getting detailed information about data quality, SNP-like probes, excluding user-specified probes and/or samples, principal component regression analysis, low quality or outlier sample or probe removal, sex prediction, cell type proportion estimation, methylation age estimation, differentially methylated region analysis and surrogate variables, etc.

```
> library(ENmix)
> #read in data
```

```
> path <- file.path(find.package("minfiData"),"extdata")
> rgSet <- readidat(path = path,recursive = TRUE)
> #attach some phenotype info for later use
> sheet <- read.metharray.sheet(file.path(find.package("minfiData"),
+          "extdata"),pattern = "csv$")
> rownames(sheet)=paste(sheet$Slide,sheet$Array,sep="_")
> colData(rgSet)=as(sheet[colnames(rgSet),],"DataFrame")
> #generate internal control plots to inspect quality of experiments
> plotCtrl(rgSet)
> #generate quality control (QC) information and plots,
> #identify outlier samples in data distribution
> #if set detPtype="negative", recommand to set depPthre to <= 10E-6
> #if set detPtype="oob", depPthre of 0.01 or 0.05 may be sufficient
> #see Heiss et al. PMID: 30678737 for details
> qc<-QCinfo(rgSet,detPtype="negative",detPthre=0.000001)
> ###data preprocessing
> #background correction and dye bias correction
> mdat<-preprocessENmix(rgSet, bgParaEst="oob", dyeCorr="RELIC",
+                       QCinfo=qc, nCores=6)
> #between-array normalization
> mdat<-norm.quantile(mdat, method="quantile1")
> #attach phenotype data for later use
> colData(mdat)=as(sheet[colnames(mdat),],"DataFrame")
> #probe-type bias adjustment
> beta<-rcp(mdat,qcscore=qc)
> #Search for multimodal CpGs, so called gap probes
> #beta matrix before qcfilter() step should be used for this purpose
> nmode<-nmode(beta, minN = 3, modedist=0.2, nCores = 6)
> #filter out poor quality and outlier data points for each probe;
> #rows and columns with too many missing values can be removed
> #Imputation will be performed to fill missing data if specify.
> beta <- qcfilter(beta,qcscore=qc,rmcr=TRUE, rthre=0.05,
+                     cthre=0.05,impute=TRUE)
> #If for epigenetic mutation analysis, use
> #beta <- qcfilter(beta,qcscore=qc,rmoutlier=FALSE,rmcr=TRUE, rthre=0.05,
> #                   cthre=0.05,impute=FALSE)
>
> ##Principal component regression analysis plot
> #phenotypes to be studied in the plot
> cov<-data.frame(group=colData(mdat)$Sample_Group,
+     slide=factor(colData(mdat)$Slide))
> #missing data is not allowed in the analysis!
```

```
> pcrplot(beta, cov, npc=6)
> #Non-negative control surrogate variables, which can be used in
> # downstream association analysis to control for batch effects.
> csva<-ctrlsva(rgSet)
> #estimate cell type proportions
> #rgDataset or methDataSet can also be used for the estimation
> celltype=estimateCellProp(userdata=beta,
+          refdata="FlowSorted.Blood.450k",
+          nonnegative = TRUE,normalize=TRUE)
> #predic sex based on rgDataSet or methDataset
> sex=predSex(rgSet)
> sex=predSex(mdat)
> #Methylation age calculation
> mage=methyAge(beta)
> #ICC analysis can be performed to study measurement reliability if
> # have duplicates, see manual
> #dupicc()
>
> #After association analysis, the P values can be used for DMR analysis
> #simulate a small example file in BED format
> dat=simubed()
> #using ipdmr method
> ipdmr(data=dat,seed=0.1)
> #using comb-p method
> combp(data=dat,seed=0.1)
```

5 Setting up the data

The first step is to import array raw data files (*.idat) to create an object of `rgDataSet`, which is similar to `RGChannelSetExtended` in minfi package, but with probe annotation integrated, and having smaller data size (about 1/3 smaller). Alternatively, User can also use minfi package to read in idat files and create `RGChannelSetExtended`. Most functions in ENmix support `RGChannelSetExtended` and `RGChannelSet`.

Some array types and corresponding manifestfiles can be guessed by the program based on the number of probes per array. However, we recommand to provide correct manifest file directly, which can be easily downloaded from Illumina website, see below for some examples. This option also allows the function to read in data from newer array, such as mouse array.

```
> library(ENmix)
> rgSet <- readidat(path = "directory path for idat files",
+                   recursive = TRUE)
```

```
> #Download manifestfile for HumanMethylation450 beadchip
> system("wget ftp://webdata2:webdata2@ussd-ftp.illumina.com/downloads/Prod
+ /HumanMethylation450/HumanMethylation450_15017482_v1-2.csv")
> mf="HumanMethylation450_15017482_v1-2.csv"
> rgSet <- readidat(path = "path to idat files",manifestfile=mf,recursive =
```

When methylation IDAT raw data files are not available, such as in many publicly available datasets, users can use methylated (M) and unmethylated (U) intensity data to create an object of `MethylSet`. `MethylSet` is also supported by most functions in ENmix.

```
> 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)
```

6 Quality Control

6.1 Internal control probes

Both Illumina 450k and EPIC array incorporated 15 different types of internal control probes (total of 848 probes for 450K and 635 probes in EPIC). The control plots from ENmix function `plotCtrl` are similar to the control plots generated by Illumina GenomeStudio software. See Illumina Infinium HD Methylation Assay for detailed description on how to interpret these control figures.

```
> plotCtrl(rgSet)
```

Below is a list of internal control types and number of probes for each type.

Control types	450K	EPIC
Sample-Independent Controls		
STAINING	4	6
EXTENSION	4	4
HYBRIDIZATION	3	3
TARGET REMOVAL	2	2
RESTORATION	1	1
Sample-Dependent Controls		
BISULFITE CONVERSION I	12	10
BISULFITE CONVERSION II	4	4
SPECIFICITY I	12	12
SPECIFICITY II	3	3
NON-POLYMORPHIC	4	9
NORM_A	32	27
NORM_C	61	58
NORM_G	32	27
NORM_T	61	58
NEGATIVE	613	411

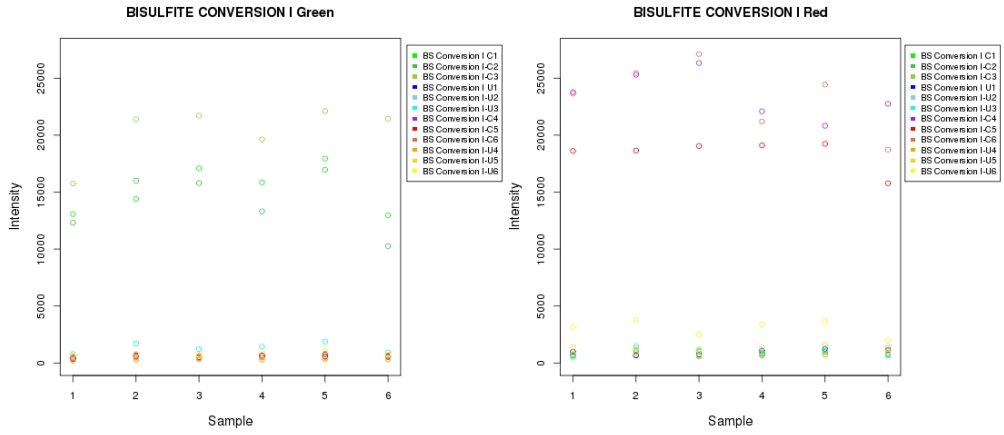

Figure 1: Bisulfite conversion controls for type I probes

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

```
> path <- file.path(find.package("minfiData"),"extdata")
> rgSet <- readidat(path = path,recursive = TRUE)
> sheet <- read.metharray.sheet(file.path(find.package("minfiData"),
+           "extdata"), pattern = "csv$")
> rownames(sheet)=paste(sheet$Slide,sheet$Array,sep="_")
> colData(rgSet)=as(sheet[colnames(rgSet),],"DataFrame")
```

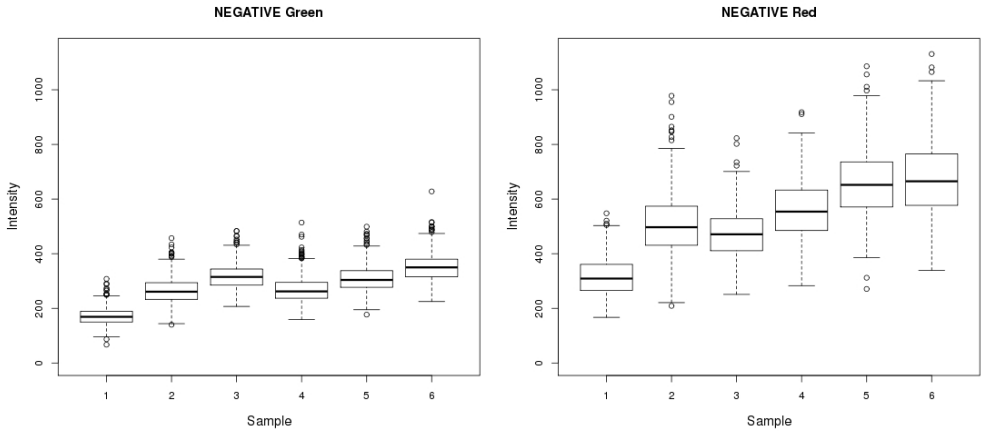

Figure 2: Negative control probes

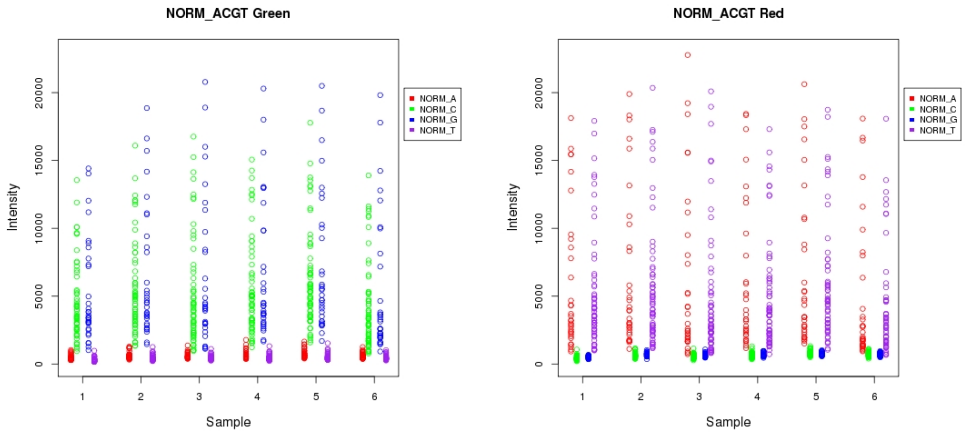

Figure 3: NORM ACGT control probes

```
> #control plots
> IDorder=rownames(colData(rgSet))[order(colData(rgSet)$Slide,
+          colData(rgSet)$Array)]
> plotCtrl(rgSet,IDorder)
```

6.2 Data distribution plots

Methylation intensity or beta value distribution plots are very useful for data summary, visual inspection and identification of outlier samples. Density plot is routinely generated using R function `multidensity`. However, the function is computationally intensive, and can take several hours to produce density plots for a large methylation dataset.

Frequency polygon plot is an alternative figure for inspection of data distribution. Similar to histogram, it can accurately reflect data distribution and easy to understand. The function is also much faster, and only takes a few minutes to produce a distribution plot for >1000 samples.

```
> mraw <- getmeth(rgSet)
> #total intensity plot is userful for data quality inspection
> #and identification of outlier samples
> multifreqpoly(assays(mraw)$Meth+assays(mraw)$Unmeth,
+              xlab="Total intensity")
> #Compare frequency polygon plot and density plot
> beta<-getB(mraw)
> anno=rowData(mraw)
> beta1=beta[anno$Infinium_Design_Type=="I",]
> beta2=beta[anno$Infinium_Design_Type=="II",]
> library(geneplotter)
> jpeg("dist.jpg",height=900,width=600)
> par(mfrow=c(3,2))
> multidensity(beta,main="Multidensity")
> multifreqpoly(beta,main="Multifreqpoly",xlab="Beta value")
> multidensity(beta1,main="Multidensity: Infinium I")
> multifreqpoly(beta1,main="Multifreqpoly: Infinium I",
+ xlab="Beta value")
> multidensity(beta2,main="Multidensity: Infinium II")
> multifreqpoly(beta2,main="Multifreqpoly: Infinium II",
+ xlab="Beta value")
> dev.off()
```

See the following figures (Figure 4) generated from the above code. When type I and type II probes are plotted separately (Fig 4 bottom 4 panels) the difference in modes between type I and II probes can be appreciated. But when all probes are plotted together (Fig 4 top panels), the multidensity plot obscures these differences, while they remain readily apparent in the multifreqpoly plot. In

addition, the multidensity plots appear to suggest that probes range in value from <0 to >1, whereas multifreqpoly correctly show the range from 0 to 1.

6.3 QC information, outlier samples, low quality samples and probes

Data quality measures, including detection P values, average intensities for bisulfite conversion probes and number of beads for each methylation read can be extracted or estimated using the function `QCinfo` from an object of `rgDataSet` or `RGChannelSetExtended`. Based on default or user specified quality score thresholds, the `QCinfo` will identify a list of low quality samples and CpG probes as wells as outlier samples based on total intensity or beta value distribution. These samples and probes should be excluded before further analysis. Data quality score figures "qc_sample.jpg" and "qc_CpG.jpg" from `QCinfo` can be used to guide the selection of quality score thresholds. Low quality samples and probes can be filtered out using `preprocessENmix` if the QCinfo object is provided. These samples and probes are listed as "badsample" and "badCpG" in QCinfo object, and also marked in red in figure "freqpolygon_beta_beforeQC.jpg".

```
> qc<-QCinfo(rgSet)
> #exclude badsample and bad CpG before backgroud correction
> mdat<-preprocessENmix(rgSet, QCinfo=qc, nCores=6)
```

6.4 Filtering outliers, low quality data points, missing values and imputation

After excluding low quality samples and probes, as well as outlier samples. we can still have many low quality and outlier data points for many CpGs. These small percentage of data points can have big impact on downstream association statistical tests for individual CpGs. Function `qcfilter` can be used to filter out these data points and replace them with missing value NA. Outliers are defined as values smaller than 3 times IQR from the lower quartile or larger than 3 times IQR from the upper quartile. Some statistical methods do not allow missing values,(e.g. principal component analysis), so argument *impute=TRUE* in the function can be specified to impute missing data using k-nearest neighbors method.

```
> #filter out outliers only
> b1=qcfilter(beta)
> #filter out low quality and outlier values
> b2=qcfilter(beta,qcscore=qc)
> #filter out low quality and outlier values, remove rows
> #and columns with too many missing values
> b3=qcfilter(beta,qcscore=qc,rmcr=TRUE)
> #filter out low quality and outlier values, remove rows and
> #columns with too many (rthre=0.05,cthre=0.05, 5% in default) missing val
```

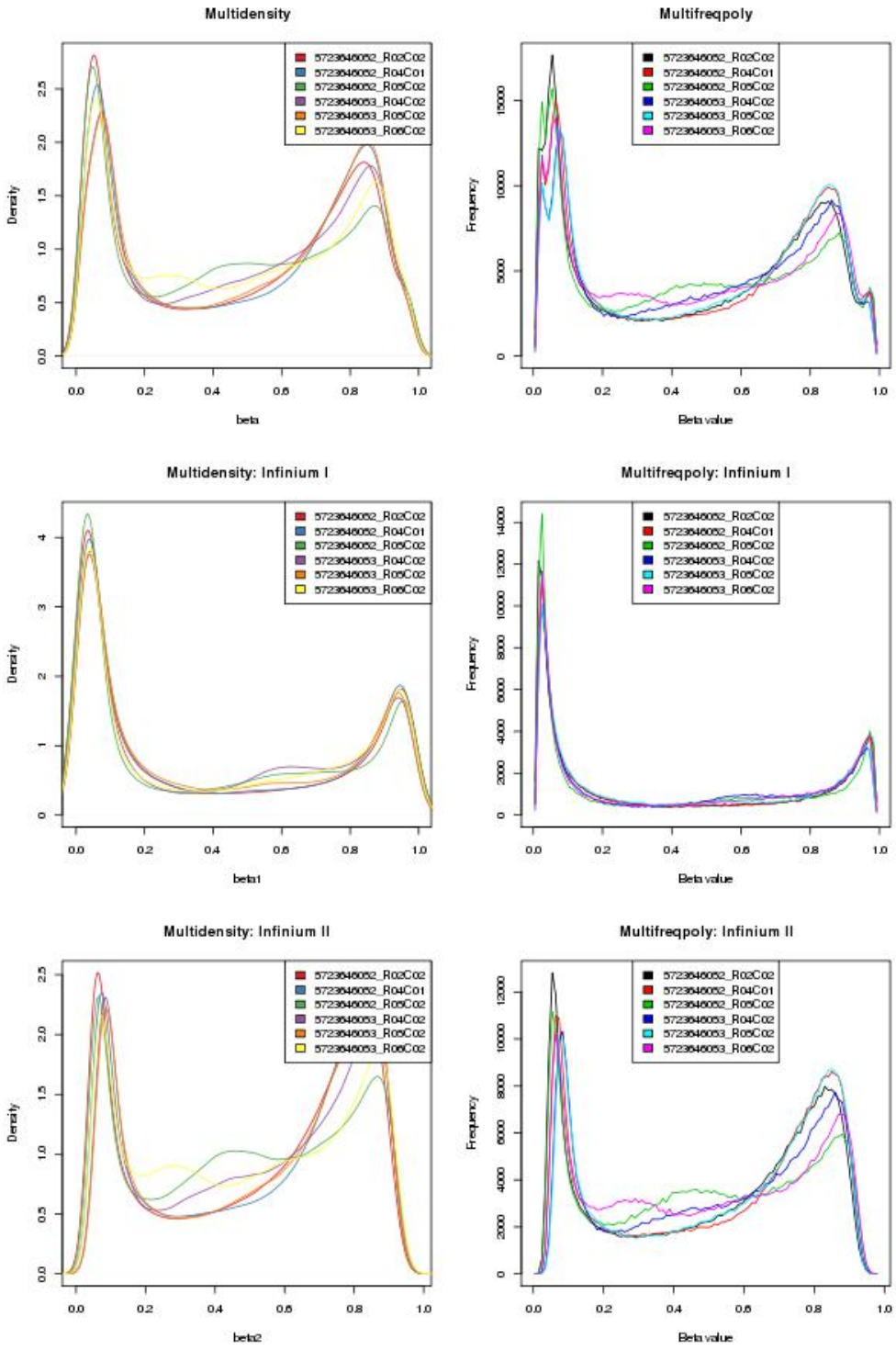

Figure 4: Methylation beta value distribution plots for all probes (top 2 panels) and for type I (middle panels) and II (bottom panels) probes separately. The smoothing function in multidensity plots (panels on left) results in misleading range and mode information which are more accurately depicted in the multifreqpoly plots (panels on right)

```
> # and then do imputation
> b3=qcfilter(beta,qcscore=qc,rmcr=TRUE,rthre=0.05,
+                cthre=0.05,impute=TRUE)
```

7 Background correction and dye-bias correction

Function `preprocessENmix` incorporates 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. Argument "dyeCorr" can be used to specify a method for dye-bias correction, the default is RELIC.

See the following papers for the detailed description of related methods:

Zongli Xu, et. al. ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip, Nucleic Acids Research, 2015

Xu Z, Langie SA, De Boever P, Taylor JA, Niu L. RELIC: a novel dye-bias correction method for Illumina Methylation BeadChip. BMC Genomics. 2017

If argument QCinfo is specified, the low quality samples and probes identified by function `QCinfo` will be excluded before ENmix background correction. Using argument *exSample* and *exCpG*, User can also specify a list of samples or probes to be excluded before background correction.

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

8 Inter-array normalization

Function `norm.quantile` can be used to perform quantile normalization on methylation intensity values.

```
> mdat<-norm.quantile(mdat, method="quantile1")
```

9 Probe-type bias adjustment

The majority of probes on Illumina 450K and EPIC BeadChips are type II probes. Although type II probes facilitate increased array genome coverage, they were shown to have decreased dynamic range and reproducibility compared to type I probes. Taking advantage of the high spatial correlation of DNA methylation levels along the human genome, The RCP (Regression on Correlated

Probes) method utilizes nearby (<25 bp) type I and II probe pairs to derive the quantitative relationship between probe types and then recalibrates type II probe measurements using type I probes as referents.

```
> beta<-rcp(mdat)
```

See the following publication for the detailed description of the method:

Niu L, Xu Z, Taylor JA. RCP: a novel probe design bias correction method for Illumina Methylation BeadChip. Bioinformatics, 2016

10 Batch effect correction

Function `ctrlsva` can be used to estimate surrogate variables for batch effects and unknown experimental confounders using intensity data for non-negative internal control probes. These surrogate variables can then be modeled as covariables in downstream association analysis to adjust for experimental variation.

```
> sva<-ctrlsva(rgSet)
```

11 Principal component regression analysis plot

First, principal component analysis will be performed on the standardized beta value matrix (standardized probe by probe), and then the specified number of top principal components (that explain most data variation) will be used to perform linear regression with each specified variables, such as batch or phenotype variables. Regression P values will be plotted to explore methylation data variance structure and to identify possible confounding variables to guide association statistical analysis. Principal components are ordered according to the percentage of variance they explained from large to small. PCA will not allow for missing values, so specify impute=true in preprocessing if there is missing data.

```
> require(minfiData)
> mdat <- preprocessRaw(RGsetEx)
> beta=getBeta(mdat, "Illumina")
> group=pData(mdat)$Sample_Group
> slide=factor(pData(mdat)$Slide)
> cov=data.frame(group,slide)
> pcrplot(beta,cov,npc=6)
```

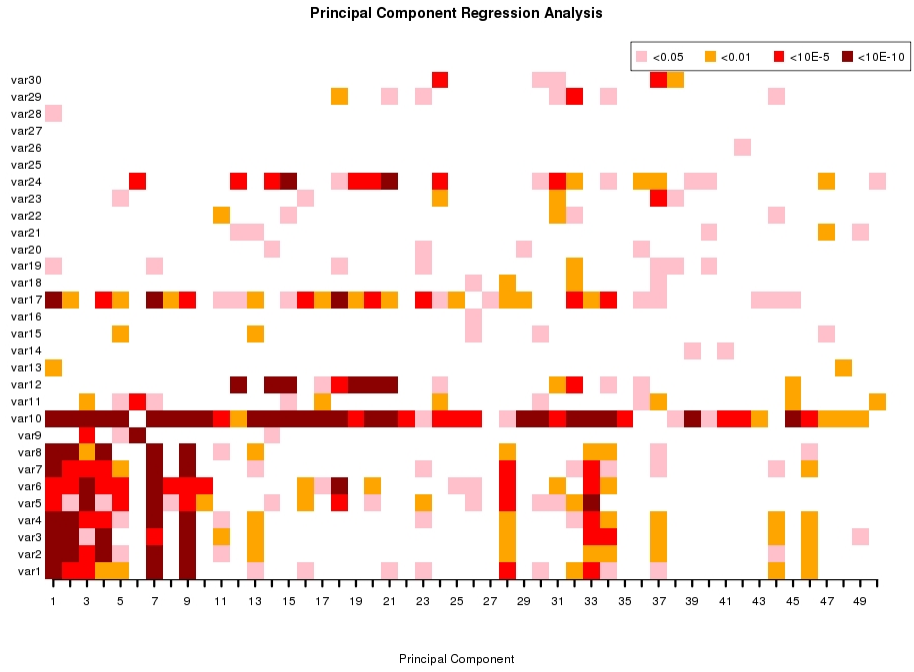

Figure 5: Example principal component regression p value plot of raw data generated using 450K methylation data from a published study

12 Multimodal CpGs or gap probes

Function `nmode` uses an empirical approach to identify CpGs whose methylation values are multimodal distributed (SNP-like or gap probes). 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 may have multimodal distributions of methylation values with sizeable differences between modes and large between-person variation.

These multimodal distributed data are usually caused by SNP effect in the probe region, problematic probe design or other unknown artifacts rather than by actual difference in methylation level and thus should be excluded from DNA methylation analysis. Researchers have often excluded CpGs based simply on SNP annotation information (e.g. reported SNPs in or near probe sequences). However, because SNP frequence and annotation always depends on population origin, we find that this approach alone may exclude many well-distributed (unimodal) CpGs, while still failing to identify other problematic multi-modal CpGs. We developed an empirical approach to identify CpGs that are obviously not uni-modally distributed, so that researchers can make more informed decisions about whether to exclude them in their particular study populations and analyses.

See online supplementary materials of the following paper for an evaluation of the method using

published EWAS data.

Zongli Xu, Liang Niu, Leping Li and Jack A. Taylor ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip, Nucleic Acids Research, 2015

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

13 Cell type proportion estimation

Whole blood samples are often used in EWAS, however, whole blood has mixed cell types and methylation level can be widely different between different cell types. Function `estimateCellProp` can be used to estimate cell types based on reference dataset. Users should select a reference that is most resemble to their own study samples. Currently there are 6 publicly deposited reference datasets: "FlowSorted.Blood.450k", "FlowSorted.DLPFC.450k","FlowSorted.CordBlood.450k", "FlowSorted.CordBloodCombined.450k", "FlowSorted.CordBloodNorway.450k" or "FlowSorted.Blood.EPIC".

```
> require(minfidata)
> path <- file.path(find.package("minfiData"),"extdata")
> #based on rgDataset
> rgSet <- readidat(path = path,recursive = TRUE)
>        celltype=estimateCellProp(userdata=rgSet,
+          refdata="FlowSorted.Blood.450k",
+          nonnegative = TRUE,normalize=TRUE)
```

The input data can be in one of the follow format: `rgDataSet`, `methDataSet`, `MethylSet`, `RGChannelSet` or methylation beta value matrix.

14 Methylation age estimation

Methylation age can reflect a person's biological age, which may be more related to the person's health status than chronological age. Three different types of methylation age can be estimated using `methyAge`: Horvath, Hannum and PhenoAge.

```
> require(minfidata)
> path <- file.path(find.package("minfiData"),"extdata")
> #based on rgDataset
> rgSet <- readidat(path = path,recursive = TRUE)
> meth=getmeth(rgSet)
> beta=getB(meth)
> mage=methyAge(beta)
```

15 Differentially methylated regions (DMRs)

EWAS is usually done probe by probe. However, in the human genome, nearby CpGs often have similar methylation status, so combined statistics from nearby CpGs can potentially increase association signal. Functions `ipdmr` and `combp` can utilize P values from EWAS to identify differentially methylated regions (DMRs). The argument "seed" in these two functions indicate the FDR threshold for initial selection of DMR regions, but we note that same value of seed do not appear to have the same effect in these two functions. It is often more stringent in ipdmr than combp.

There is no true DMR in the following example data, we set less strengent seed threshold is only for demostration purpose to show what the output looks like when DMRs were detected.

```
> dat=simubed()
> names(dat)
> ipdmr(data=dat,seed=0.1)
> combp(data=dat,seed=0.1)
```

16 Intraclass correlation coefficient (ICC) reliability measures

Several studies have showen that a large percentage of CpGs on Illumina arrays have poor reliability, i.e. they have low correlation between replicate measures in a same set of samples. Measurement of the reliability of individual CpGs can be assessed by calculating intraclass correlation coefficients (ICC) using methylation from duplicate samples. Function `dupicc` can be used for this purpose.

```
> require(minfiData)
> path <- file.path(find.package("minfiData"),"extdata")
> rgSet <- readidat(path = path,recursive = TRUE)
> mdat=getmeth(rgSet)
> beta=getB(mdat,"Illumina")
> #assuming list in id1 are corresponding duplicates of id2
> dupidx=data.frame(id1=c("5723646052_R02C02","5723646052_R04C01",
+                         "5723646052_R05C02"),
+                   id2=c("5723646053_R04C02","5723646053_R05C02",
+                         "5723646053_R06C02"))
> iccresu<-dupicc(dat=beta,dupid=dupidx)
```

17 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC)

Studies showed that 5-Methylcytosine and 5-hydroxymethylcytosine can have very different function. However, traditional bisulfite (BS) treatment commonly used in genome-wide methylation studies can not distinguish 5mC from 5hmC. Sequencing or array data from paired DNA samples (bisulfite and oxidative bisulfite conversion) are needed for the separate estimation. Bisulfite converted DNA can be used to estimate the proportion of 5mC+5hmC and oxidative bisulfite (oxBS) converted DNA can be used to estimate the proportion of 5mC. Function `oxBS.MLE` can be used to generate the maximum Likelihood Estimate (MLE) of 5mC and 5hmC using sequencing or array data from paired experiments (Xu et al. Bioinformatics 2016).

```
> load(system.file("oxBS.MLE.RData",package="ENmix"))
> resu<-oxBS.MLE(beta.BS,beta.oxBS,N.BS,N.oxBS)
> dim(resu[["5mC"]])
> dim(resu[["5hmC"]])
```

18 Compatibility with other related R packages

Most functions in the ENmix support the data structure provided by minfi package. The same data structures were also used by several other R packages, such as ChAMP and wateRmelon, so the output from ENmix functions can be easily utilized in these packages for further analysis. Here are some examples:

Example 1: mixed use of minfi and ENmix functions

```
> library(ENmix)
> #minfi functions to read in data
> sheet <- read.metharray.sheet(file.path(find.package("minfiData"),
+  "extdata"), pattern = "csv$")
> rgSet <- read.metharray.exp(targets = sheet, extended = TRUE)
> #ENmix function for control plot
> plotCtrl(rgSet)
> #minfi functions to extract methylation and annotation data
> mraw <- preprocessRaw(rgSet)
> beta<-getB(mraw, "Illumina")
> anno=getAnnotation(rgSet)
> #ENmix function for fast and accurate distribution plot
> multifreqpoly(beta,main="Data distribution")
> multifreqpoly(beta[anno$Type=="I",],main="Data distribution, type I")
> multifreqpoly(beta[anno$Type=="II",],main="Data distribution, type II")
> #ENmix background correction
> mset<-preprocessENmix(rgSet, bgParaEst="oob", dyeCorr="RELIC", nCores=6)
```

```
> #minfi functions for further preprocessing and analysis
> gmSet <- preprocessQuantile(mset)
> bumps <- bumphunter(gmSet, design = model.matrix(~ gmSet$status), B = 0,
+ type = "Beta", cutoff = 0.25)
```

19 SessionInfo

- R version 4.0.5 (2021-03-31), `x86_64-w64-mingw32`
- Locale: `LC_COLLATE=C`, `LC_CTYPE=English_United States.1252`, `LC_MONETARY=English_United States.1252`, `LC_NUMERIC=C`, `LC_TIME=English_United States.1252`
- Running under: `Windows Server 2012 R2 x64 (build 9600)`
- Matrix products: default
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Loaded via a namespace (and not attached): compiler 4.0.5, tools 4.0.5

20 References

Zongli Xu, Liang Niu, Leping Li and Jack A. Taylor, ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip. Nucleic Acids Research, 2015

Liang Niu, Zongli Xu and Jack A. Taylor, RCP: a novel probe design bias correction method for Illumina Methylation BeadChip. Bioinformatics 2016

Zongli Xu, Jack A. Taylor, Yuet-Kin Leung, Shuk-Mei Ho and Liang Niu, oxBS-MLE: An efficient method to estimate 5-methylcytosine and 5-hydroxymethylcytosine in paired bisulfite and oxidative bisulfite treated DNA. Bioinformatics 2016

Zongli Xu, Sabine A. S. Langie, Patrick De Boever, Jack A. Taylor1 and Liang Niu, RELIC: a novel dye-bias correction method for Illumina Methylation BeadChip. BMC Genomics 2017

Zongli Xu, Jack A Taylor. Reliability of DNA methylation measures using Illumina methylation BeadChip. Epigenetics 2020

Zongli Xu, Changchun Xie, Jack A. Taylor, Liang Niu, ipDMR: Identification of differentially methyl-ated regions with interval p-values. Bioinfomatics 2020

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.

Horvath S. DNA methylation age of human tissues and cell types. Genome biology 2013 14:R115.

Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. Molecular cell 2013 49:359-367.

Levine ME, Lu AT, Quach A, Chen BH, Assimes TL, Bandinelli S, et al. An epigenetic biomarker of aging for lifespan and healthspan. Aging (Albany NY) 2018 10:573-591.

Pedersen BS1, Schwartz DA, Yang IV, Kechris KJ. Comb-p: software for combining, analyzing, grouping and correcting spatially correlated P-values. Bioinfomatics 2012

EA Houseman, WP Accomando, DC Koestler, BC Christensen, CJ Marsit, HH Nelson, JK Wiencke and KT Kelsey. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC bioinformatics (2012) 13:86.

Heiss JA, Allan C Just. Improved filtering of DNA methylation microarray data by detection p values and its impact on downstream analyses. Clinical Epigenetics, 2019

Xu Z, Niu L, Taylor JA. ENmix: a comprehensive R package for the analysis of Illumina DNA methylation arrays. in review