A vignentte of Illumina Methylation Analyzer (IMA)

Dan Wang, Li Yan, Qiang Hu, Dominic J Smiraglia, Song liu
dan.wang@roswellpark.org

September 5, 2011

Contents

1 Overview

The recently released Infinium HumanMethylation450 BeadChip provides an unprecedented opportunity to quantify the methylation level of over 450,000 CpG sites within human genome. IMA (Illumina Methylation Analyzer) is a package designed to automate the pipeline for analyzing site-level and region-level methylation changes in epigenetic studies utilizing the 450K DNA methylation microarray. The pipeline loads the data from Illumina platform and provides user-customized functions commonly required to perform differential methylation analysis for individual sites as well as annotated regions. The user can either run the pipeline with default option setting or specify alternative routes in the function arguments. In this vignette, we briefly described the functions implemented in IMA.

Preprocessing: IMA takes as input the beta values representing the methylation levels of individual site reported by Illumina BeadStudio or GenomeStudio software. It allows user to choose several filtering step or modify filtering criteria for specific quality control purpose. By default, IMA will filter out loci with missing beta value, from the X chromosome or with median detection P-value greater than 0.05. The option for sample level quality control is also provided. Although the raw beta values will be analyzed as recommended by Illumina, the user can choose Arcsine square root transformation. Quantile normalization options is available for cross sample normalization.

Methylation Index Calculation: The 450k BeadChip provides broad coverage throughout gene regions including 1500 bp or 200 bp upstream of transcription start site, 5'UTR, 1st exon, gene body and 3'UTR, as well as CpG islands and surrounding shelves and shores for a comprehensive view of methylation level. For each specific region (e.g., 1st exon) of a gene, IMA will collect all the the probed loci within it and derive an index of overall region methylation value. Currently, there are three different index metrics implemented in

Sample name	group
DSRP651	g1
DSRP652	g1
DSRP653	g1
DSRP655	g2
DSRP656	g2
DSRP684	g2

Table 1: **Phenotype data:** The first column lists the sample names, and the second column lists the corresponding phenotype.

IMA: mean, median, and Tukey's Biweight robust average. By default, the median beta values will be used as the region's methylation index for further analysis.

Differential Methylation Analysis: For each specific region, Wilcoxon rank-sum test (default), Student's t-test and empirical Bayes statistics (G. K. Smyth, 2004) are available for inference in differential testing. Robust linear regression is available as an option to infer methylation change associated with continuous variable (e.g., age). A variety of multiple testing correction algorithms is available, including conservative Bonferroni correction and more liberal false discovery rate control. Users can specify the significance criteria in parameter file. The same statistical inference and multiple test correction procedures described above will also be applied to each single site to obtain site-level differential methylation inference.

Output: Detailed output files are provided for each of the three modules above. For preprocessing module, the output contains a matrix of methylation value for qualified loci across qualified samples. For methylation index calculation module, there is a matrix of methylation index across the samples for each region category of interest (e.g., promoter). For differential methylation analysis module, the differential methylation values (e.g., Delta-Beta) together with both raw and adjusted P-values of each region (or site) of interest will be provided.

2 Region definition

Compared with previously released Illumina DNA methylation platforms, the recently launched Infinium HumanMethylation450 BeadChip 450K microarray represents a significant increase in the CpG site density for quantifying methylation events. At the gene level, the 450K microarray covers 99% of RefSeq genes with multiple sites in the annotated promoter (1500 bp or 200 bp upstream of transcription start site), 5'UTR, 1st exon, gene body and 3'UTR. From the CpG context, it covers 96% of CpG islands with multiple sites in the annotated CpG Island, north or south shores (regions flanking island), and north or south shelves (regions flanking shores). The package makes use of Illumina methylation annotation for region definition. We thus adopted the 11 categories of region annotation described above.

3 An example of Workflow

3.1 Loading Data

The input for the package consists basically of two files containing beta-value methylation data (including annotation) produced by BeadStudio or GenomeStudio, and sample phenotype data prepared by the user. One example of sample phenotype file is shown in Table 1. `IMA.methy450R` load the input files with a single command described below and a `exprmethy450` object will be created, whcih includes the following features: β value matrix, locus annotion, detection Pvalue, sample phenotype information. Besides, baisc quality control information will be outputed in the QC.pdf, which include unsupervised sample clustering using all loci, boxplot for beta value of each sample, and barplot showing the percent of loci with detection Pvalue smaller than 1e-5 in each sample(Figure 1).

```
>MethyFileName = "SampleMethFinalReport.txt" ###File produced by the GenomeStudio
>PhenoFileName = "SamplePhenotype.txt"
>data =IMA.methy450R(file = MethyFileName,columnGrepPattern=list(beta=".AVG_Beta",
detectp=".Detection.Pval"),groupfile = PhenoFileName)
```

```
dimension of the input methylation data 485577 66
Slot names of x.methy450: bmatrix annot detectP groupinfo
```

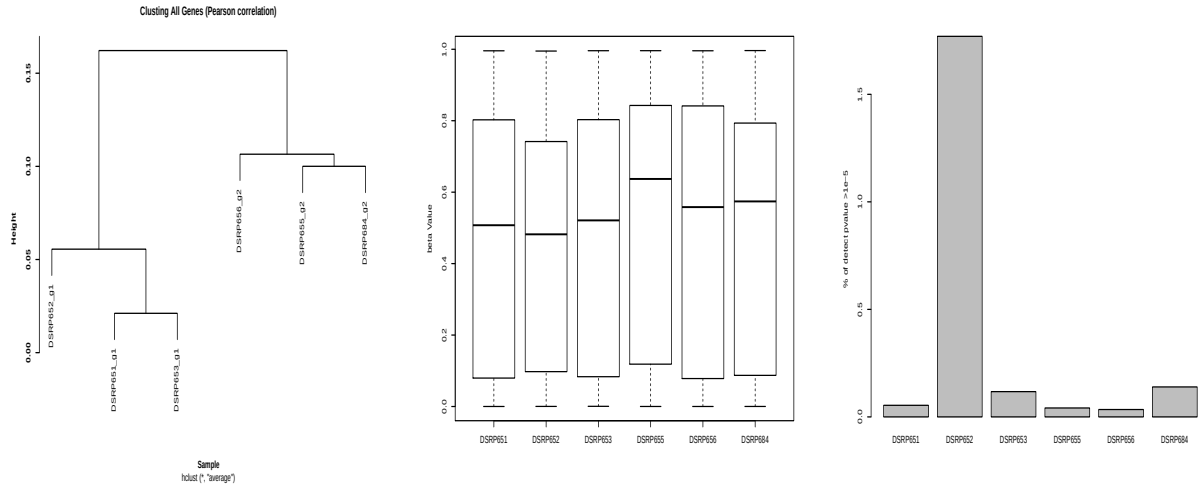

Figure 1: For left to right: **Clustering dendrogram**, Unsupervised sample clustering using all the loci; **Box plot**, the distribution of beta value for each sample; **Bar plot**, the percent of loci with detection pvalues smaller than 1e-5 in each sample.

3.2 Data preprocessing

Data preprocessing is carried out using the `IMA.methy450PP` function which includes the following arguments: *filterdetectP*, whether or not to filter probes based on detetion Pvalue; *Xchrom*, whether or not to remove the loci from the X chromosome; *transfm*, whether or not to transfer the raw beta vlaue using square arcsine; *normalization*, whether or not to perform quantile normalization; *na.omit*, whether or not to remove the loci containing missing beta values. The user can choose the preprocessing routes and corresponding cutoffs in the argument of this function. The output of this function includes beta value matrix for each of the 11 region categories.

```
> dataf = IMA.methy450PP(data,na.omit = TRUE,normalization=FALSE,transfm = FALSE,filtersample = TRUE,
filterdetectP = TRUE, detectPcut = 0.75,locidiff = FALSE,locidiffcut=0.01, Xchrom = TRUE)
```

```
Total samples: DSRP651 DSRP652 DSRP653 DSRP655 DSRP656 DSRP684
 Kept samples: DSRP651 DSRP652 DSRP653 DSRP655 DSRP656 DSRP684
5482 sites containing missing value and be removed
11093 sites on chrX
Kept loci 468982 from orignial  485577
.......
```

```
 split the annotation file to 14 annotated region list
 .......
TSS1500 region: 19583
TSS200 region: 16995
5'UTR region: 13497
1st Exon region: 14901
Gene body region: 18327
3'UTR region: 12548
PROMOTER region: 20115
Gene region: 20357
Island region: 25913
N_Shore region 24230
S_Shore region 21732
N_Shelf region 17744
S_Shelf region 16695
UCSC CPG island region 26398
Slot names of methy450batch:
bmatrix annot detectP groupinfo
TSS1500Ind TSS200Ind UTR5Ind EXON1Ind GENEBODYInd UTR3Ind PROMOTERInd GENEInd
ISLANDInd NSHOREInd SSHOREInd NSHELFInd SSHELFInd UCSCInd
```

3.3 Differential Statistical Analysis

For each annotated region, limma, Wilcoxon rank-sum test (default) and Student's t-test are available for differential methylation inference in case-control study. Robust linear regression is available to infer region-level methylation change associated with continuous variable (e.g., age). A variety of multiple testing correction algorithms, as implemented the *stats* library of R, is available, including conservative Bonferroni correction and more liberal false discovery control. Users can specify the significance criteria in the function arguments. The same statistical inference and multiple test correction procedures will also be applied to each single site to obtain site-level differential methylation analysis result.

3.3.1 Site level analysis

The `sitetest` function provide single site level testing by using either limma, two sample t-tests(either pooled or satterthwaite), wilcox rank-sum test, or robust linear regression on each site. By the default, the output includes a table containing full testing results for all loci and the their annottion information. The user can choose to output the signifiant loci by specifying the significance criteria in the arguments.

```
>sitetestALL = sitetest(dataf,gcase="g2",gcontrol="g1",test ="limma" ,Padj="BH",
outputDES = FALSE,rawpcut = NULL,adjustpcut =NULL,betadiffcut = NULL)
>sitetest = outputDESfunc(sitetestALL,outputDES = TRUE,rawpcut = 0.05,adjustpcut =0.05,betadiffcut = 0.1
>sitetest[1:10,]
                P-Value Adjust Pval beta-Difference
cg00000165 1.409300e-06 0.001409300       0.5462600
cg00000292 1.523588e-04 0.008406543       0.2498033
cg00000321 3.845818e-03 0.037338042       0.2155633
cg00001583 5.308322e-04 0.011539830       0.5278767
cg00001747 2.427103e-05 0.003033879       0.6307633
cg00001809 3.907709e-03 0.037574123      -0.3278700
cg00002033 4.810922e-03 0.041834107       0.3947300
cg00002719 3.221156e-06 0.001610578       0.7160300
cg00003287 5.826247e-03 0.047756125       0.3177867
```

```
cg00003298 2.586462e-04 0.009052337       0.6129567
```

3.3.2 Region level analysis

For each specific region, IMA will collect all the targeted loci within it and derive an index of overall region methylation value. Currently, there are three different index metrics implemented in IMA: mean, median, and Tukey's Biweight robust average. By default, the median β values will be used as the regionâĂŹs methylation index for further analysis.
For example, to study the overall methylation change on the 1st exon region of each gene, we first use the `indexregionfun` function to obtain the beta value index for the 1st exon of each gene, then the `testfunc` function could be used to obtain the statistical testing result. Further, the `outputDESfunc` could be used to output the differential methylated genes/cpG sites by user defined cut off.

```
> beta = dataf@bmatrix;
> betar = indexregionfun(indexlist=dataf@TSS1500Ind,beta=beta,sumregion="median")
> group = dataf@groupinfo
> grouplev = group$group[match(colnames(beta),group$samplename)]
> print(grouplev)
[1] g1 g1 g1 g2 g2 g2
Levels: g1 g2
>TSS1500test = testfunc(eset = betar,testmethod="limma",Padj="BH",concov=OFF,
grouplev = grouplev,gcase = "g2",gcontrol="g1")
>TSS1500DES = outputDESfunc(TSS1500test,outputDES = TRUE,rawpcut = 0.05,adjustpcut =0.05,
betadiffcut = 0.14)
```

Alternative, the user could use `regionswrapper` function to output all the statistical testing results for the 11 region categories separately in an excel file.

```
> regionswrapper(dataf,sumregion=c("mean","median","tbrm"),gcase = "g2",gcontrol="g1",
testmethod = c("wilcox","limma","pooled","satterthwaite"),Padj="BH",concov = c("OFF","ON"),
list11excel= list11excel,list11Rdata = list11Rdata,outputDES = FALSE,rawpcut = NULL,
adjustpcut=NULL,betadiffcut = NULL)
```

To compare the difference between the site level and region level analysis, one can classify each region of interest based on its site-level and region-level differential testing results. "Region only" means the region is significant but none of its site is significant, while "Site only" means the region is not significant but at least one of its sites is significant. âĂIJBothâĂİ means the region is significant and it contains at least one site-level significance event. For the example data used here, there are a number of region-level differential methylation events in CpG island and south shore categories. For other regions, the differential methylation events are mostly site-level (Figure 2).

3.4 Analysis of the methylation on user-selected CpG sites or Genes

In some cases, users may want to check the methylation on several genes/CpG sites. Below give an example to show how to detect the methylation on the selected genes: "BRCA1", "MLH1", "CCNE1", "PTEN", "PALB2". In the example,the `fullannotInd` data is the full region level annotation list without any filtering. Users also chose use the annotation list with filtering returned by the `IMA.methy450PP`.

```
>data("fullannotInd")
>indlists = c("BRCA1", "MLH1", "CCNE1", "PTEN", "PALB2")
>annot = fullannot[[match("TSS1500Ind",names(fullannot))]]
```

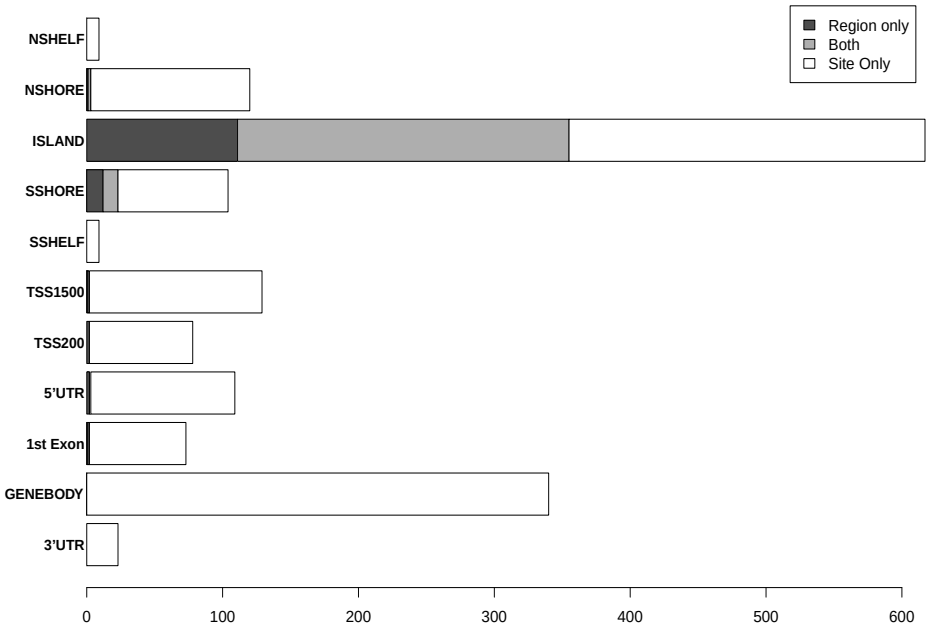

Figure 2: The counts of region belonging to each of the three classes described above. Both CpG island and South Shore contains a number of region-level differential methylation events.

```
>indexlist = annot[match(indlists,names(annot))]
>eset = sumregionfun(indexlist,data@bmatrix,"mean");
>testfunc(eset,concov = "OFF",testmethod="limma",Padj="BH",gcase ="g2",gcontrol="g1",
grouplev=grouplev)
        P-Value Adjust Pval beta-Difference
BRCA1 0.85022427   0.8502243    -0.007932301
MLH1  0.24266752   0.5660762    -0.060689000
CCNE1 0.02154944   0.1077472    -0.012437037
PTEN  0.33964569   0.5660762     0.002431453
PALB2 0.51089784   0.6386223     0.002415556
```

If users interested to see the methylation on all sites targeted one specific gene on some region, e.g. ”BRCA1” on the 1st exon, the following code can be used:

```
>data("fullannotInd")
>indlists = c("BRCA1")
>annot = fullannot[[match("EXON1Ind",names(fullannot))]]
>indexlist = annot[match(indlists,names(annot))]
> testfunc(eset=data@bmatrix[unlist(indexlist),],concov = "OFF",testmethod="limma",Padj="BH",gcase ="g2"
              P-Value Adjust Pval beta-Difference
cg04110421 0.20042839   0.6630926    0.0127166667
cg04658354 0.38317767   0.7387675    0.0042366667
cg08993267 0.68709473   0.9687583    0.0023400000
cg09441966 0.16430892   0.6630926    0.0105566667
cg13782816 0.83085193   0.9687583    0.0222200000
```

```
cg15419295 0.96875826   0.9687583    0.0004200000
cg16630982 0.90948218   0.9687583   -0.0005466667
cg16963062 0.27628859   0.6630926    0.0089166667
cg17301289 0.23485572   0.6630926    0.0096300000
cg20187250 0.43094770   0.7387675    0.0033666667
cg21253966 0.06624047   0.6630926    0.0102700000
cg24806953 0.81047709   0.9687583   -0.0013933333
```

4 Conclusion

We have introduced an R pipeline, IMA, which automates the tasks commonly required for the differential analysis of epigenetic data sets utilizing the 450K DNA methylation microarray. The package makes use of Illumina methylation annotation for region definition, as well as several Bioconductor packages for various preprocessing and differential testing steps (Gentleman et al., 2004). Written in open source R environment, it provides the flexibility for users to adopt, extend and customize the functionality for their specific needs. It can be used as an automatic pipeline to analyze specific regions as well as specific sites for downstream functional exploration and hypothesis generation. For example, the matrix of methylation index of shore regions produced by IMA can be used as the input for model-based clustering (Houseman et al., 2008) to identify clustered shores associated with the phenotype of interest.