A vignette of Illumina Methylation Analyzer (IMA)

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

February 23, 2012

Contents

1 Overview **1**

2 Region definition **2**

3 An example of workflow **2**
3.1 Loading data . . . 2
3.2 Data preprocessing . . . 3
3.3 Differential statistical analysis . . . 4
3.3.1 Site level analysis . . . 5
3.3.2 Region level analysis . . . 5
3.4 Annotation . . . 6
3.5 Analysis of the methylation on user-selected CpG sites or regions . . . 7
3.6 Adjusting confounding factors . . . 8

4 Conclusion **8**

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 exploratory analysis and summarization of 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 exploratory differential methylation analysis and summarization for individual sites as well as annotated regions. The user can either run the pipeline with default option settings or specify alternative routes in the function arguments. Note that instead of providing recommendations about which specific analysis method should be used, the main purpose of developing the IMA package is to provide a range of commonly used Infinium methylation microarray analysis options for users to choose for their exploratory analysis and summarization in an automatic way. Therefore, it is the best interest for the users to consult experienced bioinformatician/statistician about which specific analysis option/route should be chosen for their 450k microarray data. 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 sites reported by Illumina BeadStudio or GenomeStudio software. It allows user to choose several filtering steps or modify filtering criteria for specific quality control purposes. 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. As probe containing SNP(s) at/near the targeted CpG site might not be sufficient to measure DNA methylation level (but rather genomic variation), users can choose to filter out the loci whose methylation level are measured by probes containing SNP(s) at/near the targeted CpG site. The option for sample-level quality control is also provided [2]. Although the raw beta values will be analyzed

as recommended by Illumina, users can choose Arcsine square root transformation when modeling the methylation level as the response in a linear model [7][8]. Note that Arcsine transformation might not be sufficient for the use of empirical Bayes statistics. Logit transformation, as proposed by Kuan *et al.* [6], is also available as an option. The default setting of IMA is that no normalization will be performed, and quantile normalization is available as an alternative preprocessing option. It has been shown that quantile normalization is not sufficient for removing all unwanted technical variation across samples [10]. The development of normalization strategy for DNA methylation study is an active area of ongoing research [1].

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. The promoter, 5′ UTR, 1st exon, gene body and 3′ UTR are gene-based regions. The CpG island and its surrounding shore and shelve regions are not necessary gene-based, depending on their distance to the nearest genes. For each specific region (e.g., 1st exon), IMA will collect the 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 beta value 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 [9] are available for inference in differential testing. General linear models are available as an option to infer methylation change associated with continuous covariate (e.g., age), as well as to adjust confounding factors (e.g., batch). A variety of multiple testing correction algorithms are available, including stringent Bonferroni correction and widely used false discovery rate control. Users can specify the significance criteria in the parameter file. The same statistical inference and multiple test correction procedures described above can 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 the preprocessing module, the output contains a matrix of methylation value for qualified loci across qualified samples. For the methylation index calculation module, there is a matrix of methylation index across the samples for each region category of interest (e.g., South Shore). For the 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 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, with the number of regions for each category listed in Table 2.

3 An example of workflow

3.1 Loading data

The input information for the package consists basically of two files containing beta-value methylation data (including annotation) produced by BeadStudio or GenomeStudio software, and sample phenotype data prepared by the user. Exemplary sample phenotype files are shown in Table 1. `IMA.methy450R` loads the input files with a single command described below and an `exprmethy450` object will be created, which includes the following features: β value matrix, locus annotation, detection P-value and sample phenotype information. Besides, basic quality control information will be outputted in the QC.pdf, which include unsupervised sample clustering dendrogram using all the CpG loci, boxplot for beta value distribution of each sample, and barplot showing the percent of loci with detection P-value larger than $1e-5$ in each sample (Figure 1).

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

Sample name	age
DSRP651	60
DSRP652	50
DSRP653	55
DSRP655	72
DSRP656	60
DSRP684	43

Sample name	group	pair
DSRP651	g1	1
DSRP652	g1	2
DSRP653	g1	3
DSRP655	g2	1
DSRP656	g2	2
DSRP684	g2	3

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

```
>MethyFileName = "SampleMethFinalReport.txt" ###Data file produced by the GenomeStudio.
>PhenoFileName = "SamplePhenotype.txt" ###Phenotype file as shown in Table 1.
>data =IMA.methy450R(fileName = MethyFileName,columnGrepPattern=list(beta=".AVG_Beta",
detectp=".Detection.Pval"),groupfile = PhenoFileName)

For a desktop with 2GB memory and 7200RPM hard disk,
the estimated time of this process is 1-2 mins for a data with 10 samples and 6-7 mins
for a data with 200 samples
......Extracting the beta matrix.....
......Extracting the pvalue matrix......
......Extracting the annotation matrix......
Read phenotype data...
Matching the orders of samples between phenotype data and beta value matrix.
Total CpG sites without any filtering are: 485577
Total samples are: 6
....Starting Quality Control...
A exprmethy450 class is created and the slotNames are:
 bmatrix annot detectP groupinfo
Basic Quality Control information can be found in QC.pdf file
```

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 detection P-value; *Xchrom*, whether or not to remove the loci from the X,Y chromosome or both; *peakcorrection*,whether or not to perform peak correction[3];*transfm*, whether or not to transfer the raw beta value using arcsine square root or logit; *normalization*, whether or not to perform quantile normalization; *na.omit*, whether or not to remove the loci containing missing beta values; *snpfilter*, whether or not to filter out loci whose methylation level are measured by probes containing SNP(s) at/near the targeted CpG site. The user can choose the preprocessing routes and corresponding cutoffs in the argument of this function. The output of this function includes the beta value matrix for each of the 11 region categories, as well as the beta value matrix for all qualified loci. For each specific region(e.g. promoter), IMA will collect the loci within it and return the corresponding loci ID and position ID.

```
>dataf = IMA.methy450PP(data,na.omit = TRUE,peakcorrection = FALSE,normalization=FALSE,transfm = FALSE,
samplefilterdetectP = 1e-5,samplefilterperc = 0.75,sitefilterdetectP = 0.05,
sitefilterperc = 0.75,locidiff = FALSE,locidiffgroup = list(c("g1","g3"),"g2"), XYchrom = c(FALSE,"X","Y

0 samples removed with at least 75 percentage sites having pvalue greater than 1e-05
5482 sites contain missing value and are removed
11093 sites on chrX are removed
```

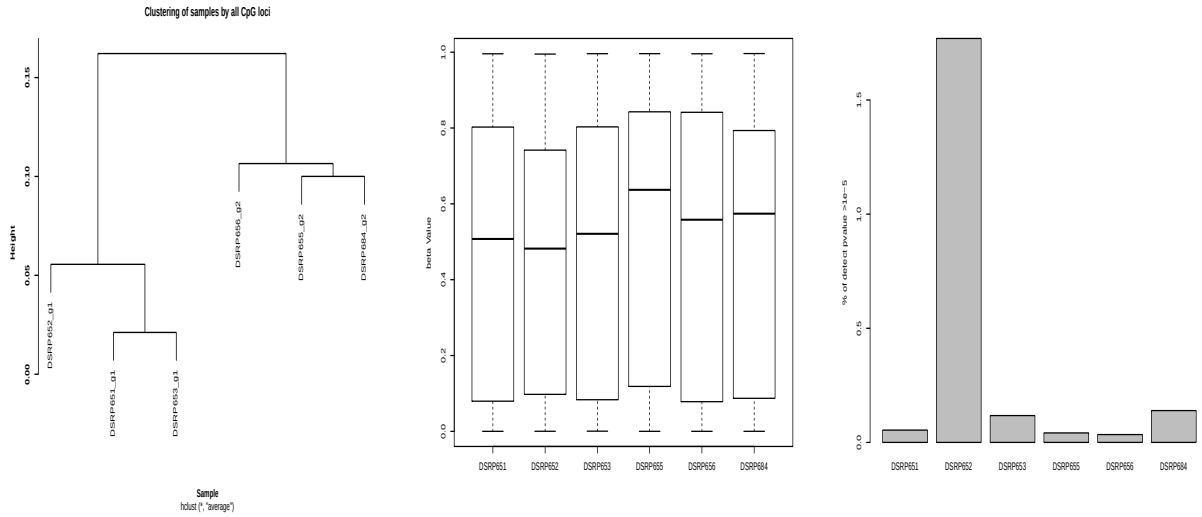

Figure 1: For left to right: **Clustering dendrogram**, Unsupervised sample clustering using all the loci; **Box plot**, the distribution of beta value for each sample; **Bar plot**, the percent of loci with detection P-values greater than 1e-5 in each sample.

```
207 sites had at least 75 % samples with pvalue great than 0.05 and are removed
468804 sites were retained from the original 485577 sites
.......
Split the annotation file to 11 annotated region categories
.......

TSS1500 region contains: 19580 UCSC REFGENE regions
TSS200 region contains: 16993 UCSC REFGENE regions
5'UTR region contains: 13496 UCSC REFGENE regions
1st Exon region contains: 14901 UCSC REFGENE regions
Gene body region contains: 18326 UCSC REFGENE regions
3'UTR region contains: 12540 UCSC REFGENE regions
Island region contains: 25913 UCSC CPG ISLAND regions
N_Shore region contains 24227 UCSC CPG ISLAND regions
S_Shore region contains 21731 UCSC CPG ISLAND regions
N_Shelf region contains 17734 UCSC CPG ISLAND regions
S_Shelf region contains 16685 UCSC CPG ISLAND regions

A methy450batch class is returned and the slotNames are:
 bmatrix annot detectP groupinfo TSS1500Ind TSS200Ind UTR5Ind EXON1Ind GENEBODYInd UTR3Ind
ISLANDInd NSHOREInd SSHOREInd NSHELFInd SSHELFInd
```

3.3 Differential statistical analysis

For each annotated region, Wilcoxon rank-sum test (default), Student′s t-test and empirical Bayes statistics are available for exploratory differential methylation inference in a case-control study. Both unpaired and paired test statistics are available. General linear models are available to infer region-level methylation change associated with continuous variable (e.g., age), as well as to adjust confounding factors (e.g., batch). A variety of multiple testing correction algorithms implemented in the *stats* library of R are available, including conservative Bonferroni correction and more liberal false discovery control. It is the best interest for the users to consult experienced

bioinformatician/statistician about which differential test statistics and multiple test correction option should be chosen for their 450k microarray data. Users can specify the significance criteria in the function arguments. The statistical inference and multiple test correction procedures described above will also be available for users to obtain site-level exploratory differential methylation analysis result.

3.3.1 Site level analysis

The `sitetest` function provides site-level testing by using either Wilcoxon rank-sum test (or Wilcoxon signed-rank test for paired design), two sample t-tests (either pooled or satterthwaite), limma, or general linear regression on each site. By default, the output includes a table containing the full testing results for all loci and the their annotation information. The user can choose to output only the significant loci by specifying the significance criteria in the arguments.

```
>sitetestALL = sitetest(dataf,gcase="g2",,gcontrol=c("g1","g3"),test ="limma" ,Padj="BH",
rawpcut = NULL,adjustpcut =NULL,betadiffcut = NULL,paired = FALSE)

Performing pooled t.test...
Keep the full comparison result.Please specify the significance criteria if you are only
interested in the differentially methylated regions/sites

>sitetest = outputDMfunc(sitetestALL,rawpcut = 0.05,adjustpcut =0.05,betadiffcut = 0.14)
>sitetest[1:10,]

                P-Value Adjust Pval Beta-Difference
cg00000165 1.417849e-06 0.001459283       0.5462600
cg00000292 1.530266e-04 0.006278067       0.2498033
cg00000321 3.844155e-03 0.038905114       0.2155633
cg00001583 5.290704e-04 0.011814843       0.5278767
cg00001747 2.419241e-05 0.002963918       0.6307633
cg00001809 3.899693e-03 0.039270346      -0.3278700
cg00002033 4.799448e-03 0.045032295       0.3947300
cg00002719 3.215022e-06 0.001712100       0.7160300
cg00003298 2.576746e-04 0.008097525       0.6129567
cg00003305 1.772885e-04 0.006719432       0.5487733
```

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-level methylation value. Table 2 summarizes the quantity of the 11 region categories without any filtering.

Currently, there are three different index metrics implemented in IMA: mean, median, and Turkey's Bi-weight robust average. By default, the median β value will be used as the region's methylation index for further analysis.

For example, to study the overall methylation change on the 1st exon 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. Furthermore, the `outputDMfunc` could be used to output the differential methylated genes/CpG sites satisfying user-defined significance cut off.

```
>beta = dataf@bmatrix;
>betar = indexregionfunc(indexlist=dataf@TSS1500Ind,beta=beta,indexmethod="median")
>TSS1500testALL = testfunc(eset = betar,testmethod="limma",Padj="BH",concov="OFF",groupinfo
 = dataf2@groupinfo,gcase ="g2",gcontrol=c("g1","g3"),paired = FALSE)
>TSS1500test = outputDMfunc(TSS1500testALL,rawpcut=0.05,adjustpcut=0.05,betadiffcut=0.14)
>TSS1500test[10:20,]
```

Region Category	# of regions
TSS1500	20406 UCSC REFGENE regions
TSS200	17731 UCSC REFGENE regions
5′ UTR	14148 UCSC REFGENE regions
1st EXON	15588 UCSC REFGENE regions
GENEBODY	19076 UCSC REFGENE regions
3′ UTR	13074 UCSC REFGENE regions
ISLAND	26662 UCSC CPG regions
NSHORE	24991 UCSC CPG regions
SSHORE	22444 UCSC CPG regions
NSHELF	18417 UCSC CPG regions
SSHELF	17337 UCSC CPG regions

Table 2: **The number of regions for each category** The first column lists the name of each region category, and the second column lists the corresponding quantities.

```
            P-Value Adjust Pval beta-Difference
ABCC11 3.286721e-04 0.010682861       0.2482067
ABCC3  2.044557e-04 0.008352806      -0.4950500
ABCC8  7.088950e-04 0.015948965      -0.3349467
ABCG4  3.461875e-03 0.039377382      -0.1637417
ABHD4  3.587270e-03 0.040242901       0.2009967
ACAN   4.949051e-03 0.049095931      -0.3909900
ACCSL  4.460786e-03 0.046089517       0.1993800
ACE    4.525757e-03 0.046431668      -0.2961967
ACER3  8.162105e-06 0.002978667      -0.4017950
ACOT12 1.067272e-03 0.020140708      -0.4056400
ACOX2  1.466321e-03 0.024425909      -0.3870400
```

Alternative, the user could use `regionswrapper` function to perform the statistical testing for all the 11 region categories. The results for each region category will be outputted to a separate sheet of an excel file.

```
regionswrapper(dataf,indexmethod ="mean",gcase = "g2",gcontrol=c("g1","g3"),testmethod = "limma",
Padj="BH",concov = "OFF",list11excel="list11result.xls",list11Rdata="list11result.Rdata",
rawpcut = NULL,adjustpcut = NULL,betadiffcut = NULL,paired = FALSE)
```

3.4 Annotation

This function provides annotation information for a list of site/region IDs of interest. The `sitetest`, `testfunc` and `regionswrapper` function will return a matrix including site/region IDs, testing pvalues and beta-value difference. It might take over 2GB space to save the annotation information for the all site/region level comparisons. The advantage of using the `annotfunc` function is that user could specify the list of site or region IDs and extract their corresponding annotation. For example, "# of org_Designed sites" tells the total number of sites designed for this region, "# of sites After Filtering" tells how many sites were left after filtering step, "Designed Probes" tells the site names designed for this region, and "CHR" tells the chromosome name where the target locus is located, *etc.*

```
>load(``fullannotInd.rda'')
>listtoannot = rownames(TSS1500test)[10:20]
>fullannotInd = fullannot
>fullIndexannot = TSS1500Ind
>filteredannot = dataf2@annot
```

```
>filteredIndexannot = dataf2@TSS1500Ind
>TSS1500sig = annotfunc(listtoannot,fullannot,filteredannot,fullIndexannot,filteredIndexannot,
category = "region")
>colnames(TSS1500sig)
```

```
 [1] "# of Ori_Designed sites"     "# of sites After Filtering"
 [3] "Designed Probes"             "NAME"
 [5] "ADDRESSA_ID"                 "ALLELEA_PROBESEQ"
 [7] "ADDRESSB_ID"                 "ALLELEB_PROBESEQ"
 [9] "INFINIUM_DESIGN_TYPE"        "NEXT_BASE"
[11] "COLOR_CHANNEL"               "FORWARD_SEQUENCE"
[13] "GENOME_BUILD"                "CHR"
[15] "MAPINFO"                     "SOURCESEQ"
[17] "CHROMOSOME_36"               "COORDINATE_36"
[19] "STRAND"                      "PROBE_SNPS"
[21] "PROBE_SNPS_10"               "RANDOM_LOCI"
[23] "METHYL27_LOCI"               "UCSC_REFGENE_NAME"
[25] "UCSC_REFGENE_ACCESSION"      "UCSC_REFGENE_GROUP"
[27] "UCSC_CPG_ISLANDS_NAME"       "RELATION_TO_UCSC_CPG_ISLAND"
[29] "PHANTOM"                     "DMR"
[31] "ENHANCER"                    "HMM_ISLAND"
[33] "REGULATORY_FEATURE_NAME"     "REGULATORY_FEATURE_GROUP"
[35] "DHS"                         "Index"
```

```
TSS1500sig[1:2,1:3]
```

```
      # of Ori_       # of sites       Designed Probes
      Designed sites  After Filtering
ABCC11 "3"                    "3"     "cg04388863/cg08404739/cg09147400"
ABCC3  "3"                    "3"     "cg05599550/cg23340875/cg27222669"
```

3.5 Analysis of the methylation on user-selected CpG sites or regions

In some cases, users may want to examine the methylation level changes only on selected regions/CpG sites of interest. Below is an example to analyze the methylation change of 1st exon on the selected genes: "BRCA1", "MLH1", "CCNE1", "PTEN", and "PALB2". In this example, the `fullannotInd` data is the full region-level annotation data without any filtering. Users can also choose to use the filtered region-level annotation data created by the `IMA.methy450PP` preprocessing function.

```
>load("fullannotInd")
>indlists = c("BRCA1", "MLH1", "CCNE1", "PTEN", "PALB2")
>annot = fullannot[[match("EXON1Ind",names(fullannot))]]
>indexlist = annot[match(indlists,names(annot))]
>eset = indexregionfunc(indexlist,data@bmatrix,"mean");
>testfunc(eset,concov = "OFF",testmethod="limma",Padj="BH",gcase ="g2",gcontrol="g1",
groupinfo=dataf@groupinfo)
        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 are interested to examine the methylation change for every single site within a given region of selected genes, e.g., all the probed sites within the 1st exon of "BRCA1", the following code can be used:

```
>load("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",gcontrol="g1",groupinfo=dataf@groupinfo)
```

	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

3.6 Adjusting confounding factors

In differential methylation analysis users may want to adjust the effect of confounding factors such as age, gender, batch effect, etc. One way of adjustment is to incorporate them as covariates in the general linear regression model. Below is an example to compare the methylation difference between the cases and controls, with the age covariate adjusted.

```
>beta = dataf@bmatrix;
>betaind <- indexregionfunc(indexlist=dataf2@TSS1500Ind,beta=beta,indexmethod="mean")
>age = c(65,60,45,40,55,46)
>betares <- apply(betaind,1,function(x){residuals(lm(unlist(x)~age))})##This function returns
 ## the residuals of the linear regression between methylation value and age
>TSS1500testALL = testfunc(eset = betares,testmethod="limma",Padj="BH",concov="OFF",groupinfo =
dataf2@groupinfo,gcase ="g1",gcontrol="g2",paired = TRUE)
TSS1500test = outputDMfunc(TSS1500testALL,rawpcut=0.05,adjustpcut=0.05,betadiffcut=0.14)
```

4 Conclusion

We have introduced an R pipeline, IMA, which automates the tasks commonly required for exploratory 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 [4]. The major differences between IMA and existing R packages for Infinium methylation analysis are that IMA pro-vides a pipeline which automates the tasks commonly required for the exploratory analysis and summarization of 450K DNA methy-lation data at both site-level and region-level.

The main purpose of developing the IMA package is to provide a range of commonly used analysis options for potential users to perform exploratory analysis and summarization of 450K microarray data in an automatic way. It is the best interest for the users to consult their bioinformatician/statistician about which specific analysis option should be chosen for their 450k microarray data. Written in the open source R environment, IMA 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 [5] to identify clustered shores associated with the phenotype of interest.

References

[1] M.J. Aryee, Z. Wu, C. Ladd-Acosta, B. Herb, A.P. Feinberg, S. Yegnasubramanian, and R.A. Irizarry. Accurate genome-scale percentage dna methylation estimates from microarray data. *Biostatistics*, 12(2):197, 2011.

[2] B.C. Christensen, A.A. Smith, S. Zheng, D.C. Koestler, E.A. Houseman, C.J. Marsit, J.L. Wiemels, H.H. Nelson, M.R. Karagas, M.R. Wrensch, et al. Dna methylation, isocitrate dehydrogenase mutation, and survival in glioma. *Journal of the National Cancer Institute*, 103(2):143, 2011.

[3] S. Dedeurwaerder, M. Defrance, E. Calonne, H. Denis, C. Sotiriou, and F. Fuks. Evaluation of the infinium methylation 450k technology. *Epigenomics*, 3(6):771–784, 2011.

[4] R.C. Gentleman, V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80, 2004.

[5] E.A. Houseman, B. Christensen, R.F. Yeh, C. Marsit, M. Karagas, M. Wrensch, H. Nelson, J. Wiemels, S. Zheng, J. Wiencke, et al. Model-based clustering of dna methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. *Bmc Bioinformatics*, 9(1):365, 2008.

[6] P.F. Kuan, S. Wang, X. Zhou, and H. Chu. A statistical framework for illumina dna methylation arrays. *Bioinformatics*, 26(22):2849, 2010.

[7] C.J. Marsit, D.C. Koestler, B.C. Christensen, M.R. Karagas, E.A. Houseman, and K.T. Kelsey. Dna methylation array analysis identifies profiles of blood-derived dna methylation associated with bladder cancer. *Journal of Clinical Oncology*, 29(9):1133–1139, 2011.

[8] D.M. Rocke. On the beta transformation family. *Technometrics*, pages 72–81, 1993.

[9] G.K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1):3, 2004.

[10] A.E. Teschendorff, U. Menon, A. Gentry-Maharaj, S.J. Ramus, S.A. Gayther, S. Apostolidou, A. Jones, M. Lechner, S. Beck, I.J. Jacobs, et al. An epigenetic signature in peripheral blood predicts active ovarian cancer. *PLoS One*, 4(12):e8274, 2009.