BioMethyl: An R package for Biological Interpretation of DNA Methylation Data

Yue Wang[1], Jennifer M. Franks[1,2], Michael L. Whitfield[1,2] and Chao Cheng[1,2,3,4*],

[1]Department of Molecular and Systems Biology, Geisel School of Medicine at Dartmouth, Hanover, New Hampshire 03755, USA, [2]Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, Lebanon, New Hampshire 03756, USA, [3] Norris Cotton Cancer Center, Lebanon, New Hampshire 03756, USA, [4]Department of Medicine, Baylor College of Medicine, Houston, TX 77030, USA.

*To whom correspondence should be addressed.

Abstract

Motivation: The accumulation of publicly available DNA methylation data sets has resulted in the need for tools to interpret the specific cellular phenotypes in bulk tissue data. Current approaches use either single differentially methylated CpG sites or differentially methylated regions that map to genes. However, these approaches may introduce biases in downstream analyses of biological interpretation, because of the variability in gene length. There is a lack of approaches to interpret DNA methylation effectively. Therefore, we have developed computational models to provide biological interpretation of relevant gene sets using DNA methylation data in the context of The Cancer Genome Atlas (TCGA).

Results: We illustrate that biological interpretation of DNA methylation (BioMethyl) utilizes the complete DNA methylation data for a given cancer type to reflect corresponding gene expression profiles and performs pathway enrichment analyses, providing unique biological insight. Using breast cancer as an example, BioMethyl shows high consistency in the identification of enriched biological pathways from DNA methylation data compared to the results calculated from RNA sequencing data. We find that 12 out of 14 pathways identified by BioMethyl are shared with those by using RNA-seq data, with a Jaccard score 0.8 for estrogen receptor (ER) positive samples. For ER negative samples, three pathways are shared in the two enrichments with a slight lower similarity (Jaccard score=0.6). Using BioMethyl, we can successfully identify those hidden biological pathways in DNA methylation data when gene expression profile is lacking.

Availability: BioMethyl R package is freely available in the GitHub repository (https://github.com/yuewangpanda/BioMethyl).

Contact: Chao.Cheng@dartmouth.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Epigenetic modification of DNA plays an important role in regulating gene activity and transcript levels without directly changing the gene sequence. DNA methylation is one of the most common epigenetic mechanisms and has been shown to impact multiple biological processes (Amir, et al., 1999; Bender, 2004; Costello and Plass, 2001; Laird, 2003). Consequently, aberrant DNA methylation has been associated with multiple human cancers including prostate cancer (Goering, et al., 2012; Goessl, et al., 2000), breast cancer (Silva, et al., 1999; Szyf, et al., 2004) and liver cancer (De Zhu, 2005; Wong, et al., 2000). Moreover, a strong relationship between DNA methylation and cancer development has been found, which has resulted in DNA methylation being used as a prognostic marker in many cancer types (Gyparaki, et al., 2013; Heyn and Esteller, 2012; Maeda, et al., 2003; Ng, et al., 2002; Sandoval, et al.,

2013). Therefore, array-based or sequencing-based approaches have been developed to measure large-scale DNA methylation profiles (Laird, 2010; Plongthongkum, et al., 2014). The Illumina methylation array platform provides an opportunity to generate genome-wide human DNA methylation profile, and the Illumina HumanMethylation450 BeadChip is one of the most commonly utilized platforms for investigating DNA methylation in a comprehensive manner. It contains 450,000 CpG probes that cover 99% of Refseq genes, encompassing promoters, gene bodies, UTRs, and intergenic regions (Sandoval, et al., 2011).

To interpret the underlying biology, differentially methylated CpG sites are routinely associated with particular phenotypes including cancer survival and biological pathway enrichment. Since gene ontology terms and pathways are generally represented as sets of genes, it is critical to accurately map DNA methylation sites to gene annotations. Mapping DNA methylation sites to genes usually involves either 1) identifying the genes with a single differentially methylated CpG site in their promoters (Kim, et al., 2012; Kriebel, et al., 2016; Li, et al., 2009; Shaknovich, et al., 2010) or 2) identifying genes cover differentially methylated regions (DMR) of the genome (Marsit, et al., 2011; Rijlaarsdam, et al., 2014; Wang, et al., 2012). Additionally, several web-based tools can be used to analyze methylation data including Annotation-Modules (Hackenberg and Matthiesen, 2008), EpiExplorer (Halachev, et al., 2012), GREAT (McLean, et al., 2010) and Galaxy (Goecks, et al., 2010). However, these methods have four main limitations. First, it is hard to capture the directionality of gene expression that results from DNA methylation. Generally, hypermethylation of the promoter causes repression (Jones and Takai, 2001), while hypermethylation in the gene body is correlated with activation (Bell, et al., 2011; Jones, 1999). Therefore, it is difficult to predict the changes in gene expression based simply upon DNA methylation results. Second, it is difficult to precisely define the extent of gene promoter methylation due to variability in the size of canonical promoters and the presence of distal enhancers, which introduces biases into the association of methylated regions with gene models. Third, the longer length of a gene, the higher the probability that this gene could be selected due to the nearby differentially methylated CpG sites. Lastly, for the web-based studies, specialized tools are needed to reformat the methylation data to genomic region formats (i.e., BED and WIG), which increases the difficulty of usage. To our knowledge, tools utilizing complete large-scale DNA methylation data to infer the directionality of gene expression and provide a framework for the interpretation of biological impact, is lacking.

We have developed an R package that we have named Biological Interpretation of DNA methylation (BioMethyl), which identifies biologically meaningful trends from a complete DNA methylation profile by using all available CpG sites. By integrating DNA methylation and RNA sequencing (RNA-seq) profiles for 37 cancer types from The Cancer Genome Atlas (TCGA), we have developed linear regression models to analyze the relationship between any single gene's expression profile and its corresponding CpG sites methylation sites for each cancer type. We inferred the contribution of each single CpG site to the expression of a gene using the coefficient of linear regression calculated by the model. Therefore, BioMethyl captures the expression values for a gene, based on its association with CpG methylation sites. We compare our results to RNA-seq profiles and show that BioMethyl accurately estimates gene expression from DNA methylation data. We have integrated Gene Set Enrichment Analysis (GSEA) (Subramanian, et al., 2005) into BioMethyl to automatically identify enriched pathways. As a result, we show that the enriched pathways identified by BioMethyl were highly consistent with those identified by RNA-seq. The BioMethyl R package is freely available on GitHub accessing by https://github.com/yuewangpanda/BioMethyl.

2 Methods

2.1 Data Collection

We downloaded RNA-seq and DNA methylation data for all 37 TCGA cancer types from Firehose (https://gdac.broadinstitute.org/, Nov, 2016). We fit models to the cancer types that contained more than 50 samples with both RNA-seq and DNA methylation profiles. We excluded ovarian cancer, lymphoid neoplasm diffuse large B-cell lymphoma, and cholangiocarcinoma due to their low sample numbers (Supplementary Table S1). To train a non-cancer model, TCGA normal samples with paired RNA-seq and methylation profiles were collected (Supplementary Table S2). To validate the non-cancer model, DNA methylation data (GSE42861) (Liu, et al., 2013) and gene expression data (GSE15573) (Teixeira, et al., 2009) were downloaded for rheumatoid arthritis (RA) samples.

2.2 Development of BioMethyl Models

To train BioMethyl models for each cancer type, we first preprocessed the RNA-seq and DNA methylation data. For RNA-seq data, we log2-transformed the data and removed genes if expression values were zero in more than half of the samples. Then, a z-transformation was further applied to the RNA-seq profile across samples. For DNA methylation data, we removed CpG sites if methylation levels were missing values in more than half of sample size. Then, R package "ENmix" (Xu, et al., 2016) was applied to methylation data to filter out outliers and to replace missing values using k nearest neighbors algorithm.

Then, we trained each BioMethyl model using linear regression to capture the association between gene expression and DNA methylation for each cancer types. For $gene_i$, $\boldsymbol{E} = \{e_1, e_2, ..., e_n\}$ is the gene expression across N samples, and $\boldsymbol{M}$ is the corresponding methylation matrix, containing all CpG sites associated to $gene_i$, where $cpg_{n,j}$ is the beta value of j-th CpG in sample n. By calculating the correlations between beta values of each CpG site and $\boldsymbol{E}$, we only selected the CpG sites whose beta values are correlated (Pearson correlation coefficient>|0.05|) with gene expression to build a model for $gene_i$ using the following function:

$$\boldsymbol{E} = \alpha + \boldsymbol{B}*\boldsymbol{M},$$

where $\boldsymbol{B} = \{\beta_1, \beta_2, ..., \beta_n\}$ is the vector of coefficients estimated by linear regression model which explains the contribution on gene expression for each CpG site. We built cancer-specific models for all TCGA cancer types and recorded those models into BioMethyl. The non-cancer model was trained with the same way instead of using matched data of TCGA normal samples.

2.3 Pathway enrichment analysis

GSEA and Fisher's exact test were applied to conduct pathway enrichment analyses in this study with the Molecular Signature Database (MSigDB) C2 dataset (c2.all.v5.2.symbols.gmt, 2016) (Subramanian, et al., 2005). Statistical significance of enriched pathways was set to false discovery rate (FDR) < 0.01. For Fisher's exact test, FDR was calculated with Benjamini & Hochberg method. To automatically identify enriched pathways, GSEA R script method "GSEA.1.0.R" was deployed in BioMethyl package to perform enrichment analysis with default settings (Subramanian, et al., 2005). The C2 dataset was set as the default gene sets.

2.4 BioMethyl Validation

To validate our model, we applied 10-fold cross-validation to test the quality of the gene expression inferred by the linear regression model. Namely, for each validation, we used 9/10 samples as training dataset to train the model. Then, by integrating the DNA methylation data and trained model, we calculated a gene expression profile for the rest 1/10 samples. After 10-fold cross-validation, we merged those profiles to a gene expression profile containing all samples and compared it with the RNA-seq data to conduct downstream validations.

2.5 Statistical analysis

To identify differentially methylated CpG sites and differentially expressed genes in two phenotypes (in our case is ER+ vs. ER-), we applied the Student t test to calculate the t scores and corresponding p values. The FDR was further calculated using the Benjamini-Hochberg multiple hypothesis testing correction method. By ranking t scores in a decreasing order, top-ranked CpG sites/genes are differentially methylated/expressed in ER+ samples and bottom-ranked CpG sites/genes are differentially methylated/expressed in ER- samples. For p values in boxplots, ANOVA test was used for comparison in multiple groups and two-sample Wilcoxon test was applied to those two groups comparisons.

3 Results

3.1 The limitations of previous approaches

Previous studies (Kim, et al., 2012; Kriebel, et al., 2016; Li, et al., 2009; Shaknovich, et al., 2010) have used gene promoter regions that contained differentially methylated CpG sites to perform pathway enrichment analyses. We tested the hypothesis that gene length might introduce biases in the downstream analyses. First, we examined the number of CpG sites associated with each gene according to the HumanMethylation450 platform mapping information. We found that 43.3% of genes (10865 out of 25094) map to less than 10 CpG sites, whereas more than 20% genes map to over 20 CpG sites (Fig. 1A). Moreover, only 7.4% of genes (1860 out of 25094) map to one CpG site. Notably, 0.7% of genes (163 out of 25094) map to more than 100 CpG sites. Broadly, genes with a longer length cover more CpG sites in the HumanMethylation450 platform (Fig. 1A). For example, PTPRN2, a receptor-type tyrosine-protein phosphatase N2 gene which occupies 1,048,741 bases, is associated with many diseases (Schmidli, et al., 1998; Smyth, et al., 2014; Sorokin, et al., 2015) and covers 1,288 CpG sites in the platform. Furthermore, by ranking genes based on the number of covered CpG sites in decreasing order, we found that the accumulation of CpG sites is derived by genes with longer length (Fig. 1B). In fact, the top 5,000 genes in length collectively occupy more than 50% of measured CpG sites.

For further biological interpretation of study results, differentially methylated CpG sites are often translated to genes. Because these mappings are not linear, it is theoretically possible for a gene to be implicated in both up- and down-regulated CpG groups in an analysis. To demonstrate this phenomenon, using TCGA breast cancer (BRCA) as an example, we identified the significantly ($p<0.01$) differentially methylated CpG sites for estrogen receptor positive (ER+) and estrogen receptor

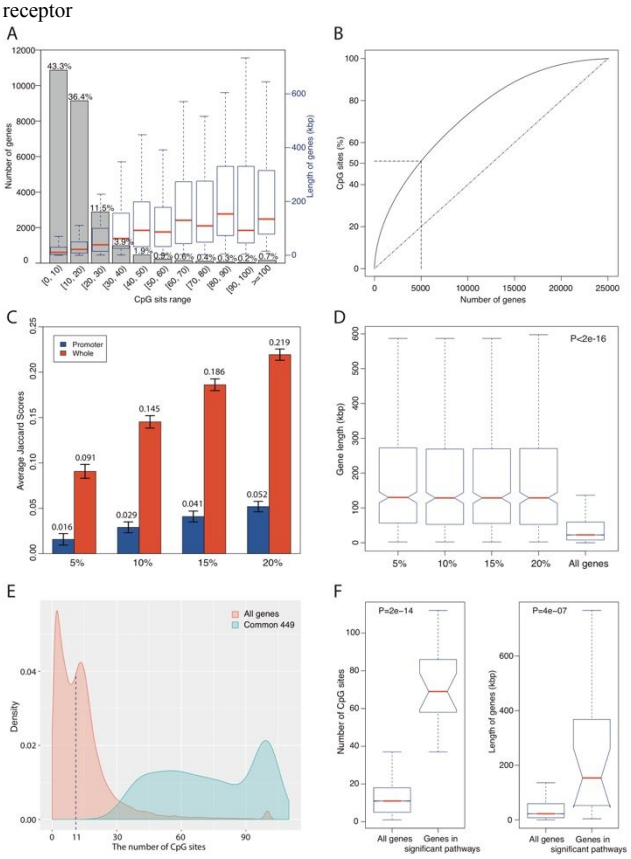

Fig. 1 Simply translating CpGs to genes confuses downstream results because of gene length diversity. (A) The mapping between number of genes and different CpG sites range (left y axis). The fraction above each bar shows the percentage of whole genome genes which located in this CpG range. Right y axis shows the relationship between the distribution of gene lengths in each CpG range. **(B)** The accumulation of CpG sites with gene length. **(C)** Barplot showing the average Jaccard scores from 10000 times simulations. Error bars show standard deviation of Jaccard scores. Blue is only using promoter CpG sites and red is using whole CpG sites. **(D)** Boxplot showing the distribution of gene lengths for each simulation and all genes. ANOVA p value is showed. **(E)** The distributions of covered CpG sites for common genes in simulations and all genes. **(F)** Boxplots for comparing gene length and covered CpG sites numbers between genes (subset of common genes) in significant pathways and all genome genes. Wilcoxon Rank Sum test p values are showed.

negative (ER-) samples which resulted in two distinct, non-overlapping lists of CpG sites. We then randomly selected 5/10/15/20% of the sites from each of the groups, mapped these sites to genes, and calculated the Jaccard score between the groups of genes. This simulation was repeated 10,000 times. When using sites only in promoter region (within 1 kbp of

the transcription start site) of genes, we found that the average number of genes overlapping between the two lists increased as the number of random CpG sites were selected. We found the same trend when considering CpG sites that map to whole gene regions (Fig. 1C). Moreover, the number of overlapping genes was at least 4.2 times larger when using all CpG sites compared to using only promoter sites. These observations suggest that simply translating CpG sites to genes based on location may confuse the overall interpretation of results. This may be in part due to the effect of gene length because longer regions generally contain more CpG sites.

Next, we identified the 500 genes that most frequently appeared in each simulation using whole region CpG sites and found that the average lengths of these genes are nearly identical between simulations but are much longer than the average length of all genes in the genome (Fig. 1D, ANOVA P<2e-16). Furthermore, we found that 449 of the 500 genes are shared by these four different simulations, and those genes tend to cover more CpG sites compared to all genes in the genome (Fig. 1E). These 449 common genes were significantly involved in sensory system (olfactory transduction), signaling molecules and interaction (cytokine receptor interaction), genetic information processing (ubiquitin mediated proteolysis, spliceosome), signaling molecules and interaction (neuroactive ligand receptor interaction), neurodegenerative diseases (Huntington's disease and Alzheimer's disease), nucleotide metabolism (purine metabolism and pyrimidine metabolism). Moreover, we found that genes involved in those significant pathways cover more CpG sites (P=2e-14) and have a longer gene length (P=4e-07) compared to all genes in the genome (Fig. 1F). Again, these results suggest that longer genes could be preferentially selected, and the length of genes might dramatically impact the biological interpretation of previously studies which use direct mapping approaches.

3.2 Overview of our analyses

To overcome these limitations, we developed a simple and powerful method called BioMethyl to interpret DNA methylation data in a wide range of cancer types. We built a linear regression model for each gene to capture the association of its expression and corresponding CpG methylation sites by integrating RNA-seq and DNA methylation profiles for each TCGA cancer type. The coefficients of each model explain the contributions of CpG methylation sites associated with each gene's expression. To validate the model, we used 10-fold cross validation to compare the biological outcome of the inferred gene expression profiles and the original RNA-seq profile (Fig. 2A). We found a high concordance between those two profiles, both in terms differentially expressed genes and enriched pathways. Therefore, our method takes advantage of the complete profiles of DNA methylation and RNA-seq to build more accurate models for each cancer type. By assembling our framework and GSEA method, we developed a user-friendly R package, called BioMethyl, which is freely available at GitHub (https://github.com/yuewangpanda/BioMethyl). When a user inputs a new DNA methylation dataset of interest for a given cancer type, BioMethyl utilizes the corresponding models to uncover relevant biological pathways hidden in the data (Fig. 2B).

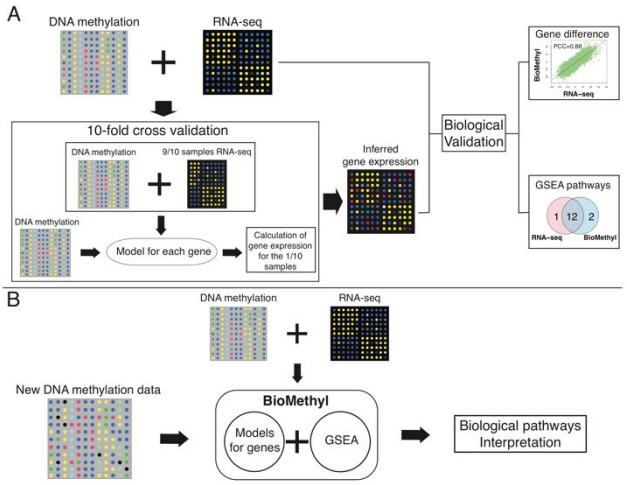

Fig. 2. Workflow of our computational framework. (A) Validation of models. Using a 10-fold cross validation manner, BioMethyl trains models and calculates a new gene expression matrix for samples. To validate our models, we compare the inferred gene expression matrix with RNA-seq data in terms of gene difference and involved pathways. **(B)** Application of models. Using the complete DNA methylation and RNA-seq profiles, we build models for each cancer. Further by integrating GSEA analysis, we develop the R package BioMethyl to reveal the relevant pathways in a new DNA methylation data of interest.

3.3 Validation of the BioMethyl model

To validate BioMethyl, we first examined whether the linear regression model could accurately depict the gene expression profile through DNA methylation data. We trained the models with cancers in TCGA that had both RNA-seq and DNA methylation profiles using 10-fold cross validation (see Methods). Comparing the pseudo gene expression estimated by the model and the RNA-seq profile, we found that BioMethyl model could explain the relationship between RNA-seq and DNA methylation profiles

of 94.1% (32 out of 34) cancers in TCGA with median Spearman correlation coefficients (SCC) greater than 0.3 (Supplementary Fig. S1).

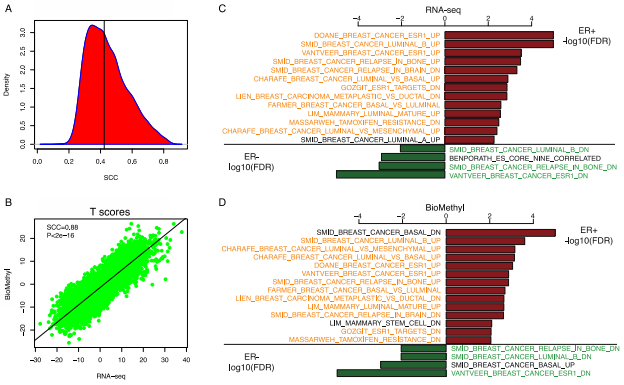

Fig. 3. Validation of BioMethyl in the context of breast cancer. (A) Density plot for spearman correlation coefficient of genes by comparing gene expression inferred by BioMethyl and RNA-seq data. **(B)** Scatter plot of t scores (ER+ samples vs. ER- samples) for genes between gene expression inferred by BioMethyl and RNA-seq data. Pathway enrichment results of GSEA are showed for **(C)** RNA-seq data and **(D)** gene expression inferred by BioMethyl by comparing ER+ to ER- samples. For pathways enriched in ER+ samples, -log10(FDR) are showed (red). The orange pathways are pathways shared by two results for ER+ samples. For pathways enriched in ER- samples, log10(FDR) are showed (green), in which green pathways are shared pathways.

We tested Biomethyl on the TCGA breast cancer (BRCA) dataset since it contained 785 tumors, the highest number of tumors in TCGA,

with both RNA-seq and DNA methylation profiles. BioMethyl was able to recapitulate tumors' gene expression using their DNA methylation data alone with a median SCC 0.423 when compared to their RNA-seq profiles (Fig. 3A). Moreover, the majority (more than 88%) of genes have SCCs greater than 0.3 (Supplementary Fig. S2). By dividing genes into low, intermediate, and high variance groups based on the original RNA-seq dataset, we found that genes with high variance had significantly higher SSCs than the other two groups (Supplementary Fig. S3, ANOVA test P<2e-16). To further test the reliability of our model, we compared the expression difference of each gene between ER+ and ER- patients for the estimated gene expression profile and RNA-seq profile. By calculating the correlation of t-scores, the result showed that the estimated gene expression profile inferred from DNA methylation data is highly consistent with the RNA-seq data (Fig. 3B, SCC=0.88) which is only slightly lower than the comparison between TCGA microarray and RNA-seq profiles (Supplementary Fig. S4, SCC=0.94). These observations suggest that BioMethyl is able to accurately infer gene expression through DNA methylation data compared to RNA-seq data.

To further compare the similarity of biological findings identified by BioMethyl and RNA-seq analyses, we performed GSEA analysis (Subramanian, et al., 2005) (Supplementary Table S3). By dividing TCGA BRCA samples into ER+ and ER- groups, we identified pathways significantly enriched in those two groups of samples. When using RNA-seq data, there were 13 pathways significantly (FDR<0.01) enriched in ER+ samples and four pathways significantly (FDR<0.01) enriched in ER- samples (Fig. 3C). When using the gene expression profile inferred by BioMethyl, there were 14 pathways significantly (FDR<0.01) enriched in ER+ samples and four pathways significantly (FDR<0.01) enriched in ER- samples (Fig. 3D). We found that 12 pathways for ER+ samples are shared by those two methods (Jaccard score=0.8). For ER- samples, three pathways are shared in the two enrichments with a slight lower similarity (Jaccard score=0.6). We also investigated Jaccard scores using different FDR thresholds (from 0 to 0.25) for shared pathways in either ER+ or ER- samples (Supplementary Fig. S5). In our experience, setting a more conservative FDR threshold (<0.05) is important to achieve high similarity (Jaccard scores>0.5) between the results of BioMethyl and RNA-seq.

3.4 Validation of Fisher's exact test enrichment results

Next, we compared the enriched pathways of genes covering differentially methylated CpG sites and the genes mostly differentially expressed in gene expression profiles. We identified the top 500 hypermethylated CpG sites (FDR<1e-30) in ER+ samples and mapped those sites to genes, resulting in 270 genes. We identified 348 genes for ER- samples through the top 500 hypermethylated CpG sites (FDR<4e-38) in ER- samples. Notably, there were 12 genes shared between these two gene sets (Fig. 4A). Next, we identified the top 270 and 348 differentially expressed genes in ER+ and ER- patients, respectively, using both the RNA-seq profile and the profile inferred from BioMethyl. We found very little overlap in the number of genes identified by CpG-mapped genes and the top differentially expressed genes identified by RNA-seq or BioMethyl for both ER+ (Fig. 4B) and ER- (Fig. 4C) samples. Then, we conducted Fisher's exact test to those six differentially expressed gene sets to identify enriched biological pathways, respectively. When using gene sets collected via differentially methylated CpG sites, there is only one pathway enriched in ER+ samples (compared to 23 identified by RNA-seq and 28 identified by BioMethyl) and no pathways enriched in ER- samples (compared to 143 identified by RNA-seq and 16 identified by BioMethyl) (data not shown). We found that 21 out of 28 significantly enriched pathways (FDR<0.01) using gene set from BioMethyl profile are shared with those from RNA-seq in the context of ER+ samples (Fig. 4D). The common pathways are highly related to ER+ context including upregulated by ESR1, genes upregulated in luminal-B breast cancer and downregulated in invasive breast cancer (Supplementary Table S4). Similar, for ER- samples, 10 out of 16 pathways (FDR<0.01) are shared in the comparison (Fig. 4E) including downregulation by ESR1, genes upregulated in basal breast cancer and downregulated in luminal-B breast cancer (Supplementary Table S4). These observations suggest that using the gene expression values inferred by the BioMethyl method is able to identify common biological pathways using Fisher's exact test.

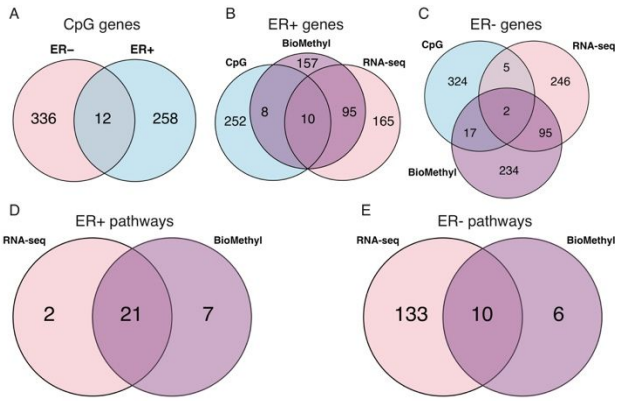

Fig. 4. Validation of BioMethyl using Fisher's exact test. Venn diagrams for **(A)** Differentially expressed genes selected by hypermethylated CpG sites in ER+ and ER- samples; **(B)** Differentially expressed genes in ER+ samples selected by hypermethylated CpG sites, RNA-seq and BioMethyl; **(C)** Differentially expressed genes in ER- samples selected by hypermethylated CpG sites, RNA-seq and BioMethyl; **(D)** Pathways enriched in ER+ samples between RNA-seq data and BioMethyl; **(E)** Pathways enriched in ER- samples between RNA-seq data and BioMethyl.

3.5 Application of BioMethyl to non-cancer diseases

To test whether the BioMethyl could be applied to non-cancer diseases as well, we selected RA, an autoimmune disease, as an example. By integrating the non-cancer model and the DNA methylation data, we estimated gene expression for samples in GSE42861. GSEA and Fisher's exact test were applied to both the estimated gene expression of GSE42861 and the real gene expression dataset GSE15573. When comparing the GSEA results, neither of them contains significant (FDR<0.25) pathways. When comparing the results of Fisher's exact test, differentially expressed genes (FDR<0.01) in our model were significantly enriched in immune-related pathways (Supplementary Table S5) such as T cell receptors, immune response, and antigen response. This result is consistent with both the characteristics of RA (Choy, 2012) and the enrichment analysis reported from the RA gene expression (GSE15573) (Teixeira, et al., 2009). This validation suggests

that our non-cancer model could capture the common biological pathways for non-cancer diseases, such as RA.

3.6 Implementation of BioMethyl package

For a user friendly and straightforward usage, we compiled models for 37 cancer types into our BioMethyl R package including five functions (Table 1). When a user inputs DNA methylation data, BioMethyl will produce the relevant biological pathways as final output. Here, we briefly introduce the procedure of BioMethyl package. BioMethyl preprocesses the data with filterMethyData() function and removes CpG sites that have missing values in more than half samples and imputes the rest missing values by integrating "ENmix", an specialized R package for DNA methylation data (Xu, et al., 2016), with default parameters. Next, calExpr() function is applied to the filtered methylation data to infer the gene expression profile for a given disease type. In this step, the inferred gene expression could be saved as a text file as an option for other customized applications. Then, differentially expressed genes (DEGs) are identified via calDEG() function in order to optimize gene expression for GSEA analysis. As well, the list of differentially expressed genes could be saved as a text file for gene set based pathway or GO term enrichment test. Lastly, using the inferred gene expression, BioMethyl integrates GSEA R code to perform pathway enrichment analysis using GSEA default settings. In this step, the parameter of cutoff for DEGs is a numeric vector in which the first element is the cutoff for t score (default is 0) and the second is for p value (default is 0.01). Moreover, BioMethyl package has a friendly recommendation function so that it helps users select the best model for their DNA methylation data. By applying a centroid manner, referCancerType() function can suggest a suitable cancer type model having the best similarity with TCGA cancers when it is not clear. The BioMethyl package and demo code are freely available at GitHub (https://github.com/yuewangpanda/BioMethyl).

Table 1. Brief introduction of functions in BioMethyl R package.

Function	Application	Function Examples
filterMethyData()	Pre-process methylation data	mydat <- filterMethyData(RawData)
calExpr()	Calculation of gene expression based on methylation data	myexpr <- calExpr(MethyData, CancerType, Example=FALSE, SaveOut=FALSE, OutFile)
calDEG()	Identification of differentially expression genes	myDEG <- calDEG(ExprData, Sample_1, Sample_2, SaveOut=FALSE, OutFile)
calGSEA()	GSEA pathway enrichment	mypath <- calGSEA(ExprData, DEG, DEGthr=c(0, 0.01), Sample_1, Sample_2, OutFile, GeneSet="C2")
referCancerType()	Recommendation of cancer type	myType <- referCancerType(MethyData)

4 Discussion

Since DNA methylation plays important roles in multiple biological processes, more and more efforts have been put on generating DNA methylation data. Attempts at investigating enriched pathways using DNA methylation profile has been an active area study. Previous studies used either single differentially methylated CpG sites or DMRs as an assumed proxy to identify the differentially expressed genes between samples. However, our results suggest that using the direct mapping method results in a pronounced overlapping of genes between opposing biological groups which could introduce bias to downstream analyses – pathway/genes associated with more CpG sites are more likely to be identified (Fig. 1 and Fig. 4A). Previous work has tried to correct this bias by modeling the probability of a gene to be selected by chance as a function of the number of CpG sites it associated with (Geeleher, et al., 2012). In this sense, all CpG sites associated with a gene are assumed to contribute equally to the transcriptional regulation of the gene. In our work, we developed the BioMethyl method to more reasonably map CpG sites to genes by assigning different weights to sites according to their relative contributions to gene expression. Our results showed that the enriched pathways determined by BioMethyl are highly consistent with those interpreted directly from RNA-seq.

Due to the internal relationship between DNA methylation and gene expression (Razin and Cedar, 1991), several studies have developed computational methods to infer gene expression from DNA methylation data in the context of a certain cancer (Li, et al., 2015; Schlosberg, et al., 2017). BioMethyl applies linear regression models to capture the association between the expression of a gene and its CpG sites methylation levels for all TCGA cancer types. We found that the gene expression profile inferred from DNA methylation data is highly like RNA-seq profile (Fig. 3 and Supplementary Table S2). Moreover, we found that using the inferred gene expression profile can classify ER+ from ER- breast cancer samples as well as using RNA-seq profile (Supplementary Fig. S6). These results suggest that BioMethyl captures the overall and true directions of gene expression via linear regression models. Therefore, the inferred gene expression profile could be applied to other downstream analyses (e.g. pathway enrichment using Fisher's exact test, identifying differentially expressed genes) when a study only has DNA methylation data available.

In this study, we took advantage of the TCGA cancer data to estimate the contribution of individual CpG site to gene expression. However, the weights of CpG sites might change in different tissues or under different physiological conditions. Therefore, lower performance might be expected when the models trained from the TCGA data are directly applied to DNA methylation data for other diseases. Nevertheless, the proposed framework can be used to re-calculate the weights of CpG sites under various contexts given matched gene expression and DNA methylation data, which has becoming more and more readily available in the future.

In summary, BioMethyl makes use of the whole DNA methylation profile and captures the association between gene expression and DNA methylation in a highly sensitivity way. In our models, we assigned coefficients to those CpG sites really associated with changes in gene expression. The models contained within BioMethyl span a large number of diverse cancers. Our freely available R package is easy to install and use. Moreover, our methods represent significant contributions to data interpretation when only DNA methylation is available. With the improvement of methylation platform (i.e. HumanMethylation850), BioMethyl hopefully could achieve a higher accuracy in terms of biological interpretation directly from DNA methylation data.

Funding

This work was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR001086 and the Geisel School of Medicine at Dartmouth College start-up funding package provided to C.C.

Conflict of Interest: none declared.

References

Amir, R.E., *et al.* Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat Genet* 1999;23(2):185-188.

Bell, J.T., *et al.* DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol* 2011;12(1):R10.

Bender, J. DNA methylation and epigenetics. *Annu Rev Plant Biol* 2004;55:41-68.

Choy, E. Understanding the dynamics: pathways involved in the pathogenesis of rheumatoid arthritis. *Rheumatology (Oxford)* 2012;51 Suppl 5:v3-11.

Costello, J.F. and Plass, C. Methylation matters. *J Med Genet* 2001;38(5):285-303.

De Zhu, J. The altered DNA methylation pattern and its implications in liver cancer. *Cell Res* 2005;15(4):272-280.

Goecks, J., *et al.* Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* 2010;11(8):R86.

Goering, W., Kloth, M. and Schulz, W.A. DNA methylation changes in prostate cancer. *Methods Mol Biol* 2012;863:47-66.

Goessl, C., *et al.* Fluorescent methylation-specific polymerase chain reaction for DNA-based detection of prostate cancer in bodily fluids. *Cancer Res* 2000;60(21):5941-5945.

Gyparaki, M.T., Basdra, E.K. and Papavassiliou, A.G. DNA methylation biomarkers as diagnostic and prognostic tools in colorectal cancer. *J Mol Med (Berl)* 2013;91(11):1249-1256.

Hackenberg, M. and Matthiesen, R. Annotation-Modules: a tool for finding significant combinations of multisource annotations for gene lists. *Bioinformatics* 2008;24(11):1386-1393.

Halachev, K., *et al.* EpiExplorer: live exploration and global analysis of large epigenomic datasets. *Genome Biol* 2012;13(10):R96.

Heyn, H. and Esteller, M. DNA methylation profiling in the clinic: applications and challenges. *Nat Rev Genet* 2012;13(10):679-692.

Jones, P.A. The DNA methylation paradox. *Trends Genet* 1999;15(1):34-37.

Jones, P.A. and Takai, D. The role of DNA methylation in mammalian epigenetics. *Science* 2001;293(5532):1068-1070.

Kim, J.H., *et al.* LRpath analysis reveals common pathways dysregulated via DNA methylation across cancer types. *BMC Genomics* 2012;13:526.

Kriebel, J., *et al.* Association between DNA Methylation in Whole Blood and Measures of Glucose Metabolism: KORA F4 Study. *PLoS One* 2016;11(3):e0152314.

Laird, P.W. The power and the promise of DNA methylation markers. *Nat Rev Cancer* 2003;3(4):253-266.

Laird, P.W. Principles and challenges of genomewide DNA methylation analysis. *Nat Rev Genet* 2010;11(3):191-203.

Li, J., *et al.* Using epigenomics data to predict gene expression in lung cancer. *BMC Bioinformatics* 2015;16 Suppl 5:S10.

Li, M., *et al.* Integrated analysis of DNA methylation and gene expression reveals specific signaling pathways associated with platinum resistance in ovarian cancer. *BMC Med Genomics* 2009;2:34.

Liu, Y., *et al.* Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat Biotechnol* 2013;31(2):142-147.

Maeda, K., *et al.* Hypermethylation of the CDKN2A gene in colorectal cancer is associated with shorter survival. *Oncol Rep* 2003;10(4):935-938.

Marsit, C.J., *et al.* DNA methylation array analysis identifies profiles of blood-derived DNA methylation associated with bladder cancer. *J Clin Oncol* 2011;29(9):1133-1139.

McLean, C.Y., *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* 2010;28(5):495-501.

Ng, C.S., *et al.* Tumor p16M is a possible marker of advanced stage in non-small cell lung cancer. *J Surg Oncol* 2002;79(2):101-106.

Plongthongkum, N., Diep, D.H. and Zhang, K. Advances in the profiling of DNA modifications: cytosine methylation and beyond. *Nat Rev Genet* 2014;15(10):647-661.

Razin, A. and Cedar, H. DNA methylation and gene expression. *Microbiol Rev* 1991;55(3):451-458.

Rijlaarsdam, M.A., *et al.* DMRforPairs: identifying differentially methylated regions between unique samples using array based methylation profiles. *BMC Bioinformatics* 2014;15:141.

Sandoval, J., *et al.* Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* 2011;6(6):692-702.

Sandoval, J., *et al.* A prognostic DNA methylation signature for stage I non-small-cell lung cancer. *J Clin Oncol* 2013;31(32):4140-4147.

Schlosberg, C.E., VanderKraats, N.D. and Edwards, J.R. Modeling complex patterns of differential DNA methylation that associate with gene expression changes. *Nucleic Acids Res* 2017;45(9):5100-5111.

Schmidli, R.S., *et al.* Antibodies to the protein tyrosine phosphatases IAR and IA-2 are associated with progression to insulin-dependent diabetes (IDDM) in first-degree relatives at-risk for IDDM. *Autoimmunity* 1998;28(1):15-23.

Shaknovich, R., *et al.* DNA methylation signatures define molecular subtypes of diffuse large B-cell lymphoma. *Blood* 2010;116(20):e81-89.

Silva, J.M., *et al.* Presence of tumor DNA in plasma of breast cancer patients: clinicopathological correlations. *Cancer Res* 1999;59(13):3251-3256.

Smyth, L.J., *et al.* DNA hypermethylation and DNA hypomethylation is present at different loci in chronic kidney disease. *Epigenetics* 2014;9(3):366-376.

Sorokin, A.V., *et al.* Aberrant Expression of proPTPRN2 in Cancer Cells Confers Resistance to Apoptosis. *Cancer Res* 2015;75(9):1846-1858.

Subramanian, A., *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005;102(43):15545-15550.

Szyf, M., Pakneshan, P. and Rabbani, S.A. DNA methylation and breast cancer. *Biochem Pharmacol* 2004;68(6):1187-1197.

Teixeira, V.H., *et al.* Transcriptome analysis describing new immunity and defense genes in peripheral blood mononuclear cells of rheumatoid arthritis patients. *PLoS One* 2009;4(8):e6803.

Wang, D., *et al.* IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. *Bioinformatics* 2012;28(5):729-730.

Wong, I.H., *et al.* Frequent p15 promoter methylation in tumor and peripheral blood from hepatocellular carcinoma patients. *Clin Cancer Res* 2000;6(9):3516-3521.

Xu, Z., *et al.* ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip. *Nucleic Acids Res* 2016;44(3):e20.