Statistics and Its Interface Volume 10 (2017) 279–289

Correcting length-bias in gene set analysis for DNA methylation data

Shaoyu Li*, Tao He, Iwona Pawlikowska, and Tong Lin

The enrichment analysis of pre-defined gene sets is a widely used tool to extract functional information in association studies. However, traditional methods give biased results on genome-wide DNA methylation data due to the different number of probes in genes. In this article, we present MethylSet, a novel two-step procedure which combines gene based association analysis with logistic regression model for enrichment analysis to correct bias induced by gene size. The adjustment of gene size effect is crucial because irrelevant gene sets may be identified otherwise. Our simulation studies showed that MethylSet has a well-controlled type I error rate and promising statistical power. When applied to a real DNA methylation data set, MethylSet was able to obtain meaningful gene sets associated with the studied disease outcome.

Keywords and phrases: Epigenome-wide Association Study (EWAS), Length bias, Logistic kernel machine regression, Gene set analysis.

1. INTRODUCTION

DNA methylation (DNAm) modifications are heritable and have been long speculated to play important roles in regulating gene expression. Recent development in high-throughput biotechnologies has facilitated the genome-wide DNA methylation profiling and made epigenome-wide association studies (EWASs) feasible, thereby provides a great opportunity to systematically identify DNA methylation variations associated with human diseases. Multiple recent EWASs have demonstrated that DNA methylation variations can be valuable diagnostic and prognostic biomarkers for human diseases [1, 2, 3, 4]. Analogous to that of genome-wide association studies (GWAS), EWAS usually generates a list of hundreds or even thousands of CpG sites, whose methylation status is significantly associated with the studied phenotype. Unfortunately, how to extract biologically interpretable information from the list is not straightforward.

Gene set analysis (GSA) is the most popular way to study how the identified associated genes relate biologically, with respect to pre-defined gene sets such as Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG), in determining phenotypic outcomes. The type of analysis is very important to improve understanding of the underlying molecular mechanisms of the susceptibility to human diseases. Especially for translational research, GSA could be very helpful in terms of identifying novel target pathways for potential treatments. Although the used database of gene sets might vary, almost every single report of EWAS relies on gene set analysis to obtain additional functional interpretation. GSA was generally implemented in two steps: (1) find genes associated with the studied phenotype and, (2) apply a statistical GSA approach, for example, Fisher's Exact Test (FET), to test the over/under-representation of a pre-defined gene set among the associated genes. Traditional approaches identify associated genes by testing individual probes followed by a post-hoc aggregation procedure. However, single probe based methods overlook the higher order interaction between probes within a gene and therefore could loose power. Besides, one arbitrary chosen post-hoc criterion for the aggregation procedure would not be suitable for all genes. More importantly, many platforms for genomewide DNA methylation profiling were designed such that the number of probes per gene varies very much. For example, on the Illumina HumanMehtylation450 BeadChip, the number of probes per gene varies from 1 to 1289. Then, in single probe based analysis, genes with one probe are tested only once and genes with hundreds/thousands of probes are tested hundreds/thousands of times. Considering a widely used criterion for aggregation: "call a gene significantly associated when at least one probe in the gene is significantly associated with the studied phenotype", apparently, genes with more probes are more likely to be called significantly associated just by chance. This phenomena violates a key assumption of many existing GSA methods that have been developed for gene expression data: every single gene is equally likely to be associated with the phenotype by chance. There are other more sophisticated post-hoc aggregation criteria that have been used to define significantly associated genes [5, 6], in all cases, genes with more probes are more likely to meet the criteria employed. In addition, the mean number of probes per gene is also very different between gene sets. For example, on the Illumina HumanMethylation450K BeadChip, mean number of probes per gene of genes annotated to the KEGG and GO terms are positively skewed (Figure 1), especially for GO terms (left panel, Figure 1). The mean number of probes

[*] Corresponding author.

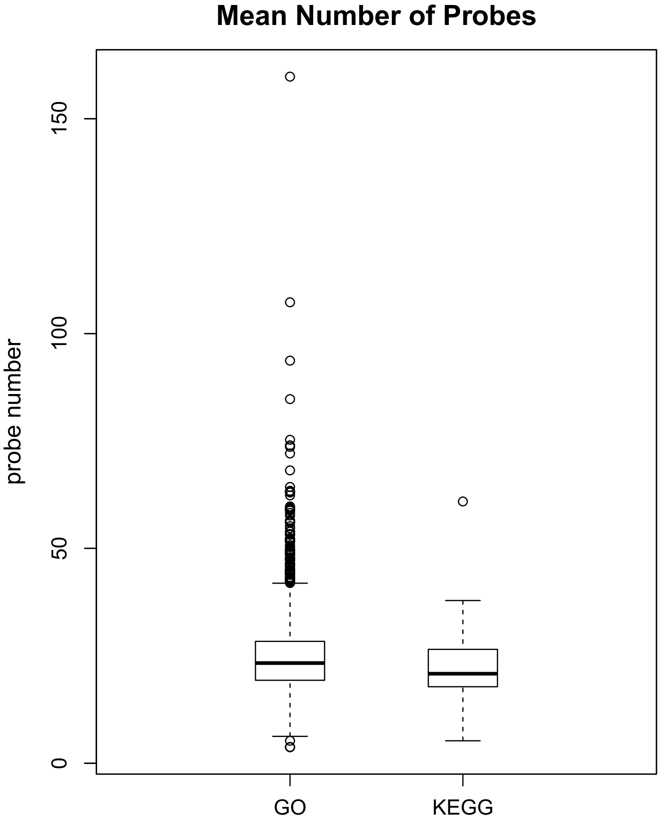


Figure 1. The distribution of mean number of probes per gene for GO terms (left) and KEGG pathways (right).

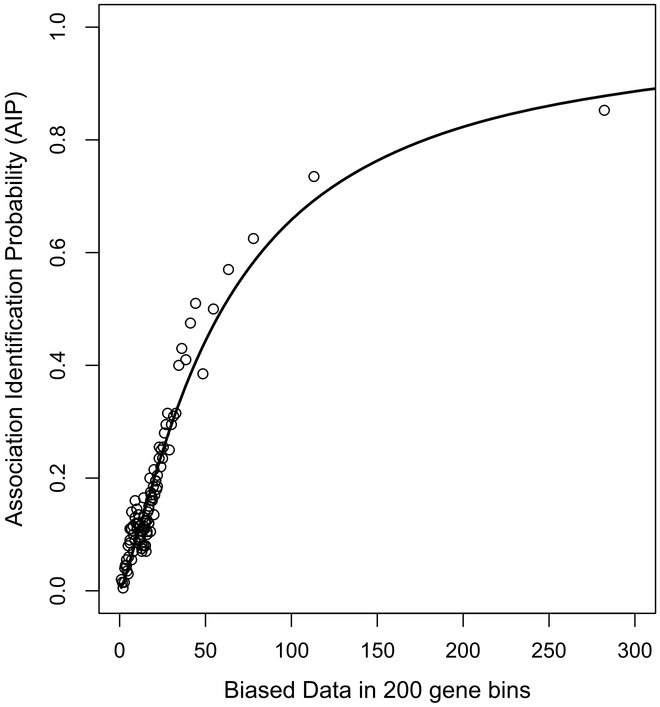


Figure 2. **Association Identification Probability (AIP) as a function of gene size.** *Each point represents the proportion of genes called significant in a bin of 200 genes plotted against the average gene size of the bin. The fitted line is the AIP estimated by fitting a logistic regression model. The pattern that longer genes have higher chance to be identified is clearly indicated.*

per gene varies from 3.7 to 160 for GO terms and from 5.2 to 60.9 for the KEGG pathways. It is not hard to imagine that gene sets that have more long genes, which might not be biologically relevant, could be falsely identified under the current analysis scheme.

The correction of gene size/length effect in GSA has been studied for some other types of "-omics" data. For example, RNA sequencing [7, 8, 9, 10] and ChIP sequencing data [11], gene size/length was referred to as physical length of a gene in these studies, while, in this article, we refer gene size/length (interchangeably) as the number of probes in a gene. The definition is similar to the number of SNPs within a gene for genotype data [12]. Some previous works have reported pitfalls for current gene set analysis procedures that overlook the gene size effect in DNA methylation data analysis [13, 11]. We also observed strong empirical evidence of the positive relationship between the association identification probability (AIP: the probability of a gene being called significantly associated) and gene size as shown in Figure 2 in the real data set that we considered in this study. The proportion of significant genes in a bin of 200 genes increases as the average gene size of the bin increases and approaches to 1.

In this study, we proposed a novel two-step procedure, MethylSet, which explicitly accounts for the gene size information in GSA for DNA methylation data. A gene based analysis model was used to identify significant genes. The model takes care of the joint function of probes within a gene and is expected to have better performance than a randomly picked post-hoc aggregation procedure. The gene size effect in GSA was further addressed by incorporating gene size as a covariate in a logistic regression model. Detailed description of the procedure is given in the following section.

2. METHODS

2.1 Logistic kernel machine model for gene-based EWAS

Suppose we have genome-wide DNA methylation profiled for n samples, including n_0 normal and n_1 tumor tissue samples. Consider one gene at a time, we denote DNA methylation measurements of L probes in the gene by a $L \times 1$ vector, $z_i = (z_{i,1}, z_{i,2}, \cdots, z_{i,L})$, for the i^{th} sample, and other covariates by a vector x_i, including the intercept. Let the disease outcome $y_i = 1$ if tumor tissue and $y_i = 0$ if normal tissue. Model the relationship between $\pi_i = P(y_i = 1|x_i, z_i)$ and the DNA methylation of the gene and the covariates via the following logistic regression:

$$(1) \qquad log\frac{\pi_i}{1-\pi_i} = x_i'\beta + f(z_i), i = 1, 2, \cdots, n$$

where β is a vector of unknown parameters associated with the covariates and function $f(\cdot)$ is a nonparametric function that captures the effect of DNA methylation. The null hypothesis that DNA mehtylation variations have no effect on the disease outcome can be formulated as $H_0 : f = 0$ and alternative hypothesis is $H_a : f \neq 0$. Assuming function f lies in a reproducing kernel Hilbert space (RKHS), it has been shown earlier by Liu et al. [14, 15] that testing the null hypothesis $H_0 : f = 0$ is equivalent to testing the variance component $H_0^* : \tau^2 = 0$ in a generalized linear mixed effects model:

$$(2) \qquad log\frac{\pi}{1-\pi} = X\beta + b$$

where $b \sim N(0, \tau^2 K)$, K is a $n \times n$ kernel matrix calculated based on the DNA methylation measurements of the gene. Specifically, element $k_{i,j} = \phi(z_i, z_j)$ measures the similarity between sample points i and j, where ϕ is a semi-positive definite kernel function associated with the RKHS. We used the linear kernel function for analysis in this work. A score type test statistic can be applied to test the null hypothesis H_0^* within the framework of linear mixed effects model [15, 16]:

$$(3) \qquad Q = (y - \hat{\pi})K(y - \hat{\pi})$$

where $\hat{\pi}$ is the estimate of π under the null hypothesis $H_0^* : \tau^2 = 0$.

2.2 Gene set analysis

Results from the previously described gene-based association study can then be summarized by a binary variable $D_g, g = 1, 2, \cdots, N$, where N is the total number of annotated genes. $D_g = 1$ if gene g is statistically significant (after multiple testing correction) and $D_g = 0$ if gene g is not significant. A membership variable S was defined based on the membership of a given gene set. And $S_g = 1$ if a gene is in the set, and $S_g = 0$ otherwise. Let L_g denote gene size. The binary outcome variable D_g depends on the membership variable S_g and the gene size L_g through the following logistic regression model:

$$(4) \qquad log\frac{\mu_g}{1-\mu_g} = \alpha_0 + \alpha_1 S_g + \alpha_2 log_{10}(L_g)$$

where $\mu_g = P(D_g = 1|S_g, L_g)$, α_0 is the intercept, α_1 is the effect of the given gene set, which is the coefficient of interest, and α_2 denotes the effect of gene size. The linear relationship between the logarithm of the gene size with base 10 and the log odds is well supported by the real data set (Figure 2). However, if nonlinear gene size effect is suggested, then a more general model which replaces the linear term by a nonparametric smooth function can be built.

$$(5) \qquad log\frac{\mu_g}{1-\mu_g} = \alpha_0 + \alpha_1 S_g + m(L_g)$$

We conducted inference based on the parametric as well as a nonparametric model and the obtained results were pretty similar (data no shown). However, the linear model is about seven times faster computationally.

The enrichment of a gene set among the associated genes can be detected by testing the null hypothesis that $H_0 : \alpha_1 = 0$, which means the membership of the gene set is not related to the odds of genes to be significantly associated with the phenotypic outcome. Multiple existing test statistics could be used for the purpose, including likelihood ratio test and Wald's test. For the analyses included in this article, the Wald's test was applied.

3. SIMULATION AND REAL DATA ANALYSIS

3.1 Simulation analysis

MethylSet adjusts the effect of gene size in GSA in two aspects: it applies gene based analysis to identify associated genes and includes gene size as a covariate in a logistic regression model to further correct its effect. We therefore conducted simulation studies to evaluate the statistical performance of the two steps. Due to the unobservable underlying correlation structure between CpG sites, we did not try to simulate genome-wide DNA methylation profile. Instead, we simulated DNA methylation data for one gene and compared the empirical type I error rate and power of the gene and single probe based methods regarding the identification of associated genes. Besides, we simulated the association identification probability (AIP) while considering different effect sizes of gene size and compare the performance of the proposed method with two other widely used methods: the Fisher's Exact Test (FET) and GOseq (a popular GSA method proposed for RNA sequencing data [7]). Detailed set up is given in the following section.

Simulate DNA methylation data

Consider one gene with gene size denoted as L, we first generated M-values, which are the log_2 ratio of the intensities of methylated and unmethylated signals from a HumanMethylation450 BeadChip, from a multivariate normal distribution with mean vector μ and autoregressive (order 1) covariance structure, $AR(1; \rho)$. The simulated data (M-values) were then transformed to β-values, that is the methylation scores. Specifically,

$$M_{0i} \sim MN(\mathbf{0}, AR(1; \rho)), i = 1, 2, \cdots, n_0;$$
$$M_{1i} \sim MN(\mu, AR(1; \rho)), i = 1, 2, \cdots, n_1;$$

and

$$\beta_{ci} = \frac{2^{M_{Ci}}}{1 + 2^{M_{Ci}}}, c = 0, 1; i = 1, 2, \cdots, n_c$$

Elements of the mean vector μ were set to be zeros for controls and nonzero for cases. We considered six different scenarios under the alternative hypothesis: only 1, 1%, 2%, 5%, 10%, and 20% of the L probes are associated with the phenotype, that is, only the corresponding number/proportion of elements of the mean vector μ were set to be nonzero (=0.5) and others remain as zero. For example, if the gene size $L = 100$, then 1, 1, 2, 5, 10, and 20 probes were set to be nonzero, respectively, under the six scenarios considered. The associated probes were randomly designed among the L CpG sites in the gene. DNA methylation data for 100 samples, 50 case samples and 50 control samples, were simulated. And seven different values of $L(=10, 20, 30, 40, 50, 60, 100)$ were considered. For single probe based approach, individual p-values were obtained for all probes in the gene, and a gene was claimed to be significant as long as at least one individual p-value was less than the preset level, as people normally would do in practice. Both single probe and gene based association test were applied to analyze the simulated data and empirical type I error rate (based on 1,000,000 replicates) and statistical power (based on 1000 replicates) were summarized.

Simulate gene set data
We simulated pseudo gene sets that contain M annotated genes. Three different cases considering various average number of probes per gene were investigated. Specifically, genes were ranked by their sizes and then M genes from the lower quartile, middle 50%, and upper quartile, respectively, were randomly picked to form a gene set. So that gene set simulated from the lower quartile/middle 50% has smaller mean number of probes per gene than a gene set selected from the middle 50%/the upper quartile. We denoted the three gene size level as lower, middle, and upper. And we considered six different values of M $(10, 20, 40, 100, 500, 1000)$ at each size level. Therefore, a total of 18 different settings for every scenario that we have considered. Empirical type I error rate and power were summarized based on 1,000 replicates.

Simulate the association identification probability
For simplicity, instead of generating genomewide DNA methylation data, we simulated the association identification probability (AIP) of every gene via a logistic regression model:

$$log\Big(\frac{\mu_g}{1-\mu_g}\Big) = \alpha_0 + \alpha_1 S_g + \alpha_2 log_{10}(L_g), g = 1, 2, \cdots, 20261 \qquad (6)$$

Here, 20261 was used in consistent with the total number of annotated genes in the real data set we used. Coefficient α_1 denotes the effect of a given gene set. By setting $\alpha_1 = 0$, the association identification probabilities were simulated under the null hypothesis of no gene set effect. Setting $\alpha_1 \neq 0$ indicates the gene set affects the AIPs, which implies over/under-representation of the gene set among the identified associated genes. We set $\alpha_1 = 1$ as the alternative in our simulation studies. Values of α_0, α_2 were set by mimicking the coefficient estimates we observed in a real data set. We considered two scenarios corresponding to (1) gene based association test; and (2) single probe based association analysis + "at least one probe in the gene is significant" criterion for aggregation. The corresponding α_2 values were set to be 1.03 and 3, respectively, under the two scenarios as shown in Table 1. Not surprisingly, the gene size effect was set to be more severe in the single probe case as what we observed in a real data analysis. For every single one of the 20261 annotated genes on the platform, the significance status was determined through a binomial distribution, with the probability of success set to be the association identification probability calculated via the logistic regression model (Equation (6)).

Table 1. Values of α (coefficients in Equation (6)) used in the simulation studies

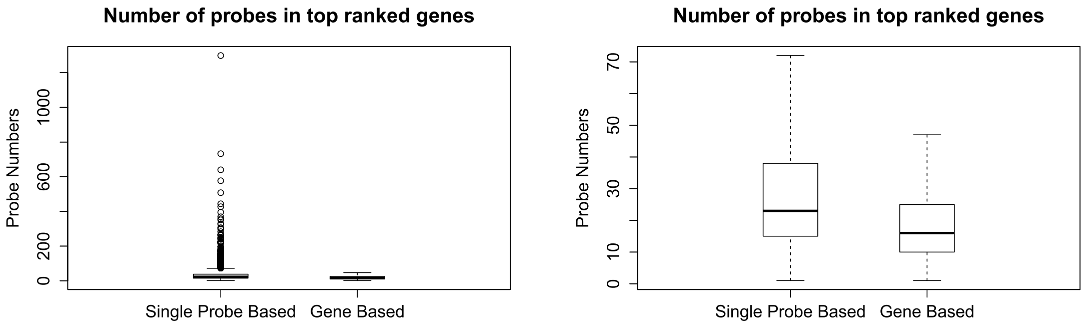
Scenarios	(1)	(2)
Null	(-3.28, 0, 1.03)	(-5.76, 0, 3)
Alternative	(-3.28, 1, 1.03)	(-5.76, 1, 3)

3.2 Real data analysis

We further applied the proposed method, MethylSet, to a data set from Gene Expression Omnibus (GSE29290) [17]. The data set contains Genome-wide DNA methylation profiling of HCT116 WT, HCT116 DNMT1 and DNMT3B double KO, and eight archival fresh frozen breast cancer tissue samples (BC) and eight normal breast tissue samples (N) by the Infinium Methylation 450K BeadChip. Detailed information about the samples and DNA methylation profiling can be found in an earlier report by Dedeurwaerder et al. [17]. In this study, we used only the data from the eight breast cancer tissues and the eight normal tissues, a total of sixteen samples. The raw intensity data, fluorescent signals of unmethylated and methylated denoted as U and M respectively, were used to calculate the β-values via

$$\beta = \frac{M}{M + U + 100} \qquad (7)$$

The obtained β values (methylation scores) for each CpG site ranged from 0 to 1 on a continuous scale. The constant offset of 100 added to the denominator was recommended by Illumina to regularize β-value when both methylated and unmethylated intensities are low. Data pre-processing were done using the Bioconductor package **ChAMP**. Probes from sex chromosomes were excluded from the analysis. We carried out gene set analysis using the GO and KEGG database. The bioconductor packages **org.Hs.eg.db** and **KEGG.db** were used to extract GO term and KEGG pathway information. GO terms with fewer than 10 genes were discarded from the analysis.



(a)

(b)

Figure 3. Boxplots of number of probes in top ranked genes identified by single probe and gene based methods with (left) and without (right) outliers.

We conducted both single probe based and gene based association analysis in the first step. For the single probe based analysis, we called genes with at least one significant probe as associated genes. For fair comparison, we chose a threshold that ended up with approximately the same number of significant genes as by the gene based method. As we expected, the list of significant genes called by the two methods were significantly different in terms of number of probes per gene (p-value $< 2.2 \times 10^{-16}$ by Wilcoxon Test; Box-plots of number of probes in top ranked genes by the two approaches (Figure 3) also indicates the difference). Less than half of the genes identified by the two methods overlapped. Therefore, only results from the gene based association analysis were used in the second-step for the enrichment analysis. The Fisher's Exact Test, GOseq, and the proposed MethylSet were applied for the purpose. The Fisher's exact test and the GOseq analysis were implemented by using the GOseq package (version 1.16.2) [7].

4. RESULTS

4.1 Simulation results

4.1.1 Comparing gene and single probe based association test

The empirical type I error rates for association test using gene and single probe based method were summarized as in Table (2). The single probe based method led to inflated type I error rate that bias towards longer genes. For example, at nominal level 1.0×10^{-5}, the empirical type I error rates obtained by single probe based test were 1.3×10^{-4} and 1.2×10^{-3}, respectively, when gene size was set to be 10 and 100. These numbers are roughly 10 and 100 times as large as the nominal level. On the other hand, the gene based method kept the type I error rate well controlled at various levels. However, when the gene size gets sufficiently large, we did see some inflation (2.2×10^{-5}). And that motivated the further correction of gene size in the second step.

Table 2. Empirical type I error rates for the single probe and gene based analysis at different nominal levels

Nominal Level (τ)	Single Probe Based		Gene Based	
	10	100	10	100
0.05	0.3875	0.9925	0.0518	0.0567
0.01	0.0946	0.6284	0.0107	0.0120
0.0001	0.0011	0.0108	0.00013	0.00016
1.0×10^{-5}	1.3×10^{-4}	1.2×10^{-3}	1.4×10^{-5}	2.2×10^{-5}

We also invested the empirical power of the kernel machine regression for testing association with different gene sizes and numbers of truly associated probes. The obtained results were shown as in Figure 4. When there was only one truly associated probe, the power decreases as the gene size increases. While, when the number of signals are proportional to the gene size, empirical power actually increases as the proportion of true signals increases, even for large genes. The fluctuated pattern shown in panel (III) and (IV) was due to we rounded up the number of truly associated probes to integer numbers. For example, 5% of 10, 20, and 30 are 0.5, 1, and 1.5, exactly. When simulated the data, we actually rounded up these numbers to be 1, 1, and 2, respectively. So, it is reasonable that the power for the case with 1 true signal out of 10 probes is higher than the case when 1 true signal out of 20 probes (panel (III) in Figure 4). The result was consistent with the results under scenario I.

For the effect of correlation between CpG sites on the empirical power, we saw consistent negative relationship between the correlation and power of the test under all circumstances. That is to say, when the probes in a gene are more positively correlated (larger ρ value), the test has lower statistical power.

However, we would like to point out that DNA methylation data was only simulated by assuming the AR(1) covariance structure, which could be different from the unknown true correlation pattern between DNA methylation

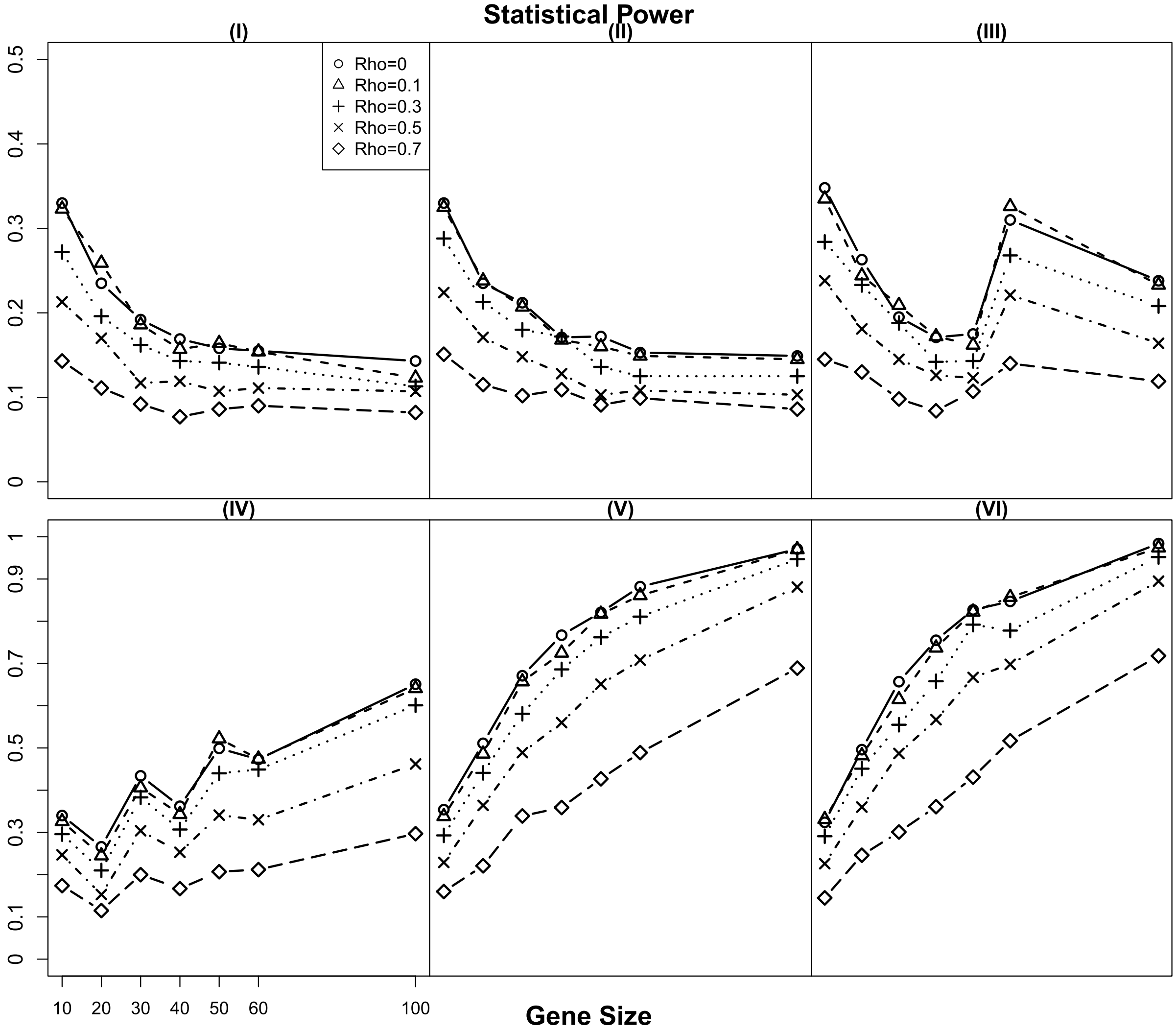


Figure 4. Empirical power of gene based association test. Scenarios considered were (I) 1, *(II)* 1%, *(III)* 2%, *(IV)* 5%, *(V)* 10%, *and (VI)* 20% *of probes in a gene are truly associated with the phenotypic outcome.*

probes. Besides, only linear kernel was used in our analysis. An "optimal" kernel which could reflect the true underlying similarity structure might have better performance. How to select the "optimal" kernel could be a very interesting topic for our future work, while it is beyond the scope of this paper.

4.1.2 Results for gene set analysis

The empirical type I error rates obtained under the two scenarios, (1) and (2), were summarized as in Figures 5–6. For scenario (2), where the AIPs were simulated to mimic the single probe based association test, the empirical type I error rates obtained by the FET and GOseq depend on gene size very much. Gene sets that contain longer genes (selected from the upper quartile) were more likely to be detected as significantly enriched. While, gene sets with genes selected from the lower quartile, the observed empirical type I error rates were conservative when there were only few genes in the set (<20), but became liberal when gene set size became bigger. For example, when the gene set size reached 1000, it was almost for sure that the gene set would be falsely detected as enriched by FET. Although less severe compared to that of the FET, the empirical type I error rates obtained by GOseq showed similar pattern: liberal for larger gene sets with longer genes and too conservative for smaller sets with shorter genes. Overall, besides the effect of gene size, we also observed the effect of gene set size affecting the empirical type I error rate for FET and GOseq.

While, under scenario (1), where the simulated AIPs mimic the gene based association analysis results, the inflation in type I error rates for the FET, especially GOseq reduced dramatically. However, deviations from the nominal level of 0.05 were still observed, especially when gene set size was large (Figure 6).

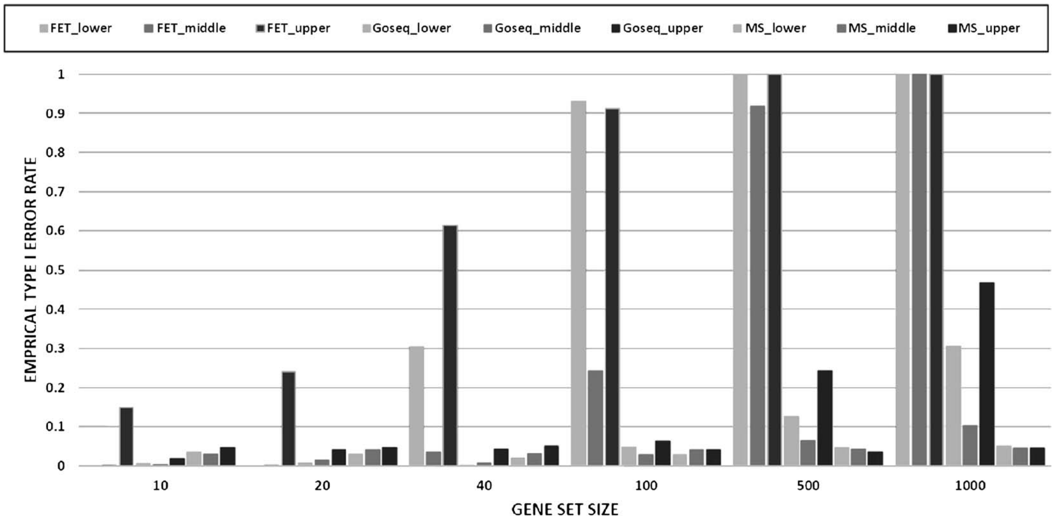


Figure 5. Empirical type I error rate ($\alpha = 0.05$) under scenario (2) (single probe based association analysis + enrichment analysis).

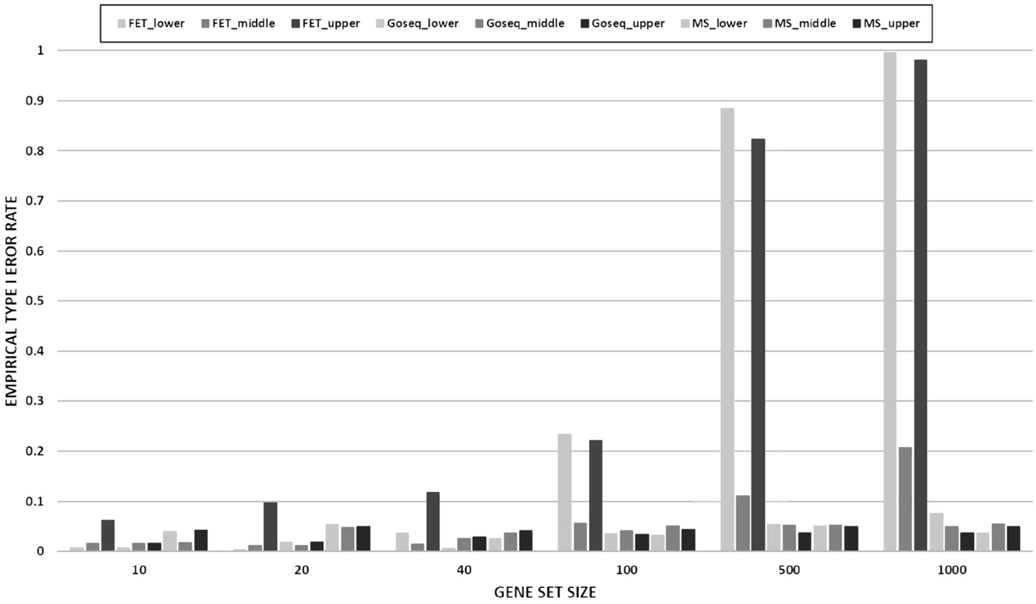


Figure 6. Empirical type I error rate ($\alpha = 0.05$) under scenario (1) (gene based association analysis + enrichment analysis).

In contrast, the empirical type I error rates for our proposed MethylSet procedure were well controlled around the nominal level 0.05 under all settings. Overall, our proposed MethylSet approach is robust to gene size and gene set size for enrichment analysis.

Only empirical power for GOseq and MethylSet were summarized and compared since FET led to so biased type I error rate as we have discussed in the previous session.

The empirical statistical power increases as gene set size increases for both methods (Figures 7–8). Generally, MethylSet obtained higher power than GOseq, especially where gene set size were relatively small, and comparable power when gene set size became large. When gene set size was 20, for example, the empirical powers were 0.433, 0.555, and 0.540 for MethlSet comparing to 0.26, 0.355, and 0.42 for GOseq under scenario (1), at the three gene size levels (lower, middle, upper) respectively. For scenario (2), the corresponding empirical powers were 0.207, 0.443, and 0.576 by MethlSet compared to 0.052, 0.297, and 0.559 for GOseq, at the three gene size levels (lower, middle, upper) respectively. The gene based analysis in the first step also helped boost the power for GOseq in the second step of analysis. The empirical powers obtained under scenario (1) were greater than the corresponding ones under scenario (2), which indicated

Table 3. List of significantly enriched KEGG pathways identified by MethylSet at level 0.05 *after multiple testing adjustment*

ID	PATHWAY NAME	p-value	Set Size	MPN
04080	Neuroactive ligand-receptor interaction	4.07×10^{-16}	253	19.43
04514	Cell adhesion molecules (CAMs)	2.58×10^{-6}	125	25.86
04740	Olfactory transduction	1.64×10^{-4}	372	5.22
04060	Cytokine-cytokine receptor interaction	1.71×10^{-4}	236	13.88
00970	Aminoacyl-tRNA biosynthesis	3.08×10^{-4}	41	22.51
04270	Vascular smooth muscle contraction	5.45×10^{-4}	113	25.73
00860	Porphyrin and chlorophyll metabolism	6.84×10^{-4}	41	15.98
04614	Renin-angiotensin system	8.75×10^{-4}	15	15.27
04120	Ubiquitin mediated proteolysis	1.80×10^{-3}	127	21.85
03022	Basal transcription factors	2.03×10^{-3}	32	21.00
00534	Glycosaminoglycan biosynthesis - heparan sulfate	2.04×10^{-3}	25	24.56
04974	Protein digestion and absorption	2.06×10^{-3}	75	31.33
05320	Autoimmune thyroid disease	2.21×10^{-3}	41	27.89
03010	Ribosome	2.47×10^{-3}	83	16.51
00512	Mucin type O-Glycan biosynthesis	2.71×10^{-3}	26	31.04
04020	Calcium signaling pathway	2.92×10^{-3}	166	27.89
03013	RNA transport	2.92×10^{-3}	133	17.50
05332	Graft-versus-host disease	3.10×10^{-3}	37	27.49

MPN: Mean Probe Number

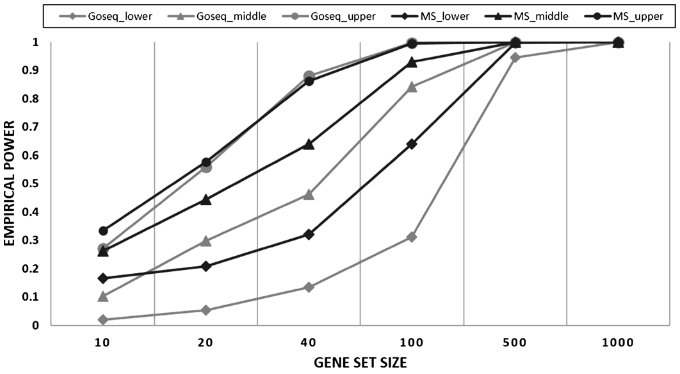


Figure 7. Empirical power under scenario (2), where analyses were conducted by single probe based association analysis + gene set enrichment analysis.

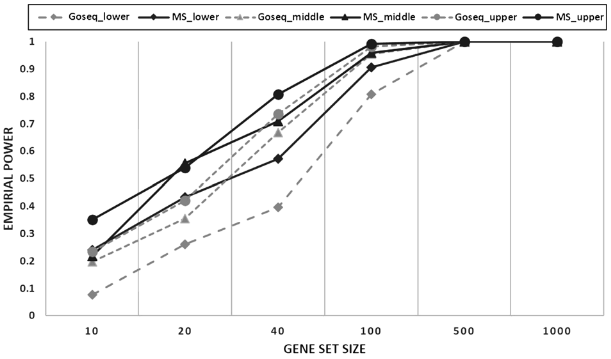


Figure 8. Empirical power under scenario (1), where analyses were conducted by gene based association analysis + gene set enrichment analysis.

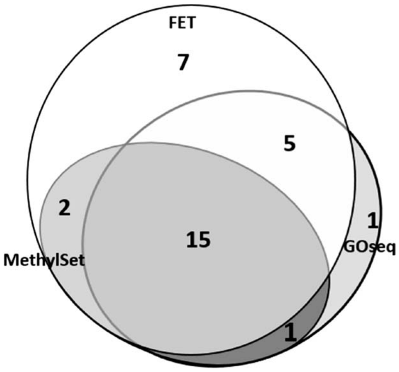


Figure 9. The Venn Diagram of number of enriched KEGG pathways identified by FET (white), GOseq (gray), and MethylSet (dark grey). Common pathways were indicated by the overlap parts.

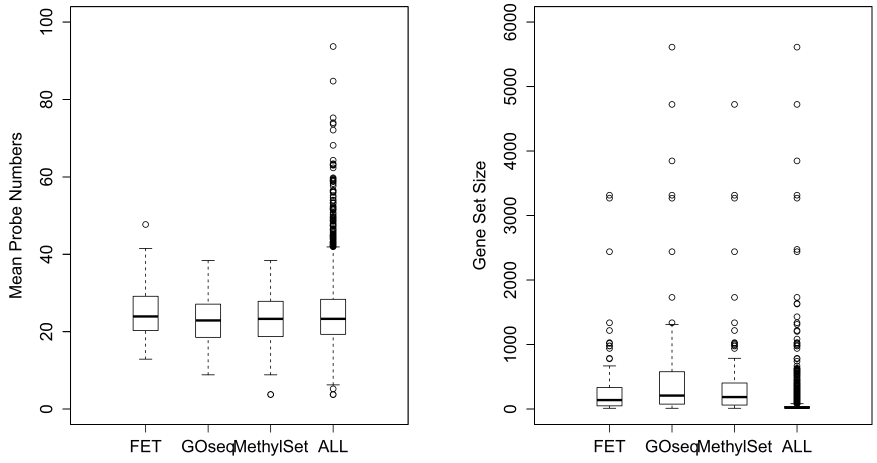
the advantage of gene based association study in terms of controlling gene size effect in gene set analysis.

To summarize, our simulation studies demonstrated good statistical properties of our proposed MethylSet method.

4.2 Real data analysis results

KEGG pathway enrichment analysis results

For the breast cancer data, MethylSet identified 18 enriched KEGG pathways at level $\alpha = 0.05$ (Table 3) after multiple testing correction using Benjamini and Hochberg's approach [18]. FET and GOseq identified 29 and 22 pathways, respectively, at the same significance level. The overlap between the results obtained by the three methods is shown in Figure 9. 15 pathways were detected by all the three ap-



(a)

(b)

Figure 10. Comparision of the GO term enrichment analysis. (a) Boxplot of the average number of probes per gene for GO terms identified by the FET, GOseq, MethylSet, and all GO terms. (b) Boxplot of the set size of GO terms identified by the FET, GOseq, MethylSet, and all GO terms.

proaches. Many of the KEGG pathways listed in Table 3 were reported in the literature to be related with human cancer or subtypes of breast cancer [19, 20, 21, 22]. In a recent study by Huan et al., neuroactive ligand-receptor interaction, cytokine-cytokine receptor interaction, and cell adhesion molecules (CAMs) pathway were found to show different gene expression in breast cancer cell line MCF-7 treated with 17β-estradiol [23]. The two pathways were at the top our list (Table 3). For those pathways identified by FET only (7 of the 29), the mean number of probes is significantly larger than other identified pathways ($p-value = 0.0018$ by Wilcoxon Rank Sum test).

GO term enrichment analysis results
2905 GO terms which contain at least 10 genes were used in our analysis. 102, 83, and 94 GO terms were identified by the FET, GOseq, and MethylSet, respectively, at level $\alpha = 0.01$ after multiple testing correction using Benjamini and Hochberg's approach. The average number of probes per gene of the identified GO terms were compared (Figure 10 (a)). The distribution of the average number of probes per gene for GO terms identified by FET is larger than that of GOseq and MethylSet. We also looked at the gene set size between the top ranked GO terms by FET, GOseq, and MethylSet (Figure 10 (b)). We compared the mean number of probes per gene of pathways identified only by FET and those overlapped with at least one other method, significantly larger mean probe number per gene were observed in pathways identified by FET only ($p-value = 6.96 \times 10^{-6}$, by Wilcoxon Rank Sum test). Similar analysis for comparing the gene set size of GO terms identified by GOseq and others also revealed significant difference ($p-value = 0.002841$ by Wilcoxon Rank Sum test). These results confirmed what we have observed in the simulation studies.

5. DISCUSSION

In this study we proposed a two-step procedure, MethylSet, for gene set analysis especially for DNA methylation data. To correct the bias caused by calling significantly associated genes via single probe based test followed by a post-hoc aggregation procedure, the kernel machine regression method was applied for gene based association analysis. DNA methylation measurements within a gene were considered simultaneously as a unit to study their joint effect on the disease outcome. We then tested enrichment of predefined gene sets using traditional logistic regression model, however, the gene size effect were further adjusted by incorporating gene size as a covariate in the model. Simulation studies shown the merits of MethylSet in term of correcting bias caused by gene size. Beside, MethylSet achieved higher power than GOseq, which is a widely used method in GSA for RNA sequencing data analysis nowadays.

Since our focus was gene set analysis in this study, we emphasized on gene based analysis association study to have significance results for all annotated genes ready for the use in the enrichment analysis. However, if other genomics features are of interest, for example, methylation island, the kernel machine frame work can be generalized without too much difficulty.

We would also want to point out that the sensitivity of different kernel functions to gene size varies. Some kernel functions could be affected by the gene size and make the test conservative for longer genes (e.g. quadratic kernel). The linear kernel used in our analysis led to well controlled type I error rates, but not necessary the best kernel function for EWAS. Therefore, the selection of an "optimal" kernel function which can represent the true function of the gene and is invariant to gene size is an interesting area and worth to be investigated for genomewise DNA methylation data analysis. There are other existing challenges in epigenome-wide association study, such as batch effect [24] and cellular heterogeneity issue [25, 26, 27]. It remains a very active research area where researchers are making efforts to [28]. It is for sure that more accurate results from the association analysis would reduce the false discovery rate in the enrichment analysis.

Despite the challenges, many EWASs are already underway. And we believe that these studies will shed new lights on the causes of human diseases. Especially, could be combined with other types of data, such as gene expression, genotype, and microRNA expression, for a better understanding of the disease etiology and the potential development of novel therapeutics and diagnostic in the future.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no competing interests.

ACKNOWLEDGMENTS

The authors thank two reviewers for their constructive comments, which helped us improve the manuscript.

Received 18 September 2015

REFERENCES

[1] Y. Liu, M. Aryee, L. Padyukov, M. Fallin, E. Hesselberg, A. Runarsson, L. Reinius, N. Acevedo, M. Taub, M. Ronninger, K. Shchetynsky, A. Scheynius, J. Kere, L. Alfredsson, L. Klareskog, T. Ekström, and A. Feinberg, "Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis," *Nat Biotechnol.*, vol. 31, pp. 142–147, Feb 2013.

[2] V. Rakyan, H. Beyan, T. Down, M. Hawa, S. Maslau, D. Aden, A. Daunay, F. Busato, C. Mein, B. Manfras, K. Dias, C. Bell, J. Tost, B. Boehm, S. Beck, and R. Leslie, "Identification of type 1 diabetes-associated DNA methylation variable positions that precede disease diagnosis," *PLoS Genet.*, vol. 7, p. e1002300, Sep 2011.

[3] K. Dick, C. Nelson, L. Tsaprouni, J. Sandling, D. Aïssi, S. Wahl, E. Meduri, P. Morange, F. Gagnon, H. Grallert, M. Waldenberger, A. Peters, J. Erdmann, C. Hengstenberg, F. Cambien, A. Goodall, W. Ouwehand, H. Schunkert, J. Thompson, T. Spector, C. Gieger, D. Trégouët, P. Deloukas, and N. Samani, "DNA methylation and body-mass index: a genome-wide analysis," *Lancet.*, vol. 383, pp. 1990–8, Jun 2014.

[4] J. Hass, E. Walton, C. Wright, A. Beyer, M. Scholz, J. Turner, J. Liu, M. Smolka, V. Roessner, S. Sponheim, R. Gollub, V. Calhoun, and S. Ehrlich, "Associations between DNA methylation and schizophrenia-related intermediate phenotypes – a gene set enrichment analysis," *Prog Neuropsychopharmacol Biol Psychiatry.*, vol. 59, pp. 31–9, Jun 2015.

[5] T. Dunwell, L. Hesson, T. A. Rauch, L. Wang, R. E. Clark, A. Dallol, D. Gentle, D. Catchpoole, E. R. Maher, G. P. Pfeifer, and F. Latif, "A genomewide screen identifies frequently methylated genes in haematological and epithelial cancers," *Mol Cancer*, vol. 9, doi:10.1186/1476-4598-9-44, Feb 2010.

[6] S. Kalari, M. Jung, K. H. Kernstine, T. Takahashi, and G. P. Pfeifer, "The dna methylation landscape of small cell lung cancer suggests a differentiation defect of neuroendocrine cells," *Oncogene*, vol. 32, pp. 3359–68, Jul 2012.

[7] M. Young, M. Wakefield, G. Smyth, and A. Oshlack, "Gene ontology analysis for RNA-seq: accounting for selection bias," *Genome Biol*, vol. 11, p. R14, Feb 2010.

[8] L. Gao, Z. Fang, K. Zhang, D. Zhi, and X. Cui, "Length bias correction for RNA-seq data in gene set analyses," *Bioinformatics*, vol. 27, pp. 662–9, Mar 2011.

[9] G. Mi, D. Y, S. Emerson, J. S. Cumbie, and J. H. Chang, "Length bias correction in gene ontology enrichment analysis using logistic regression," *PLoS One*, vol. 7, p. e46128, Oct 2012.

[10] P. Jia, L. Wang, A. Fanous, X. Chen, K. Kendler, I. S. Consortium, and Z. Zhao, "A bias-reducing pathway enrichment analysis of genome-wide association data confirmed association of the MHC region with schizophrenia," *J Med Genet*, vol. 49, pp. 96–103, Feb 2012.

[11] R. Welch, C. Lee, P. Imbriano, S. Patil, T. Weymouth, R. Smith, L. Scott, and M. Sartor, "ChIP-Enrich: gene set enrichment testing for ChIP-seq data," *Nucleic Acids Res.*, vol. 42, p. e105, Jul 2014.

[12] Q. Yan, H. Tiwari, N. Yi, W. Lin, G. Gao, X. Lou, X. Cui, and N. Liu, "Kernel-machine testing coupled with a rank-truncation method for genetic pathway analysis," *Genet. Epi.*, vol. 38, pp. 447–456, May 2014.

[13] P. Geeleher, L. Hartnett, L. Egan, A. Golden, A. R. Raja, and S. C, "Gene-set analysis is severely biased when applied to genome-wide methylation data," *Bioinformatics*, vol. 29, pp. 1851–7, Aug 2013.

[14] D. Liu, X. Lin, and D. Ghosh, "Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models," *Biometrics*, vol. 63, pp. 1079–1088, Dec 2007. MR2414585

[15] D. Liu, D. Ghosh, and X. Lin, "Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models," *BMC Bioinformatics*, vol. 9, doi:10.1186/1471–2105–9–292, Jun 2008.

[16] M. Wu, A. Maity, S. Lee, E. Simmons, Q. Harmon, X. Lin, S. Engel, J. Molldrem, and P. Armistead, "Kernel machine SNP-set testing under multiple candidate kernels," *Genet Epidemiol.*, vol. 37, pp. 267–75, Apr 2013.

[17] S. Dedeurwaerder, M. Defrance, E. Calonne, H. Denis, C. Sotiriou, and F. Fuks, "Evaluation of the infinium methylation 450K technology," *Epigenomics.*, vol. 3, pp. 771–84, Dec 2011.

[18] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society Series B*, vol. 57, no. 1, pp. 289–300, 1995. MR1325392

[19] L. Lanzetti and F. P. Di, "Endocytosis and cancer: an 'insider' network with dangerous liaisons," *Traffic.*, vol. 9, pp. 2011–21, Dec 2008.

[20] G. Vinson, S. Barker, and J. Puddefoot, "The renin-angiotensin system in the breast and breast cancer," *Endocr Relat Cancer*, vol. 19, pp. R1–19, Feb 2012.

[21] L. Fishchuk and N. Gorovenko, "Genetic polymorphisms of the renin-angiotensin system in breast cancer patients," *Exp Oncol*, vol. 35, pp. 101–4, Jun 2013.

[22] F. Tian, Y. Wang, M. Seiler, and Z. Hu, "Functional characterization of breast cancer using pathway profiles," *BMC Med Genomics*, vol. 7, doi:10.1186/1755-8794-7-45, Jul 2014.

[23] J. Huan, L. Wang, L. Xing, X. Qin, L. Feng, X. Pan, and L. Zhu, "Insights into significant pathways and gene interaction networks underlying breast cancer cell line MCF-7 treated with 17β-estradiol (E2)," *Gene.*, vol. 533, pp. 346–55, Jan 2014.

[24] A. E. Teschendorff, U. Menon, A. Gentry-Maharaj, S. J. Ramus, S. A. Gayther, S. Apostolidou, A. Jones, M. Lech-

ner, S. Beck, I. J. Jacobs, and M. Widschwendter, "An epigenetic signature in peripheral blood predicts active ovarian cancer," *PLoS One*, vol. 4, p. e8274, Dec 2009.

[25] A. Jaffe and R. A. Irizarry, "Accounting for cellular heterogeneity is critical in epigenome-wide association studies," *Genome Biology*, vol. 15, doi:10.1186/gb-2014-15-2-r31, Feb 2014.

[26] L. Liang and W. Cookson, "Grasping nettles: cellular heterogeneity and other confounders in epigenome-wide association studies," *Hum. Mol. Genet.*, doi:10.1093/hmg/ddu284, Jun 2014.

[27] J. Zou, C. Lippert, D. Heckerman, M. Aryee, and J. Listgarten, "Epigenome-wide association studies without the need for cell-type composition," *Nat Methods*, vol. 11, pp. 309–11, Mar 2014.

[28] S. D. Paul and S. Beck, "Advances in epigenome-wide association studies for common diseases," *Trends Mol Med.*, vol. 20, pp. 541–543, Oct 2014.

Shaoyu Li
Department of Mathematics and Statistics
University of North Carolina at Charlotte
Charlotte, NC 28270
USA
E-mail address: sli23@uncc.edu

Tao He
Department of Mathematics
San Francisco State University
San Francisco, CA 94132
USA
E-mail address: hetao@sfsu.edu

Iwona Pawlikowska
Department of Biostatistics
St. Jude Children's Research Hospital
Memphis, TN 38105
USA
E-mail address: iwona.pawlikowska@stjude.org

Tong Lin
Department of Biostatistics
St. Jude Children's Research Hospital
Memphis, TN 38105
USA
E-mail address: tong.lin@stjude.org