RESEARCH ARTICLE

Genetic Epidemiology
OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

A Novel Method for Detecting Association Between DNA Methylation and Diseases Using Spatial Information

Wai-Ki Yip,[1]* Heide Fier,[2] Dawn L. DeMeo,[3] Martin Aryee,[4] Nan Laird,[5] and Christoph Lange[5]

[1]*Harvard University, Boston, Massachusetts, United States of America;* [2]*Department of Genomic Mathematics, University of Bonn, Bonn, Germany;* [3]*Channing Division of Network Medicine, Brigham and Women Hospital, Boston, Massachusetts, United States of America;* [4]*Massachusetts General Hospital, Boston, Massachusetts, United States of America;* [5]*Harvard School of Public Health, Boston, Massachusetts, United States of America*

Received 10 June 2014; Revised 19 July 2014; accepted revised manuscript 29 July 2014.
Published online in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21851

ABSTRACT: DNA methylation may represent an important contributor to the missing heritability described in complex trait genetics. However, technology to measure DNA methylation has outpaced statistical methods for analysis. Taking advantage of the recent finding that methylated sites cluster together, we propose a Spatial Clustering Method (SCM) to detect differentially methylated regions (DMRs) in the genome in case and control studies using spatial location information. This new method compares the distribution of distances in cases and controls between DNA methylation marks in the genomic region of interest. A statistic is computed based on these distances. Proper type I error rate is maintained and statistical significance is evaluated using permutation test. The effectiveness of the SCM we propose is evaluated by a simulation study. By simulating a simple disease model, we demonstrate that SCM has good power to detect DMRs associated with the disease. Finally, we applied the SCM to an exploratory analysis of chromosome 14 from a colorectal cancer data set and identified statistically significant genomic regions. Identification of these regions should lead to a better understanding of methylated sites and their contribution to disease. The SCM can be used as a reliable statistical method for the identification of DMRs associated with disease states in exploratory epigenetic analyses.
Genet Epidemiol 00:1–8, 2014. © 2014 Wiley Periodicals, Inc.

KEY WORDS: DNA methylation; spatial analysis; genetic analysis; Genome-Wide Association Test (GWAS); differentially methylated regions (DMRs)

Introduction

For the past decade, the Genome-Wide Association Test (GWAS) for population-based genetic analysis has identified many genetic factors associated with complex diseases [Manolio et al., 2008]. However, DNA polymorphisms have explained only a portion of inheritance patterns in many of these diseases, leaving a large amount of heritability still to be accounted for [Maher, 2008]. Some of the missing heritability might be explained by epigenetic mechanisms, such as DNA methylation, that control gene expression by means other than changes to the DNA sequence [Manolio et al., 2009]. Most human DNA methylation marks are likely stable, making them useful as disease biomarkers. Analyzing data from cancer patients, we observe a widespread disruption of the human DNA methylation profile. The human DNA methylation profile plays a role in the etiology of numerous other complex diseases including asthma, coronary heart disease, and bipolar disorder [Foley et al., 2009]. However, there is currently no "gold-standard" statistical approach to identify human DNA methylation profiles that are associated with diseases. The need for better analytical methods for studying DNA methylation has grown, as recent advances have made large-scale measurements much more tractable.

Recent advances in biomedical technology make it possible to perform large-scale measurements of DNA methylation across the human genome. As a result, the methylated state of each CpG site and its location in the human genome can be identified [Laird, 2010]. Thus, some of the methods used to detect disease susceptibility loci for genetic variants can be applied to detect disease susceptibility CpG sites. However, there are numerous challenges in analyzing DNA methylation data: measurement errors and batch effects are common for DNA methylation data, methylation effects may not be captured completely if surrogate tissue samples are used, and contamination from cellular heterogeneity in samples has led to spurious conclusions. Furthermore, methylation levels are clustered [Hackenberg, 2010]. If we can take advantage of distance information, we can devise a more powerful test to detect association.

Supporting Information is available in the online issue at wileyonlinelibrary.com.

*Correspondence to: Wai-Ki Yip, 855 Huntington Avenue, Building II, 4th Floor, BioStatistics Department, Boston, MA 02115, USA. E-mail: wkyip@hsph.harvard.edu

Instead of trying to identify CpG sites individually, we propose screening the genome for potential regions worthy of detailed analysis by utilizing spatial location information of CpG sites. Distances between CpG sites, measured in DNA base pairs, provide us with crucial information about where DNA methylation events occur. There are two reasons why this information is important. First, CpG sites tend to cluster together in promoter regions that affect gene expression. Second, studies have shown gene expression levels are associated with genomic regions of variable methylation instead of a single CpG site. For example, the level of methylation in the promoter region of a gene may be inversely associated with the level of gene expression of that gene [Eckhardt et al., 2006; Hansen et al., 2011; Lister et al., 2009]. Jaffe et al. proposed a comprehensive framework for a DNA methylation analysis pipeline [Jaffe et al., 2012]. In their approach, the user specifies a statistical model. Our method is similar to Jaffe's method in that we are "bumping hunting" along the genome with the goal of detecting differentially methylated regions (DMRs). The main difference is that we are using specific spatial information to come up with a robust statistic.

Our proposed approach is adapted from an algorithm which successfully detects rare genetic variants associated with diseases using spatial information [Fier et al., 2012]. The algorithm is based on genetic distances between polymorphic single nucleotide polymorphism variants and their distribution. We believe the distributions of genetic distances between CpG sites are similar. In epidemiology, identification and quantification of patterns in disease occurrence provide the first steps toward increased understanding of that particular disease. As stated, "The location where an event happens may provide some indication as to why that particular event occurs. Thus, spatial statistical methods offer a means for researchers to use locational information to detect and quantify patterns in public health data, and to investigate the degree of association between potential risk factors and diseases" [Waller and Gotway, 2004]. Spatial location information for CpG sites may be the additional data needed for a robust statistic for DNA methylation.

We call our method the Spatial Clustering Method (SCM). By creating a statistic that incorporates spatial location information and methylation measurement at each CpG site, our method can identify DMRs in the genome that are candidates for association with diseases. Since the location is measured from the beginning of each chromosome, the analysis is restricted to one chromosome at a time. In the following sections, we describe the SCM in detail and highlight its potential. We evaluate the SCM with a simulation study demonstrating that it has good power to detect DMRs while maintaining the prespecified type I error rate. To test our approach, we apply the SCM to a publicly available clinical data set with colorectal cancer as the phenotype. In doing this, we identify significant DMRs of the human genome potentially linked to this disease. By reliably and specifically identifying methylated regions associated with disease, rather than individual sites, SCM may facilitate the study of causal associations between methylation and disease.

Methods

Conceptual Development

By observing patterns of genetic events in our genome, we can identify and quantify patterns in disease occurrence. We adapt some of the statistical methods developed for spatial analysis to help our investigation of association between potential genetics factors and diseases. There are two categories of data in spatial data analysis: feature data consisting of geographical information that uniquely identifies the location where events occur, and attribute data consisting of counts or measured information about the events. To frame genetic investigation into the spatial data analysis framework, we need a "map" with relevant feature information including location and distance. Then, genetic events can be identified together with the location of occurrence and related with one another based on topological constraints.

In our model, CpG methylation marks can be viewed as spatial information arranged as points along the chromosome when it is stretched into a line. Thus, an analysis that explicitly uses spatial information can be very informative. In particular, we wish to detect whether the set of locations observed contains clusters of events reflecting areas with associated increases in the likelihood of occurrence (e.g., unusual aggregations of cases of a particular disease).

Array-based platforms to assay DNA methylation result in a percent methylation at a single CpG site; for our method we need to transform these percent methylation values to methylation units. The transformation is done by using a weight for each CpG site, developed based on the percent methylation values of controls and the site's distance from the nearest neighboring CpG site. The weight is then applied to each CpG site transforming the percent methylation value to methylation units for that site regardless of whether it is a case or control. Each methylation unit is counted as an occurrence of an event.

With the basic entities defined, the next step is to characterize location patterns. The traditional spatial descriptions of the location patterns such as the intensity functions or the K-functions are based on geographical map information and may not be suitable for genetic analysis. So, we adapted an idea that has been applied successfully for the development of a nonparametric association test for genetic rare variants to our method [Fier, 2012]. The location pattern is characterized by the distance distribution between each consecutive event on the line. A distance vector is a vector of distance between events from all the methylation measurements in the genome region of interest. In our current adaptation of the rare variant method, we create two distance vectors—one is for methylation units from the cases and one for methylation units from controls. These two distance vectors represent samples from the distance distribution for events in case and control groups.

With case-control point data, we can compare the control locations which provide baseline information on spatial patterns of the population at risk and the case locations which provide spatial patterns of the disease. As done in many other epidemiology studies, an association test that identifies clusters of CpG sites and examines their association with diseases can now be constructed based on differences between the control and the case distance distributions. The idea is to develop a test statistic that captures the differences between these two distance distribution functions. If the differences are statistically significant, we can reject the null hypothesis that distributions are the same, and conclude that the genomic region is associated with the disease phenotype.

With the distance vectors developed, we can use the Ansari-Bradley statistic to test if there is a difference in dispersion [Ansari-Bradley, 1960]. Since the choice of creating the weight based on the control group is arbitrary, we repeat the same procedure using a weighting scheme based on percent methylation values and nearest neighboring CpG site from cases. The final statistics is computed as the maximum of the two derived statistics. Since we assumed that the CpG sites are clustered with the same association effect in the same genomic region, the differences between the distribution functions are magnified for the small genomic distances. Therefore, the test captures information for both distances and methylation association effects.

As in most other spatial tests, the distribution of the difference in distribution functions does not have a closed analytic form. Thus, the statistical significance of the proposed test statistic is obtained by Monte Carlo/permutation method that randomly assigns case/control status to the study population while maintaining the total number of cases and controls and the DNA methylation structure of the genomic region of interest.

For exploratory analysis, we can set up a genomic window which contains a fixed number of CpG sites. Then, we scan each of the chromosomes from the beginning to the end using a sliding genomic window to look for regions that are significantly different for further analyses. However, it is important to apply multiple testing corrections for this approach [Kuan et al., 2012].

Detailed Algorithm

In order to explain our idea fully, we present the detailed algorithm using the following notation.

We assume that a defined genomic region has been assayed or sequenced in N subjects in the context of a case-control study, recording the physical positions in DNA base pairs from the beginning of each chromosome of all the CpG sites in the region, and the methylation levels at each site for all subjects. Assume that there are a total of K CpG sites under consideration. The sorted position vector for the CpG sites is denoted by $P = (p_1, \dots, p_i, \dots, p_K)$ and index by $1 \le i \le K$. The subjects in the study are indexed by $1 \le j \le N$. To identify cases and controls, we define the indicator functions $I_{case}(s_j) = 1$ if the subject s_j is from the case group and =0 otherwise and $I_{control}(s_j) = 1$ if the subject s_j is from the control group and =0 otherwise. The methylation signal vector corresponding to the position vector, P, for subject, j, and is denoted by $M_j = (m_{1,j}, \dots, m_{i,j}, \dots, m_{K,j})$.

Computing the Test Statistic

First, we developed a weight for each CpG site based on information obtained from the control group. We denoted the weight based on the control for the ith CpG site by $w_{0,i}$. The derivation of the formula for the weights is explained later in the section. By applying this weight to each CpG site, we transform the percent methylation value for cases and control into methylation units for the ith CpG site denoted by $u_{case\ or\ control,i}$ given by taking the largest integer from the following formula:

$$u_{case,i} = w_{0,i} \times \sum_j \left(m_{i,j} \times I_{case}(s_j)\right) \text{ and}$$
$$u_{control,i} = w_{0,i} \times \sum_j \left(m_{i,j} \times I_{control}(s_j)\right). \quad (1)$$

Since each methylation unit is considered as an occurrence of an event, we have an ordering of events based on the position order of the CpG sites. Next, we create a sequence of distances, du_{case}, between each event for cases and a sequence of distances, $du_{control}$, between each event for controls. Each of these two sequences of distances can be treated as samples from the corresponding distance distribution.

$$du_{case} = \{\dots, \text{ distance in base - pair between the } i\text{th event and } i+1\text{th event for cases}, \dots\}, \text{ and}$$
$$du_{control} = \{\dots, \text{ distance in base - pair between the } i\text{th event and } i+1\text{th event for control}, \dots\}. \quad (2)$$

The applied weighting schemes only influence the skewness of the derived distance distribution functions, and have no impact on the values of the observed nonzero distances. The null and alternative hypotheses are:

H_0 : *the distribution functions of cases and controls are the same (Or no association of methylation pattern and disease)*

H_A : *they are different* (3)

To test our hypothesis, we used the Ansari-Bradley test which is a nonparametric, two-sample test on the variability of the distance distribution functions. If the distribution functions were different, we expect to see one of the samples to be more "spread-out" than the other.

We applied the test to our two samples, du_{case} and $du_{control}$. Since the samples are obtained using a weighing scheme based on subjects from the control group, we denoted the resulting Ansari-Bradley statistics by $A_{control}$.

The weighting scheme described above was based on subjects from the control group. We also derived the weight based on subjects from the case group. Using the same notation, we obtained the weight based on the cases for the ith CpG site by $w_{1,i}$ and transformed the percent methylation values to methylation units for the *i*th CpG site denoted by v_i :

$$v_{case,i} = w_{1,i} * \sum_j (m_{i,j} * I_{case}(s_j)) \text{ and}$$

$$v_{control,i} = w_{1,i} * \sum_j (m_{i,j} * I_{control}(s_j)). \quad (4)$$

With the implicit ordering of events based on CpG site positions, we created a sequence of distances, dv_{case} , between each event for cases and a sequence of distances, $dv_{control}$, between each event for controls.

$$dv_{case} = \{\ldots, \text{distance in base - pair between the } i\text{thevent and } i + 1\text{thevent for cases}, \ldots\}, \text{ and}$$

$$dv_{control} = \{\ldots, \text{distance in base - pair between the } i\text{thevent and } i + 1\text{thevent for control}, \ldots\}. \quad (5)$$

Similarly, we applied the Ansari-Bradley test and obtain the statistics, A_{cases}. This tests for the differences between the distance distribution functions for cases and controls when weighted by the subjects from cases.

The final statistic is the maximum of the two:

$$A_{final} = \max\left(A_{cases},\ A_{control}\right). \quad (6)$$

A rejection of the null hypothesis implies an association of the tested methylation sites in the region with affection status. Because of the likely presence of tied observations in the distance distribution functions, we used an implementation of the standard Streitberg/Roehmel shift algorithm to retrieve exact distribution functions for our test statistics [Streitberg-Roehmel, 1986]. To obtain the significance of the test statistics, we used permutation testing. For each specified genomic region, case/control status was randomly assigned to each individual so that the total number of cases and controls of the original study is maintained. As a result, the methylation structure of the study-population samples was kept constant. The *P*-value was estimated as the proportion of permutation test statistics which were more "extreme" than the actual observed test statistic for the data.

Computing the Weight

In order to incorporate the importance of spatial proximity between CpG sites and to control for an uneven distribution of the percent methylation values, we defined weights that depend on both cluster and percent methylation values. We took the distance to the nearest neighbor for each of the CpG sites and the shortest distance vector, $D = \{d_i\}$ for all $2 \le I \le K - 1$ where

$$d_i = \min\left(|p_i - p_{i-1}|, |p_i - p_{i+1}|\right)$$

$$d_1 = |p_2 - p_1|$$

$$d_K = |p_K = p_{K-1}|. \quad (7)$$

Then, we computed the mean percent methylation value, ms_{case}, of all subjects from the case group and mean percent methylation value, $ms_{control}$, of subjects from the control group separately:

$$ms_{case} = \sum_i \sum_j \left(m_{i,j} * I_{case}\left(s_j\right)\right) / \sum_i \sum_j \left(I_{case}\left(s_j\right)\right)$$

$$ms_{control} = \sum_i \sum_j \left(m_{i,j} * I_{control}\left(s_j\right)\right) / \sum_i \sum_j \left(I_{control}\left(s_j\right)\right). \quad (8)$$

We standardized each methylation site's percent methylation value with respect to the mean of the cases. Our weighting scheme for the *i*th CpG site based on the subjects from cases was given by

$$w_{1,i} = 1 + \left(\left[ms_{case}\right] \Big/ \left[\sum_j \left(m_{i,j} * I_{case}\left(s_j\right) + 1\right]\right) \Big/ \log\left(d_i + 1\right), \text{ and}$$

based on the subjects from controls, was given by

$$w_{0,i} = 1 + \left(\left[ms_{control}\right] \Big/ \left[\sum_j \left(m_{i,j} * I_{control}\left(s_j\right) + 1\right]\right) \Big/ \log\left(d_i + 1\right). \quad (9)$$

The weights are created to more highly weight close by CpG sites and lower methylation values.

Simulation Study

In order to evaluate the proposed SCM, we simulated the null population where the disease phenotype was independent of the percent methylation value at each CpG site. Then, we simulated the disease populations where certain sites were selected to associate with the disease status. The proposed test statistic was computed on 1,000 simulated samples so that we could evaluate both the type I error rate and examine the power of the test to detect associations with the disease phenotype.

For exploratory purposes, we varied the size of the genomic region of interest centered on the chosen CpG sites in our simulation study. For most of our reported results, we used a genomic region of interest of 81 consecutive sites, denoted by *K*, centered with the chosen CpG site. For this simulation study, we used the es_ICGN data set, which contains 325 controls and 620 cases with chronic obstructive lung disease (COPD). The es_ICGN data set contains data collected for the International COPD Genetics Network with 1,085 Caucasian subjects using the Illumina infinium27K beadchip [Qiu et al., 2012]. Es_ICGN is a family-based study of subjects with ages between 45 and 65 with at least 5 pack-years of cigarette smoking. In our analysis, we treated all subjects in the data set as independent cases and all unaffected siblings as controls. This should have negligible impact on our simulation study. The data set has 26,486 CpG sites after data cleaning and a total of 1,085 subjects. We only used data from the 945 subjects with clean data for our analysis.

Simulation of the Null Population

Currently, there is no standard paradigm to simulate methylation marks in the human genome. In order to simulate the null population, we used the following approach. A random small genomic region might provide the null region (region that is not associated with the phenotype) that we were looking for. Using the methylation measurements for a specific CpG site for all subjects, we approximated the distribution of the methylation percent values for that site.

We picked three CpG sites from chromosomes from the beginning, middle, and the end of the human genome (the 1,300th CpG site of chromosome 1, the 200th CpG site of chromosome 14, and the 800th CpG site of chromosome 19) to be the center of genomic region of interest for our study. We simulated samples using a normal distribution $N(\mu, \sigma^2)$ with μ and σ^2 estimated from the data and then truncated any values outside of 0 and 1. Thus, the active range is much smaller than [0, 1]. The es_ICGN data set provided the genomic positional information and the empirical distribution of the percent methylation values of all the CpG sites.

Using this approach, we generated a methylation percent value for each subject. This was applied to all the sites in the genomic region that we were interested in. Since the CpG sites are relatively sparse in this data set, we treated sites within subjects as independent. Since the null model stipulates that there was no effect of methylation on the disease outcome, in order to achieve maximum power, we randomly allocated the 945 subjects to controls (473) and cases (472). We generated 1,000 replicates of each null sample.

Simulation of the Disease Population

There is also no standard approach on how to simulate methylation marks on the DNA molecule that are associated with the disease. To evaluate our test statistics, we chose a simple disease model. By assuming that the mean of the percent methylation values of CpG sites selected as associated with the disease is shifted by the same constant percentage of their standard deviation (SD) from the null, we simulated methylation data for case subjects based on the null distribution with the mean shifted upward. Since the SCM detects the difference in methylation level, this is the same as shifting the mean upward for control subjects.

Following the same convention established in the null population simulation, we considered a genomic region with K number of CpG sites without knowing in advance how many of the sites are associated with the disease and their effects. Thus, the number of disease-related CpG sites and their effects are being varied in this simulation and are controlled by the following parameters: D denotes the number of disease-related CpG sites in the genomic region of interest while S denotes the percentage of SD shift in the mean of cases from the null distribution (assumed to be the same for all disease-related CpG sites).

To simplify the simulation, we assumed the disease-related CpG sites were adjacent to one another and at the center of the genomic region of interest while the rest of the CpG sites under investigation were not associated with the disease. This is designed for maximum power. Future research can be done to test against other alternatives with disease-related CpG sites not centered or not adjacent to one another. We first generated the null percent methylation values using null empirical distributions based on the estimates from the es_ICGN data set for all K CpG sites. To obtain our case sample, we replaced the percent methylation value for the selected disease site for each subject from cases by percent methylation value generated using a distribution that was based on the null empirical distribution but with a varying shift in the mean which was S percent of the SD. We repeated the same procedure for all the other disease-related CpG sites, where the percent methylation values at each site were independently selected. Similar to our null population simulation, we generated 1,000 replicates of each case sample.

Application to a Colorectal Cancer Data Set

To illustrate the feasibility of the SCM to real data, we applied it to an exploratory analysis of a colorectal data set, which is publicly available from the Cancer Genome Atlas (TCGA) website [TCGA Network, 2012]. The data were processed using Illumina Infinium 450K chip. There are a total of 329 samples from cancer patients. To avoid subject heterogeneity, we used only 76 matched samples from the same patients—one sample from solid tumor and a corresponding sample from normal tissue. For simplicity and illustrative purposes, we treated them as independent samples. The SCM can be extended to handle matched pair but is beyond the scope of this study. These samples were processed as three separate batches. We used the ComBat function in the R package sva [Leek et al., 2012] to correct for batch effect.

To illustrate our method, we applied it to all the methylation marks (methylome) on chromosome 14 in our data set using a fixed sliding window approach. Each sliding window, representing a genomic region of interest, contained W CpG sites. The first window of chromosome 14 started at the 1st CpG site and runs through the Wth CpG site; the second window of chromosome 14 started at the 2nd CpG site and runs through the $W + 1$th site; and so on till the end of the chromosome.

Results

Evaluation of Type I Error Under the Null

We applied the SCM to all the simulated null samples. Using an α-level of 0.05, the test maintained a type I error rate of about 5%. Using the notation K for the total number of consecutive CpG sites included in the genomic region of interest, we experimented with varying the values of K from 3 to 81 in our studies. The type I error rates were 0.052, 0.048, and 0.054 for the three chosen regions with $K = 81$. Similar

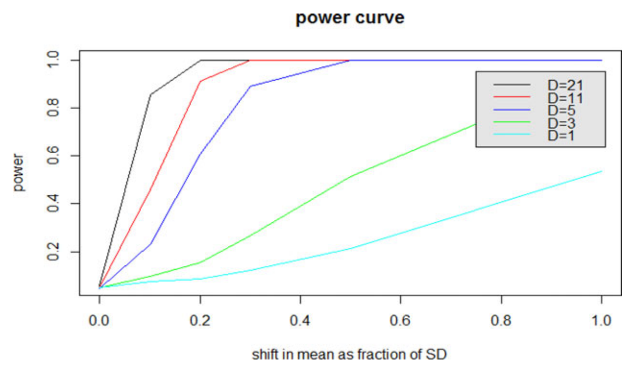

Figure 1. Power curve as the mean of percent methylation value is shifted. The color indicates the number of disease sites within the window (of 51 CpG sites) of investigation.

results for type I error rate were obtained for values of K as small as 3. Based on the simulation results, we concluded that the SCM captured the type I error rate properly at the 0.05 significance level regardless of the size of selected region.

Power Estimates

To estimate power, we used the sites around the 1,300th CpG site of chromosome 1. We chose K to be 51 and let D, the number of disease-related CpG sites, vary from 1 to 21. To be precise, with $K = 51$ and $D = 1$, there are 25 null sites followed by one disease-related site, the chosen site, and followed by 25 more null sites. We shifted the mean of the percent DNA methylation value as a percentage (0%, 10%, 20%, 30%, 50%, and 100%) of the SD.

The result is summarized in Table 1 and plotted in Figure 1. If there is only one disease CpG site among the 51 sites that we were investigating, the power to detect the association is very small. As expected, this parameter combination did not differ much from the null. However, as the number of disease sites increases, the power to detect the association increased significantly even if the amount of shift in the mean methylation value was small. Note that this was a simple model that assumed the sites were sampled independently and no correlation structure of the disease-associated sites was taken into account. So, for complex diseases with a few disease sites (<5) and a small difference in percent methylation values (<50% from the SD), the power was low (i.e., less than 80%). We demonstrated that this method is good at detecting association if there are a number of disease CpG sites (>5) clustered together affecting the outcome of the disease.

Results: Application of the SCM to Chromosome 14 of a Cancer Data Set

For illustrative purposes, we applied the SCM to chromosome 14 of a colorectal cancer data set from TCGA [2012]. The study characterizes somatic alterations in colorectal carcinoma and identified 32 somatic recurrently mutated genes. For methylation patterns, the paper reported the identification of four subgroups based on unsupervised clustering of the promoter DNA methylation profiles of 236 colorectal tumors but not direct association. Since DNA methylation profiles were disrupted extensively by cancer, we expected our method to show diverse number of DMRs. As we have demonstrated our result for null sample work with values of K ranging between three sites and 81 sites, we chose the number of consecutive CpG sites, K, as 51, somewhere in the middle of our studied range. We scanned the entire chromosome 14 from beginning to the end.

The scanning result showed that there were 2,079 windows with P values $< 10^{-5}$, which is the Bonferroni corrected significance level. Many of these significant windows were contiguous forming 67 clusters of statistically significant regions as shown in Figure 3. Two windows belong to a window cluster if they have four or less nonsignificant windows in between. The two largest window clusters comprised 176 and 144 windows. These two window clusters are shown as the

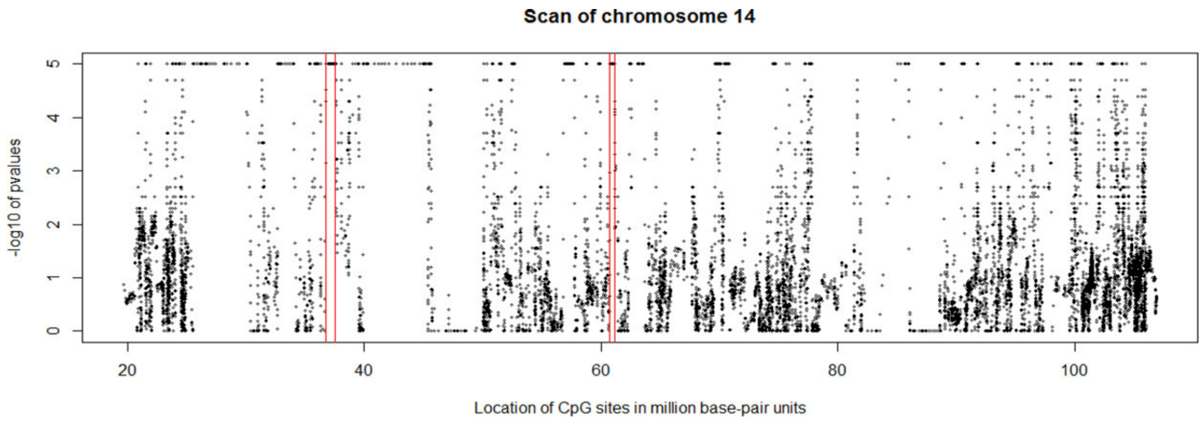

Figure 2. Using windows of 51 CpG sites, the sliding window scan shows regions of chromosome 14 which have P values of $<10^{-5}$ by using the SCM.

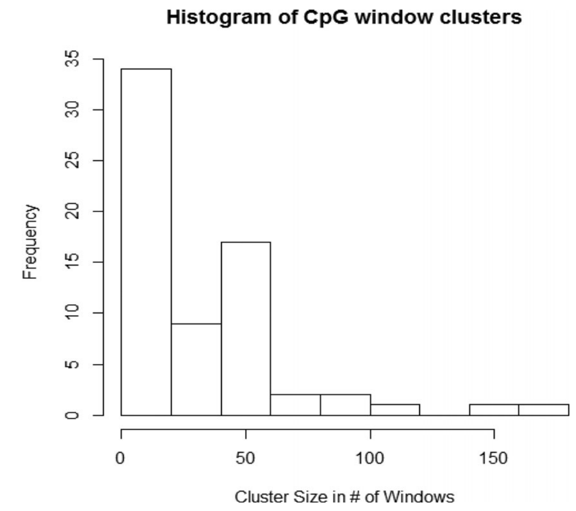

Figure 3. Histogram of the size of CpG window clusters on chromosome 14 with P values $<10^{-5}$.

Table 1. Power estimates for detecting disease sites within 51 CpG sites around chromosome 1, 1,300th CpG site

S (%)	*D* = 1	*D* = 3	*D* = 5	*D* = 11	*D* = 21
0	0.052	0.052	0.048	0.051	0.057
10	0.075	0.096	0.231	0.459	0.854
20	0.087	0.155	0.608	0.912	0.100
30	0.124	0.266	0.891	0.100	1.000
50	0.214	0.512	0.999	1.000	1.000
100	0.537	0.951	1.000	1.000	1.000

"red" vertical lines in Figure 2. The following was the distribution of the window cluster sizes.

This exploratory analysis showed that there were two big window clusters with sizes >140 CpG sites on chromosome 14 which were associated with the disease. To show direct association, detailed modeling and analysis work should be done on these two particular DMRs.

One window here consists of 51 consecutive CpG sites which usually cover several genes. For example, the biggest window cluster of DMRs that was detected on chromosome 14 started at location 22958402 and ended at 23398747 on chromosome 14 with a width of 440,345 base pair units in the region of 14q11.2. It comprises the following 10 genes: AJUBA, PSMB5, ACIN1, CEBPE, SLC7AB, BCL2L2, PABPN1, IL25, CMTM5, and MYH6. So, the SCM used in this exploratory investigation helped to identify genomic regions that were associated with colorectal cancer.

A similar analysis is done on chromosome 10 and the results are available in the supplementary material section.

Discussion and Conclusion

It is well known that human DNA methylation CpG sites are spatially clustered [Hansen et al., 2011; Lister et al., 2009]. Thus, spatial location information of the CpG sites should help to identify the associations between disease and DMRs. Our proposed SCM combines the use of physical spatial location information and percent methylation values from CpG sites of a genomic region to create a statistic. This statistic is then used to assess the significance of the relationship. By incorporating spatial location information, the SCM improves the likelihood of finding real significant associations. As shown in the simulation result, the SCM has reasonable power to detect methylated regions which had differences in the mean level of percent methylation values.

The SCM is a simple and easy to use statistical test for comparing genomic regions. No modeling is needed. As demonstrated in the simulation study, the SCM maintains proper type I error rate. Compared with a simple GWAS approach which treats all CpG sites as independent and a linear mixed model with estimated correlation, one may end up with an inflated type I error. And if the disease-associated CpG sites are spatially clustered and the shift in mean percent methylation values is associated with the disease, the SCM method will have good power to detect the association. The SCM can also be used for screening association across the entire genome. By applying a sliding window approach, we can locate genomic regions of high significance quickly. In the colorectal cancer data set, we used the SCM to identify two regions with large clusters in chromosome 14 of a colorectal data set. Once the SCM identifies DMRs, rigorous follow-up such as biological analysis on the identified regions can be used to find possible causative agents for disease. Our simulation is based on a simple shift in the mean percent methylation values and if the disease process causes such events to occur, the SCM will have good power to detect these DMRs. But, one must be cautious to interpret the *P* values from a genomic region scan since a single CpG site can belong to a number of these sliding windows. Multiple testing adjustments must be applied. Thus, using the SCM for screening is suitable at this point as an exploratory tool and identifies DMRs for further investigation using additional biological methods. Further research should be done to examine the power of SCM in other measures besides a simple shift in the mean.

Although the SCM does not allow the user to adjust for batch effects directly, this can be corrected by using a number of standard software packages such as sva [Leek et al., 2012] to the data before applying the SCM. In order to avoid confounding, one can match samples by the confounding covariate or by a propensity score. However, if the experimental design is not under your control, one can adjust for potentially confounding covariates, such as gender, by applying less powerful statistical technique such as stratification. The *P*-value is computed using the permutation approach. It requires a significant amount of computing time. On an Intel i7 2.90GHz CPU, it has taken about 40 hr to complete a scan on the entire chromosome 14 with 10,000 permutation using a sliding window size of 51 CpG sites on the TCGA data set. So, this is not yet practical to do a genome-wide scan for sliding windows sequentially. However, since the computations are independent of one another, one can break up the computations into parallel processes to get the result faster. The actual performance depends on the computing platform, processors available, amount of memory available, and the precision one would like to obtain for the *P*-value. Furthermore, we expect to continue fine-tuning the algorithm by taking advantage of parallelization and improving efficiency. These preprocessing works are common for statistical analysis and should not incur a big burden on the user.

Before this can be used as a general association test, we need to determine how population and family substructure affect DNA methylation profiles in future studies. If we ignore these effects, it can lead to spurious conclusions. Furthermore, since DNA methylation profiles can change over time, we cannot conclude definitively that DNA methylation are in the causal pathway to the disease. The reverse could also be true—the disease could have caused the variability in DNA methylation profile. Thus, for now, the SCM is best used as an exploratory screening method to locate potential associated DMRs for further analysis.

However, there are some limitations imposed by the current technology for providing full information on the spatial distribution of CpG locations as only predefined CpG sites are on the chip. The SCM is best suited for high coverage chip or sequencing data where more location information of CpG sites is available. For example, data from the denser Illumina Infinium 450K chip provide more information than data from a sparser Illumina Infinium 27K chip.

Additional extensions to the basic algorithm can be made to further enhance the SCM's power. For example, different kinds of window patterns can be used for screening analysis, in addition to the sliding window pattern. The advantages and disadvantages of using different kinds of window patterns will need to be studied in a future research. Another extension is to study when it is appropriate to treat small percent methylation values as totally "unmethylated." The state of methylation at a particular CpG site is binary in nature, just on or off, but the percent methylation value for a CpG site is continuous. Due to the uncertainties associated with the signal, a small value is reported even if the methylated state is off. Thus, small percent methylation values could be an artifact of the measuring process. The distribution of distances to neighboring CpG sites will change if there are more zero value methylation sites. The power to detect association may increase if we can treat these CpG sites with very small percent methylation values as totally unmethylated. By expanding our approach, we can further increase the power and utilities of the SCM.

By incorporating spatial location information into the analysis, we create a novel and simple association test that locate biologically relevant DNA methylation segments in the genomic region of interest. It captures type 1 error rate properly and has power to detect if there is a shift in the mean methylation value in the genomic region of interest. Locating the DMR is the essential step that can lead to the discovery of the underlying biological process between DNA methylation and diseases.

Acknowledgments

We would like to thank Dr. Emily Wan for her assistance in the preparation of the COPD data set and the TCGA Research Network for making the COAD data set publicly available. Generation of the methylation data included in the simulation (es_ICGN) was supported by NIH/NHLBI R01 HL089438. DLD is additionally supported by P01 HL 105339 and P01 HL 083069. We would also like to acknowledge the support of Wai-Ki Yip by the Clinical Epidemiology of Lung Diseases Training Grant (T32 HL 007427).

International COPD Genetics Network (ICGN) investigators: Edwin K. Silverman, Brigham & Women's Hospital, Boston, MA, USA; David A. Lomas, Cambridge Institute for Medical Research, University of Cambridge, Cambridge, UK; Barry J. Make, National Jewish Medical and Research Center, Denver, CO, USA; Alvar Agusti and Jaume Sauleda, Hospital Universitari Son Dureta, Fundación Caubet-Cimera and Ciber Enfermedades Respiratorias, Spain; Peter M.A. Calverley, University of Liverpool, UK; Claudio F. Donner, Division of Pulmonary Disease, S. Maugeri Foundation, Veruno (NO), Italy; Robert D. Levy, University of British Columbia, Vancouver, Canada; Peter D. Paré, University of British Columbia, Vancouver, Canada; Stephen Rennard, Section of Pulmonary & Critical Care, University of Nebraska Medical Center, Omaha, NE, USA; Jørgen Vestbo, Department of Cardiology and Respiratory Medicine, Hvidovre Hospital, Copenhagen, Denmark; Emiel F.M. Wouters, University Hospital Maastricht, The Netherlands.

References

Ansari A, Bradley R. 1960. Rank-sum tests for dispersions. *Am Math Statist* 31:1174–1189.

Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox TV, Davies R, Down TA and others. 2006. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet* 38:1378–1385.

Fier H, Won S, Prokopenko D, AlChawa T, Ludwig KU, Fimmers R, Silverman EK, Pagano M, Mangold E, Lange C. 2012. 'Location, Location, Location': a spatial approach for rare variant analysis and an application to a study on non-syndromic cleft lip with or without cleft palate. *Bioinformatics* 28(23):3027–3033.

Foley DL, Craig JM, Morley R, Olsson CJ, Dwyer T, Smith K, Saffery R. 2009. Prospects for epigenetic epidermiology. *Am J Epidermol* 169:389–400. doi:10.1093/aje/kwn380

Hackenberg M, Barturen G, Carpena P, Luque-Escamilla PL, Previti C, Oliver JL. 2010. Prediction of CpG-island function: CpG clustering vs. sliding-window methods. *BMC Genomics* 11:327–341.

Hansen KD, Timp W, Bravo HC, Sabunciyan S, Langmead B, McDonald OG, Wen B, Wu H, Liu Y, Diep D and others. 2011. Increased methylation variation in epigenetic domains across cancer types. *Nat Genet* 43:768–775.

Jaffe AE, Murakami P, Lee H, Leek JT, Fallin MD, Feinberg AP, Irizarry RA. 2012. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int J Epidemiol* 41:200–209.

Kuan PF, Chiang DY. 2012. Integrating prior knowledge in multiple testing under dependence with applications to detecting differential DNA methylation. *Biometrics* 68(3):774–783.

Laird P. 2010. Principles and challenges of genome-wide DNA methylation analysis. *Nat Rev Genet* 11:191–203.

Leek, JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. 2012. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28(6):882–883.

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo Q-M and others. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462:315–322.

Maher B. 2008. Personal genomes: the case of the missing heritability. *Nature* 456:18–21. doi:10.1038/456018a

Manolio TA, Brooks LD, Collins FS. 2008. A Hapmap harvest of insights into the genetics of common disease. *J Clin Invest* 118(5):1590–1605.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A and others. 2009. Finding the missing heritability of complex diseases. *Nature* 461(7265):747–753. doi:10.1038/nature08494

Qiu W, Baccarelli A, Carey VJ, Boutaoui N, Bacherman H, Klanderman B, Rennard S, Agusti A, Anderson W, Lomas DA and others. 2012. Variable DNA methylation is associated with chronic obstructive pulmonary disease and lung function. *Am J Respir Crit Care Med* 185(4):373–381.

Streitberg B, Roehmel J. 1986. Exact distribution for permutation and rank tests: an introduction to some recently published algorithms. *Statist Software Newslett* 12:10–17.

TCGA Network. 2012. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 487(7407):330–337.

Waller LA, Gotway CA. 2004. *Applied Spatial Statistics for Public Health Data*. Chapter 1 p1. Hoboken, NJ: John Wiley and Sons Inc.