Biostatistics and Biometrics
Open Access Journal
ISSN: 2573-2633

Research Article
Volume 9 Issue 3 - March 2019
DOI: 10.19080/BBOAJ.2019.09.555763

Biostat Biometrics Open Acc J

Detecting Differentially Methylated Genes Associated with Drug Response

Hongyan Xu*, Fengjiao Hu, Santu Ghosh, Sunil Mathur and Varghese George

Department of Population Health Sciences, Augusta University, USA

Submission: July 25, 2018; **Published:** March 22, 2019

***Corresponding author:** Hongyan Xu, Department of Population Health Sciences, Medical College of Georgia, Augusta University, USA

Abstract

DNA methylation has long been involved in inter-individual variations in drug response. In this study, we focused on the methylation changes associated with the response in terms of triglyceride changes before and after the treatment with fenofibrate using the real data set. We analyzed samples that are independent (founders and marry-ins) from each pedigree. Subjects were categorized into responders and non-responders according to percent changes in triglyceride. We then applied a novel spatial scan statistic to identify genes that are differentially methylated between the responders and non-responders. All the CpG sites within a gene were analyzed together. The spatial scan statistic approach uses a mixed-effects model to incorporate correlations of methylation rates among CpG sites. We analyzed the methylation data at visit 2, accounting for the effects of age, sex, and smoking status as covariates. Methylation levels at 312 genes from 22 autosomes were significantly associated with drug response with p< 0.01.

Abbreviations: DMRs: Differentially Methylated Regions; MLE: Maximum Likelihood Estimator

Introduction

Drug response is a complex trait involving multiple genetic and epigenetic factors. In particular, DNA methylation, which is an important regulator of gene expression, has been shown to be involved in inter-individual variation in drug response [1]. In the past, most such studies took a candidate gene approach in which only a few genes are studied. With the advent of high-throughput genomic technologies, we can now survey DNA methylation information across genome-wide CpG sites. The methylation information could be analyzed with a single-marker approach. However, this may lead to many false-positives because of the huge number of CpG sites genome-wide. It has been found that methylation levels at close-by CpG sites could be highly correlated. The single marker approach also ignores this correlation. Therefore, a better approach is to jointly analyze the CpG sites in a genomic region and identifying differentially methylated regions (DMRs) between different drug response groups. In this study, we take a region-based approach and treat each gene as a genomic region to identify DMRs between responders and non-responders in terms of triglyceride changes before and after the fenofibrate treatment. We applied a novel scan statistic approach based on normal distribution to the real data provided by GAW20.

Method

In this section, we describe the scan statistic to detect DMRs based on the difference in methylation rates between two groups (responders and non-responders) [2].

Adjusting for correlation between CpG sites

We assume p_{kij} is the true methylation rate at CpG site j for individual i in group k, $k = A,U$, $i = 1,2,\ldots,n_k$, $j = 1,2,\ldots s$. Here $k = A$ for responders and $k = u$ for non-responders.

To account for the correlation of methylation rates among nearby CpG sites, a random slope and intercept logistic regression model is considered to model methylation rate at each CpG site for every individual. A random slope and intercept logistic regression has the following form,

$$\log\left(\frac{p_{kij}}{1-p_{kij}}\right) = \beta_0 + \beta_1 s_j + \beta_2 x_{ki} + v_{0ki} + v_{1ki} s_j, \quad (1)$$

Where s_j represents the distance of CpG site j from the start point. In the mixed-effect model setting, the random effect $v_{ki} = \begin{pmatrix} v_{0ki} \\ v_{1ki} \end{pmatrix}$ is assumed to vary independently across individuals, with $v_{ki} \sim N\left(0, \begin{pmatrix} \sigma^2_{v_{0ki}} & \sigma_{v_{0ki}}\sigma_{v_{1ki}} \\ \sigma_{v_{0ki}}\sigma_{v_{1ki}} & \sigma^2_{v_{1ki}} \end{pmatrix}\right)$. By adding x_{ki} in the mixed-effect model (1), we can also adjust for the covariates.

The fitted odds of methylation rates can be calculated for CpG site j of individual i in group k, and can be used to get the corresponding adjusted expected methylation rate $\hat{p}_{kij}$, with its logit transformation $\text{logit}(\hat{p}_{kij}) = \log\left(\frac{\hat{p}_{kij}}{1-\hat{p}_{kij}}\right)$. Then we can calculate the adjusted logit transformation of methylation rates (residuals) as, $y_{kij} = \text{logit}(p_{kij}) - \text{logit}(\hat{p}_{kij})$.

Since the methylation rates are independent, the rate at each CpG site j in group k is given by

$$y_{kj} = \frac{1}{n_k}\sum_{i\in k} y_{kij}$$

with variance,

$$\sigma_{kj}^2 = \frac{1}{n_k}\sum_{i\in k}\left(y_{kij} - y_{kj}\right)^2$$

Scan statistic based on normal distribution

We assume $y_{Aj} \sim N\left(\mu_A, \sigma_{Aj}^2\right)$ and $y_{Uj} \sim N\left(\mu_U, \sigma_{Uj}^2\right)$, with known σ_{Aj}^2 and σ_{Uj}^2, where μ_A and μ_U are the true methylation rates in cases and controls, respectively. Considering $y_{kj} \sim N\left(\mu_k, \sigma_{kj}^2\right)$, then the likelihood of y_{kj} is given by

$$f\left(y_{kj}\right) = \frac{1}{\sqrt{2\pi\sigma_{kj}}}\exp\left(-\frac{2}{2\sigma_{kj}^2}\left(y_{kj}-\mu_k\right)^2\right)$$

$$= \frac{1}{\sqrt{2\pi\sigma_{kj}}}\exp\left(\frac{1}{\sigma_{kj}^2}\left(\mu_k y_{kj} - \frac{\mu_k^2}{2} - y_{kj}^2/2\right)\right)$$

It is evident from this likelihood that the distribution of adjusted methylated rate follows a one-parameter exponential family $EXP(\eta,\phi,T,B_e,a)$ where $T(y_{kj}) = y_{kj}$, $\eta = \mu_k$, $B_e(\eta) = \mu_k^2/2$, $\phi_j = \sigma_{kj}^2$, and the log-likelihood $l(\eta;y) = \left(\eta T(y) - B_e(\eta)\right)/\phi$ after ignoring an additive constant that does not depend on η.

Based on this likelihood function, we can find the maximum likelihood estimator (MLE) of parameter η in the one-parameter exponential family $EXP(\eta,\phi,T,B_e,a)$ as $\hat{\eta} = g_e\left(T(y)\right)$, where $g_e(T) = \left(B_e'\right)^{-1}(T)$ [3].

For a specific genomic region, after adjusting for correlation between CpG sites by using the mixed-effect model, $(y_{k1}, y_{k2}, \ldots, y_{k2})$ are assumed to be independent for the s consecutive CpG sites. Then,

$$\frac{1}{\phi^*} = \sum_{j=1}^{s}\frac{1}{\sigma_{kj}^2},$$

$$v_j = \phi^*/\phi_j,$$

and

$$T\left(y_k\right) = \sum_{j=1}^{s} v_j T\left(y_{kj}\right) = \frac{\sum_{j=1}^{s} y_{kj}/\sigma_{kj}^2}{\sum_{j=1}^{s} 1/\sigma_{kj}^2}$$

In order to test the hypotheses $H_0 : \mu_A = \mu_U$ versus $H_1 : \mu_A \neq \mu_U$, the ratio of the likelihood under H_1 versus H_0 can be used as a test statistic. More conveniently, we can use the log of this likelihood ratio as our test statistic, which we refer to as the scan statistic. It is given by

$$\Delta = k\left(T_A, \Phi_A\right) + k\left(T_U, \Phi_U\right) - k\left(T, \Phi\right), \quad (2)$$

Where $k(x,y) = \left(xg_e(x) - B_e\left(g_e(x)\right)\right)/y$ and $\frac{1}{\Phi} = \frac{1}{\Phi_A} + \frac{1}{\Phi_U}, T = b_A T_A + (1-b_A)T_U$ with $b_A = \frac{1}{\Phi_A}\Big/\left(\frac{1}{\Phi_A} + \frac{1}{\Phi_U}\right)$.

Here we have

$$\Phi_A = \left(\sum_{j=1}^{s}\frac{1}{\sigma_{Aj}^2}\right)^{-1} \text{ and } \Phi_U = \left(\sum_{j=1}^{s}\frac{1}{\sigma_{Uj}^2}\right)^{-1},$$

$$T_A = \frac{\sum_{j=1}^{s} y_{Aj}/\sigma_{Aj}^2}{\sum_{j=1}^{s} y_{Aj}/\sigma_{Aj}^2} \text{ and } T_U = \frac{\sum_{j=1}^{s} y_{Uj}/\sigma_{Uj}^2}{\sum_{j=1}^{s} y_{Uj}/\sigma_{Uj}^2}$$

for the two groups.

Application to the GAW20 real data

In our analysis, we first classified a subject as either a responder or a non-responder. A subject was classified as a responder if the percent change between pre and post treatments values is more than 35% [4,5]. Using this criteria, there are 42 responders and 45 non-responders in our sample. We then applied the scan statistic method to the methylation data at visit 2 for chromosomes 1 to 22. All the CpG sites within a gene region are jointly analyzed. Based on the annotation file provided, for each chromosome a gene region is defined as the continuous region within same gene name.

Results

We further performed functional annotation of the significant genes using DAVID 6.8 (https://david.ncifcrf.gov/). Eleven genes, HMGCR [6,7], KLF10 [8], RBMS1 [9], THADA [10], CRY2 [11], FADS1 [12], PTER [13], STK11 [14], TSPAN8 [15], TFRC [16], and IARS2 [17], were found to be involved in type 2 diabetes, in which lipid levels including triglyceride have been shown to be significant risk factors.

We have performed further analysis including single CpG analysis and another region-based analysis with IMA [18], for comparison purpose. The single CpG analysis was carried out using limma package in Bioconductor for all the 4,565 CpG sites annotated to the 312 genes identified through our approach. We also adjusted for the effect of age, sex, and smoking status as we did in our approach. Single CpG analysis identified 836 CpG sites from 112 genes with $p < 0.01$. We performed region-based analysis using IMA package in Bioconductor for the 312 genes identified through our approach. We also adjusted for the effect of age, sex, and smoking status. IMA detected 11 genes with $p < 0.01$.

Discussion

In this study, we applied a novel method based on scan statistic to detect differentially methylated genes between responders and non-responders to the treatment with fenofibrate. We treat each gene as a genomic region and our method accounts for the correlation of methylation levels between CpG sites within a gene. By doing so, we can utilize the information across multiple CpG sites to boost the statistical power of identifying difference in methylation levels between the two groups. The method is based on regression approach. Therefore, it is natural to account for the effect of covariates by including them in the model. We applied our method to the GAW20 real data set and was able to identify genes with biological relevance. One of the limitations is that the statistical significance is based on 1,000 permutations because of the constraints on computational speed. Therefore, the smallest p-value we can get is 0.001 due to the discreteness of the test statistic distribution from permutation.

How to cite this article: Hongyan X, Fengjiao H, Santu G, Sunil M, Varghese G. Detecting Differentially Methylated Genes Associated with Drug Response. Biostat Biometrics Open Acc J. 2019; 9(3): 555763. DOI: 10.19080/BBOAJ.2019.09.555763.

From our comparison with the single CpG site analysis approach and the region-based analysis using IMA, these two approaches detected less significant genes at 0.05 level. It should be noted that this is not a strict comparison in statistical sense because we only performed the analysis in the subset of 312 genes detected with our approach. In summary, we proposed a new region-based method to detect genes whose methylation levels are associated with triglyceride responses to drug treatment. We applied our method to the real data set from GAW20. Our method could identify some genes related to obesity and lipid in the literature, which suggests that this is a reasonable approach.

References

1. Maier S, Dahlstroem C, Haefliger C, Plum A, Piepenbrock C (2005) Identifying DNA Methylation Biomarkers of Cancer Drug Response. Am J PharmacoGenomics 5(4): 223-232.
2. Hu F, Xu H, Ryu D, Ghosh S, Shi H, et al. (2015) Detection of differentially methylated regions using kernel distance and scan statistics. Hum Genet. Under review.
3. Agarwal D, Phillips JM, Venkatasubramanian S (2006) The hunting of the bump: on maximizing statistical discrepancy. Society for Industrial and Applied Mathematics 1: 1137-1146.
4. Giacco R, Costabile G, Della Pepa G, Anniballi G, Griffo E, et al. (2014) A whole-grain cereal-based diet lowers postprandial plasma insulin and triglyceride levels in individuals with metabolic syndrome. Nutr Metab Cardiovasc Dis 24(8): 837-844.
5. Skulas-Ray AC, Kris-Etherton PM, Harris WS, Heuvel JPV, Wagner PR, et al. (2011) Dose-response effects of omega-3 fatty acids on triglycerides, inflammation, and endothelial function in healthy persons with moderate hypertriglyceridemia. Am J Clin Nutr 93(2): 243-252.
6. Ference BA, Robinson JG, Brook RD, Catapano AL, Chapman MJ, et al. (2016) Variation in PCSK9 and HMGCR and Risk of Cardiovascular Disease and Diabetes. N Engl J Med 375(22): 2144-2153.
7. Lotta LA, Sharp SJ, Burgess S, Perry JRB, Stewart ID, et al. (2016) Association Between Low-Density Lipoprotein Cholesterol-Lowering Genetic Variants and Risk of Type 2 Diabetes: A Meta-analysis. JAMA 316(13):1383-1391.
8. Gutierrez-Aguilar R, Benmezroua Y, Balkau B, Marre M, Helbecque N, et al. (2007) Minor contribution of SMAD7 and KLF10 variants to genetic susceptibility of type 2 diabetes. Diabetes Metab 33(5): 372-378.
9. Qi L, Cornelis MC, Kraft P, Stanya KJ, Linda Kao WH, et al. (2010) Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes. Hum Mol Genet 19(13): 2706-2715.
10. Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, et al. (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet 40(5): 638-645.
11. Dupuis J, Langenberg C, Prokopenko I, Saxena R, Soranzo N, et al. (2010) New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat Genet 42(2): 105-116.
12. Hicks AA, Pramstaller PP, Johansson A, Vitart V, Rudan I, et al. (2009) Genetic determinants of circulating sphingolipid concentrations in European populations. PLoS Genet 5(10): e1000672.
13. Ghosh S, Watanabe RM, Valle TT, Hauser ER, Magnuson VL, et al. (2000) The Finland-United States investigation of non-insulin-dependent diabetes mellitus genetics (FUSION) study. I. An autosomal genome scan for genes that predispose to type 2 diabetes. Am J Hum Genet 67(5): 1174-1185.
14. Keshavarz P, Inoue H, Nakamura N, Yoshikawa T, Tanahashi T, Itakura M. Single nucleotide polymorphisms in genes encoding LKB1 (STK11), TORC2 (CRTC2) and AMPK alpha2-subunit (PRKAA2) and risk of type 2 diabetes. Mol Genet Metab 93(2): 200-209.
15. Grarup N, Andersen G, Krarup NT, Albrechtsen A, Schmitz O, et al. (2008) Association testing of novel type 2 diabetes risk alleles in the JAZF1, CDC123/CAMK1D, TSPAN8, THADA, ADAMTS9, and NOTCH2 loci with insulin release, insulin sensitivity, and obesity in a population-based sample of 4,516 glucose-tolerant middle-aged Danes. Diabetes 57(9): 2534-2540.
16. Fernández-Real JM, Mercader JM, Ortega FJ, Moreno-Navarrete JM, López-Romero P, et al. Transferrin receptor-1 gene polymorphisms are associated with type 2 diabetes. Eur J Clin Invest. 40(7): 600-607.
17. Hägg S, Ganna A, Van Der Laan SW, Esko T, Pers TH, et al. (2015) Gene-based meta-analysis of genome-wide association studies implicates new loci involved in obesity. Hum Mol Genet 24(23): 6849-6860.
18. Wang D, Yan L, Hu Q, Sucheston LE, Higgins MJ, et al. (2012) IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. Bioinformatics 28(5): 729-730.

DOI: 10.19080/BBOAJ.2019.09.555763

How to cite this article: Hongyan X, Fengjiao H, Santu G, Sunil M, Varghese G. Detecting Differentially Methylated Genes Associated with Drug Response. Biostat Biometrics Open Acc J. 2019; 9(3): 555763. DOI: 10.19080/BBOAJ.2019.09.555763.