Gene-based Methods for DNA Methylation Analysis

Qi Yan[1], Juan C. Celedón[1], and Wei Chen[1,2]

[1]Division of Pulmonary Medicine, Allergy and Immunology; Department of Pediatrics, Children's Hospital of Pittsburgh of UPMC, University of Pittsburgh;
[2]Departments of Human Genetics and Biostatistics, University of Pittsburgh Graduate School of Public Health

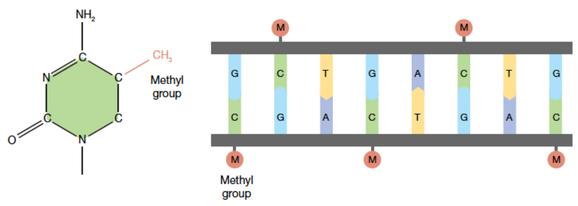
ABSTRACT

With the advances of microarray and next-generation sequencing technologies, genome wide epigenomic data at a high resolution are available for several hundreds and even thousands subjects. Among these epigenomic data, the investigation of methylation that is adding methyl groups to cytosines in the CpG dinucleotide context can serve as a complement to genomic studies. DNA methylation has been studied in various biological processes, cancer and some other diseases. Unlike genetic markers that are static, DNA methylated loci are high dynamic across cell types and over time. Therefore, it requires the development of appropriate study designs and statistical methods. Although, in rare situations, the single methylated locus is associated with diseases, researchers are more interested in the association of a set of methylated loci, such as CpG islands, CpG shores and UTRs. Therefore, we aim to use gene-based statistics to test the association between diseases and methylated loci where "gene" is defined as its physical location. In the simulation studies, we evaluated the performance of two gene-based methods, Sequence Kernel Association Test and Functional Linear Model. Finally, we illustrate our proposed approaches by analyzing whole-genome DNA methylation data from an asthma study as well as a rheumatoid arthritis study.

BACKGROUND

DNA Methylation

- A methyl group added to cytosine or adenine DNA nucleotide.
- A critical epigenetic modification affects gene transcription and expression and associates with many biological processes, e.g. cell differentiation, cancers, asthma…
- Unlike genetic markers that are static, DNA methylated loci are high dynamic across cell types and over time.
- A set of methylated loci, such as CpG islands, CpG shores and UTRs, are more of interest to researchers than single methylated locus.



MOTIVATION

- Most studies have focused association test for a single site. Researchers are more interested in the association of a set of methylated loci, such as CpG islands, CpG shores and UTRs
- In genome-wide association studies, various methods have been developed to study the association between a group of rare variants and complex diseases

AIM

- Develop a region-based test for epigenome-wide association studies of complex diseases
- "Region" is defined as the physical location of this gene plus several thousand base pairs (e.g. 10kb) on either side of this gene or a set of CpG sites

METHODS

Data characteristics for methylated loci from one gene

	M locus 1	*M locus 2*	*M locus 3*	*M locus 4*	*M locus 5*	*M locus 6*	*M locus 7*	*M locus 8*
Sub1:	0.4717472	0.50728724	0.71213225	0.78337668	0.84951523	0.8829686	0.1187730	0.09706991
Sub2:	0.5861663	0.56063761	0.73616749	0.81304153	0.87074355	0.8533281	0.1380723	0.27017744
Sub3:	0.4848768	0.48497147	0.67084446	0.74025466	0.75619173	0.7625518	0.0949787	0.08249194

Method 1. Kernel machine regression approach

Motivated by Sequence Kernel Association Test (Wu et al., 2011), we assume that the n × 1 vector of the quantitative trait y follows a linear model,

$$y = XC + M\beta + \varepsilon$$

When the trait is binary, consider a logistic model,

$$\text{logit}P(y = 1) = XC + M\beta$$

$$\beta \sim N(0,\ \tau W)$$

$$\varepsilon \sim N(0,\ \sigma_E^2 I)$$

The null hypothesis we are interested in is
The p-value can be calculated by variance component test.

Method 2. Functional Linear Model

Motivated by (Fan et al., 2013), for a quantitative trait, we still consider a linear model,

$$y_i = X_i'C + \int_0^1 M_i(t)\,\beta(t)dt + \varepsilon_i$$

$$M_i(t) = \left(M_i(t_1), \ldots, M_i(t_q)\right)\Phi[\Phi'\Phi]^{-1}\phi(t) \qquad \beta(t) = \theta'(t)(\beta_1, \cdots, \beta_K)'$$

Therefore, after some algebra,

$$y_i = X_i'C + R_i'\beta + \varepsilon_i$$

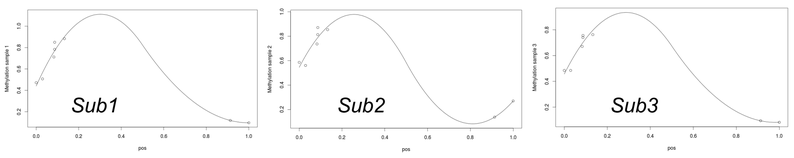
Similarly, when the trait is binary, consider a logistic model,

$$\text{logit}P(y_i = 1) = X_i'C + R_i'\beta$$

	M locus 1	*M locus 2*	*M locus 3*	*M locus 4*	*M locus 5*	*M locus 6*	*M locus 7*	*M locus 8*
Sub1:	0.4717472	0.50728724	0.71213225	0.78337668	0.84951523	0.8829686	0.1187730	0.09706991
Sub2:	0.5861663	0.56063761	0.73616749	0.81304153	0.87074355	0.8533281	0.1380723	0.27017744
Sub3:	0.4848768	0.48497147	0.67084446	0.74025466	0.75619173	0.7625518	0.0949787	0.08249194
POS:	0.0000000	0.02961208	0.08306189	0.08661534	0.08691146	0.1326621	0.9130885	1.00000000

$$R_i = \left(M_i(t_1), \ldots, M_i(t_q)\right)\Phi[\Phi'\Phi]^{-1}\int_0^1 \phi(t)\theta'(t)\,dt$$

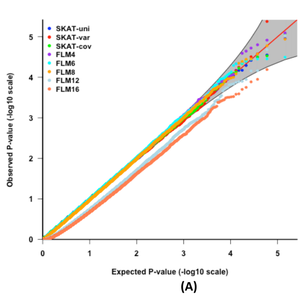
	R 1	R 2
Sub1:	0.2632807	0.1585255
Sub2:	0.2899638	0.1958969
Sub3:	0.2422441	0.1420319



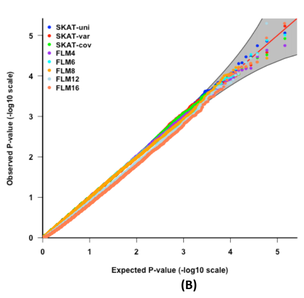
RESULTS

Simulation studies

Type I error rate

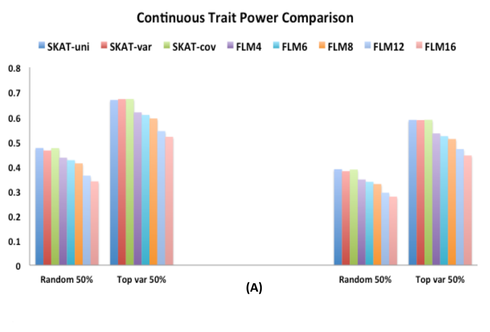


(A)

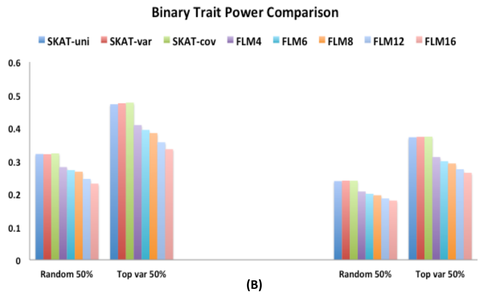


(B)

Power comparison



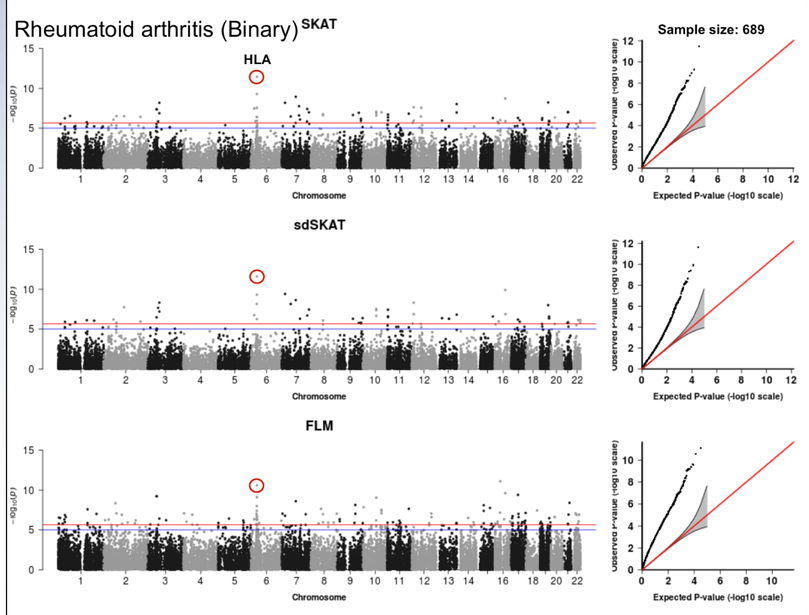
(A)



(B)

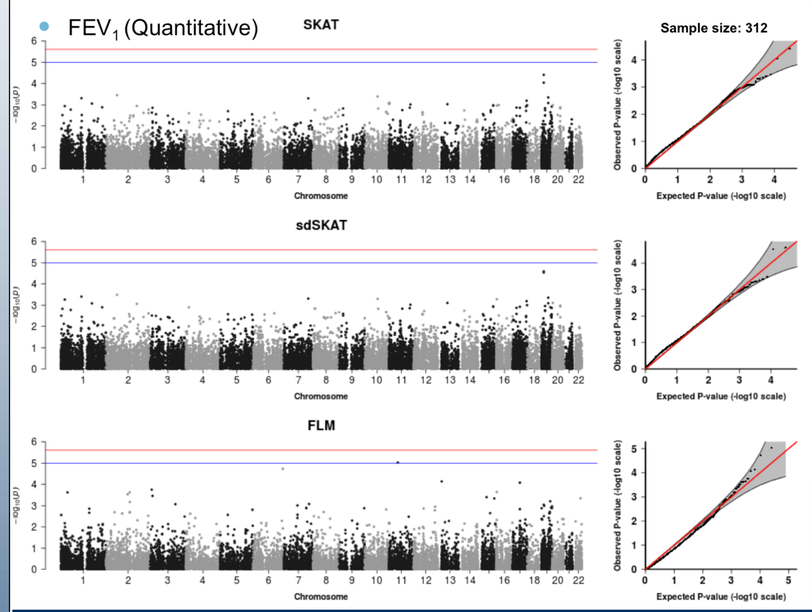
Rheumatoid Arthritis (RA) Data

- The DNA methylation data are from Illumina's 450k arrays on whole blood cells from a RA study
- The data contain 354 cases with citrullinated protein antibodies and 335 controls, a Swedish population-based case-control study
- The data include 298,109 CpG positions
- We assigned methylated loci to a gene if they were located within a 10kb flank of the gene on either side
- We adjusted for age, gender, smoking status, 5 principal components and estimated cell type composition



PRGOAL Asthma Data

- The DNA methylation data are from Illumina's 450k arrays on white blood cells from an asthma study
- The data contain 312 Puerto Rican subjects with the information of cell type composition of white blood cell
- 143,376 CpG positions are included after filtering
- We assigned methylated loci to a gene if they were located within a 10kb flank of the gene on either side.
- We adjusted for age, gender and asthma status for FEV1 (continuous) test (with or without cell type adjustment does not affect the type I error); adjusted for age, gender and cell types for Asthma status (binary) test (without cell type adjustment inflates the type I error).



CONCLUSION

- We proposed two statistical approaches for region-based association tests
- Two approaches are comparable and may be both used in real studies
- Need more comprehensive simulations with different set definitions

REFERENCES

1. Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet, 89*(1), 82-93.
2. an, R., Wang, Y., Mills, J. L., Wilson, A. F., Bailey-Wilson, J. E., & Xiong, M. (2013). Functional linear models for association analysis of quantitative traits. Genet Epidemiol, 37(7), 726-742.