ARIC Manuscript Proposal #2265

PC Reviewed: 12/10/13	**Status: A**	**Priority: 2**
SC Reviewed: _________	**Status: _____**	**Priority: ____**

1.a. Full Title: Evaluation of microarray-based DNA methylation measurement using technical replicates

b. Abbreviated Title (Length 26 characters):

2. Writing Group: ARIC Epigenetics Working Group

Working group members:

Weihua Guan
Maitreyee Bose
Chong Wu
James Pankow
Ellen Demerath
Jan Bressler
Myriam Fornage
Megan Grove
Tom Mosley
Chindo Hicks
Kari North
Linda Kao
Eric Boerwinkle

Other interested investigators are welcome to join the writing group

I, the first author, confirm that all the coauthors have given their approval for this manuscript proposal. _WG_ **[please confirm with your initials electronically or in writing]**

First author: **Weihua Guan**
Address: Division of Biostatistics
University of Minnesota
A460 Mayo Bldg., MMC 303
420 Delaware St., S.E.
Minneapolis, MN 55455

Phone: 612-626-4765 Fax: 612-624-0660
E-mail: wguan@umn.edu

ARIC author to be contacted if there are questions about the manuscript and the first author does not respond or cannot be located (this must be an ARIC investigator).

Name:
Address:

Phone: Fax:
E-mail:

3. Timeline:

Hybridization of the DNA samples to the HM450 methylation array was completed in February 2012. The preliminary dataset was available for distribution to the ARIC Coordinating Center and the ARIC Epigenetics Working Group in March 2012. We anticipate a draft ready to submit for Publications Committee review in Nov 2013.

4. Rationale:

Epigenetics is the study of mitotically heritable modifications in chromatin structure (i.e., modifications not involving the germline DNA sequence), and their impact on the transcriptional control of genes and cellular function. Epigenetic variation includes post-translational modifications of histone proteins, non-coding RNAs, and DNA methylation, the latter primarily occurring at cystosine-guanine dinucleotides (CpGs).

Recent technological advances have provided multiple platforms for systematically interrogating DNA methylation variation across the genome (Laird, 2010). This has paved the way for epigenome-wide association studies (EWASs), analogous to genome-wide association studies, to evaluate regions of the genome in which variation in DNA methylation may influence gene expression and ultimately disease risk (Raykan, 2011). Like GWASs, EWASs are based on an agnostic approach in which epigenetic marks can be investigated across the epigenome without prespecifying the genes or regions in which inter-individual variation in DNA methylation is thought to be important for phenotypic variation. However, unlike inherited changes to the genetic sequence, variation in site-specific methylation varies by tissue, stage of development, disease state, and may be impacted by aging and exposure to environmental factors such as diet or smoking (Raykan, 2011).

Arrays to efficiently profile DNA methylation have only recently become commercially available (Laird, 2010). In ARIC, the Illumina HumanMethylation450 BeadChip (HM450) is being used to measure DNA methylation in peripheral blood obtained from ~3000 African American participants at visit 2 (and a small number at visit 3). The array includes 485,577 assays and provides coverage of 98.9% of RefSeq genes with a global average of 17.2 probes per gene region (Bibikova, 2011; Dedeurwaerder, 2011).

However, unlike measuring SNP genotypes, which are discrete values, methylation levels are measured in a continuous scale and are less tolerant to measurement errors. A wellknown source of bias for EWASs is the so-called “batch effect”, which is largely caused by measurement errors. Although many statistical methods have been proposed to correct for batch effects (Bock, 2012; Chen et al., 2011; Maksimovic et al., 2012; Leek et al., 2010; Sun et al., 2011), no approach has been universally accepted. Alternatively, it will be useful to consider statistical measures that can quantify the extent to which the measured methylation level at a specific CpG site is affected by measurement errors. In experiments, technical replicates are often included which can be used to evaluate the consistency of measurement. Meng et al. (2010) demonstrated a method to estimate the proportion of non-variable CpG sites, i.e. sites which showed no variation among individuals studied. Their data came from the Illumina GoldenGate methylation assay and consisted of 311 samples assayed at 1505 sites. However, no comprehensive work has been done for the newly available 450K chip.

When the ARIC samples were assayed with the HM450 array, technical replicates were included on the plates for 130 samples (total n = 265 with 5 samples replicated 3 times). All replicate pairs were distributed to different plates except the 5 triplicate samples which were also assayed within a single plate to test intra-plate variability. Using this rich set of technical replicates, we will be able to evaluate the reproducibility of each CpG site assayed on this array. We hope that our results can add to the guideline for inclusion/exclusion of CpG sites in the subsequent EWASs.

5. Main Hypothesis/Study Questions:

This paper will primarily examine the reproducibility of methylation measures on the HM450 array using technical replicates, and classify the CpG sites into groups based on the their intra-class correlation. We will demonstrate the performance of this grouping when applied to the study of association between methylation levels and smoking status of individuals.

6. Design and analysis (study design, inclusion/exclusion, outcome and other variables of interest with specific reference to the time of their collection, summary of data analysis, and any anticipated methodologic limitations or challenges if present).

Study design: Analysis of DNA methylation by Illumina Infinium HumanMethylation450 (HM450) BeadChip (Illumina Inc., San Diego, CA) has been adapted for methylation profiling by exploiting technology previously developed for SNP genotyping. The assay requires using sodium bisulfite to convert unmethylated cytosine residues to uracil under conditions in which 5-methylcytosine remains unreactive. This difference is then detected as a C/T nucleotide polymorphism at each CpG site. Data analysis is performed using proprietary Genome Studio software (Illumina Inc.) that includes algorithms to obtain the relative level of methylation as a beta value, a continuous variable ranging between 0 and 1. The beta value is calculated as the ratio of methylated signal intensity to the sum of methylated and unmethylated signals for each probe after first subtracting the background signal intensity of negative controls included on the array. Several different controls were included on each 96-well plate of DNA samples that was processed for hybridization to the HM450 arrays. These consisted of four replicate DNAs, a commercially available positive control DNA (Universal Methylated Human DNA Standard, Zymo Research Corporation; Irvine, CA), and a whole-genome amplified DNA sample from an ARIC study participant used as an unmethylated negative control. A series of blind duplicates were also analyzed on the arrays in accordance with ARIC study policy.

Inclusion/exclusion criteria: A cross sectional selection of African American study participants at either visit 2 or 3 was included on the array if the individual had not restricted use of their DNA, if there was 1 ug or more of DNA available for methylation analysis, and if there was genotyping array data available from either the Affymetrix Human SNP Array 6.0, the Illumina HumanCVD BeadChip, the Illumina HumanCardio-MetaboChip, or the Illumina HumanExome BeadChip. Individuals will be excluded from analysis if a pass rate for the DNA sample for the study participant was less than 99% (probes with a detection p-value >0.01/all probes on the array). Probes on the HM450 array for which the detection p-value is >0.01 will not be analyzed.

(1) To characterize the consistency/reproducibility of methylation measurement on each CpG site included on the HM450 array, we will calculate the intra-class correlation coefficient (ICC) for each site using the n=265 technical replicates. We will describe the distribution of ICC values across all sites, and the relationship between ICC and variation of methylation levels.

(2) We will statistically model the distribution of ICC using a mixture model approach, and classify the CpG sites into multiple components based on the posterior probability calculated from the mixture model. Specifically, we will consider two mixture models: 1) a two-component model: $f(x)=\pi_1 f_1 (x)+\pi_2 f_2 (x)$, where π_i are the mixing proportions that sum to one, $f_1 (x)$ a censored normal distribution with ICC censored at 0, representing low reproducibility CpG sites, and $f_2 (x)$ a normal distribution representing high reproducibility sites; 2) a three-component model: $f(x)=\pi_0 I(x=0)+\pi_1 f_1 (x)+\pi_2 f_2 (x)$, where π_i summing to one, $I(x=0)$ a point mass at 0 for CpG sites with ICC of 0, $f_1 (x)$ a truncated normal distribution with ICC truncated at 0, and $f_2 (x)$ a normal distribution. The second model is approximately equivalent to a two-component mixture of truncated normal and normal distributions, fitting to CpG sites with ICC strictly greater than 0. The means and variances of normal distributions will be estimated using the expectation-maximization (EM) algorithm (Dempster et al., 1977; Lee and Scott, 2012).

(3) We will apply a linear mixed model (LMM) to correct for potential batch (chip) effects. Using residuals from LMM, which can be considered as the methylation measure with batch effects subtracted, we will repeat step (1) and (2) and compare the results.

(4) We will first use the basic normalization approach suggested by Illumina for the methylation measurement, but will also explore the impact of different normalization approaches.

(5) We will demonstrate the impact of measurement reproducibility on EWAS results. As an example, we will carry out EWAS between DNA methylation and smoking status (current vs. former and never smoking), adjusting for other covariates including age, gender, visit, center, PCs from GWAS. Batch effects will be adjusted through LMM. Significant results ($p < 10^{-7}$) will be summarized across sites, and grouped according to classification results in (2). We will compare the number of significant associations by each classification group.

7.a. Will the data be used for non-CVD analysis in this manuscript? ____ Yes __x__ No

b. If Yes, is the author aware that the file ICTDER03 must be used to exclude persons with a value RES_OTH = "CVD Research" for non-DNA analysis, and for DNA analysis RES_DNA = "CVD Research" would be used? _x___ Yes ____ No
(This file ICTDER03 has been distributed to ARIC PIs, and contains the responses to consent updates related to stored sample use for research.)

8.a. Will the DNA data be used in this manuscript? ___x_ Yes ____ No Limited to ancestry information obtained from AIMs or GWAS markers

8.b. If yes, is the author aware that either DNA data distributed by the Coordinating Center must be used, or the file ICTDER03 must be used to exclude those with value RES_DNA = "No use/storage DNA"?
__x__ Yes ____ No

9. The lead author of this manuscript proposal has reviewed the list of existing ARIC Study manuscript proposals and has found no overlap between this proposal and previously approved manuscript proposals either published or still in active status. ARIC Investigators have access to the publications lists under the Study Members Area of the web site at: http://www.cscc.unc.edu/ARIC/search.php

___x___ Yes _______ No

10. What are the most related manuscript proposals in ARIC (authors are encouraged to contact lead authors of these proposals for comments on the new proposal or collaboration)?

11.a. Is this manuscript proposal associated with any ARIC ancillary studies or use any ancillary study data? __x__ Yes ____ No

11.b. If yes, is the proposal
___ A. primarily the result of an ancillary study (list number* _________)

_x_ B. primarily based on ARIC data with ancillary data playing a minor role (usually control variables; list number(s)* __________ __________ __________)

2007.02 (CARe, genotyping in African Americans)

*ancillary studies are listed by number at http://www.cscc.unc.edu/aric/forms/

12a. Manuscript preparation is expected to be completed in one to three years. If a manuscript is not submitted for ARIC review at the end of the 3-years from the date of the approval, the manuscript proposal will expire.

12b. The NIH instituted a Public Access Policy in April, 2008 which ensures that the public has access to the published results of NIH funded research. It is **your responsibility to upload manuscripts to PUBMED Central** whenever the journal does not and be in compliance with this policy. Four files about the public access policy from http://publicaccess.nih.gov/ are posted in http://www.cscc.unc.edu/aric/index.php, under Publications, Policies & Forms. http://publicaccess.nih.gov/submit_process_journals.htm shows you which journals automatically upload articles to Pubmed central.

References:

Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, Schroth GP, Gunderson KL, Fan JB, Shen R. High density DNA methylation array with single CpG site resolution. *Genomics* 2011; 98:288-95.

Bock, C., Analysing and interpreting DNA methylation data. Nat Rev Genet, 2012. 13(10): p. 705-19.

Chen, C., et al., Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. PLoS One, 2011. 6(2): p. e17238.

Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F. Evaluation of the Infinium Methylation 450K technology. *Epigenomics* 2011; 3:771-84.

Dempster, A.P., N. Laird, and D. Rubin, Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 1977. 39: p. 1-38.

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

Lee, G. and C. Scott, EM algorithms for multivariate Gaussian mixture models with truncated and censored data. Computational Statistics & Data Analysis, 2012. 56(9): p. 2816-2829.

Leek, J.T., et al., Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet, 2010. 11(10): p. 733-9.

Maksimovic, N., et al., Health-related quality of life in patients with atopic dermatitis. J Dermatol, 2012. 39(1): p. 42-7.

Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. *Nat Rev Genet* 2011;12:529-41.

Sandoval J, Heyn H, Moran S, Serra-Musach J, Pujana MA, Bibikova M, Esteller M. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* 2011; 6:692-702.

Sun, Z., et al., Batch effect correction for genome-wide methylation data with Illumina Infinium platform. BMC Med Genomics, 2011. 4: p. 84.

Zhu ZZ, Hou L, Bollati V, Tarantini L, Marinelli B, Cantone L, Yang AS, Vokonas P, Lissowska J, Fustinoni S, Pesatori AC, Bonzini M, Apostoli P, Costa G, Bertazzi PA, Chow WH, Schwartz J, Baccarelli A. Predictors of global methylation levels in blood DNA of healthy subjects: a combined analysis. *Int J Epidemiol* 2010 (published online).