1.a. Full Title: Genome-wide DNA methylation profiling in peripheral blood: quality control and association with demographic characteristics

b. Abbreviated Title (Length 26 characters):

2. Writing Group: ARIC Epigenetics Working Group

Working group members:

James Pankow
Ellen Demerath
Haitao Chu
Jan Bressler
Myriam Fornage
Tom Mosley
Chindo Hicks
Weihua Guan
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. __JP__ [please confirm with your initials electronically or in writing]

First author: Jim Pankow
Address: Division of Epidemiology and Community Health
University of Minnesota
1300 South Second Street, Suite 300
Minneapolis, MN 55454

Phone: 612-624-2883 Fax: 612-624-0315
E-mail: pankow@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:
3. **Timeline:**

Hybridization of the DNA samples to the HM450 methylation array was completed in February 2012. The preliminary dataset is expected to be 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 September 2012.

4. **Rationale:**

Epigenetics is the study of mitotically heritable modifications in chromatin structure (i.e., modifications not involving the underlying 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 cytosine-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 interindividual 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). Because DNA methylation patterns can change over time, EWASs (in contrast to GWASs) are subject to many of the same threats to validity that affect traditional epidemiologic investigations, including reverse causality and confounding by non-genetic factors that may affect both methylation and risk of disease.

Arrays to efficiently profile DNA methylation have only recently become commercially available (Laird, 2010). In ARIC, the recently released Illumina 450K Infinium Methylation BeadChip is being used to measure DNA methylation in peripheral blood obtained from over 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). Before beginning efforts to relate epigenetic variation to disease outcomes, it is important to first characterize the distribution and patterns of DNA methylation obtained in large, population-based samples and to assess associations between DNA methylation patterns and basic demographic factors such as age and sex that may be important confounders in future EWASs. Other studies have found that the relative distribution of cell types within whole blood is associated with the pattern of
DNA methylation (Zhu, 2010; Talens, 2010), suggesting that this source of variability may be an important consideration for epidemiologic studies that rely on blood samples.

5. **Main Hypothesis/Study Questions:**

This initial paper will primarily provide descriptive information about the ARIC epigenetics project and basic distributions in the study population. The Working Group intends to submit separate manuscript proposals for analytic studies relating DNA methylation patterns to relevant exposures (e.g., diet, physical activity) and disease outcomes (e.g., CHD, stroke, diabetes).

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 completely methylated 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 was also analyzed on the arrays in accordance with ARIC study policy.

Inclusion/exclusion criteria: DNA extracted from blood collected from 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 genome-wide genotyping data available either using the Affymetrix Human SNP Array 6.0 or the ITMAT-Broad-CARE (IBC) Illumina iSelect custom array. 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.
To characterize the ARIC methylation study design and methods, we will describe:

a) Sample selection procedures
b) Laboratory methods used to process the samples
c) Plate-to-plate variability and reproducibility, measured through use of internal control samples
d) Distribution of missing values across samples and sites

As part of this paper, we may also describe:

e) Distribution of population-level methylation by:
   i. number of CpGs within probes
   ii. probe length
   iii. overall GC content of the source sequence;
   iv. melting temperature of the source sequence;
   v. location of CpG sites (island, shore, shelf, other);
   vi. functional location of sites (promoter, body, 3’ UTR, intergenic);
   vii. type of assay (Infinium I vs. Infinium II)
   viii. length of sample storage

To better understand the demographic correlates of methylation, we will evaluate and quantify associations between array-wide and site-specific methylation patterns and phenotypes such as:

a) White blood cell count
b) Percent distribution of blood cell types (e.g., neutrophils, lymphocytes, monocytes). This will be limited to Forsyth County African Americans, as Jackson participants did not have differentials performed at visit 2.
c) Age at the visit in which blood was collected
d) Gender (separately evaluating sites for autosomal and X/Y chromosomes)
e) Study center (Forsyth or Jackson)
f) Measures of socioeconomic status (education, income)
g) Menopausal status in women
h) Genetic ancestry, estimated by AIMs and/or principal components of ancestry derived from GWAS markers

For the analyses in part (2) above, we will utilize standard regression techniques. For example, to evaluate associations between DNA methylation and gender, we will use linear regression to regress percent methylation (beta, 0 to 100%) at each of the 450,000 methylation sites on gender (male, female), and summarize results across sites through q-q plots, volcano plots, or other techniques. An a priori genome-wide significance threshold will be set at \( p = 2 \times 10^{-6} \) after applying an expected false positive rate of 1 divided by the number of tests conducted (1/450,000). Analyses may require different approaches to account for the unique features of Illumina 450K Infinium Methylation
BeadChip data, such as variance-stabilizing techniques (Du, 2010) and weighting site-specific analyses by probe-specific detection p-values across samples (Kuan, 2010).

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?  ____ 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”?  ____ 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.cscce.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

_ x_  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:


