1.a. **Full Title**: Supervised principal component analysis in epigenome-wide association studies

b. **Abbreviated Title (Length 26 characters)**: SPCA in EWASs

2. **Writing Group**: ARIC Epigenetics Working Group

Working group members:

Weihua Guan
Chong Wu
James S. Pankow
Ellen W. Demerath
Jan Bressler
Megan L. Grove
Myriam Fornage

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:**

We anticipate a draft ready to submit for Publications Committee review by end of 2016.

4. **Rationale:**

Recent technological advances have provided multiple platforms for systematically interrogating DNA methylation variation across the genome (Laird, 2010; Bock. 2012). 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). Failing to control for these non-genetic factors in epigenome-wide association studies (EWASs) may lead to false discoveries and loss of statistical power. These factors, if available, can be directly included in a regression model, or can be inferred through decomposition of the methylation matrix, e.g., using principal component analysis (PCA).

Direct adjustment of confounders in regression analysis can only remove the effect of known factors. In practice, the confounders are sometimes not collected, or we do not have a complete knowledge about confounding effects. Therefore, PCA or singular vector-based methods have been proposed (e.g., Leek, 2007, Zou, 2014; Rahmani, 2016) to remove all “unwanted” variation associated with large-scale methylation data. The standard PCA extract surrogate variables using all CpG sites available from a genome-wide array, which may not efficiently capture factors that are only associated with a small proportion of sites. Rahmani et al. (2016) proposed a sparse PCA-based approach to select a small subset of CpG sites, which can hopefully distinguish between different biological conditions the most. From the selected sites, the principal components are computed and are included as surrogate variables in EWASs. Because the sparse PCA approach still selects the CpG sites from the entire genome-wide array, it ignores the correlation between methylation and phenotype of interest. Here we propose to use a supervised PCA method (Barshan, 2011) which derives the principal components having maximal dependence on the phenotype of interest.

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

This paper will develop a supervised and standard principal component analysis based approach to adjust the confounders and important covariates. We will demonstrate the performance of the proposed method by studying the 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).**

Bisulfite-treated DNA extracted from blood collected from 2,905 ARIC African-American study participants at Visit 2 (1990-92; n=2,504) or Visit 3 (1993-95; n=441) was included on the HM450 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 Genome-Wide
Human SNP Array 6.0, the Illumina HumanCVD Genotyping BeadChip (also named the
Illumina IBC BeadChip), or the Illumina HumanExome BeadChip.

We develop a principal component based method for adjusting confounders in
methylation data, using the fact that methylation levels at some CpG sites are likely to be
predictive for unobserved confounders. Specifically, we propose to run a supervised
principal component analysis (SPCA) to the significant CpG sites selected by the linear
regression model and run a standard PCA to all of the CpG sites. The idea is that SPCA
will be able to capture the unobserved confounders, leading to an acceptable type 1 error
rate, while standard PCA can help to control the impact of the important covariates (both
confounders and independent factors), leading to improved power.

We will use computer simulations to investigate the performance of the proposed method
in terms of type 1 error rate and statistical power. We will evaluate the impact of
unobserved confounder patterns. We will also compare the results to alternative
approaches, such as ReFACTor (Rahmani, 2016) and FaST-LMM-EH (Zou, 2014).

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
      ___x__ 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.cscd.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)?

Genome-wide DNA methylation profiling in peripheral blood: quality control and association with demographic characteristics (manuscript number #1929)

Genome-wide methylation analyses of cardiovascular disease (CVD) and its risk factors (manuscript number #1928)

11.a. Is this manuscript proposal associated with any ARIC ancillary studies or use any ancillary study data? __ Yes  __x__ 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)*__________ __________ __________)

*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:


