ARIC Manuscript Proposal #2364

PC Reviewed: 5/13/12	**Status: A**	**Priority: 2**
SC Reviewed: _____	**Status: _____**	**Priority: ____**

1.a. Full Title: Efficient multiple imputation for missing phenotype using genome-wide DNA methylation data

b. Abbreviated Title (Length 26 characters):

2. Writing Group: ARIC Epigenetics Working Group

Working group members:

Weihua Guan
Maitreyee Bose
Chong Wu
Baolin Wu
James Pankow
Ellen Demerath
Megan L. Grove
Jan Bressler
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]**

F**irst 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 in summer 2014.

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). While these wide-range correlations pose analytical challenges including reverse causality and confounding by non-genetic factors in epigenome-wide association studies (EWASs), it brings opportunities to infer missing phenotype values using the rich methylation data.

Missing data is a common problem in applied research. Restricting the analysis to complete cases will reduce the size of data and the power of association tests, which is important for EWASs with stringent genome-wide significance level. On the other hand, leaving covariates with large proportion of missing values from the regression model may lead to false positive findings due to confounding. For example, the relative distribution of white blood cell (WBC) types is well known to be associated with the pattern of DNA methylation (Adalsteinsson et al., 2012), along with many disease states and their risk factors (e.g, de Jong et al., 1997). It is therefore important to control for WBC types in EWASs, which may not be available to the investigators. Other examples include missing WBC counts and sample ethnicity.

Statistical methods have been developed to impute missing data and draw appropriate association inference. The multiple imputation method (Little and Rubin, 2002) involves multiple draws of plausible values for missing data. The imputed datasets can then be analyzed separately using standard complete-case analysis, with the point estimates and corresponding covariance matrix being combined subsequently for final inference. Given the large-scale methylation data, such as that provided by the Illumina Infinium HumanMethylation450 (HM450) BeadChip (Illumina Inc., San Diego, CA), imputation accuracy can be greatly improved compared to standard approach (based on complete cases only). Imputation can be made from multiple CpG sites that are associated with the phenotype with missing values.

5. Main Hypothesis/Study Questions:

This paper will develop a multiple imputation-based approach to impute missing phenotype values using DNA methylation data. 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 multiple imputation methods for analysis of methylation data, using the fact that methylation levels at multiple CpG sites are likely to be predictive for missing value of phenotypes. Let M denote the methylation level, Y the trait of interest, X the phenotype with missing data, and Z other covariates without missing values. At CpG site j, the M-X association can be tested in the following regression model:

$$M_j = Y\alpha + X\beta + Z\gamma + \varepsilon.$$

We first assume X is continuous. To impute missing values in X, we first determine a panel of CpG sites that are associated with X. To do so, we run an EWAS on complete cases with X being the outcome, at each CpG site included on the HM450 chip:

$$X = Y\alpha^* + M_j\beta^* + Z\gamma^* + \varepsilon$$

The sites with genome-wide significance or the K sites (say K = 500) with the smallest p-values will be selected for imputation purpose. In standard imputation approach, we can draw $X^{(i)}$, one set of imputed data based on methylation level at the ith CpG site, from a posterior predictive distribution. Using epigenome wide methylation data, we can improve the imputation precision by combining the imputations from the K selected sites. Specifically, we propose to use the mean imputation averaged across the K sites, and derive the its variance from a multivariate normal distribution. If the covariate X is discrete, other type of regression models can be applied to obtain the prediction. For example, for a binary X, logistic regression will be used. When multiple covariates contain missing values, a sequential imputation method (Raghunathan, 2001) can be applied.

The imputation will be performed multiple times (typically 5-10 times, but more may be needed for large proportion of missing data). EWAS of trait Y will then be carried out on the imputed datasets separately. The association results at each CpG site will be combined using standard multiple imputation formula (Little and Rubin, 2002).

We will use computer simulations to investigate the performance of the proposed imputation method in terms of type 1 error rate and statistical power. We will evaluate the impact of missing data proportion and number of CpG sites used for imputation. We will also compare the results to alternative approaches, such as complete-case analysis, and principal component-based approaches.

In the ARIC methylation data, there are only 175 samples with measured WBC subtypes, and WBC counts were not obtained for samples from Visit 3. To control for potential confounding by WBC count and WBC subtypes, we will apply the multiple imputation technique to the EWAS of smoking status in the ARIC study.

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? ____ 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:

Adalsteinsson, B.T., et al., Heterogeneity in white blood cells has potential to confound DNA methylation measurements. PLoS ONE, 2012. 7(10): p. e46705.

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

de Jong, J.W., et al., Peripheral blood lymphocyte cell subsets in subjects with chronic obstructive pulmonary disease: association with smoking, IgE and lung function. Respir Med, 1997. 91(2): p. 67-76.

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

Little, R.J.A. and D.B. Rubin, Statistical analysis with missing data. 2nd ed. Wiley series in probability and statistics2002, Hoboken, N.J.: Wiley. xv, 381 p.

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