ARIC Manuscript Proposal #2084

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

1.a. Full Title:
DNA Sequence Variation and the Human Metabolome in African Americans from the Atherosclerosis Risk in Communities (ARIC) Study

b. Abbreviated Title (Length 26 characters): DNA Sequence & Metabolome

2. Writing Group:
Writing group members: Bing Yu, Alexander H. Li, Yan Zheng, Donna Muzny, Danny Alexander, Josef Coresh, Richard Gibbs, Eric Boerwinkle

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

First author: **Bing Yu**
Address: Department of Epidemiology, Human Genetics and Environmental Sciences,
University of Texas Health Science Center at Houston
1200 Hermann Pressler, Suite E-435
Houston, TX, 77030
Phone: 713-500-9892 Fax: 713-500-0900
E-mail: Bing.Yu@uth.tmc.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: Eric Boerwinkle
Address: Department of Epidemiology, Human Genetics and Environmental Sciences,
University of Texas Health Science Center at Houston
1200 Hermann Pressler, Suite E-447
Houston, TX, 77030
Phone: 713-500-9816 Fax: 713-500-0900
E-mail: Eric.Boerwinkle@uth.tmc.edu

3. Timeline:
Analysis is to start as soon as approval is obtained. Manuscript is to be prepared as soon as analysis is available. We expect that the manuscript will be prepared within six months from approval of the analysis plan.

4. Rationale:
Understanding the role of genetic and environment interaction is essential to advance our knowledge on diseases etiology, pathophysiology and prevention. Human

metabolome represents downstream changes in the genome, transcriptome and proteome,[1,2] and thus, provides a functional feature of the physiological and pathophysiological status of the body. Recent genome-wide association studies (GWAS) in Caucasians have identified several common genetic loci for human blood metabolites; however, the underlying causality is not clear.[3-6] Whole-exome sequencing comprehensively identifies common and rare protein-coding changes in the human genome providing novel functional insights into the biology and epidemiology of disease and risk factor associations.[7-9]

The ARIC study is one of the largest cohorts in the United States with a multi-ethnic sampling framework. Over 90% of study participants have GWAS genotypes already available. Thus, the goal of this study was to use the human metabolome in ARIC African-Americans to identify novel pathways underlying disease etiology and possible novel avenues of prevention and treatment.

5. Main Hypothesis/Study Questions:

Aims:

1. To evaluate the common genetic determinants of metabolomic profiles in African-Americans using GWAS.
2. To evaluate the association between genes, metabolites and diseases identified in Aim 1.
3. To explore loss-of-function (LOF) variant effects on the human metabolome using whole-exome sequencing.

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 samples and laboratory measurements

This is a cross-sectional study that consists of African American ARIC participants at the baseline (visit 1) with serum metabolites and genotypes measured. In total, detection and quantification of 602 metabolites were measured using an untargeted, gas chromatography-mass spectrometry and liquid chromatography-mass spectrometry (GC-MS and LC-MS)-based metabolomic quantification protocol. Metabolites will be excluded if: 1) more than 50% of the samples had values below the detection limit; and 2) they had unknown chemical structures.

SNPs were genotyped on the Affymetrix 6.0 chip and were imputed to ~2.5 million SNPs based on a panel of cosmopolitan reference haplotypes from HapMap CEU and YRI. Exomes were captured on Nimblegen's VChrome2.1 (Roche NimbleGen, Madison, WI, USA) and the captured exomes were sequenced on Illumina HiSeq 2000 (Illumina, San Diego, CA, USA). Sequence reads were aligned to the hg19 reference genome using Burrows–Wheeler Aligner.[10] SNPs were called using AtlasSNP[11] and variants were annotated using Cassandra.[12]

Statistical Analysis:

A total of 308 metabolites will be included in this study. Metabolite levels below the detectable limit of the assay will be imputed with the lowest detected value for that metabolite in all samples, and all metabolites values will be natural log-transformed prior to the analyses. In a secondary analysis, metabolite levels below the detectable limit of the assay will be imputed to zero to see if improved signal will be gained.

For the GWAS analyses, linear regressions and an additive genetic model will be applied to each metabolite, adjusting for age, sex and the first 10 principal components. Genome-wide significance will be defined as a p-value $< 1.6\times10^{-10}$ based on Bonferroni correction. If more than one significant SNP clustered at a locus, the SNP with the smallest p-value will be reported as the sentinel marker.

For whole-exome sequencing analyses, all autosomal LOF variants (defined as stopgain or stopgain/splice variants) will be identified, and two lines of analysis will be pursued. First, two collapsing tests, a single summary T5 test and a Sequence Kernel Association Test (SKAT)[13] will be conducted to evaluate the effects of rare LOF alleles (defined as MAF < 5%) on serum metabolite levels. Statistical significance for the DNA sequence analysis will be defined as a p-value $< 2.2\times10^{-7}$ using Bonferroni correction. Second, a recessive genetic model will be used to analyze the impact of homozygous and compound heterozygous LOF alleles on the serum metabolome. In particular, we will determine if these individuals reside in the extreme tails of the phenotypic distributions.

References:

1. German JB, Hammock BD, Watkins SM. Metabolomics: Building on a century of biochemistry to guide human health. *Metabolomics : Official journal of the Metabolomic Society*. 2005;1:3-9
2. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, et al. Hmdb: The human metabolome database. *Nucleic acids research*. 2007;35:D521-526
3. Gieger C, Geistlinger L, Altmaier E, Hrabe de Angelis M, Kronenberg F, Meitinger T, et al. Genetics meets metabolomics: A genome-wide association study of metabolite profiles in human serum. *PLoS genetics*. 2008;4:e1000282
4. Illig T, Gieger C, Zhai G, Romisch-Margl W, Wang-Sattler R, Prehn C, et al. A genome-wide perspective of genetic variation in human metabolism. *Nature genetics*. 2010;42:137-141
5. Suhre K, Shin SY, Petersen AK, Mohney RP, Meredith D, Wagele B, et al. Human metabolic individuality in biomedical and pharmaceutical research. *Nature*. 2011;477:54-60
6. Kettunen J, Tukiainen T, Sarin AP, Ortega-Alonso A, Tikkanen E, Lyytikainen LP, et al. Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nature genetics*. 2012;44:269-276
7. Bilguvar K, Ozturk AK, Louvi A, Kwan KY, Choi M, Tatli B, et al. Whole-exome sequencing identifies recessive wdr62 mutations in severe brain malformations. Nature. 2010;467:207-210
8. Musunuru K, Pirruccello JP, Do R, Peloso GM, Guiducci C, Sougnez C, et al. Exome sequencing, angptl3 mutations, and familial combined hypolipidemia. The New England journal of medicine. 2010;363:2220-2227
9. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature. 2012;485:237-241
10. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics. 2009;25:1754-1760
11. Shen Y, Wan Z, Coarfa C, Drabek R, Chen L, Ostrowski EA, et al. A snp discovery method to assess variant allele probability from next-generation resequencing data. Genome research. 2010;20:273-280

12. Bainbridge MN, Wiszniewski W, Murdock DR, Friedman J, Gonzaga-Jauregui C, Newsham I, et al. Whole-genome sequencing for optimized patient management. Science translational medicine. 2011;3:87re83
13. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. American journal of human genetics. 2011;89:82-93

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

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

8.c. If yes, is the author aware that the participants with RES_DNA = 'not for profit' restriction must be excluded if the data are used by a for profit group?
_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

Yes. There is no overlap between this proposal and current proposals/published manuscripts.

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)?

Nettleton J, Follis J L, Alonso A, et al. "Metabolomic predictors of incident heart failure: a case-control study nested within the Atherosclerosis Risk in Communities (ARIC) Study". Poster section presented at American Heart Association(AHA) Epidemiology Council meeting in Atlanta, GA; March 2011.

MS#1847 Zheng Y, et al. Role of the Human Metabolome in Incident Heart Failure Etiology among African Americans in the Atherosclerosis Risk in Communities (ARIC) Study

MS#1853 Yu B, et al. Genome-Wide Association Study of Heart Failure-related Human Metabolite Profiles among African Americans in the Atherosclerosis Risk in Communities (ARIC) Study
MS#1882 Yu B, et al. A longitudinal Study of Metabolomics and Kidney Function among African Americans in the Atherosclerosis Risk in Communities (ARIC) Study
MS#1918 Zheng Y, et al. Metabolomics and Incident Hypertension among African Americans The Atherosclerosis Risk in Communities (ARIC) Study

11. a. Is this manuscript proposal associated with any ARIC ancillary studies or use any ancillary study data? Yes
11.b. If yes, is the proposal
___ A. primarily the result of an ancillary study (list number* 2008.16 **)**
___ B. primarily based on ARIC data with ancillary data playing a minor role (usually control variables; list number(s)* 2008.16 "Metabolomics & Heart Failure: A Novel Approach to Biomarker Discovery"**)**
*ancillary studies are listed by number at http://www.cscc.unc.edu/aric/forms/

12. 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.
Yes, the lead author is aware that manuscript preparation is expected to be completed in 1-3 years, and if this expectation is not met, the manuscript proposal will expire.