ARIC Manuscript Proposal #2792

PC Reviewed: 712/16	**Status: A**	**Priority: 2**
SC Reviewed: _____	**Status: _____**	**Priority: ____**

1.a. Full Title: Whole Genome Sequence and Metabolomics for Gene Discovery in the Atherosclerosis Risk in Communities (ARIC) Study

b. Abbreviated Title (Length 26 characters): ARIC WGS and Metabolomics

2. Writing Group:

Writing group members:

Bing Yu, Paul S. de Vries, Elena V. Feofanova, Azam M. Yazdani, Xiaoming Liu, Zhe Wang, Ginger A. Metcalf, Lynne E. Wagenknecht, Richard A. Gibbs, Alanna C. Morrison and 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: Human Genetics Center
1200 Pressler Street, Suite E-641

Phone: 713-500-9285 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: Human Genetics Center
1200 Pressler Street, Suite W114A

Phone: 713-500-9058 Fax: 713-500-9020
E-mail: Eric.Boerwinkle@uth.tmc.edu

3. Timeline:

The data are available, and analysis is to start as soon as approval is obtained. Because of the large and complex nature of the metabolomics data, multiple manuscripts may emerge from this work and this proposal. The 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:

Most contemporary genomic studies have achieved adequate power by increasing the size of the discovery sample to tens or hundreds of thousands of individuals. An alternative approach for detecting novel genes with variants of functional effect is to measure phenotypes that more immediately reflect genome function (*1*). By focusing on proximal measures of cellular, physiologic and metabolic processes, we optimize the size of a gene's effect relative to corresponding risk factor level or disease endpoint. An illustrative example is the effect of rare, loss-of-function (LoF) mutations in APOC3 on triglycerides levels (TG) (with ~40% decrease per allele) detected by sequencing 3,734 individuals; however, in the same study over 100,000 individuals were studied to reveal the association between these LoF mutations and coronary heart disease (CHD) (*2*).

The human metabolome is a collection of small molecules reflecting a variety of cellular and physiologic processes. The metabolome may be risk factors for future disease or biomarkers of current disease processes. A few common variants have been reported by GWASs on human metabolites (*1, 3-6*). With the advance in sequencing technology, low-frequency variations with more marked functional consequences demonstrated a large cumulative effect on phenotypic variation. Demirkan et al (*7*) reported additional variants that modulate metabolite levels independently of the GWAS hits using candidate genes based exome-sequencing approach. To date, no study has assessed the low-frequency variations captured by whole genome sequence (WGS) on human metabolites in a biracial population.

5. Main Hypothesis/Study Questions:

1. To identify novel loci associated with inter-individual differences in levels of individual metabolites. Special attention will be given to rare variants and regulatory regions.

2. To identify novel loci associated with inter-individual differences in the network of metabolites.

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).

This is a cross-sectional study that consists of 3,200 ARIC participants at the baseline (visit 1) with serum metabolites and WGS data. Metabolites were measured using an untargeted, gas chromatography-mass spectrometry and liquid chromatography-mass spectrometry (GC-MS and LC-MS)-based metabolomic quantification protocol. WGS data was generated by the BCM HGSC using the HiSeq X (Illumina, Inc.; San Diego, CA).

We will primarily focus on 245 metabolites detected in both European Americans and African Americans. Metabolites to be analyzed will be: a) excluded when more than 75% samples had missing values or values below the detection limit (BDL); b) missing/BDL less than 25% will be analyzed as a continuous variable; where missing/BDL <25% will impute the lowest value to those with missing data; and c) metabolites with missing data between 25%-75% will be analyzed as an ordinal variables. A second strategy will be multivariable imputation using the k nearest neighbor method of Verboven et al (*8*).

Whole genome annotation: Our annotation pipeline is the result of decades of work with clinical/Mendelian genomes and large epidemiologic (e.g. CHARGE) whole genome sequencing projects. It is broadly divided into two parts: variant-centric and gene/region-centric. Variant-centric annotation includes (but is not limited to): coding, regulatory, splicing, and disease-related information. Gene/region-centric annotation includes (but is not limited to): gene, transcript and cell type specific epigenetic annotation from Encode, Roadmap and FANTOM5.

Single variant analyses: All gene-based nonsynonymous variants with MAF > 5% will be evaluated individually for association with the metabolomics measures. Within the genomic subset of non-coding variation, the primary focus will be hypothesized regulatory variants (e.g. RegulomeDB) with MAF > 5%. This approach will, in part, improve the power and facilitate interpretation of any significant results. Given the focus on common variation for single variant analyses, standard regression approaches will be applied adjusting for age, sex, and principal components accounting for population substructure.

Burden tests in annotated domains: A sliding window approach across the genome aggregates variants within a physical window (defined as 4kb in length beginning at position 0 bp for each chromosome with a skip length of 2kb) (*9*) and relates them to individual metabolite levels and metabolomic patterns. A T5 burden test (*10*) and the Sequence Kernel Association Test (SKAT) (*11*) will be performed on rare variants (defined as MAF $\leq$ 5%) within each window. For genomic coding regions, the priority will be nonsynonymous and loss-of-function variants. Within genomic subset of non-coding variations, the primary focus will be annotated miRNA targets, miRNA genes and regulatory elements (e.g. enhancers). The feature-based analyses will, in part, improve the power and facilitate interpretation of any significant results. All the analyses will adjust for age, sex, and principal components.

Analysis of metabolomics patterns: Two types of metabolomics patterns will be considered for these analyses. The first is based on a priori pathways, such as those available in Kyoto Encyclopedia of Genes and Genomes (KEGG). The second is based on a Bayesian network analysis and resulting directed acyclic graphs (DAG) (*12*).

Significance thresholds: Different significance thresholds will be applied for each analysis to account for the differing number of tests using Bonferroni correction. For regions of the genome that have already been implicated by GWAS or a priori biologic information, the null hypothesis is not whether a modifier locus is present, but rather the biologic nature of the locus.

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 ICTDER 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

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)?

MS#2084 Yu B, et al. DNA Sequence Variation and the Human Metabolome in African Americans from 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? __X__ Yes ____ No

11.b. If yes, is the proposal

_X_ A. primarily the result of an ancillary study (list number* AS#2014.20)

___ 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.

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.

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.

Yes, the lead author is aware of the policy.

13. Per Data Use Agreement Addendum, approved manuscripts using CMS data shall be submitted by the Coordinating Center to CMS for informational purposes prior to publication. Approved manuscripts should be sent to Pingping Wu at CC, at pingping_wu@unc.edu. I will be using CMS data in my manuscript ____ Yes __X__ No.

References:

1. B. Yu *et al.*, Genetic determinants influencing human serum metabolome among African Americans. *PLoS Genet* **10**, e1004212 (2014).
2. Tg *et al.*, Loss-of-function mutations in APOC3, triglycerides, and coronary disease. *N Engl J Med* **371**, 22-31 (2014).
3. S. Y. Shin *et al.*, An atlas of genetic influences on human blood metabolites. *Nat Genet* **46**, 543-550 (2014).
4. E. P. Rhee *et al.*, A genome-wide association study of the human metabolome in a community-based cohort. *Cell Metab* **18**, 130-143 (2013).
5. J. Kettunen *et al.*, Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nat Genet* **44**, 269-276 (2012).
6. K. Suhre *et al.*, Human metabolic individuality in biomedical and pharmaceutical research. *Nature* **477**, 54-60 (2011).
7. A. Demirkan *et al.*, Insight in genome-wide association of metabolite quantitative traits by exome sequence analyses. *PLoS Genet* **11**, e1004835 (2015).
8. S. Verboven, K. V. Branden, P. Goos, Sequential imputation for missing values. *Comput Biol Chem* **31**, 320-327 (2007).
9. A. C. Morrison *et al.*, Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. *Nat Genet* **45**, 899-901 (2013).
10. B. Li, S. M. Leal, Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* **83**, 311-321 (2008).
11. M. C. Wu *et al.*, Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* **89**, 82-93 (2011).
12. A. Yazdani, A. Yazdani, A. Samiei, E. Boerwinkle, Generating a robust statistical causal structure over 13 cardiovascular disease risk factors using genomics data. *J Biomed Inform* **60**, 114-119 (2016).