ARIC Manuscript Proposal # 3324

PC Reviewed: 1/8/19                      Status: _____                      Priority: 2
SC Reviewed: __________                      Status: _____                      Priority: ____

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

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

2. Writing Group:
Writing group members:
Bing Yu, Adrienne Tin, Rachel Ostroff, Richard A. Gibbs, Josef Coresh and Eric Boerwinkle. Other authors will be invited as the work takes shape. Exact authorship will be determined once we are clear of the paper structure.

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

First author: Bing Yu
Address: Human Genetics Center
         1200 Pressler Street, Suite E-641
         Houston, TX 77030

         Phone: 713-500-9285                      Fax: 713-500-0900
         E-mail: Bing.Yu@uth.tmc.edu

First author: Adrienne Tin
Address: Department of Epidemiology, W6017
         Johns Hopkins Bloomberg School of Public Health
         Welch Center for Prevention, Epidemiology, and Clinical Research
         Baltimore, MD 21205

         Phone: 443-287-4740                      Fax:
         E-mail: atin1@jhu.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
         Houston, TX 77030

         Phone: 713-500-9058                      Fax: 713-500-9020
3. **Timeline:**
   The first tranche of Somalogic data has been delivered already and is undergoing QC. The analysis will start once approval is obtained and the data are ready. Because of the large and complex nature of the proteomics 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). Proteomes, characterized by large protein-abundance differences, cell type and time-dependent expression patterns and post-translational modifications (2), are promising intermediate phenotypes for genetic association to provide functional evidence to map disease associations and translate them into clinical applications. Recent genome-wide association studies (GWASs) have identified scores of protein quantitative trait loci (pQTLs) by integrating proteomics, which provides insights into genes, proteins, and pathways that may be causally associated with disease and can serve as therapeutic targets for treatment and prevention (3-5). With advances in sequencing technology, low-frequency variations with more marked functional consequences demonstrated a large cumulative effect on phenotypic variation. Solomon et al. identified cis and trans genetic variation associated with the serum levels of 20 proteins and utilized these pQTLs to study molecular mechanisms underlying disease and physiological phenotypes (6). Most recently, INTERVAL investigators reported 1,927 associations between regions of the genome and 1,478 proteins, where 89% of which were previously unidentified (7). AGES investigators revealed 27 different network modules of proteins – those the protein modules were controlled by cis- and trans-acting genetic variants, which in many cases were also associated with complex disease (8). To date, no study has assessed the low-frequency variations captured by whole genome sequence (WGS) on human proteomes in a biracial population.

5. **Main Hypothesis/Study Questions:**
   1. To identify novel loci associated with levels of individual proteins. Special attention will be given to rare variants, annotated loss of function variants, and regulatory regions.
   2. Use genetic pathway analysis (e.g. from Ingenuity or KEGG) to identify novel sets of loci associated with levels of individual proteins. Follow-up enrichment analysis will ask whether the affected proteins lie in the same or different pathways as the genes.
   3. Identify novel loci (see #1) or sets of loci (see #2) associated with protein levels across a pathway (e.g. from Ingenuity or KEGG).
   4. Identify novel loci (see #1) or sets of loci (see #2) associated with protein levels within a data-driven cluster of correlated proteins and leverage protein interaction databases, e.g. IntAct and STRING to reveal potential protein interactions underlying the correlations.
   5. Leverage functional gene annotation (e.g. kinases) to identify novel loci associated with levels of similarly modified or otherwise acted-on proteins (e.g. phosphoproteins).
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 ~6,000 ARIC participants at visit 5 with plasma proteomes and WGS data. Proteomes were measured by SOMAscan assay (Somalogic Inc., Boulder, CO) and WGS data was generated by the BCM HGSC using the HiSeq X (Illumina Inc., San Diego, CA). We will also explore and validate pQTL using other available genetic information in ARIC (i.e. whole exome sequencing, exonchip and/or TOPMed imputation).

We will primarily focus on the proteins that pass the quality control check and have coefficient of variation ≤ 10%. We will investigate the appropriate transformation of the relative fluorescence units (RFU) of protein levels, such as log and inverse normal transformation.

Whole genome annotation: Our annotation pipeline is the result of decades of work with clinical/Mendelian genomes and large epidemiologic (e.g. CHARGE and TOPMed) WGS 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: The focus is annotated functional single nucleotide variants, indels and structural variants. All gene-based nonsynonymous variants with MAF ≥ 1% will be evaluated individually for association with the proteomics measures. Within the genomic subset of non-coding variation, the primary focus will be hypothesized regulatory variants (e.g. RegulomeDB) with MAF ≥ 1%. 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 protein levels and proteomic patterns. A T1 burden test (10) and the Sequence Kernel Association Test (SKAT) (11) will be performed on low-frequency variants (defined as MAF < 1%) 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.

Burden tests in functional annotation classes: The statistical methods are the same as the above. In contrast to the above, the annotation will not focus on physical proximity but rather functional
similarity. Functional classes of proteins to be considered will included, but are not limited to GO and InterPro.

Analysis of proteomic pathways and functional classes: There are a myriad of ways to organize proteins into related sets, including presumed function and being members of the same biologic pathway. Using the statistical genetic analysis methods described above, we will analyze the vector protein levels within a pathway or functional class.

Analysis of data-driven clusters of correlated proteins: Besides the functional similarities mentioned above, one can organize the proteins into directed acyclic graphs (DAG) (12) or clusters based on structure within the data itself. Using the statistical genetic analysis methods described above, we will analyze the vector protein levels within a graph’s domain or cluster.

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 ___ 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#2792 Yu B, et al. Whole Genome Sequence and Metabolomics for Gene Discovery in 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*  __AS2017.27__)
  ____  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/](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/](http://publicaccess.nih.gov/) are posted in [http://www.cscc.unc.edu/aric/index.php](http://www.cscc.unc.edu/aric/index.php), under Publications, Policies & Forms. [http://publicaccess.nih.gov/submit_process_journals.htm](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.

References:
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).