ARIC Manuscript Proposal #1530B

PC Reviewed: 8/14/12	**Status: A**	**Priority: 2**
SC Reviewed: _____	**Status: _____**	**Priority: ____**

1.a. Full Title: Evaluation of whole genome sequence in relation to quantitative traits

b. Abbreviated Title (Length 26 characters):

2. Writing Group: CHARGE-S WGS Working Group

Writing group members: Alanna Morrison*, Arand Voorman*, Andrew Johnson*, Xiaoming Liu, Jin Yu, Alexander Li, Donna Muzny, Fuli Yu, Kenneth Rice, Gerardo Heiss, Christopher O'Donnell, Bruce Psaty, Adrienne Cupples, Richard Gibbs, Eric Boerwinkle

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

First author: Alanna C. Morrison
Address: University of Texas Health Science Center at Houston
1200 Herman Pressler; Suite 453E; Houston, TX 77030

Phone: 713-500-9913 Fax: 713-500-0900
E-mail: alanna.c.morrison@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: University of Texas Health Science Center at Houston
1200 Herman Pressler; Suite 453E; Houston, TX 77030

Phone: 713-500-9816 Fax: 713-500-0900
E-mail: eric.boerwinkle@uth.tmc.edu

3. Timeline: Whole genome sequence available in consecutive data freezes. First freeze of data is available in Summer 2012. Second freeze of data available in Fall 2012. Analyses can begin immediately for the first freeze of data.

4. Rationale:

The allelic architecture of complex traits is currently unknown, but could involve contributions from both common and rare genomic variation. Whole genome sequence (WGS) data can be interrogated to characterize the genetic architecture of heart, lung and blood risk factor phenotypes, such as high density lipoprotein cholesterol (HDL-C), low density lipoprotein cholesterol (LDL-C), total cholesterol, and triglycerides.

5. Main Hypothesis/Study Questions:

This study will assess three major genotype-phenotype relationships in an unbiased and coordinated approach:

1) estimate the relative contribution of common and rare variation to the heritability of a quantitative trait relevant to health and disease
2) identify individuals who carry variants causing Mendelian disease and analyze the effects of these variants on phenotypes in apparently asymptomatic individuals
3) determine the relative value of regulatory and non-protein coding regions of the genome compared with the protein coding portions.

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

ARIC participants that have been previously selected for exome sequencing will be available for whole genome sequencing conducted at the Baylor College of Medicine Human Genome Sequencing Center as a part of CHARGE-S. The first freeze of data will contain 990 individuals comprised of 406 from ARIC, 238 from CHS, and 346 from FHS.

Statistical analyses

Prior to statistical analyses, rigorous quality control measures will be applied to the data. These measures will follow closely the analysis plan as proposed by the CHARGE-S WGS Working Group.

To estimate the heritability explained by common autosomal variants (minor allele frequency, MAF>0.01), a mixed linear model $\mathbf{y} = \mathbf{X\beta} + \mathbf{g_c} + \boldsymbol{\varepsilon}$ is defined, where $\mathbf{y}$ is the phenotype vector, $\boldsymbol{\beta}$ is the vector of fixed effects (these can include covariates such as age, gender, BMI, and study center) and $\mathbf{g_c}$ is the vector of cumulated genetic effects explained by the common variants. The vector $\mathbf{g_c}$ is assumed to follow $N\left(0, \mathbf{A_c^*}\sigma_g^2\right)$ where A_c^* is the GRM estimated using common autosomal variants. The heritability is estimated as $h_c^2 = \sigma_c^2/\sigma_p^2$ using the restricted maximum likelihood (REML) method[1], where σ_p^2 is the phenotypic variance. GRMs can also be estimated using variants of different MAF bins.

To characterize the distribution of Mendelian variants observed in the WGS data, we will first identify those variants previously reported to cause Mendelian disorders in the

Human Gene Mutation Database (HGMD-DM)[2] associated with heart, lung and blood diseases or risk factor phenotypes. For the HGMD-DM variants identified in the WGS data, we can assess whether they are associated with altered values the trait(s) of interest (e.g., altered HDL-C levels) based on published characteristics (www.omim.org) for these variants. Finally, the distribution of trait values for all carriers and non-carriers of these Mendelian variants will be evaluated. We expect that carriers of the alternative allele will lie in the extremes of the quantitative trait distribution.

Finally, our approach to evaluating phenotype-WGS association allows for a global survey of the entire genomic landscape, complemented by an annotation-based assessment of the genome. First, all variants with MAF>1% will be evaluated for association with the quantitative trait (e.g., HDL-C), assuming an additive genetic model. Next, a sliding window across the genome will consider the aggregate contribution of variants to the trait. Sliding windows will begin at position 0 bp for each chromosome and the skip length is 2 kb. Within each window, a T1 test generates a statistic for each person that counts the number of variant alleles across variants with MAF <1%. The summary statistic for each person is used in the linear regression model and a Wald test was used to assess significance. SKAT will also be used to perform a score test for the model that includes all variants within the window. This allows for heterogeneous effect of the variants within a window, but has lower power for homogeneous effects relative to T1.

T1 and SKAT tests will also be used to assess phenotype-WGS association among annotated domains. Annotation may include regulatory regions of the genome [e.g., as determined by the Open REGulatory ANNOtation database (ORegAnno) http://www.oreganno.org/]. Annotation may include gene-based methods such as genes annotated by the RefSeq database.[3]

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

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)?

There are currently no other ARIC manuscript proposals related to assessment of whole genome sequence (WGS) data.

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

___ A. primarily the result of an ancillary study (list number* 2009.14)

___ 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

1. Patterson H, Thompson R. Recovery of inter-block information when block sizes are unequal. *Biometrika*. 1971;58:545-554
2. Stenson PD, Ball E, Howells K, Phillips A, Mort M, Cooper DN. Human gene mutation database: Towards a comprehensive central mutation database. *J Med Genet*. 2008;45:124-126
3. Pruitt KD, Tatusova T, Klimke W, Maglott DR. Ncbi reference sequences: Current status, policy and new initiatives. *Nucleic Acids Res*. 2009;37:D32-36