ARIC Manuscript Proposal # 1794r

PC Reviewed: 7/12/11	**Status: A**	**Priority: 2**
SC Reviewed: _________	**Status: _____**	**Priority: ____**

1.a. Full Title:

Interaction between HMGCR and LIPC affects High-Density Lipoprotein Cholesterol in both European Americans and African Americans

b. Abbreviated Title (Length 26 characters):

Interaction affecting HDL-C

2. Writing Group:
Writing group members:

Li Ma, Alon Keinan, Eric Boerwinkle, Charles Sing, Ariel Brautbar, Andrew Clark

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

F**irst author**: Li Ma
Address: 102F Weill Hall, Cornell University, Ithaca, NY, 14853

Phone: 607-254-1328 Fax: 607-255-4698
E-mail: lm529@cornell.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: 1200 Hermann Pressler, Suite E447
Houston, TX 77030

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

3. Timeline:

Data are collected and the analysis plan has been tested.
Manuscript will be ready for submission by **07/2011**.

Initial target journal: Nature Genetics

4. Rationale:

Plasma lipoprotein levels are associated with risk of CHD: LDL-C is positively associated and HDL-C is inversely associated. The independent relationship with triglycerides is less clear. Previous analyses within ARIC and elsewhere have identified genes (or gene regions) associated with LDL-C, HDL-C and triglycerides. However, the combined additive effect of these loci accounts for less than 5% of the heritability of the traits. This experience is shared with other CHD risk factors and CHD itself. Using plasma cholesterol levels in ARIC as a model system, we will test whether gene-gene interaction account for some of the "missing heritability" underlying cardiovascular disease risk.

5. Main Hypothesis/Study Questions:

Can we identify significant and well-replicated gene-gene interactions affecting lipid traits in ARIC.

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).

The proposed manuscript uses ARIC Affymetrix 6.0 SNP data to test for gene-gene interactions affecting four lipid traits (total cholesterol, low-density lipoprotein cholesterol, triglyceride, and high-density lipoprotein cholesterol). ARIC European American samples are used for discovery and, ARIC African American, Framingham Heart Study, Multi-Ethnic Study of Atherosclerosis samples as the replication data sets.

Statistical model for two-locus epistatic interactions

A quantitative trait and two SNPs are considered in our two-locus epistasis test. Assume that Y is the trait of interest and G_i is the genotype of SNP i (i=1,2). G_i is coded as 0, 1, and 2 respectively according to the number of the reference allele. Two indicator variables x_i and z_i are defined below for each SNP as

$$x_i = \begin{cases} 1, & G_i = 0 \\ 0, & G_i = 1 \\ -1, & G_i = 2 \end{cases} \qquad z_i = \begin{cases} -0.5, & G_i = 0 \\ 0.5, & G_i = 1 \\ -0.5, & G_i = 2 \end{cases}$$

Two linear models with and without epistasis effects are fitted as,

$$Y = Z_0\beta_0 + x_1a_1 + z_1d_1 + x_2a_2 + z_2d_2 + \varepsilon \quad (1)$$

$$Y = Z_0\beta_0 + x_1a_1 + z_1d_1 + x_2a_2 + z_2d_2 + x_1x_2i_{aa} + x_1z_2i_{ad} + z_1x_2i_{da} + z_1z_2i_{dd} + \varepsilon \quad (2)$$

Here, β_0 denotes a vector of intercept and covariates including gender, age, age squared, body mass index (BMI) and principle components for controlling the effects of population stratification. a_i and d_i denote the additive and dominance effects of SNP *i*. i_{aa}, i_{ad}, i_{da}, and i_{dd} are the four interaction effects between the two SNPs. An F test is used to compare models (1) and (2) for detecting epistatic interactions with four degrees of freedom. This epistasis test is similar to the --epistasis option in PLINK except that only additive effects and their interaction are considered in PLINK which results in a one-degree-of-freedom interactive test.

To avoid unreliable results due to small subgroup sizes, we only test SNP pairs where the smallest two-SNP genotype group has no fewer than 20 individuals. Assuming equal allele frequency and linkage equilibrium for the two SNPs, this requirement is equivalent to minor allele frequency no less than 5%.

Epistasis searching strategy

Although we are only focusing on simple two-locus analysis, the total number of tests is still huge, about $5\text{x}10^{11}$ from 1 million SNPs. Due to the stringent multiple comparison correction and low power of epistasis testing, we need to restrict the number of tests at a reasonable level. In addition, we want to enrich epistasis signals in our limited number of tests through different filtering processes using marginal effects, protein-protein interactions (PPIs) and pathway information.

1. Epistasis testing based on marginal effects

In total 95 loci were reported to be associated with the four lipid traits in a recent GWAS meta analysis. We exhaustively test the pairwise interactions among the 125 significant SNPs ($P < 5\text{x}10^{-8}$) in the 95 loci on the four cholesterol related traits and the total number of epistasis tests is <7750 for each trait.

2. Epistasis testing based on PPIs

Over 3000 high-confidence human PPIs are collected for PPI based analysis. For each PPI, we test all the pairwise interactions between SNPs in the first gene and SNPs in the second gene. Assume n_1 and n_2 are the numbers of SNPs in the first and second gene respectively and then the number of epistasis tests is $n_1\text{x}n_2$ for the PPI. Gene information (HG18) from UCSC genome browser is used to map SNPs to genes and SNPs located 5kb upstream and downstream of gene regions are also included for each gene. The total number of pairwise interaction tests is about 6.2 million.

3. Epistasis testing based on pathway information

A gene enrichment test was performed using the genes reported in the meta analysis GWAS paper. The METABOLISM OF LIPIDS AND LIPOPROTEINS pathway is the most significant pathway with p-value almost 0. There are 228 genes in this pathway and

12,716 SNPs are mapped to the genes. All of the pairwise interactions among the 12,716 SNPs are tested and the total number of tests is about 27 million.

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 no related manuscripts that tested for gene-gene interactions

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* _________)
___ 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/

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.

Agree.