ARIC Manuscript Proposal # 1522

PC Reviewed: 6/9/09	**Status: A**	**Priority: 2**
SC Reviewed: _____	**Status: _____**	**Priority: ____**

1.a. Full Title: Mining Gold Dust Under the Genome Wide Significance Level: A Two-Stage Approach

b. Abbreviated Title (Length 26 characters): A Two-Stage GWA Analysis

2. Writing Group:

Gang Shi, Eric Boerwinkle, Alanna Morrison, C. Charles Gu, Aravinda Chakravarti, DC Rao

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

First author: Gang Shi
Address: Washington University School of Medicine
Division of Biostatistics, Campus Box 8067
660 S. Euclid Avenue, St. Louis, MO 63110
Phone: 314.362.3643; Fax: 314.362.2693
E-mail: gang@wubios.wustl.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: DC Rao (rao@wubios.wustl.edu)

3. Timeline: Analysis to begin immediately.

4. Rationale: There is convergence of knowledge that common complex traits derive from etiologic factors with multiple loci and the genetic main effect of each locus is typically small.[1] On the other hand with millions of SNPs in current genome-wide association (GWA) studies, large number of loci with small main effects may fail to pass stringent genome-wide significance level that is set for controlling false positive error rate. With more and more cohorts being genotyped with dense genome-wide SNP markers, pooling evidence through meta-analysis or validating results among studies are now possible, reported association results are however still limited by the genome-wide significance level. This makes it necessary to consider alternative statistical strategies that focus on controlling false discovery rate[2] (FDR) at study-level analysis. We will consider a two-stage approach, in which stage 1 will control false discoveries and stage 2 will control false positives. FDR adjusted threshold will be used in the first stage and LASSO regression[3] in the second. Quantitative traits will be simulated with ARIC genotype data and "causal" SNPs will be selected with varying minor allele frequencies and simulated with different effect sizes. Hopefully, this methodology research will yield some guidance for GWAS analysis in mining "gold dust" under the genome-wide significance level.

5. Main Hypothesis/Study Questions: We hypothesize that genetic loci with small effect sizes can be discovered with a two-stage approach, in which the first controls false discoveries and second controls false positives.

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).

General Analysis Approach:
Subjects*:* European-American in ARIC population.

Exposure*:* 1 million genotyped SNPs and 2.5 million imputed HapMap SNPs

Outcome: simulated quantitative traits.

Primary statistical approach*:* False discovery rate (FDR) approach, LASSO regression analysis.

Secondary statistical approach*:* Ridge regression, elastic net, or other multi-marker approaches will be evaluated if LASSO regression does not yield satisfactory performance.

Statistical significance*:* FDR adjusted genome-wide significance level

Validation and Replication: Since the phenotype data will be simulated, it will not need external replication. Results will be simply checked against the simulation parameters.

Major Phenotypes to Analyze: quantitative phenotypes will be simulated based on ARIC genotype data.

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

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.
___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 manuscript proposals in ARIC since this proposal deals with methodology research based on simulated phenotypes using the GWA genotype 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
__X_ A. primarily the result of an ancillary study (list number* __2006.03__)
___ B. primarily based on ARIC data with ancillary data playing a minor role (usually control variables; list number(s)* _____ _____ _____)

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.**

References

1. Human Molecular Genetics. 2008;17:Review Issue 2.
2. Benjamini Y and Hochberg Y. Controlling the dales discovery rate: A practical and powerful approach to multiple testing. *J Roy Stat Soc B*. 1995;57:289-300.
3. Tibshirani R. Regression shrinkage and selection via the Lasso. *J Roy Stat Soc B.* 1996;58:267-288.