1.a. Full Title: Integration of Multifactor Dimensionality Reduction and Gene Ontology to determine interaction of nonsynonymous SNPs and cardiovascular risk factors on cardiovascular disease

b. Abbreviated Title (Length 26 characters): INTEGRATION OF MDR AND GO

2. Writing Group:
Corinne Aragaki, Kathy Klos, Kelly Volcik, Eric Boerwinkle

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

First author: Corinne Aragaki
Address: Human Genetics Center
UT School of Public Health
University of Texas Health Science Center at Houston,
1200 Herman Pressler Drive, Suite RAS-E511
Houston, Texas 77030

Phone: 713-500-9915 Fax: 713-500-0900
E-mail: Corinne.C.Aragaki@uth.tmc.edu

Corresponding/senior author (if different from first author correspondence will be sent to both the first author & the corresponding author):
Eric Boerwinkle
Human Genetics Center
UT School of Public Health
University of Texas Health Science Center at Houston,
1200 Herman Pressler Drive, Suite RAS-E447
Houston, Texas 77030

Phone: 713-500-9800 Fax: 713-500-0900
E-mail: Eric.Boerwinkle@uth.tmc.edu

3. Timeline:
Summer 2007: Analysis by Corinne Aragaki (she has signed a data distribution agreement)
Fall 2007: Manuscript submission
4. **Rationale:**
Coronary heart disease (CVD) is the leading cause of death in the US (60% of overall mortality). An estimated 34% of all Americans, or 70 million, have CVD with an estimated 927,000 deaths per year at an estimated cost of $393 billion for 2005 (American Heart Association 2005). Identified risk factors for CVD include hypertension and lipid metabolism.

In order to elucidate the pathways involved in complex disease, new analytic strategies need to be developed to integrate the human genome and environmental factors. In particular, the sheer number of potential combinations and integration of the multiple biologic pathways are two problems which need to be solved by new methods. One analytic strategy is to combine current differing analytic techniques to minimize each problem. In the following approach, we propose to use Moore's multifactor dimensionality reduction (MDR) within a biological pathway on intermediate endpoints to determine combinations that impact disease and then integrate multiple pathways in a Hierarchical Bayes approach.

5. **Main Hypothesis/Study Questions:** Are there pathways as defined by biological processes in Gene Ontology that predict cardiovascular disease and stroke in the ARIC cohort?

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).**

Inclusion: ARIC cohort
Intermediate outcome(s) of interest: Hypertension and dyslipidemia at baseline
Outcome(s) of interest: time to first CHD and stroke event.

Exposure(s) of interest:
SNPs: Nonsynonymous SNPs in the lipid metabolism and renin-angiotensin pathways previously typed in the overall ARIC cohort. Namely, we will use the genes: ACE, ADD1, ADRB2, ADRB3, AGT, ANP, APOC1, APOE, AT1R, CETP, CYBA, CYP11B, EDNRA, FBN1, GNB3, HL, LPL, LTA, MMP3, MMP9A, NOS2A, NOS3, PCSK9, PON1A, PPARA, PPARG.
Which correspond to the SNPs: rs4961, rs5050, rs5063, rs11568822, rs429358, rs7412, rs1042713, rs1042714, rs4994, rs5186, rs4673, rs7175654, rs5343, rs5334, rs5333, rs1137933, rs11591147, rs679620, rs5051, rs2297518, rs4364, rs5442, rs12721054, rs708272, rs1799998, rs4544, rs6489738, rs1800588, rs328, rs1041981, rs909253, rs17576, rs1799983, rs28362263, rs28362286, rs28362270, rs662, rs1800206, rs3892755, rs6008259, rs9615784, rs1801282, rs1805192, and rs3856806. There were 4 SNPs without rs#: there ABI Id#s were: PCSK9_AFD1033060, PCSK9_H391N, PCSK9_R469W, PCSK9_Y142X.

Environmental factors: Physical activity (as measured by work physical activity index and sports physical activity index at baseline), Keys score at baseline, BMI at baseline, and smoking at baseline.
Potential confounding factors: age, race, gender, center, education, anti-hypertensive medication use, energy intake, smoking status, and alcohol consumption measured at baseline and at Visit 4.

METHODS: Using candidate nonsynonymous SNPs and cardiovascular risk factors measured in the Atherosclerosis Risk in Communities cohort study, we will determine SNP-environmental factor pathway combinations that impact risk for hypertension and hyperlipidemia. Based on the designations of biological processes from Gene Ontology, we obtained only traceable human biological processes and used only nonsynonymous SNPs within a biological process. If there is more than one gene within a biological process, we will use MDR to reduce the number of SNP-environment interactions that would be used. We will integrate the resulting score matrix on time to cardiovascular event by ethnicity in a hierarchical Bayes Cox regression.

Measurement error: Due to the self-report on many environmental factors (such as diet and physical activity), possible substantial measurement error could be an issue in analysis. To minimize the effect of measurement error, we will use only indicator variables of dietary and physical activity scores.

Multiple comparisons: The issue of multiple comparisons will be minimized due to 1) we will only use pathways as defined by traceable author statements in Gene Ontology for biological processes, 2) we will refine the number of interactions by their effect on intermediate endpoints, and 3) we will use only one comparison by disease endpoint.

7.a. Will the data be used for non-CVD analysis in this manuscript?  No

b. If Yes, is the author aware that the file ICTDER02 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
   (This file ICTDER02 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?  Yes

8.b. If yes, is the author aware that either DNA data distributed by the Coordinating Center must be used, or the file ICTDER02 must be used to exclude those with value RES_DNA = “No use/storage DNA”?  Yes

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.csecc.unc.edu/ARIC/search.php
   Yes
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)?

11. a. Is this manuscript proposal associated with any ARIC ancillary studies or use any ancillary study data? Yes

11.b. If yes, is the proposal
   A. primarily the result of an ancillary study 1995.07

*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.