ARIC Manuscript Proposal # 1198

1.a. Full Title: Genes, environment, and their interactions: Exploring determinants of metabolic risk factors in the ARIC study

b. Abbreviated Title (Length 26 characters): Genetics of metabolic risk

2. Writing Group: Keri Monda, Kari North, Linda Adair, Karen Mohlke, Molly Bray, Keiko Asao, Jim Pankow, Linda Kao, Eric Boerwinkle, and other interested ARIC investigators

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: Keri Monda, PhD
Address: Department of Epidemiology
School of Public Health
University of North Carolina at Chapel Hill
E-mail: monda@unc.edu

Corresponding/senior author (if different from first author correspondence will be sent to both the first author & the corresponding author): Kari North, PhD
Address: Same as above
Phone: (919) 966-2148 Fax: (919) 966-9800
E-mail: kari_north@unc.edu

3. Timeline: All genotyping of cardiovascular disease (CVD) candidate genes has been completed. Work on the manuscript will begin immediately upon approval by the ARIC publications committee. We anticipate that a period of one year will be adequate for analysis and manuscript preparation.

4. Background and Rationale:
The primary objective of the proposed study is to utilize the single nucleotide polymorphism (SNP) genotype data from 36 CVD candidate genes (typed as part of approved ancillary study 2002.06. See attached appendix for full list of genes and SNPs available for study) to investigate the relationship of genetic polymorphism to the clustering of individual metabolic risk factors (e.g. obesity, atherogenic lipid profiles, high fasting glucose and insulin levels, hypertensive blood pressure levels). We are further interested in investigating gene-environment interactions between genetic polymorphism in the candidate genes and physical activity to clustering of these risk factors.
The clustering of metabolic risk factors:
The constellation of metabolic derangements such as obesity, insulin resistance, dyslipidemia, and hypertension that cluster within individuals has lead researchers and organizations to propose the existence of a condition termed the “metabolic” or “insulin resistance” syndrome [1, 2]. Individuals diagnosed with the syndrome have been found to be at increased risk of type 2 diabetes [3, 4], cardiovascular disease [5-7], and premature mortality [8-10]. Obesity and central adiposity are key factors in the development of the metabolic syndrome. Although evidence for the relationship between level of obesity and the presence of multiple risk factors is robust, there is considerable heterogeneity in the presence of the metabolic syndrome within body mass index (BMI) categories, an important limitation in using a standardized definition of the metabolic syndrome. Clustering of metabolic risk factors has been shown previously in the ARIC cohort as a potentially useful method for predicting incident diabetes and other chronic conditions [11, 12]. However, we feel that the use of data-mining methods (such as cluster analysis) to define population-specific metabolic risk factor groupings will provide interesting and potentially useful categorizations for genetic associations.

Genetic basis of metabolic risk:
There are several lines of evidence that support a genetic component to the development of metabolic risk factors. Studies of related individuals have found evidence of significant familial aggregation for individual components of the metabolic syndrome, including obesity [13-15], abdominal obesity [16, 17], blood lipids [18], blood pressure [19-21], and blood glucose/insulin levels [22, 23]. Further studies have investigated the co-occurrence of risk factors and found evidence that pleiotropy (shared gene effects) may underlie the clustering [24, 25]. Results from a number of genome-wide screen linkage analyses provide further evidence of common genetic factors on multiple individual components of the metabolic syndrome [26-28], as well as evidence for clustering [24, 25, 29-32].

Genetic interactions with physical activity:
There is an extensive literature devoted to the study of the effects of physical activity on obesity, glucose tolerance, lipids, and blood pressure, and it is well established that there is great interindividual variation in response to activity. This interindividual variability in response to lifestyle change is likely to be partly determined by genetics and provides a rationale for studying genetic and environmental factors simultaneously. The term gene-environment interaction refers to the idea that one’s genotype may influence the response to the effects of the environment [33], and in its absence, the phenotypic response to an environmental effect is similar across genotypes. A recent study concluded that variants in the leptin and leptin receptor genes modified the effects of regular physical activity on glucose homeostasis in white men and women [34]. Similar gene-environment interactions have been studied in the PPARγ [35] and hepatic lipase genes [36]. Many researchers agree that the development of obesity is dependent upon the presence of not only specific genetic factors but also certain environmental conditions [13].

5. Main Hypothesis/Study Questions:
a. Using a representative subset of the African American and white members of the ARIC cohort (ARIC CRS N = 1065), create distinct groups of individuals via cluster analysis methods based on their metabolic profiles [HDL-C, LDL-C, triglycerides, fasting glucose and insulin,
blood pressure, and anthropometrics (e.g. BMI, waist/hip circumferences]). Group membership will be used as the dependent variable in subsequent analyses.

b. Using the same sample, identify genes important in conferring metabolic risk or benefit by investigating the relationship between the main effects of polymorphisms in 36 candidate CVD genes and the metabolic risk factor clusters. Where we find evidence of an association, we will investigate the relationship between the genotype and individual metabolic risk factor to see if the relationship holds.

c. Investigate the relationship between the main effects of physical activity and the metabolic risk factor clusters in the same population as specified above.

d. Investigate the joint effects of genes and environment (interaction) using polymorphisms of genes found to be important (part b) and physical activity on metabolic risk factor clusters in the same population as specified above.

**Analytical strategy to address study questions:**
Because we have multiple SNPs for each of our 36 candidate genes, we will first test for redundancy between SNPs thus limiting the multiple comparisons necessary. We will also test for deviations from Hardy Weinberg Equilibrium. Prior to modeling we will also investigate the call rate of each individual gene. For all SNPs with a call rate of less than 90% we will review the data and consider exclusion.

**Development of metabolic risk factor clusters:** To create metabolic risk factor clusters we will use cluster analysis to create groups of individuals who have similar metabolic profiles, maximizing intergroup and minimizing intragroup differences. Cluster analysis is distinct from data reduction methods such as factor analysis and principal components analysis in that it is a data mining method used to find groups of individuals similar in specified characteristics (e.g. metabolic risk characteristics). We will use k-means clustering algorithms (PROC FASTCLUS in SAS) to group individuals based on Euclidean distance measures between observations. 1000 iterations of cluster procedures will be conducted to identify the optimal specification for initial cluster centers, with initial group centers randomly generated. Iterations with the largest overall $r^2$ values indicate the maximum inter- to intra-variability ratio, i.e. that individuals differ more from individuals in other clusters than from those in the same cluster. This method allows us to capture the heterogeneity in the clustering of metabolic risk factors, rather than being restricted to an *a priori* definition of the metabolic syndrome. We can then estimate the association between membership in a given cluster and a given genotype.

Clusters will be created using measures of the risk factor variables HDL-C, LDL-C, triglycerides, fasting glucose and insulin, blood pressure (systolic and diastolic), and anthropometrics. Anthropometric variables that will be considered include BMI (from weight and height), and waist and hip circumferences. While some of these variables are not generally used clinically as defining characteristics of the metabolic syndrome, we are interested in investigating multiple risk factors available to us that have been shown to confer cardiovascular risk without being limited by a prescribed definition. In preliminary analyses, clustering will be done both with dichotomous (e.g. presence/absence of hypertension) and continuous values of metabolic risk variables. Clinically derived definitions of metabolic syndrome components will be used to generate the dichotomous variables. Many of these definitions take into account whether an individual is currently being medicated for a given condition (e.g. on anti-hypertensive medication). However,
when clustering using continuous values, we will run analyses including medicated individuals, controlling for medication use, and excluding the individuals to help determine the best model fit for the data. The consideration of diabetics is particularly important considering that treatment for diabetes can affect other cardiovascular risk factors such as obesity measures and cholesterol levels which we are also using as clustering variables. Thus, initially we will include diabetics and then exclude them to determine differential effects and decide upon the best modeling strategy.

Modeling strategy: Our purpose is (1) to determine the main effects of genetic polymorphisms in (up to) 36 candidate genes (study question b) and physical activity (study question c) as well as (2) to explore the potential gene-environment interaction effects (study question d) on metabolic risk factor clusters. For each aim, a general model (2 degree of freedom test) will be utilized to analyze the gene-disease association. All analyses will be stratified by race to account for population stratification. Haplotype analysis will be performed as described below.

We are particularly interested in physical activity as a potential modifier of the genetic – metabolic risk relationship while acknowledging that activity data is measured with more error than the other components of our analysis. Nonetheless, the physical activity assessment used in the ARIC study shows high reliability and good validity in assessing heavy intensity activity, and has been shown to accurately assess certain light intensity activities (e.g. walking and bicycling for pleasure and/or for transportation [37, 38]. In these respects the questionnaire compares well to other physical activity survey instruments. We further acknowledge that because the ARIC population is quite sedentary the limitations in measuring certain light and moderate intensity activities may be compounded.

Cross-sectional models will be used to estimate the associations; specifically, multivariable multinomial logistic regression will be used to estimate the effect of genotype and activity on membership in risk factor cluster. Potential confounders specific to each end-point and identified from the literature will be evaluated and included in models if necessary. Effect measure modification of the association of SNP/haplotypes of the candidate genes with each continuous phenotype will be assessed using a product interaction term.

Haplotype Analysis: An important aspect of this project will be the ability to test multiple genotypes using haplotype analysis methods to consider multiple genotypes within the gene (here we will consider five or more SNPs). The effects of haplotypes will then be tested using strategies described above. We will infer haplotypes using the method developed by Stephens, Smith, and Donnelly for population-based data [39]. This method utilizes Markov chain-Monte Carlo techniques in which unknown haplotypes are derived from a conditional distribution that assumes genealogical relationships between individuals based on the neutral coalescent theory. Lin et al. compared haplotypes estimated using the SSD algorithm to empirically-derived haplotypes and found the SSD method to be over 95.2% accurate for polymorphisms with a minor allele frequency >0.2 over 100 kb of DNA sequence [40]. Because coalescent theory does not consider recombination, errors are introduced using the SSD method when recombination is present. Within blocks of high linkage disequilibrium (i.e., no recombination), the SSD algorithm was demonstrated to be over 98.6% accurate compared to empirical haplotypes. The PHASE software, available from the Oxford Mathematical Genetics Group, with suggested modifications from Lin
et al. will be used to perform the haplotype analyses for this study. This software provides both the estimated haplotypes as well as the probabilities associated with their accuracy. Haplotypes can then be incorporated into the analyses described in the current proposal in place of genotypes, weighted by their posterior probability. Several strategies can be used for testing the effects of haplotypes. Rare haplotypes may be collapsed into more common haplotype groups, which may help to remove the statistical noise that may be introduced by error in haplotype estimation. Presence or absence of individual haplotypes is most easily tested, with systematic modification of sites within the haplotype being added to the model as alternate "alleles."

Multiple comparisons: While progress has been made in recent years in the field of multiple comparisons, an ideal solution remains elusive. The association mapping described in this proposal will involve estimating single-locus models separately for each candidate marker and then evaluating statistical significance. As expected, a large number of dependent tests will be performed, necessitating a correction for multiple comparisons. Often, statistical significance corrections involve the over-conservative Bonferroni correction, which is not appropriate for the correlated SNP data described here.

To address the issue of multiple comparisons, we will apply the false discovery rate (FDR). This method evolved out of the work by Benjamini and Hochberg [41], which focuses on controlling the false discovery rate (described as the percent of tests considered significant that are actually false-positives). Adequately controlling the FDR permits researchers to ensure only a given percentage, \( \alpha \), of all positive results are false. Briefly, the FDR orders the tests by significance level, imposing increasingly rigid alpha levels serially, while adjusting for the number of tests.

6. **Data (variables, time window, source, inclusions/exclusions):**
   a. **Outcome measures (from Visit 1)**
      Metabolic risk factor cluster created from a combination of: BMI (from weight and height), obesity (BMI \( \geq 30 \)), waist and hip circumferences, lipid profile (total cholesterol, LDL cholesterol, HDL cholesterol, triglycerides), insulin, glucose, and blood pressure.
   b. **Exposure measures**
      Genotype data on 36 candidate obesity genes (see Appendix at end of proposal). Physical activity data from Visit 1.
   c. **Study population**
      Representative sample of ARIC cohort (N=1065) from ancillary study 2002.06. All individuals with incident stroke, incident CHD, and in the cohort representative sample were genotyped for the 36 candidate genes that we proposed to analyze herein.

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 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 _____ No
   (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?** _____ 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 ICTDER02 must be used to exclude those with value RES_DNA = “No use/storage DNA”?

X 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.

X 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)?

We identified a few related manuscript proposals in the ARIC study. Below we list these manuscripts and explain why overlap is not an issue for each particular manuscript proposal.

1. Manuscript proposal #274: The influence of anthropometric indices and fasting insulin on changes in the components of the multiple metabolic syndrome. Lead author: Gerardo Heiss. This proposal does not consider genetic data.

2. Manuscript proposal #848: Physical activity and CETP polymorphisms as predictors of HDL cholesterol levels. Lead author: Molly Bray; first author: Mitzi Laughlin. Different genes are being evaluated. Moreover, we have asked Molly Bray if she would like to collaborate on our proposal.

3. Manuscript proposal #1089: Effects of ADRB2 polymorphism and physical activity (and interactions) on hypertension. Lead author: Eric Boerwinkle; first author: Corinne Aragaki. We will be evaluating a different endpoint and we have asked Eric Boerwinkle if he would like to participate on our proposal.

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

X Yes

11.b. If yes, is the proposal

X A. primarily the result of an ancillary study (list number* 2002-06)

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:


Appendix: Candidate genes and number of SNPs per gene available for analysis

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Number of SNPs</th>
<th>Gene name</th>
<th>Number of SNPs</th>
<th>Gene name</th>
<th>Number of SNPs</th>
<th>Gene name</th>
<th>Number of SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADRB2</td>
<td>4</td>
<td>GLP1</td>
<td>8</td>
<td>LIPE</td>
<td>3</td>
<td>POMC</td>
<td>3</td>
</tr>
<tr>
<td>ADRP</td>
<td>6</td>
<td>GLUT</td>
<td>7</td>
<td>LPL</td>
<td>5</td>
<td>PPAR</td>
<td>9</td>
</tr>
<tr>
<td>APM1</td>
<td>5</td>
<td>GNB3</td>
<td>3</td>
<td>MC4R</td>
<td>3</td>
<td>PREF1</td>
<td>5</td>
</tr>
<tr>
<td>C/EB</td>
<td>2</td>
<td>HCRTR1</td>
<td>1</td>
<td>NPY</td>
<td>5</td>
<td>RSTN</td>
<td>3</td>
</tr>
<tr>
<td>CART</td>
<td>3</td>
<td>HL</td>
<td>4</td>
<td>OX</td>
<td>1</td>
<td>RXRA</td>
<td>3</td>
</tr>
<tr>
<td>CCK</td>
<td>3</td>
<td>IGF-1</td>
<td>5</td>
<td>OXR1</td>
<td>4</td>
<td>SREBP</td>
<td>3</td>
</tr>
<tr>
<td>CPT1</td>
<td>3</td>
<td>IGFR1</td>
<td>5</td>
<td>PC1</td>
<td>6</td>
<td>TNFA</td>
<td>2</td>
</tr>
<tr>
<td>FLJ12355</td>
<td>1</td>
<td>LEP</td>
<td>4</td>
<td>PGC1</td>
<td>39</td>
<td>UCP2</td>
<td>4</td>
</tr>
<tr>
<td>GHRL</td>
<td>4</td>
<td>LEPR</td>
<td>5</td>
<td>PLIN</td>
<td>4</td>
<td>UCP3</td>
<td>3</td>
</tr>
</tbody>
</table>