1a. Full Title:

b. Abbreviated Title: miRNA-Mediated Functional Genetic Variants and CVD

2. Writing Group:

Mariaelisa Graff
Kari E. North
Ellen Demerath
Lynne Wagenknecht
Eric Boerwinkle
Christie Ballantyne
Josef Coresh
Ron Hoogeveen

Other investigators welcome

I, the first author, confirm that all the coauthors have given their approval for this manuscript proposal. MG

First author: Mariaelisa Graff
Address: Department of Epidemiology
University of North Carolina at Chapel Hill
137 E. Franklin St, Suite 306
CB #8050
Chapel Hill, NC 27514
Phone: 919-966-8491 Fax: 919-966-9800
E-mail: migraff@unc.edu

Corresponding/senior author (if different from first author correspondence will be sent to both the first author & the corresponding author): Kari North
Address: Department of Epidemiology
University of North Carolina at Chapel Hill
137 E. Franklin St, Suite 306
CB #8050
Chapel Hill, NC 27514
Phone: 919-966-2148 Fax: 919-966-9800
E-mail: kari_north@unc.edu
3. Timeline (The following are estimates. Please see project phases and timeline below for more details):
   Individual cohort statistical analyses: August 31, 2013
   Consortium meta-analyses: November 30, 2013
   Manuscript preparation: March 31, 2014
   Manuscript submission: July 31, 2014

4. Rationale:

Abnormal plasma lipid levels and obesity are important risk factors for cardiovascular disease (CVD), and inflammatory biomarkers are also shown to be associated with CVD risk. Genetic variants associated with these CVD risk factors contribute significantly to CVD, but other genetic factors may be equally important. Specifically, human microRNAs (miRNAs) are short non-coding RNA estimated to regulate 30-60% of genes [1-3] via binding their targets (mRNAs) and have emerged as important epigenetic regulators of biological processes and human diseases, including lipid metabolism, inflammation, and CVD [4-24]. However, the role of genetic variation of miRNA-mediated regulation for CVD risk factors is almost unexplored. Recent studies have identified functional single nucleotide polymorphisms (SNPs) that alter miRNA-mediated gene repression and may potentially alter CVD-related traits [25-30]. In order to focus efforts on SNPs most likely to participate in miRNA target site modification (e.g., functional SNPs), our research group has created a genome-wide miRNA regulatory SNP database (~740,000 SNPs) by integrating miRNA targeting prediction algorithms and databases from various resources. This comprehensive database allows us to assess the genetic effect/contribution of miRNA-mediated regulation on the traits of interest. We hypothesize that SNPs predicted to interfere with miRNA regulation are more likely to be functional and have phenotypic consequences. More specifically, we proposed that miRNA-related genetic polymorphisms mediate CVD risk through altering miRNA-mRNA interactions to modulate the transcript level (mRNA) of a target gene. In addition, miRNA levels can be modulated by environmental factors; therefore, we further hypothesize that the association between miRNA-related SNPs and cardio-metabolic risk can be modified by dietary factors (primary focus), gender and body size. Moreover, we propose that miRNA mediated gene repression is a mechanism through which environmental factors modulate genetic susceptibility to CVD. The combined contributions of genetic variants and environmental factors on the roles of miRNAs in regulating lipid metabolism (primary focus: phase 1 and 2) and inflammatory processes (planning: phase 3) remain
virtually unexplored, and are the major focus of the current proposal. The general concept of research background is depicted in Figure 1.

Figure 1. General Concept

5. Main Hypotheses/Study Questions:

We propose to perform a “functional genome-wide association study (GWAS)” and a “functional genome-wide interaction study (GWIS)” of miRNA regulatory SNPs in the CHARGE cohorts using miRNA regulatory SNP genotype (genotyped and imputed) data and measured environmental factors (diet and body conditions) on three groups of CVD risk factors (as phenotype/dependent outcomes) in individual cohort. This proposal aims to identify SNPs which modify miRNA-target interactions affecting the regulation of lipid metabolism and inflammation markers and to further identify SNPs that respond to an environmental factor (i.e. gene-environment interactions, GxE) through the mechanism of miRNA regulation. This genetically targeted (hypothesis-driven genome-wide) approach is more selective than current/traditional GWAS that use a “hypothesis-free” or “agnostic” approach, so it has the advantage of lowering the multiple testing burden in addition to providing immediate biological hypotheses for SNPs achieving significance. The number of tests is estimated at around 50,000-70,000 SNPs (Bonferroni adjusted $P=1\times10^{-6}$ to $7.14\times10^{-7}$) per outcome. However, in light of the limited understanding of how environmental factors interact with miRNA targeting to affect CVD-related phenotypes, we propose a comprehensive, phased exploration of miRNA regulatory SNPs and several potential environmental modulators (primary focus on dietary macronutrients). We expect that better understanding of miRNA-mediated gene regulation induced by a modifiable environmental factor will generate preventative
strategies to reduce CVD risk through improvement of dysregulated lipid metabolism and reduction of chronic inflammation.

6. Design and Analysis:

Project Phases/Timeline

This proposal is part of a meta-analysis involving several cohorts thus extending the results to a wider population. Therefore, we incorporate the necessary time and methods involved in a genetic meta-analysis below. As this proposal pertains to ARIC study variables, those described below will be from visit 1 baseline exam (1987-1989).

We focus on plasma lipid traits in this analytic plan.

1. **Phase 1a**: Plasma lipid traits (3 GWASs); **TARGET DATE: August 30, 2013**
   - QC and meta-analysis of GWAS main effect: September 15, 2013
   - Published GWAS results will be incorporated.

2. **Phase 1b**: Plasma lipid traits (12 GWISs; 12 Environment main effect); **TARGET DATE: September 30, 2013**
   - QC and meta-analysis of GWIS for plasma lipid traits: November 18, 2013

Subjects/Cohorts

Meta-analysis includes European-descent/white adults. Thus, we propose to contribute data from ARIC European descent participants.

Exclusion Criteria

- **Subjects**:
  - Those with missing data on outcomes, exposure variables, or covariates
  - Non-European-descent; non-whites (self-reported or IBS clustering of GWAS data)

- **Phenotype Outcomes**:
  Suggestions for detecting potential outliers: In general, each cohort could exclude any phenotype outcome value > upper quartile (Q3) + 1.5*interquartile range (IQR) or < lower quartile (Q1) - 1.5*IQR. However, each cohort may apply different criteria for each outcome depending on the characteristics of each cohort. Note: IQR = Q3-Q1

1. Plasma lipid traits:
   - Taking lipid-lowering medications at the time of plasma lipid/lipoprotein measurement
• Non-fasting.
• When LDL-C is calculated using Friedewald equation [31,32], the values assigned for missing for subjects with TG levels > 400 mg/dL (4.52 mmol/L) or chylomicrons are present.

\[
\text{LDL-C (mg/dL)} = \text{TC} - \text{HDL-C} - 0.20 \times \text{TG}
\]

\[
\text{LDL-C (mmol/L)} = \text{TC} - \text{HDL-C} - 0.45 \times \text{TG}
\]

Even at TG levels 2.5 to 4.5 mmol/L, this formula is considered inaccurate [33]. If both TC and TG levels are elevated then a modified formula may be used [34]:

\[
\text{LDL-C (mg/dL)} = \text{TC} - \text{HDL-C} - 0.16 \times \text{TG}
\]

\[
\text{LDL-C (mmol/L)} = \text{TC} - \text{HDL-C} - 0.37 \times \text{TG}
\]

• TG levels > 1000 mg/dL

- Genotypes:
  o Failed subject QC: Study-specific exclusion criteria for sample call rate (individual missingness), heterozygosity, gender checks, cryptic relatedness, sample heterogeneity and ethnic group/population outlier.
  o Genotype data without variation.
  o Failed SNP QC criteria: Please upload unfiltered post-imputation data, i.e., NO FILTERS; we will perform QC/filter ourselves. The followings are the criteria we are going to use.
    ▪ Hardy-Weinberg equilibrium (HWE) \( p < 1 \times 10^{-6} \) (within population; not in admixed population).
    ▪ Genotype missingness < 0.05 (per SNP in a population; call rate < 95%).
    ▪ Poor imputation quality (\( \text{Rsqhat} < 0.3 \), MACH and BIMBAM; PROPER INFO < 0.4 IMPUTE/SNPTEST; information score (INFO) < 0.8, PLINK).
  • Since BIMBAM does not give imputation quality score, \( \text{Rsq} \) will be calculated as the ratio of observed to expected variance: \( \frac{\text{var}}{2p(1-p)} \), where \( p \) is the minor allele frequency (MAF), and \( \text{var} \) is the sample variance of the estimated dosages.
  • Imputation quality measures: Ratio of the observed variance of the allele dosage to the empirically observed variance (for imputed SNPs) and average maximal posterior probability (for imputed SNPs)

- Environmental modifiers:
Suggestions for detecting potential outliers: In general, each cohort could exclude any phenotype outcome value > Q3 + 1.5*IQR or < Q1 -1.5*IQR. However, each cohort may apply different criteria depending on the characteristics of each cohort. Note: IQR = Q3-Q1

For dietary factors, extreme or implausible energy intake and/or dietary data (cohort-specific criteria, e.g., in GOLDN: < 600kcal/day or > 6000kcal/day for males and < 550kcal/day or > 5000kcal/day for females) or failed other cohort-specific QC/criteria.

Phenotype Outcomes

All units should be in SI so that betas are in the same magnitude. All phenotype outcomes are continuous and should be checked for potential outliers and then do natural log transformation for fasting TG.

1. Plasma lipid traits:
   - Fasting TG (mmol/L): TG (natural log transformation)
   - Fasting LDL-C (mmol/L): LDL
   - Fasting HDL-C (mmol/L): HDL

Table 1. Description of Phenotype Outcomes

<table>
<thead>
<tr>
<th>Phenotype Outcome Categories</th>
<th>Phenotype Outcomes</th>
<th>Abbreviation for Model</th>
<th>Conventional Unit</th>
<th>Converting Factor</th>
<th>SI Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Plasma lipid traits</td>
<td>TG</td>
<td>TG</td>
<td>mg/dL</td>
<td>0.0113</td>
<td>mmol/L</td>
</tr>
<tr>
<td></td>
<td>LDL-C</td>
<td>LDL</td>
<td>mg/dL</td>
<td>0.0259</td>
<td>mmol/L</td>
</tr>
<tr>
<td></td>
<td>HDL-C</td>
<td>HDL</td>
<td>mg/dL</td>
<td>0.0259</td>
<td>mmol/L</td>
</tr>
</tbody>
</table>

Genotypes

Selected genotyped and imputed SNPs after applying exclusion criteria AND found in the miRNA regulatory SNP list. Genotyped SNPs are coded as 0, 1, and 2, and imputed SNPs are the expected dosages between 0 and 2 from imputation software. SNPs will be modeled additively for analysis (build 36 preferred).

Environmental Modifiers
1. Dietary factors: All dietary variables will be modeled as a continuous variable (Model E and Model GE).

- **Energy nutrients**
  - Carbohydrate intake (% of total energy)
  - Saturated fatty acid (SFA) intake (% of total energy)
  - Monounsaturated fatty acid (MUFA) intake (% of total energy)
    We would also like correlation statistics for SFA and MUFA. As depending on the population, the sources may be similar or different resulting in higher or lower correlation, respectively.
  - Polyunsaturated fatty acid (PUFA) intake (% of total energy)

### Table 2. Description of Environmental Modifiers

<table>
<thead>
<tr>
<th>Environmental Modifier Categories</th>
<th>Environmental Modifiers</th>
<th>Abbreviation for Model</th>
<th>Variable Type</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>D: Dietary factors</td>
<td>Carbohydrate intake</td>
<td>CHOpct</td>
<td>Continuous</td>
<td>% of total energy</td>
</tr>
<tr>
<td></td>
<td>Saturated fatty acid</td>
<td>SFApct</td>
<td>Continuous</td>
<td>% of total energy</td>
</tr>
<tr>
<td></td>
<td>intake</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Monounsaturated fatty</td>
<td>MUFApct</td>
<td>Continuous</td>
<td>% of total energy</td>
</tr>
<tr>
<td></td>
<td>acid intake</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Polyunsaturated fatty</td>
<td>PUFApct</td>
<td>Continuous</td>
<td>% of total energy</td>
</tr>
<tr>
<td></td>
<td>acid intake</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Model Overview (Cross-Sectional Analyses)

Genome-wide linear regression models or linear mixed effects models for family data will be run in each cohort for genetic (G) and environmental (E) main effect and GxE interaction (GE).

Each model is formatted in 4 components:

- The first part indicates the model is to test genetic main effect (G), environmental main effect (E), or gene and environment interaction (GE).
- The second part indicates which group of phenotype outcomes (1: plasma lipid traits), which should be followed by the exact phenotype abbreviation (Table 1).
- The third part indicates which group of environmental modifiers (0: when testing for genetic main effect, Model G; D: dietary factors), which should be followed by the exact phenotype abbreviation (Table 2).
- The fourth part indicates which group of covariates is used (Table 3) and followed by which group of additional dietary covariates (Tables 4, 5, and 6).

For instance, Model GE-1LDL-DSFApct-LAa is:

LDL = SNP + SFApct + SNP*SFApct + (age + sex + field center + pedigree + population structure PCs) + (ENG)

When creating a final result file, please add the name of the cohort before each model and the date uploaded (MMDDYYYY) at the end. (e.g., GOLDN.GE-1LDL-DSFApct-LAa.07262013). Please see Data Submission section.

- **Model G (Genetic Main Effect):** TRAIT = SNP (as a fixed effect) -- **Phase 1a**
- o Model G-1x:
  - Model G-1x-0-LA: Plasma lipid trait = SNP + (Covariates LA)
- **Model E (Environmental Main Effect):** TRAIT = Environmental factor -- **Phase 1b**
- o Model E-1x:
  - Model E-1x-Dx-LAx: Plasma lipid trait = Environment + (Covariates LA) + (Additional Dietary Covariate)
- **Model GE (G*E interaction):** TRAIT = SNP + Environmental factor + SNP*Environmental factor -- **Phase 1b**
- o Model GE-1x:
- Model GE-1x-Dx-LAx: Plasma lipid trait = SNP + Environment + SNP*Environment + (Covariates LA) + (Additional Dietary Covariate)
### Phase 1a Plasma lipid traits (3 GWASs)

<table>
<thead>
<tr>
<th>Model G</th>
</tr>
</thead>
<tbody>
<tr>
<td>G-1TG-0-LA</td>
</tr>
<tr>
<td>G-1LDL-0-LA</td>
</tr>
<tr>
<td>G-1HDL-0-LA</td>
</tr>
</tbody>
</table>

### Phase 1b Plasma lipid traits (12 GWISs; 12 Environment main effect)

<table>
<thead>
<tr>
<th>Model GE</th>
</tr>
</thead>
<tbody>
<tr>
<td>GE-1TG-DCHOpct-LAa</td>
</tr>
<tr>
<td>GE-1TG-DSFApct-LAa</td>
</tr>
<tr>
<td>GE-1TG-DMUFApct-LAa</td>
</tr>
<tr>
<td>GE-1TG-DPUFApct-LAa</td>
</tr>
<tr>
<td>GE-1LDL-DCHOpct-LAa</td>
</tr>
<tr>
<td>GE-1LDL-DSFApct-LAa</td>
</tr>
<tr>
<td>GE-1LDL-DMUFApct-LAa</td>
</tr>
<tr>
<td>GE-1LDL-DPUFApct-LAa</td>
</tr>
<tr>
<td>GE-1HDL-DCHOpct-LAa</td>
</tr>
<tr>
<td>GE-1HDL-DSFApct-LAa</td>
</tr>
<tr>
<td>GE-1HDL-DMUFApct-LAa</td>
</tr>
<tr>
<td>GE-1HDL-DPUFApct-LAa</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model E</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-1TG-DCHOpct-LAa</td>
</tr>
<tr>
<td>E-1TG-DSFApct-LAa</td>
</tr>
<tr>
<td>E-1TG-DMUFApct-LAa</td>
</tr>
<tr>
<td>E-1TG-DPUFApct-LAa</td>
</tr>
<tr>
<td>E-1LDL-DCHOpct-LAa</td>
</tr>
<tr>
<td>E-1LDL-DSFApct-LAa</td>
</tr>
<tr>
<td>E-1LDL-DMUFApct-LAa</td>
</tr>
<tr>
<td>E-1LDL-DPUFApct-LAa</td>
</tr>
<tr>
<td>E-1HDL-DCHOpct-LAa</td>
</tr>
<tr>
<td>E-1HDL-DSFApct-LAa</td>
</tr>
<tr>
<td>E-1HDL-DMUFApct-LAa</td>
</tr>
<tr>
<td>E-1HDL-DPUFApct-LAa</td>
</tr>
</tbody>
</table>
Main Effect Test (Model G and E)

A regression coefficient ($\beta \pm SE$) for the main effect term for each SNP (Model G) or environmental modifier (Model E) on each phenotype will be calculated in each cohort, and reported values will be meta-analyzed.

Model G: In each cohort, each mapped genotyped or imputed SNP will be tested for association with each phenotype outcome assuming an additive genetic model. Linear regression will be employed for studies of unrelated individuals (population-based), and linear mixed effects models will be used to account for familial relationship in the family-based studies. For the top hits, if SNP genotype data used are imputed, we will confirm the associations with the associations of proxy SNPs that have been directly genotyped on arrays.

Model E: All dietary factors will be modeled as a continuous variable in Model E to evaluate dietary main effect.

Model G Covariates

1. Plasma lipid traits:

Covariates LA: age (in years), sex, field center (if relevant), pedigree (as a random effect, if relevant), and population structure PCs (if relevant)

Plasma lipid trait (LA) = SNP + age + sex + field center + pedigree + population structure PCs

Table 3. Description of Covariates for Model G

<table>
<thead>
<tr>
<th>Phenotype Outcome Categories</th>
<th>Phenotype Outcomes</th>
<th>Abbreviation for Model Name</th>
<th>Covariates</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Plasma lipid traits</td>
<td>TG</td>
<td>TG</td>
<td>LA: age, sex, field center, pedigree, and population structure PCs</td>
</tr>
<tr>
<td></td>
<td>LDL-C</td>
<td>LDL</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HDL-C</td>
<td>HDL</td>
<td></td>
</tr>
</tbody>
</table>

Note: Please refer Table 6 for details for description of covariates.
**Model E Covariate (For Dietary Factors)**

Model G covariates, environmental modifier and additional dietary covariate (Model E covariate for dietary factors; a) should be included in the model E.

Plasma lipid trait (LAA) = Dietary factor + age + sex + field center + pedigree + population structure PCs + total energy

Table 4. Description of Covariates for Model E

<table>
<thead>
<tr>
<th>Environmental Modifier Categories</th>
<th>Environmental Modifiers</th>
<th>Abbreviation for Model Name</th>
<th>Covariates</th>
<th>Additional Dietary Covariate</th>
</tr>
</thead>
<tbody>
<tr>
<td>D: Dietary factors</td>
<td>Carbohydrate intake</td>
<td>CHOpct</td>
<td>LA (for plasma lipid traits): age, sex, field center, pedigree, and population structure PCs</td>
<td>a. ENG (total energy intake)</td>
</tr>
<tr>
<td></td>
<td>Saturated fatty acid intake</td>
<td>SFApct</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Monounsaturated fatty acid intake</td>
<td>MUFApct</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Polyunsaturated fatty acid intake</td>
<td>PUFApct</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Please refer Table 6 for details for description of covariates.
**Interaction Test (Model GE)**

A regression coefficient (β±SE) for the interaction term for SNP*(environmental modifier) (Model GE) on each phenotype will be calculated in each cohort, and reported values will be meta-analyzed.

- Phenotype Outcome = SNP + Dietary factor (continuous) + SNP*Dietary factor (continuous) + (Covariates) + (Additional Dietary Covariates)

In each cohort, each mapped genotyped or imputed SNP-environment interaction will be tested with each phenotype outcome assuming an additive genetic model.

All dietary factors will be modeled as a continuous variable in Model GE to evaluate gene-diet interactions (SNP*dietary factor).

**Model GE Covariates**

In addition to SNP and Model G covariates, environmental modifier and additional dietary covariate (Model E covariate for dietary factors) should be included in the model GE. Please refer Model G covariates (Table 3) and Model E covariate (Table 4) for details.

**Plasma lipid trait (LAa) = SNP + Dietary factor + SNP*Dietary factor + age + sex + field center + pedigree + population structure PCs + total energy**

**Table 5. Description of Covariates for Model GE** -- combination of Table 3 and Table 4

<table>
<thead>
<tr>
<th>Phenotype Outcome Categories</th>
<th>Covariates</th>
<th>Environmental Modifier Categories</th>
<th>Environmental Modifiers</th>
<th>Abbreviation for Model Name</th>
<th>Additional Dietary Covariate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Plasma lipid traits</td>
<td>LA:</td>
<td>D: Dietary factors</td>
<td>Carbohydrate intake</td>
<td>CHOpct</td>
<td>a. ENG</td>
</tr>
<tr>
<td></td>
<td>age, sex,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>field center,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>pedigree,</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>and population</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Saturated fatty acid intake</td>
<td>SFApct</td>
<td></td>
</tr>
</tbody>
</table>
|                             |             |                                   | Monounsaturated fatty acid intake | MUFApct |}

- Plasmolipid trait (LAa) = SNP + Dietary factor + SNP*Dietary factor + age + sex + field center + pedigree + population structure PCs + total energy.
<table>
<thead>
<tr>
<th>Covariates</th>
<th>Variable Type</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Continuous</td>
<td>Year</td>
</tr>
<tr>
<td>Sex (MALE)</td>
<td>Dichotomous</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>(1: male vs 0: female)</td>
<td></td>
</tr>
<tr>
<td>Field center</td>
<td>Cohort-specific</td>
<td>NA</td>
</tr>
<tr>
<td>Pedigree</td>
<td>Cohort-specific</td>
<td>NA</td>
</tr>
<tr>
<td>Population structure PCs</td>
<td>Cohort-specific</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>e.g. Using principle components analysis to correct for within-study population structure [35]</td>
<td></td>
</tr>
<tr>
<td>Total energy intake (ENG)</td>
<td>Continuous</td>
<td>kcal/d</td>
</tr>
</tbody>
</table>

Note: Please refer Table 6 for details for description of covariates.
**Meta-Analysis** -- FOR INFORMATIONAL PURPOSES. Analysis will be conducted by Yu-Chi Lee (GOLDN: Yu-Chi.Lee@tufts.edu) and another analyst (to be determined).

For meta-analysis, the analyses will be performed on summary statistics of main effects and interactions between SNPs and environmental factors as predictors for each trait. Each cohort will generate beta regression coefficients ($\beta$), standard errors (SE), and p values for (1) the main effects of a SNP and an environmental factor and (2) the interaction between variant and environment. The method of genomic control ($\lambda$GC) will be used to correct SE ($\text{GC corrected SE} = \text{SE}/\sqrt{\lambda_{\text{GC}}}$) for possible stratification for main-effects GWAS [36]. Results will be then summarized using meta-analytic methods. After gathering GWAS and GWIS results from each cohort, SNP QC will be performed (in the level of meta-analysis). SNP information in result files will be also examined carefully to avoid strand issue (e.g. for SNPs with alleles A/T or G/C and 40% < MAF < 50%) and ensure the quality of the results organization for meta-analysis.

We will first conduct main effects meta-analysis (Model G and E) and interaction meta-analysis (Model GE). METAL software (www.sph.umich.edu/csg/abecasis/metal) or GWAMA [37] will be used to perform meta-analysis of association studies using either weighted (signed) Z-score (combining p-values) or fixed and random effects inverse variance weighted approach (using regression coefficients) to determine an overall level of significance of the effect of SNP or environmental factor. R software using meta library and function metagen will be used for random effects inverse variance weighted meta-analysis. Both fixed and random effects models will be used and compared. Cochran’s Q test and $I^2$ index will be used to test heterogeneity [38]. If outcomes are non-homogeneous, weighted Z-score approach (fixed) will be chosen, and pooled effect estimates will not be generated. Bonferroni correction will be used for adjusting for multiple testing.

The joint meta-analysis method implemented in METAL will be also used [39]. Each cohort will generate beta regression coefficients ($\beta$), standard errors (SE), p values, and covariance estimate between the SNP $\beta$ and the environmental modifier $\beta$. The covariance estimate will be used to conduct 2 degree of freedom tests of the association of a SNP when taking interaction into account. Bonferroni correction will be used for adjusting for multiple testing.

Statistical significance threshold will be determined after SNP QC and before meta-analysis starts.
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
___ 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)?

This manuscript does not overlap any proposal other than Dr. North’s own proposals.

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 (AS #2006.03 & 2007.02)
__ 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: