1.a. Full Title: Gene-by-Environment Interaction for Type 2 Diabetes

b. Abbreviated Title (Length 26 characters): GxE for diabetes

2. Writing Group:
   Writing group members:

   Audrey Chu
   Linda Kao
   Josef Coresh
   Ingo Ruczinski
   Jim Pankow
   Eric Boerwinkle
   Keri North
   Others

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

First author: Audrey Chu
Address: 2024 E. Monument Street
         Suite 1-500K
         Baltimore, MD 21205

         Phone: 310-387-5558       Fax: 
         E-mail: achu@jhsph.edu

Corresponding/senior author (if different from first author correspondence will be sent to both the first author & the corresponding author):
Address: Linda Kao
         Department of Epidemiology
         Johns Hopkins School of Public Health
         615 N. Wolfe Street
         Rm. W6513
         Baltimore, MD 21205

         Phone: 410-614-0945       Fax: 410-955-0863
         E-mail: wkao@jhsph.edu
3. Timeline:

Will begin analysis when SNP genotyping is completed (summer 2007)

4. Rationale:

It is theorized that, in addition to the independent contribution of genes and environment, the interrelationship between genes, and genes and the environment are responsible for development of many complex diseases such as type 2 diabetes (T2D) (1). While many “environmental” risk factors for T2D, including obesity, diet, physical activity, fiber intake, alcohol consumption and smoking (2), have been established, few genetic risk factors have been identified until now.

Recent genome wide association (GWA) studies have identified genetic variants associated with type 2 diabetes (3-7). Together, with candidate genes/SNPs identified through previous candidate gene studies, there are 10 such genes that have been consistently replicated in primarily white populations and are listed Table 1 with the most significantly associated SNP and chromosome location. In addition, two additional genes have been associated with T2D in ARIC, CAPN10 (African Americans (8)) and NOS1AP (whites only, personal communication).

The “environmental” risk factors proposed for this analysis fall into two categories, 1) diet/behavior, and 2) antecedent/comorbid conditions. Dietary and behavioral factors proposed include: alcohol intake, coffee and sweetened beverage consumption, serum and dietary magnesium, fiber, as well as smoking and physical activity. The antecedent or comorbid condition is obesity.

These risk factors have either been analyzed and published for association with T2D in ARIC (which include alcohol consumption(9), coffee and sweetened beverage intake(10), serum magnesium(11), fiber intake(12) and obesity(13)) or have submitted ARIC manuscript proposals for analysis (including smoking #539B and physical activity #167).

The goal of the present proposal is to explore potential interactions between the candidate variants listed in Table 1 and some of the well established environmental risk factors for type 2 diabetes listed in Table 2 using logic regression, a regression method that can identify complex combinations of covariates which are associated with disease (14). Logic regression selects multiple genetic and environmental factors in sets and chooses best fitting models based on a scoring function (the model deviance in the case of binary outcomes). The interactions detected are more complicated than those from traditional interaction analysis and do not adhere to the traditional definition in which the presence of a covariate affects the effect of another covariate. The main advantage of using logic regression analysis to detect gene-environment interaction is that higher dimensional relationships between genes and environment can be identified since the interactions present in the data do not have to be pre-specified as in traditional regression analyses.

1 Non-genetic
The ARIC Study is a uniquely suited cohort for gene-environment interaction analysis with a large, well-phenotyped black and white study population and an extensive number of environmental and genetic variables measured before onset of diabetes.

### Table 1: Genetic Variants

<table>
<thead>
<tr>
<th>Gene</th>
<th>SNP</th>
<th>Chromosome</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDKN2A/2B</td>
<td>rs10811661</td>
<td>9p21</td>
<td>3’ intergenic</td>
</tr>
<tr>
<td>IGF2BP2</td>
<td>rs4402960</td>
<td>3q28</td>
<td>intron 2</td>
</tr>
<tr>
<td>CDKAL1</td>
<td>rs7754840</td>
<td>6</td>
<td>intron 6</td>
</tr>
<tr>
<td>HHEX</td>
<td>rs1111875</td>
<td>10q24</td>
<td>3’ intergenic</td>
</tr>
<tr>
<td>SLC30A8</td>
<td>rs13266634</td>
<td>8q24</td>
<td>non synonymous coding (R325W)</td>
</tr>
<tr>
<td>TCF7L2</td>
<td>rs7903146</td>
<td>10q25</td>
<td>intron 3</td>
</tr>
<tr>
<td>FTO</td>
<td>rs8050136*</td>
<td>16q12</td>
<td>intron 1</td>
</tr>
<tr>
<td>PPARG</td>
<td>rs1801282</td>
<td>3p25</td>
<td>non synonymous coding (P12A)</td>
</tr>
<tr>
<td>KCNJ11</td>
<td>rs5219</td>
<td>11p15</td>
<td>non synonymous coding (E23K)</td>
</tr>
<tr>
<td>Unknown</td>
<td>rs9300039</td>
<td>11p11</td>
<td>intergenic between NGL1 and AP15</td>
</tr>
<tr>
<td>CAPNJ0</td>
<td>rs3792267</td>
<td>2q37</td>
<td>intron 1</td>
</tr>
<tr>
<td>GCKR</td>
<td>rs780094</td>
<td>2p23</td>
<td>intron 16</td>
</tr>
<tr>
<td>NOS1AP</td>
<td>rs12026452</td>
<td>1q23</td>
<td>intron 2</td>
</tr>
<tr>
<td></td>
<td>rs4657139</td>
<td></td>
<td>5’ intergenic</td>
</tr>
</tbody>
</table>

5. **Main Hypothesis/Study Questions:**

This manuscript will address one main study question:

Do any of the currently known diabetes susceptibility genes interact with well established environmental risk factors for diabetes in both black and white ARIC participants?

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).**

All analyses will be stratified by race and will exclude subjects who are not white or black. Separate analyses will be performed for prevalent and incident cases, a pooled case-control analysis, matched on follow up time, will only be performed if the analyses of prevalent and incident cases are similar.

T2D will be defined by any of the following conditions: 1) fasting plasma glucose >126 mg/dL, 2) casual plasma glucose >200 mg/dL, 3) self-report of physician diagnosed diabetes, 4) self-report of diabetes medication. Prevalent T2D will include all participants with diabetes diagnosis up to visit 1; incident T2D will include all
participants who subsequently developed diabetes in visits 2-4, and will therefore exclude those with prevalent diabetes.

SNPs will be modeled by two variables: $X_{id}$ and $X_{ir}$. $X_{id}$ represents a dominant genetic trait at SNP locus $X_i$ and $X_{id}=1$ if an individual has at least one variant allele, $X_{id}=0$ if not. $X_{ir}$ corresponds to a recessive genetic locus $X_i$, where $X_{ir}=1$ if an individual has two variant alleles $X_{ir}=0$ if not. The additive model is represented as the difference between the dominant and recessive models (not exactly identical to an additive model in logistic regression, but is seen in the logic search space as the beta-coefficient when fitting the model: $\logit(p) = a + b*SNP[dom] + b*SNP[rec]$).

All environmental variables will be derived from Visit 1 data and categorized into groups identified by Table 2.

<table>
<thead>
<tr>
<th>Environment Variable</th>
<th>Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Alcohol (drinks/week)</td>
<td>0 1-21  &gt;21</td>
</tr>
<tr>
<td>Caffeinated Coffee (cups/day)</td>
<td>almost never 1-3 ≥4</td>
</tr>
<tr>
<td>Serum Magnesium (mmol/L)</td>
<td>0.25-0.70 0.71-0.89 0.90-1.30</td>
</tr>
<tr>
<td>Smoking</td>
<td>never former current</td>
</tr>
<tr>
<td>Obesity (b)</td>
<td>non-obese obese</td>
</tr>
<tr>
<td>Physical Activity Sport Leisure</td>
<td>tertiles (METs)</td>
</tr>
<tr>
<td>Dietary Magnesium Total Dietary Fiber Cereal Fiber</td>
<td>quintiles (mg/4.2kJ daily energy intake) quintiles (g/day) quintiles (g/day)</td>
</tr>
</tbody>
</table>

For logic regression analysis, variables with three categories will be indicated by the combination of two variables $X_{Ei1}$ and $X_{Ei2}$. Variables divided into quartiles/quintiles will be dichotomized into persons in the highest (or lowest) category and persons not belonging to that category. All analyses will be adjusted for age and sex.

To assess GxE interaction, we will be using logic regression analysis developed by Ruczinski et al. (14). Logic regression is a regression-based method which identifies Boolean combinations (AND-OR combinations) of binary variables which best predict disease status given a set of covariates. Variables available for analysis are selected and built into a logic trees which consist of combinations of variables in AND-OR combinations. The covariates which remain in the logic tree are determined by a series of
steps: 1) randomly adding, removing or changing covariates (leaves) and operators (AND <-> OR), 2) comparing the deviance of the model before and after the move, 3) accepting or rejecting the move according to the deviances of the different models. This process is repeated until the pre- versus post-move deviances no longer improve drastically; multiple trees may be constructed within this context. The best fitting model of a predetermined tree size is assessed for each set of logic models containing different numbers of leaves. The best fitting model overall is determined by the lowest deviance for all sizes of trees and containing different numbers of leaves.

Selection of best model size will be performed by permuting case-control status of models with different numbers of trees and leaves. In comparing histograms of deviance distribution of the models using the permuted data with the deviance of the best fit model using real ARIC data, the permutations for the best model size should be normally distributed over the deviance of the best fit model.

Additional interaction analyses will be performed with standard logistic regression and the results will be compared with the results obtained from logic regression. We will also compare results of adjusting for within-population stratification using readily available software (ie. EIGENSTRAT or other).

The major limitation of logic regression is that correlated variables will not usually be included in the same model, there is no formal verification that one variable should be selected over the other. However, a useful tool in the logic regression package to overcome this limitation is the availability of Monte Carlo logic regression which uses Markov chain Monte Carlo and logic regression to identify variables which repeatedly occur in the best fitting models for further investigation.

The major strength of interaction analyses using logic regression is that logic regression can identify complicated combinations of predictors which are associated with risk of disease, whereas traditional GxE interaction analyses would be limited by power. Also, multiple comparisons are minimized in logic regression due to permutation testing in the model selection step. In addition, the availability of Monte Carlo logic regression allows several sets of important predictor variables to be selected.

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.cscce.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)?

   Manuscript Proposal #1042 (Bray) is investigating gene-gene interaction in predicting obesity in ARIC, but we are examining gene-environment interaction in relation to type 2 diabetes

   Manuscript Proposal #796, (Brancati, et al) investigated resistin gene polymorphisms and association with insulin resistance and diabetes in ARIC and examined resistin-environmental variable interactions in relation to type 2 diabetes, but did not include variants in genes other than resistin.

   Manuscript Proposal #1198 (Monda, et al) investigated genes, environment, and their interactions to explore determinants of metabolic risk factors, but examined genes associated with cardiovascular disease and the metabolic syndrome.

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

   ___ 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.
References:

3. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661-78.