1.a. Full Title: Cumulative Association of Ten Genetic Variants with Retinopathy

b. Abbreviated Title (Length 26 characters):

2. Writing Group:
   Writing group members: Yu Yan, Ronald Klein, Barbara Klein, Gerardo Heiss, Eric Boerwinkle, James Pankow, Cynthia Girman, Ethan Lange, Suzanne West, Wei Sun, Anna Kottgen, Kari E. North

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

First author: Yu Yan
Address: Department of Epidemiology
School of Public Health
University of North Carolina at Chapel Hill

E-mail: yuyan@email.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 and Carolina Center for Genome Sciences
University of North Carolina Chapel Hill
Bank Of America Center
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 10 SNPs related to diabetes have been typed on the ARIC cohort. Approval of this manuscript proposal by the ARIC publications Committee will then enable work on this manuscript. Once started, this work will lead to a manuscript within one year.
4. Rationale:

Type 2 diabetes mellitus affects up to 150 million people worldwide\(^1\) and is one of the leading causes of morbidity and mortality in middle-aged persons\(^2,3\). Diabetic retinopathy, one of the common complications of type 2 diabetes, is a leading cause of blindness in people 20 to 74 years of age. Duration and severity of diabetes are major predictors for the development of diabetic retinopathy\(^4-6\). Retinopathy is found in people with prediabetes\(^6\) which suggests that microvascular disease may contribute to the development of type 2 diabetes\(^7,8\). Studies showed that microvascular abnormalities such as arteriolar narrowing and impaired microvascular blood flow in the skin and skeletal muscles have been noted in persons with type 2 diabetes and in persons at high risk of developing diabetes, such as those with prediabetes and first-degree relatives of persons with diabetes\(^9-12\). Previous ARIC studies showed that the retinal arteriolar narrowing is independently associated with risk of diabetes\(^12\) and that retinopathy predicts subsequent risk of clinical diabetes in individuals with a family history of diabetes\(^13\), supporting a microvascular role in the development of diabetes. Therefore, early identification of individuals with increased risk for retinopathy among diabetic and non-diabetic persons might be important for effective intervention.

Type 2 diabetes is a multifactorial disorder. Apart from conventional risk factors like obesity, smoking, and physical inactivity, genetics plays an important role in its complex etiology. There is also evidence that genetic factors may influence the calibers of retinal blood vessels\(^14\), although the exact genetic determinants are unknown. In the Beaver Dam Eye Study, the between siblings correlation for arteriolar and venular diameters were 0.23 (95% confidence interval (CI) 0.16-0.31) and 0.20 (95% CI 0.12-0.28), respectively, whereas the spousal correlations were 0.03 and 0.04, respectively (P>0.05 for spousal correlations). The ARIC Study confirmed that the T allele at the single nucleotide polymorphism (SNP) rs7903146 in the gene TCF7L2 (chromosome 10) confers risk for type 2 diabetes (Ms. #1141) and the association between SNPs in TCF7L2 and retinopathy in ARIC is under investigation (Ms. #1247). Besides TCF7L2 rs7903146, large-scale genome-wide association studies (GWAS) have also implicated the following SNPs in the etiology of type 2 diabetes\(^15-19\): CDKN2A/2B rs10811661 (chromosome 9), IGFBP2 rs4402960 (chromosome 3), CDKAL1 rs7754840 (chromosome 6), HHEX rs1111875 (chromosome 10), SLC30A8 rs13266634 (chromosome 8), FTO rs12255372 (chromosome 16), PPARG rs1801282 (chromosome 3), KCNJ11 rs5219 (chromosome 11), and rs9300039 in an intergenic region (chromosome 11). Although the mechanisms by which most of these genes/regions influence the susceptibility to type 2 diabetes are unclear, several mechanistic hypotheses have been proposed. First, the FTO gene is related to increased adiposity, which explains its association with diabetes\(^20\). The TCF7L2 gene is primarily associated with impaired insulin secretion\(^21-23\), whereas the PPARG gene is related to increased insulin resistance\(^24\). Other genes such as KCNJ11, SLC30A8, and CDKAL1 have been implicated in the regulation of beta cell function.

Each genetic variant confers only a moderate risk for type 2 diabetes (odds ratio ranging from 1.1-1.5 per copy of each risk allele), however, their combined effect may be
substantially greater. Scott et al. examined the combined risk of type 2 diabetes based on a logistic regression model containing the 10 SNPs listed above by ranking individuals according to their genetic risk in the Finland-United States Investigation of NIDDM Genetics (FUSION) Study that enrolled a large number of affected sib-pair families with type 2 diabetes to map and identify susceptibility genes for type 2 diabetes and for the intermediate quantitative traits associated with type 2 diabetes. They found a four times higher risk of type 2 diabetes in the highest predicted risk groups than in the lowest. However, the authors acknowledged that their predictions may be biased compared to predictions based on the general population. In addition, the FUSION Study did not account for gene-gene interactions and gene-environment interactions. Another analyses of ARIC data by Pankow (Ms. #1273) has begun to investigate whether the genetic risk score will provide similar levels of prediction of type 2 diabetes in population-based samples. However, with regard to the prediction of diabetic retinopathy, it is still unknown whether a combination of these 10 SNPs has a stronger association than any individual SNP. It is also unknown whether a combination of these 10 SNPs predicts risk beyond the risk predicted by traditional risk factors. Moreover, literature on diabetes-related gene–gene and gene-environment interaction assessment on the association of diabetes-related SNPs and retinopathy is very limited. Results of such analyses could have significant public health implications in patients at risk of retinopathy or diabetes, long before they develop frank diabetes or retinopathy in persons with diabetes. Given the current increasing incidence of obesity and diabetes, this could inform public health initiatives to encourage lifestyle changes in such patients at risk.

This proposed study will address these major gaps in the literature. The availability of data on SNPs in the TCF7L2 gene in ARIC, a large, bi-racial cohort, will enable analyses focused on prediabetes and retinal microvascular diseases. The ARIC Study photographed 2000 selected participants again based on the distribution of carotid IMT 10 years after initial retinal photography, which provides a good opportunity to prospectively study the association between a combination of these 10 SNPs and incidence, progression and regression of retinal vascular characteristics and abnormalities besides the cross-sectional associational analyses at visit 3. The detailed phenotypic characterization available on cohort members will permit adjustment for a range of potential confounders and evaluation on gene-environment interactions. The use of a powerful statistical approach, based on two different models (parametric approach and Classification And Regression Trees) to test for complex gene-gene and gene-environment interaction effects, will be another important strength of our study. The proposed study will potentially contribute important knowledge about the etiology of type 2 diabetes and retinopathy, and may aid in the development of screening strategies and treatment regimes utilizing genetic information.

5. Main Hypothesis/Study Questions:

1. Are individual SNPs associated with prevalent retinopathy, and with incidence, progression and regression of retinal vascular abnormalities and characteristics?
2. Will a combination of the 10 SNPs described above have a stronger association with retinal vascular abnormalities and characteristics than each SNP?

3. Are there interactions between diabetes-related SNPs in their association with the retinal vascular abnormalities and characteristics?

4. If associations are present,
   a. Do the associations vary among diabetic participants, non-diabetics (without IFG) and participants with impaired fasting glucose (IFG)?
   b. Do the associations vary among African Americans and Caucasians?
   c. To which extent are associations influenced by adjustment for or stratification on gender, center, family history of diabetes, physical activity, body mass index, smoking, hypertension, hyperglycemia, high LDL cholesterol and low HDL cholesterol?

5. If associations are present, will a combination of the 10 SNPs plus another two SNPs typed in ARIC (rs780094 in GCKR and rs3792267 in CAPN10) improve the prediction of retinal vascular abnormalities and characteristics compared to traditional risk factors?

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 methodological limitations or challenges if present).

I. STUDY DESIGN, INCLUSION/EXCLUSION, AND OUTCOMES

Study Design: A cross-sectional analysis will be conducted among all ARIC participants at the third examination (1993-1995); a prospective analysis will be conducted among ARIC participants who attended the third (1993-1995) and fifth examinations (2004-2005) (Table 1). Considering the power in the prospective analyses may be limited given the sample size of 2000 participants, especially when we evaluate the gene-gene and gene-environment interactions, the prospective analyses will be exploratory.

Inclusion/Exclusion Criteria: Participants who did not attend the third examination will be excluded from analysis. Of the 12,642 participants examined at the third examination, individuals reporting race other than African American or Caucasian and African American residents living in Minneapolis and Maryland will be excluded as well. Participants who did not consent to genotyping will be excluded (use of DNA data distributed by the Coordinating Center with confirmation by using the variable RES_DNA = “No use/storage DNA” in the file ICTDER02). Retinal photographs will be used to define retinal phenotypes, so participants who did not undergo the retinal examination or had ungradable fundus photographs at visit 3 will be excluded from the analyses. Participants with retinal microvascular diseases at visit 3 will be excluded from the incidence analysis. Individuals lacking information on diabetic status will be excluded from the stratified analysis by diabetic status. The missing number of individuals excluded due to missing genotype information will depend on the each SNP.
Table 1. Framework for the cumulative association of diabetes-related genetic variants with retinal abnormalities and characteristics

<table>
<thead>
<tr>
<th>Outcomes</th>
<th>Binary Outcomes (Yes/No)</th>
<th>Continuous Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Diabetic retinopathy</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nondiabetic retinopathy</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Focal arteriolar narrowing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A/V nicking</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CRAE</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CRVE</td>
<td></td>
</tr>
</tbody>
</table>

**Exposure**
- 12 diabetes-related SNPs*

**Potential effect modifiers**
- Hypertension, race, body mass index/obesity, gender, physical activity, low HDL

**Statistical Analyses**

**-- Model construction**
- We will construct 2 nested models following published literature:
  1. Adjusted Model: age, gender, and ARIC field center.
  2. Multivariate-adjusted model: age, gender, ARIC field center, diabetes, fasting glucose, systolic blood pressure, current smoking (yes/no), body mass index, total serum cholesterol, total serum triglycerides, and hypertension.

**-- Regression model selection**
- (1) Cross-sectional approach (Visit 3)
  - Prevalence – logistic regression
  - Prospective approach *(exploratory analyses†)* (Visits 3 & 5)
    - Incidence ‡ – proportional hazard regression
    - Progression ‡ – GEE regression
    - Regression ‡ – GEE regression
- (1) Cross-sectional approach (Visit 3)
  - Analysis of variance (ANOVA); GLM
  - Prospective approach *(exploratory analyses†)* (Visits 3 & 5)
  - GEE regression

Abbreviations: A/V, arterio-venous; CRAE, central retinal artery equivalent; CRVE, central retinal venular equivalent; GEE, Generalized Estimating Equation; GLM, Generalized Linear Model.

*Association between each individual SNP and retinal abnormalities and characteristics will be evaluated first before assessing the combinational effects of 12 SNPs.
†Due to the power limitation, the prospective approach will be exploratory analyses.
‡The incidence of retinal abnormalities and characteristics is defined as the appearance of retinal signs at visit 5 in persons without these signs at visit 3; progression as an increase in severity of signs (e.g. from mild to severe AV nicking); and regression as disappearance of these signs between visit 3 and 5.
Outcomes: The primary outcomes are retinal abnormalities or characteristics which include five markers (visit 3,5): retinopathy, focal retinal arteriolar narrowing, arteriovenous (A/V) nicking, retinal arteriolar caliber (CRAE) and venular caliber (CRVE).

1) Retinopathy: will be defined if any characteristic lesion as defined by the Early Treatment Diabetic Retinopathy Study severity scale was present in the absence of specific ocular diseases e.g., retinal vein occlusion, exudative age-related macular degeneration: microaneurysms, retinal hemorrhages (blot or flame shaped), soft exudates (cotton-wool spots), macular edema, hard exudates, intraretinal microvascular abnormalities. A retinopathy severity score will be assigned based on the presence of various lesions: level 10, none; level 20, minimal nonproliferative retinopathy (microaneurysms only or blot hemorrhages only); level 35, early nonproliferative retinopathy (microaneurysms and one or more of the following: venous loops, soft exudate or hard exudate, and questionable intraretinal microvascular abnormalities or venous beading); levels 43 to 47, moderate to severe nonproliferative retinopathy (microaneurysms and one or more of the following: intraretinal microvascular abnormalities, venous beading, hemorrhages, and microaneurysms exceeding those in standard photograph); level 60+, proliferative retinopathy. Because the pathogenesis of diabetic retinopathy is different from nondiabetic retinopathy, we will perform subgroup analyses in persons with and without diabetes.

2) Focal arteriolar narrowing and A/V nicking: were assessed in each of four quadrants of the eye. Focal narrowing will be considered definite if an arteriole estimated to be 50μm in diameter or greater had a constricted area of two thirds or less the width of the proximal and distal vessel segments. The maximum grade from the four quadrants defines two endpoints: focal narrowing on the disc or within 0.5 disc diameter of its margin and outside the disc. Both will be categorized as definite versus questionable or absent. A/V nicking will also be categorized as definite versus questionable or absent.

3) CRAE and CRVE: To detect and quantify generalized retinal arteriolar narrowing and venular widening, the diameters of all arterioles and venules coursing through a specified area surrounding the optic disc were measured with an image processor and summarized as the central retinal artery equivalent (CRAE) and the central retinal venular equivalent (CRVE), representing average calibers of retinal arterioles and venules, respectively.

For the cross-sectional approach, the primary outcomes will be defined as the presence of retinal abnormalities or characteristics; for the prospective approach, the primary outcomes will be defined as the incidence, progression and regression of retinal vascular abnormalities over 10 years. The incidence of retinal microvascular diseases will be defined as the presence of retinal signs at visit 5 in persons without these retinal signs at visit 3, progression will be defined as an increase in severity of retinal signs, and regression as disappearance of these retinal signs between visit 3 and 5.

II. OTHER VARIABLES OF INTEREST
(1) Ten diabetes-related SNPs (TCF7L2 rs7903146, CDKN2A/2B rs10811661, IGFBP2 rs4402960, CDKAL1 rs7754840, HHEX rs1111875, SLC30A8 rs13266634, FTO rs12255372, PPARG rs1801282, KCNJ11 rs5219, and rs9300039 in an intergenic region) plus two SNPs typed in the full cohort (rs780094 in GCKR and rs3792267 in CAPN10).

(2) Cardiovascular risk factors (visit 1, 2, 3, 4, 5): hypertension, systolic blood pressure, diastolic blood pressure, anti-hypertension medication use, diabetes, glucose, post-load insulin and glucose (visit 4), diabetic medication use, duration of diabetes, family history of diabetes, cigarette smoking status (ever/never, current/past, pack-years), serum HDL cholesterol, LDL cholesterol, triglycerides, total cholesterol, body mass index, waist to hip ratio, physical activity.

(3) Other covariates: age, gender, race, center, alcohol consumption, HbA1c (visit 2).

(4) Variables to variability (visit 3): the repeat measurements of retinal vascular variables such as arteriolar and venular diameters, CRAE, and CRVE from the Individual Variability Study (n = 206), and the Grader Variability Study (n = 495).

III. DATA ANALYSES

(1) Data Quality Analyses:

Tests of Hardy-Weinberg equilibrium will be performed for each SNP in the cohort. Significant deviations from Hardy-Weinberg will be assessed using a chi-square test, by comparing the observed distribution of genotypes to the Hardy-Weinberg 'expected' distribution, with degrees of freedom equal to the number of alleles (n) – 1.

The frequency distribution of all the variables in this proposed study will be examined in exploratory plots to examine their general characteristics. The two-sample t test and the $X^2$ test will be employed to detect differences in continuous and categorical characteristics between 'cases' and 'controls' at visit 3 (for retinal abnormalities). Additionally, allele frequencies will be calculated and a two-sample test for binomial proportions will be used to assess differences in allele frequencies between 'cases' and 'controls'.

(2) Association Analyses:

All association analyses will be first examined within each ethnic (African American or Caucasian) group. For each SNP, a general model (no mode of inheritance assumption) using a 2-degree-of-freedom F-test will be utilized to analyze the gene-disease association in the cross sectional analysis. If this result is found to be statistically significant, or if there is an a priori hypothesis based on the literature, further testing of the SNP effects assuming genetic modes of inheritance will then be performed. For dominant or recessive genetic transmission models, a single indicator variable (eg. taking the value 1 if an individual has genotype X0 or 00 and 0 otherwise) will be used. A variable taking on the values -1 for genotype XX, 0 for genotype X0, and 1 for genotype 00 will be used to test for additive genetic effects. Adjustment will be made for multiple testing using the false discovery rate\textsuperscript{28}. 
We will examine the cumulative effects of the 10 diabetes-related SNPs on the outcomes (retinal abnormalities and characteristics) by counting the number of diabetes-risk genotypes (on the basis of the best-fitting genetic model from single-SNP analysis) or the number of at-risk alleles in each participant. The odds ratio/hazard ratio for the outcomes carrying any combination of at-risk genotypes or alleles will be estimated by comparing them with participants carrying none of the diabetes-related SNPs with the use of logistic regression/Cox Proportional Hazards modeling. If data is sparse in the upper range (e.g. ten genotypes), we will cap the number of genotypes and alleles (e.g. >=5 genotypes).

Population attributable fraction (PAF) will be estimated for each SNP with the use of the following equation in a logistic regression:

$$PAF\% = 100 \times pd \times (odds ratio - 1)/odds ratio$$

where pd is the proportion of cases exposed to the at-risk genotype(s)\(^{29}\).

The joint PAF will be calculated based on the basis of the individual PAF of each associated SNP, with the use of the following equation:

$$Joint\ PAF\% = 1 - \prod_{i=1}^{10} (1 - PAF_i)$$

Following published work in ARIC\(^{27}\) we will construct 2 nested models to account for confounders: (1) the adjusted model: age, gender, and ARIC field center, and (2) the multivariate-adjusted model: age, gender, ARIC field center, diabetes, fasting glucose, systolic blood pressure, current smoking (yes/no), body mass index, total serum cholesterol, total serum triglycerides, and hypertension.

(i) Cross-sectional Approach (Visit 3) (Table 1)

For the association of diabetes-related SNPs with prevalent diabetic retinopathy, nondiabetic retinopathy, focal retinal arteriolar narrowing, and A/V nicking, prevalence odds ratios will be calculated from logistic regression models. Multivariable relationships between characteristics and severity of retinopathy will be examined with logistic regression in which severity of retinopathy will be categorized into four groups: level 10, level 20, level 35, and level 43 or worse. Analysis of variance (ANOVA) will be used to compare the mean CRAE and CRVE for diabetes-related SNPs; generalized linear models will be used to evaluate the association of diabetes-related SNPs with continuous CRAE and CRVE. Models will be constructed to account for potential confounders.

(ii) Prospective Approach (exploratory analyses; Visits 3 and 5) (Table 1)

Due to the nonrandom nature of the sample that was photographed at ARIC visit 5 and selected to obtain 60% of baselines participants with high carotid IMT (>85 percentile) and 40% of remaining participants with low carotid IMT(<85 percentile), we will perform analyses stratified by carotid IMT category, and perform weighted analysis, adjusting for sampling fractions. For the association of diabetes-related SNPs with the incidence of retinal vascular abnormalities, hazard ratios will be calculated from proportional hazards regression models, whereas for progression and regression of retinal vascular abnormalities odds ratios will be obtained from generalized estimating equation (GEE) regression models. The entry-age-adjusted age-scale model will be used to model the time-to-event\(^{30}\). Plots of the log (-log) survival curves and the Cox test will be utilized
to assess violations of proportional hazard assumptions. GEE regression models will be used to assess longitudinal changes in retinal vessel calibers (Visit 5 vs. Visit 3).

(3) Interaction Assessment: Three different interaction assessments will be involved: gene-gene interactions only, gene-environment interactions only, gene-gene and gene-environment interactions. For each interaction assessment, we will apply and compare two statistical methods, a parametric approach and Classification And Regression Trees (CART).

Parametric Approach
Logistic regression modeling and proportional hazard modeling will be used for cross-sectional and time-to-event analyses, respectively. For example, the logistic regression model is a parametric approach that relates one or more independent variables, $X_i$ (for example, SNPs and their interactions), to a dependent or outcome variable, $Y$ (for example, retinopathy). The logistic regression model is expressed as:

$$
\log\left(\frac{P}{1-P}\right) = \beta_0 + \sum_{i}^{p} \beta_i x_i + \sum_{i}^{p-1} \sum_{j>i}^{p} \beta_{ij} x_i x_j
$$

where $p$ is the number of independent variables, $\beta_i$ for the SNP main effects ($x_i$) and $\beta_{ij}$ for the SNP-SNP interaction ($x_i x_j$). Only two-way interactions will be considered, and adjustment will be made for multiple testing using the false discovery rate. A likelihood ratio test statistic for the $\chi^2$ values will be employed to test the significance of interaction effects. We will use a stepwise regression procedure based on forward selection to select the most significant gene-gene and gene-environment interactions. When testing multiple interactions, forward stepwise selection procedures are considered to be more efficient compared to backward stepwise selection. Odds ratios (ORs) and hazard ratios (HRs) (and 95% CIs) of outcome status on the indicator variables of predictor subcategories will be obtained by logistic regression and proportional hazard models, respectively. For example, in the $2^2*2$ table of SNP1*SNP2*SNP3, we would consider 3 ($=2^2-1$) indicator variables of the possible combinations, with SNP1 homozygous wild genotype, SNP2 homozygous wild genotype and SNP3 homozygous wild genotype as the reference category. These analyses will be performed using PROC LOGISTIC for logistic regression or PROC TPHREG for proportional hazard regression of SAS (SAS institute, Cary, NC).

Classification and Regression Trees (CART)
CART, a non-parametric approach, uses recursive partitioning and asymmetric stratification to develop tree-based models using binary predictors. CART constructs a tree by using splitting rules to stratify data into risk subgroups on which to discriminate the outcome of interest. The strength of CART compared to a traditional parametric technique, such as logistic regression, is its ability to detect high-level interactions among the risk factors. It forms subgroups, or terminal nodes, that are homogeneous with respect to the outcome of interest. Splits between terminal nodes are determined by iteratively searching for the optimal split which minimizes misclassification within and among the predictors. To set up these models, the Gini Index will be used as a splitting criterion assuming a minimum terminal node size of 10. A 10-fold cross-validation will then be applied to prune the tree to its final size. Following construction of the optimal
tree using CART, we will construct indicator variables (1, yes and 0, no) for terminal nodes, with the last terminal node as the reference category. The ORs and HRs (and 95% CIs) for the final tree will be computed by logistic regression and proportional hazard regression models, respectively. Sensitivity and specificity will be obtained by comparing the predicted outcome and the observed outcome. These analyses will be performed using the CART program.

(4) Risk Prediction: Only incident retinal abnormalities will be considered for the risk prediction analyses. A genetic risk score for each individual will be calculated by summing up the genotypic values of each SNP. Unless prior analyses suggest strong evidence of gene-gene interaction, we will assign 1 for the risk homozygote, 0 for the heterozygote and -1 for referent homozygote in the Cox Proportional Hazards analyses. The area under the curve for a model containing a genetic risk score and traditional risk factors will be compared to that for a model containing traditional risk factors alone. Based on the diabetes risk functions in the ARIC cohort, we will consider the following traditional risk factors: age, parental history of diabetes, fasting glucose, systolic blood pressure, waist circumference, HDL cholesterol, and triglycerides. Since this risk function is developed for diabetes, other potential predictors (e.g. gender, BMI, physical activity, smoking, alcohol consumption, fasting insulin) will be considered as well.

IV. METHODOLOGIC LIMITATIONS OR CHALLENGES

The major methodological limitation for this proposed study is the power in the prospective analyses. Given the sample size of 2000 participants who were photographed 10 years after the initial photography, the power in the prospective analyses may be limited, especially when we evaluate the multiple gene-gene and gene-environment interactions. Thus the prospective analyses in this proposed study will be an exploratory analysis.

An important methodological challenge in genetic epidemiologic studies is how to capture and analyze the complex gene-gene and gene-environment interactions when the number of genes is large. Two-way interactions alone may not be sufficient to identify all important gene-gene and gene-environment interactions. Logistic regression or proportional hazard regression models can detect only low-order interactions as the model complexity increases with the order of interactions. A fully saturated model with numerous terms may tend to unstable, and sparse data and multicolinearity may generate biased estimates. Moreover, large sample theory underlying the test statistic may not hold when data is sparse. CART does not require or assume any specific parametric form for the relation between independent and dependent variables. CART will examine every possible binary split on every possible variable and build a binary hierarchical tree which will point to interactions. Therefore, CART offers the advantages in that it might uncover gene-gene and gene-environment interactions that are missed by logistic regression or proportional hazard regression models. CART can also deal with sparse and high-dimension data and can account for non-linear gene-gene interactions. CART does have limitations - tree model may not fit well if outcomes are rare; sensitive to strongly correlated variables used in prediction - may result in instability. As such, CART is very
useful in analyses of gene-gene and gene-environment interactions of complex diseases, however, the use of CART models is quite limited in current genetic epidemiology studies possibly due to unfamiliarity. With more and more new genes discovered by genome-wide associations studies, CART seems to be particularly useful to handle complex gene-gene and gene-environment interactions, and our study will turn out to be a good example in that sense.

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.csec.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)?

Ms. #1141 titled "Transcription factor 7-like 2 (TCF7L2) gene and type 2 diabetes" (first author: Yu Yan; corresponding/senior author: Gerardo Heiss). This ms. did not evaluate combinations of diabetes-related genes.

Ms. #1247 titled “Transcription factor 7-like 2 (TCF7L2) gene and retinopathy” (first author: Yu Yan; corresponding/senior author: Kari E. North). This ms. will not evaluate combinations of diabetes-related genes.

Ms. #1273 titled “Genetic risk score for type 2 diabetes” (first author: Jim Pankow). This ms. will not evaluate the combined effect of SNPs on retinopathy.
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 (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:


15. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447(7145):661-78.


