ARIC Manuscript Proposal # 1470

PC Reviewed: 02/10/08	**Status: A**	**Priority: 2**
SC Reviewed: _____	**Status: _____**	**Priority: ____**

1.a. Full Title: Genome-wide association study of prevalent type 2 diabetes in the Atherosclerosis Risk in Communities (ARIC) Study

b. Abbreviated Title (Length 26 characters): GWAS of Prevalent Diabetes

2. Writing Group:

Writing group members: Man Li, Anna Kottgen, David Couper, Linda Kao, James S. Pankow, Suzette J. Bielinski, Eric Boerwinkle. Other interested investigators from ARIC are welcome. We have proposed that the DIAGRAM consortium nominate three authors in recognition of providing replication data (see section 6 below), but details are still under negotiation.

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

First author: Man Li
Address: 615 N. Wolfe St, Room 6507B
Baltimore, MD21218

Phone: (410) 502 3593 Fax: (410) 955 0863
E-mail: mali@jhsph.edu

ARIC author to be contacted if there are questions about the manuscript and the first author does not respond or cannot be located (this must be an ARIC investigator).
Name: **James S Pankow**
Address: Division of Epidemiology and Community Health
1300 South Second Street, Suite 300
Minneapolis, MN 55454

Phone: (612) 624 2883 Fax: (612) 624 0315
E-mail: pankow@umn.edu

3. Timeline: Journal submission February-March 2009

4. Rationale:

Type 2 diabetes mellitus (T2DM) is a major health burden with over 20 million Americans estimated to be living with the disease [1]. T2DM is a heterogeneous disease with a complex etiology including a strong genetic component. Genes that contribute to genetic susceptibility to T2DM function in numerous biochemical pathways. Genome wide association studies in large populations have resulted in the identification of numerous T2DM susceptibility genes [2]. To date, TCF7L2 has the strongest association with T2DM and is consistently replicated across populations [3-9]. We propose to conduct a genome-wide association study to identify variants associated with T2DM within this well characterized cohort.

5. Main Hypothesis/Study Questions:

We propose to study the association of ~ 2.5 Million genotyped and imputed SNPs from the Affy 6.0 array in ARIC Study participants and prevalent type 2 diabetes at the ARIC baseline exam.

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).

Study design and inclusion/exclusion: subjects and sample size

The following exclusion criteria will be applied; 1) subjects who did not consent to genetic research, 2) subjects who self-reported race other than "white", and 3) subjects missing measurements needed to assess diabetes at baseline.

The final baseline sample will therefore consist of successfully genotyped individuals among 7116 white ARIC participants.

Publication Strategy

Preliminary analyses indicate that there is one novel region that reaches genome-wide significance for prevalent T2DM in ARIC. We have a tentative agreement with the Diabetes Genetics Replication And Meta-analysis (DIAGRAM) consortium [2] to provide replication results for this region. DIAGRAM has assembled GWAS results on over 10,000 T2DM cases and 10,000 controls of European ancestry. In return, ARIC will exchange in silico replication results for 25 novel regions/SNPs that have been identified in DIAGRAM consortium's latest meta-analysis. Once the exchange with DIAGRAM is made, a separate ARIC manuscript proposal will be submitted to document ARIC's participation as a replication cohort in the DIAGRAM paper that is currently in development. Finally, ARIC has agreed to contribute GWAS results for prevalent T2DM if the DIAGRAM consortium decides to update its meta-analysis at some future date.

Exposure Measurements and Definitions
This manuscript proposal is concentrated on the analyses of the SNP data at ~2.5 Million genotyped and imputed SNPs.

Quality Control of Genotyping Data
Called genotype data using the Birdseed algorithm will be transferred from the Broad Institute where genotyping is conducted to the ARIC coordinating center. Quality control analyses of the ARIC genotype data will be conducted centrally by Dr. Boerwinkle's group at UT Houston, and will occur prior to the distribution of the genotype data to the ARIC study sites. Quality control procedures will include the analyses of blind duplicates and evaluation of missing data (both at the individual and SNP level) and allele/genotype distribution. Exclusions will be data-driven and include excessive missing data (likely >20%), lack of chromosomal coordinates, monomorphic SNPs, severe deviations from Hardy-Weinberg equilibrium (likely $p<10^{-6}$), excessive autosomal heterozygosity, and genetic outliers.

Outcome Measurements and Definitions
Prevalent diabetes are defined as any participant who self reported a physician diagnosis of diabetes or high blood sugar, used diabetic medication, had a fasting glucose ≥ 126 mg/dL, or a non-fasting glucose of ≥ 200 mg/dL at the baseline exam. To increase specificity, non-cases are defined subjects with fasting glucose < 110 mg/dL. Glucose was measured using a hexokinase/glucose-6-phosphate dehydrogenase procedure. Height was measured while participants were standing without shoes, heels together against a vertical mounted ruler. A Detector Platform Balance was used to measure weight. BMI was calculated as weight (kg)/height^2 (m^2). Hip and waist circumferences were measured at the maximal protrusion of the hips and at the level of the umbilicus with the participant standing erect. Medical and personal histories were ascertained via interview.

Statistical Analysis
Genotypes will be modeled using an additive genetic model in order to reduce the number of tests, because this model has been shown to fit reasonably well in prior GWAS analyses of common variants. Covariates will include age, study center, gender, and measures of adiposity and physical activity. The outcomes will be evaluated using logistic regression with adjustment for the covariates. All analyses will be conducted using the software PLINK[10] and probABEL[11]. Moreover, analyses will be conducted on combining the new findings with previous T2DM GWAS findings, such as SNP* SNP effect. Further secondary analyses will likely be necessary depending on the nature of the findings.

Population stratification can cause spurious association in the analysis of GWAS data.[12] We are currently evaluating whether we will use the IBS clustering feature provided as part of the software plink, or whether we will use principal components analyses as implemented in the software Eigenstrat.[13] An estimate of ancestry can be obtained from either method, which can then be adjusted for as a covariate in the association analyses if we found that estimates of ancestry are associated with both the outcome and the SNP of

interest. We will also estimate the genomic control parameter to obtain an overall estimate of potential stratification.

The issue of multiple testing, which is a threat to inferences from genome-wide association studies, will be addressed by replicating association findings by DIAGRAM. We will further evaluate results corrected for multiple testing applying the Bonferroni correction as well as the FDR method.[14]

Finally, we will use bioinformatics tools to obtain information about informative SNPs, both with respect to linkage disequilibrium as well as to function. SNPs will be evaluated for their location, SNP type, across-species conservation, and association with gene expression. For nonsynonymous coding SNPs, we will evaluate the predicted consequence of the amino acid change. Collaborations with additional scientists to functionally study promising SNPs may be initiated as scientifically justified and agreed upon within the working group.

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 ICTDER03 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 ICTDER03 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 ICTDER03 must be used to exclude those with value RES_DNA = "No use/storage DNA"?
__X__ Yes ____ No

8.c. If yes, is the author aware that the participants with RES_DNA = 'not for profit' restriction must be excluded if the data are used by a for profit group?
__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)?

Genome-wide Association Study of Diabetes-Related Quantitative Traits in ARIC Whites
First Author: Jim Pankow (ms#1409)

Genetic variants are associated cross-sectionally and longitudinally with multiple measures of fasting glucose: the ARIC Study; Lead: Laura Rasmussen-Torvik (ms #1468)

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

__X_ 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)* _____ _____ _____)

2006.03 (Stampede and Geneva genotype funding in Caucasians) ;

*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

1. Association, A.H. 2008. Heart Disease and Stroke Statistics - 2008 Update. American Heart Association. 2008
2. Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, de Bakker PI, Abecasis GR, Almgren P, Andersen G, Ardlie K, Bostrom KB, Bergman RN, Bonnycastle LL, Borch-Johnsen K, Burtt NP, Chen H, Chines PS, Daly MJ, Deodhar P, Ding CJ, Doney AS, Duren WL, Elliott KS, Erdos MR, Frayling TM, Freathy RM, Gianniny L, Grallert H, Grarup N, Groves CJ, Guiducci C, Hansen T, Herder C, Hitman GA, Hughes TE, Isomaa B, Jackson AU, Jorgensen T, Kong A, Kubalanza K, Kuruvilla FG, Kuusisto J, Langenberg C, Lango H, Lauritzen T, Li Y, Lindgren CM, Lyssenko V, Marvelle AF, Meisinger C, Midthjell K, Mohlke KL, Morken MA, Morris AD, Narisu N, Nilsson P, Owen KR, Palmer CN, Payne F, Perry JR, Pettersen E, Platou C, Prokopenko I, Qi L, Qin L, Rayner NW, Rees M, Roix JJ, Sandbaek A, Shields B, Sjogren M, Steinthorsdottir V, Stringham HM, Swift AJ, Thorleifsson G, Thorsteinsdottir U, Timpson NJ, Tuomi T, Tuomilehto J, Walker M, Watanabe RM, Weedon MN, Willer CJ, Illig T, Hveem K, Hu FB, Laakso M, Stefansson K, Pedersen O, Wareham NJ, Barroso I, Hattersley AT, Collins FS, Groop L, McCarthy MI, Boehnke M, Altshuler D. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet*. 2008;40:638-45.
3. Grant SF, Thorleifsson G, Reynisdottir I, Benediktsson R, Manolescu A, Sainz J, Helgason A, Stefansson H, Emilsson V, Helgadottir A, Styrkarsdottir U, Magnusson KP, Walters GB, Palsdottir E, Jonsdottir T, Gudmundsdottir T, Gylfason A, Saemundsdottir J, Wilensky RL, Reilly MP, Rader DJ, Bagger Y, Christiansen C, Gudnason V, Sigurdsson G, Thorsteinsdottir U, Gulcher

JR, Kong A, Stefansson K. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat Genet*. 2006;38:320-3.

4. Cauchi S, El Achhab Y, Choquet H, Dina C, Krempler F, Weitgasser R, Nejjari C, Patsch W, Chikri M, Meyre D, Froguel P. TCF7L2 is reproducibly associated with type 2 diabetes in various ethnic groups: a global meta-analysis. *J Mol Med*. 2007.
5. Zhang C, Qi L, Hunter DJ, Meigs JB, Manson JE, van Dam RM, Hu FB. Variant of transcription factor 7-like 2 (TCF7L2) gene and the risk of type 2 diabetes in large cohorts of U.S. women and men. *Diabetes*. 2006;55:2645-8.
6. Scott LJ, Bonnycastle LL, Willer CJ, Sprau AG, Jackson AU, Narisu N, Duren WL, Chines PS, Stringham HM, Erdos MR, Valle TT, Tuomilehto J, Bergman RN, Mohlke KL, Collins FS, Boehnke M. Association of transcription factor 7-like 2 (TCF7L2) variants with type 2 diabetes in a Finnish sample. *Diabetes*. 2006;55:2649-53.
7. Melzer D, Murray A, Hurst AJ, Weedon MN, Bandinelli S, Corsi AM, Ferrucci L, Paolisso G, Guralnik JM, Frayling TM. Effects of the diabetes linked TCF7L2 polymorphism in a representative older population. *BMC Med*. 2006;4:34.
8. Humphries SE, Gable D, Cooper JA, Ireland H, Stephens JW, Hurel SJ, Li KW, Palmen J, Miller MA, Cappuccio FP, Elkeles R, Godsland I, Miller GJ, Talmud PJ. Common variants in the TCF7L2 gene and predisposition to type 2 diabetes in UK European Whites, Indian Asians and Afro-Caribbean men and women. *J Mol Med*. 2006;84:1-10.
9. Groves CJ, Zeggini E, Minton J, Frayling TM, Weedon MN, Rayner NW, Hitman GA, Walker M, Wiltshire S, Hattersley AT, McCarthy MI. Association analysis of 6,736 U.K. subjects provides replication and confirms TCF7L2 as a type 2 diabetes susceptibility gene with a substantial effect on individual risk. *Diabetes*. 2006;55:2640-4.
10. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC. PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics*. 2007;81(3):559-75.
11. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. GenABEL: an R library for genome-wide association analysis. *Bioinformatics*. 2007;23(10):1294-6.
12. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007; 447: 661-678.
13. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.* 2006; 38: 904-909.
14. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate - A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B. 1995; 57(1): 289-300.