ARIC Manuscript Proposal #2144

PC Reviewed: 5/14/13	**Status: A**	**Priority: 2**
SC Reviewed: _____	**Status: _____**	**Priority: ____**

1a. Full Title: Correlations among metabolic traits and inflammation biomarkers raise expectations of genetic pleiotropic discoveries

b. Abbreviated Title: Correlations among metabolic syndrome traits may inform pleiotropic discoveries

2. Writing Group: ARIC - Kari North, Jim Pankow, Joseph Coresh, Weihong Tang, Linda Kao, Eric Boerwinkle, Alana Morrison, Other ARIC Investigators welcome
Outside: Aldi T. Kraja, Josée Dupuis, Daniel Chasman, Alexander P. Reiner, Lisa R. Yanek, Tuomas O. Kilpeläinen, Jennifer A. Smith, Abbas Dehghan, Andrew D. Johnson, Mary F. Feitosa, Fasil Tekola-Ayele, Audrey Y. Chu, Zari Dastani, Andrew Morris, Sarah A. Pendergrass, Yan V. Sun, Marylyn Ritchie, Ahmad Vaez, Honghuang Lin, Symen Ligthart, Letizia Marullo, Rebecca Rohde, Yaming Shao, Jerome I. Rotter, Sharon L.R. Kardia, Ruth J.F. Loos, Martin G. Larson, Yi-Hsiang Hsu, Michael A. Province, Russell Tracy, Benjamin F. Voight, Dhananjay Vaidya, Behrooz Z. Alizadeh, Emelia Benjamin, Christopher O'Donnell, Inga Prokopenko, Ingrid B. Borecki, James B. Meigs
I, the first author, confirm that all the coauthors have given their approval for this manuscript proposal. KEN

First 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

Corresponding/senior author (if different from first author correspondence will be sent to both the first author & the corresponding author):

3. Timeline:

Individual cohort statistical analyses: Now
Manuscript preparation: Now
Manuscript submission: Summer 2013

4. Rationale:

Metabolic syndrome (MetS), defined as comorbidities of central obesity, dyslipidemia, fasting glucose impairment and high blood pressure, has become prevalent in today's societies. Inflammatory / prothrombotic biomarkers likely are part of the etiology of MetS. Our long term goal is to investigate if there is overlap of genes influencing both metabolic risk factors for MetS and biomarkers. As a beginning we will look at phenotypic correlations and factor analysis as well as established on line genetic databases to inform our discovery of possible pleiotropy.

5. Main Hypotheses/Study Questions:

In this study, we will use a weighted combination of correlations among metabolic traits (which represent also risk factors for MetS), and a large number of inflammatory / prothrombotic biomarkers among several studies to inform our expectations of genetic pleitropic discoveries. This is the first paper in a series of papers on pleiotropic genetic effects. For this paper, only phenotypic data will considered.

6. Design and Analysis:

This study includes 12 human cohorts (N > 68,000) and analyses of 8 metabolic traits (also risk factors for MetS), and 12 inflammatory / prothrombotic biomarkers. Pleiotropy discovery is approached by analyzing correlations among these traits per study and constructing a weighted average covariance matrix for all studies. Factor analyses on simulated data based on above covariances, will be will use in identifying important phenotypic contributions with implications in pleiotropy. In addition, we will use NCBI-gene and GWILL BE databases to mine for putative pleiotropic genes related to these traits, to make the paper more interesting.

Phenotypic variables: ARIC variables to be considered include body mass index (BMI, kg/m^2), waist circumference (WAIST, in cm), high density lipoprotein cholesterol (HDLC, mg/dL), fasting (at least 8 hours) triglycerides (TG, mg/dl), fasting insulin (INS, in mU/L), fasting glucose (GLUC, in mg/dL), systolic and diastolic blood pressure (SBP, DBP, in mmHg, as average of three or 2-nd and 3-rd measures) for metabolic traits and any inflammatory markers from fibrinogen (FIB, in mg/dL), and white blood cell counts (WBCC, in 10e9/L) present in a study, hematocrit (in), platelet count, (PLATCOUNT, in) and hemoglobin (in).

Statistical Analyses- Three correlation analyses will be performed on (A) raw data; (B) raw, but corrected by constants for medication use; and (C) on normally distributed, corrected for medication use, adjusted for covariates and standardized N(0,1) distributions. Correlations for all variables of a study will be performed using pairwise Pearson correlation (using SAS v. 9.3 or R v. 2.15.1). Then two simulation processes will be performed to produce the average correlation matrix and final correlated simulated data for all studies (N = 61,636 subjects). Simulation 1 will be done by following these steps: using N (largest number of subjects *5 per study) and variance-covariance matrices (from above single studies) we will simulate MVN(0,1) of dimension (p-variables, n-subjects) of each study, by using R "mvrnorm" function of MASS library [1]. This large simulated set will be evaluated via Pearson pairwise correlation, which produced a full variance-covariance matrix, representing a simulated approximation of the average of correlation matrices of single studies (Figure 1). Simulation 2 will be performed by using the 1st simulation's average variance-covariance matrix, now to simulate a full set of data with no missing values based on MVN(0,1) with p = 20 variables and N ~ 300,000 subjects. Factor analyses with and without VARIMAX rotation will be performed in SAS, with the final simulated data. VARIMAX rotation creates orthogonal clusters of correlated variables. "No rotation" achieves the simplest latent factor structure, in the extreme case loading any variable in one of the factors and almost negligible loadings in the rest of the factors. On the contrary, when VARIMAX rotation is applied, the objective is to maximize the independence of the clusters of correlated variables that contribute onto specific factors. A loading of 0.4 or larger (which when the data are standardized represents a correlation of an original variable to a factor) will be considered as an important contribution. We note, that it is difficult to average correlations among studies, because they are dependent on transformations that each study may have applied to render the normality of each trait, when such transformation will be needed. Depending on transformations they may change the direction of two trait correlations from one study to the next. Hence, for simplifying combinations of correlations from each study's transformed and standardized variables, we will use the absolute values of each study variance-covariance matrices. Therefore, the signs of correlation coefficients in the average summary variance-covariance matric for all studies are meaningless, because for practical purposes we are evaluating their importance/weight and not their direction.

Selected genes analysis- Genes lists will be searched via Gene Entrez of NCBI for each indidual term, related with the traits studied "body mass index", "waist circumference", "high density lipoprotein cholesterol", "triglycerides", "insulin", "glucose", "systolic blood pressure", "diastolic blood pressure", "fibrinogen", "C-reactive protein", "plasminogen activator 1", "interleukin 6", "interleukin 10", "intercellular adhesion molecule 1", "adiponectin", "white blood cell counts", and "tumor necrosis factor alpha". Our search was conditioned only for human, mouse and rat species (http://www.ncbi.nlm.nih.gov/gene/). Identified genes will represent publication evidence of their contribution to a trait based on linkage, association, function, expression etc. All single traits gene lists will be merged by gene name and selected for most contributions among metabolic traits and inflammatory / prothrombotic markers.

For the same terms, we will implement searches at http://www.genome.gov/gwastudies/. Genes identified as possible candidates will also be checked via Association Results Browser of NCBI http://www.ncbi.nlm.nih.gov/projects/gapplusprev/sgap_plus.htm, These data represent large genome wide studies with at least 100,000 SNPs and with a high statistical significance in the overall (initial GWAS + replication) population [2]. Searches will also be performed on genes using GeneGO (http://thomsonreuters.com/products_services/science/systems-biology/) and Literature Lab of ACUMENTA (http://acumenta.com/) software.

7.a. Will the data be will use for non-CVD analysis in this manuscript?

___ Yes

_x_ No

b. If Yes, is the author aware that the file ICTDER02 must be will use to exclude persons with a value RES_OTH = "CVD Research" for non-DNA analysis, and for DNA analysis RES_DNA = "CVD Research" would be will use?

___Yes

_X__ 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 will use in this manuscript?

___ Yes

__X__ No

8.b. If yes, is the author aware that either DNA data distributed by the Coordinating Center must be will use, or the file ICTDER02 must be will use 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 & 2007.0**2_)**

___ 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.

1. Ripley, B.D., *Stochastic simulation*. 1987, New York ; Chichester: Wiley.
2. Hindorff, L.A., et al., *Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.* Proc Natl Acad Sci U S A, 2009. **106**(23): p. 9362-7.