ARIC Manuscript Proposal # 3351

PC Reviewed: 2/12/19     Status: _____     Priority: ____
SC Reviewed: _________     Status: _____     Priority: ____

1.a. Full Title:  
Sex-Specific Effects and Gene-Sex Interactions on the Serum Metabolome: The Atherosclerosis Risk in Communities Study

b. Abbreviated Title (Length 26 characters):  
Gene-sex on metabolome

2. Writing Group:  
Writing group members:  
Zhe Wang, Han Chen, Michael Swartz, Alanna Morrison, Kari North, Bing Yu, Eric Boerwinkle

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

First author:  Zhe Wang  
Address: Human Genetics Center  
1200 Pressler Street, Suite E-435  
Houston, TX 77030  
Phone: 713-500-9827  Fax: 713-500-0900  
E-mail: zhe.wang@uth.tmc.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: Eric Boerwinkle, PhD  
Address: Human Genetics Center  
1200 Pressler Street, Suite W114A  
Houston, TX 77030  
Phone: 713-500-9058  Fax: 713-500-9020  
E-mail: Eric.Boerwinkle@uth.tmc.edu

3. Timeline:  
The exome chip and metabolome data are available. No new data collection is proposed. Once this proposal is approved, analysis will start immediately. The manuscript is to be prepared as soon as analyses are finished (<6 months).

4. Rationale:
Metabolomics is a scientific approach that systematically evaluates small-molecule metabolites in biologic samples and may provide additional insights into disease pathology (1-3). Both traditional genome-wide association studies (GWAS) and sequencing analyses across the exome or whole genome have successfully identified and verified hundreds of genetic loci associated with the levels of metabolites (4-16), and many of them can be further related to complex diseases or clinically relevant risk factors (12, 17). Sex-specific differences in metabolite patterns in healthy humans have been reported in urine and plasma (18, 19), which suggests that sex-specific effects should be considered further in metabolomic studies. Limited work has identified sex-specific metabolism-related genetic polymorphisms through sex-stratified GWAS and sex-specific pathway differences in the serum metabolome (20, 21). Additional systematic studies are needed to better understand the modifying effect of sex on the human metabolome and its genetic determinants including rare and low frequency (MAF ≤ 5%) variants. Rare and low frequency variants make up the vast majority of the genetic variation in the genome (22), and may account for part of the missing heritability along with gene-environmental (G×E) interactions (23).

Emerging methods have been recently proposed for detecting rare variant G×E interactions. Su et al. proposed a novel and rigorous framework, Mixed effects Score Tests for interaction (MiSTi), to derive independent score statistics for fixed effects and the variance component that is more powerful to test G×E interaction terms of rare variants (24). A joint test that allows one to simultaneously test genetic main effects and interaction effects was proposed and successfully implemented by Chen and colleagues in the R package ‘rareGE’ (25). Compared to common variant analyses, rare variant analyses often require a larger sample size to attain comparable power. Interaction analyses also need larger sample sizes in comparison with main effect analysis (26, 27). Therefore, interaction analyses for rare genetic variants require extra attention, particularly related to consideration of statistical power.

To date, there is no study systematically utilizing methods developed for testing rare variant G×E interactions in the setting of large-scale metabolomic data. In addition, current understanding of gene-sex interactions or sex-specific genetic variants related to metabolites has solely originated from studies in Whites, and there is a need for expanding these studies to underrepresented populations, such as African-Americans (AAs). Therefore, we propose to leverage existing data from the Atherosclerosis Risk in Communities (ARIC) study that contains well-characterized AAs and European-Americans (EAs), to investigate any sex-specific differences in the genetic effects on the metabolome and to identify novel genetic loci that were not identified when considering the genetic main effect alone. We will also compare the power of two emerging approaches “rareGE” (25) and “MiSTi” (24) with standard stratified analyses followed by a test of differences of the effect sizes, “Z test”, (28) using simulated data to improve the understanding of the power of GxE interaction analyses and help interpret the observed results.

5. Main Hypothesis/Study Questions:
To investigate sex differences in the genetic effects on the metabolome and to identify novel genetic loci that were not identified when considering the genetic main effect alone. Both common and rare functional exonic variants will be considered.
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:
1) This is a study using cross-sectional exome chip and metabolome profiles data in both AAs and EAs of the ARIC study.
2) Metabolite profiling was completed in 2010 (batch 1) and 2014 (batch 2) using fasting serum samples that had been stored at -80°C since collection at the baseline examination.
3) Genotyping was performed with the Illumina HumanExome BeadChip v1.0 (“exome chip”) at the baseline examination in 11,071 EAs and 2,953 AAs in the ARIC study. To improve accurate calling of rare variants, joint calling from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium were used with details described elsewhere (29). Exome chip variant annotation was completed using the Whole Genome Sequencing Annotation (WGSA) pipeline v055 (30), including dbNSFP v2.9 (31). Functional variants and genes were determined using ANNOVAR (32) according to the reference genome GRCh37/hg19 and National Center for Biotechnology Information RefSeq.

Exclusion:
1. We exclude individuals without genotype or phenotype data and those who failed genotype quality control.
2. We will exclude individuals not providing informed consent, including consent for genetic studies.
3. Included in this analysis will be 3,540 participants with metabolite measurements and exome chip genotyped data at the baseline examination.

Variables:
Outcome variables: named metabolites primarily.
Metabolites will be excluded if:
1) Only detected in one batch;
2) More than 40% of the samples have missing values within each batch;
3) The Pearson correlation coefficient (r) between 2010 and 2014 measurements on the same stored sample is less than 0.30.
Metabolite levels will be analyzed as continuous variables, where missing values will be imputed using random forest imputation based on the remaining observed measurements (33, 34). After applying such exclusion criteria, this study will be based on an evaluation of 271 named metabolites.

Exposure variables: Exome chip data with annotated functional variants such as splicing, stop-gain, stop-loss, nonsynonymous, and indels.

Covariates: Age, first 3 ancestry specific principle components, estimated glomerular filtration rate (eGFR) and batch for the metabolomic measurements.
Data analysis:
Metabolite levels will be winsorized (99%) within each batch, respectively. Due to right-skewed distributions of many metabolite levels, natural log transformation will be applied to most metabolites prior to analyses. For metabolites that cannot be normalized using log transformation, a rank based inverse normal transformation will be used. All statistical analysis will be conducted within each race group follow an analysis flow as showed in Figure 1.

Figure 1. Data analysis flow

We will perform single nucleotide polymorphisms (SNPs) analysis for common variants (MAF >5%) and gene-based analyses for rare and low-frequency variants (MAF ≤5%). A MAF cut-off 5% was chosen in light of the relatively small available sample size.

Common variant analyses:
Standard linear regression will be applied for each metabolite level on an additive genetic model in men and women separately. To test each common variant and metabolite for difference of the effect size estimates for the variants calculated in the sex-specific analyses, we will use an approximately normally distributed test statistic, Z (28).

Gene-based analyses:
The unit-of-analysis will be an annotated gene. All annotated coding variants, such as splicing, stop-gain, stop-loss, nonsynonymous, and indels within the gene will be aggregated for the analysis. We will perform three types of test for rare variants:

1. Burden test in men and women separately. Genes with cumulative minor allele count < 3 within men or women of each race group will not be analyzed. The Z test will be used to evaluate the aggregated gene effects for heterogeneity between men and women (28).
2. A joint analysis of genetic main effects and G×E interaction effects will be conducted using rareGE (28). This joint testing approach aims to detect associated genetic effects allowing for G×E interactions. Genes with cumulative minor allele counts ≤ 6 in each race will be excluded.

3. A G×E interaction term test only using both rareGE and MiSTi (27, 28) will be conducted. The interaction-only test allows detecting G×E interactions regardless of the genetic main effect. Again, genes with cumulative minor allele counts ≤ 6 in each race will be excluded.

We will use AAs as our discovery sample and conduct replication in EAs. Using a false discovery rate (FDR) to correct for number of genes/SNPs tested while considering the 271 metabolites, we define exome-wide significant genes as those with FDR Q ≤ 5% in discovery AAs; these genes will be pursued for replication analyses in EAs. Replication is defined as those genes with FDR Q ≤ 5%, corrected for the number of genes taken forward to evaluate in EAs. All statistical analyses will be conducted in R version 3.4 (R Foundation for Statistical Computing, Vienna, Austria).

Simulation for power:
We will use 11,071 EAs with exome chip data and simulate a quantitative phenotype (metabolite levels) with varying effect size (0.5, 1, 1.5, 2) and directions of effects, as well as total sample size (from 20,000 by doubling the EA exome chip data down to 2,000) to investigate the power of rareGE, MiSTi and conventional Z test. The primary parameter of interest in the simulation studies will be the effect size.

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 ICTDER 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

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.cscu.unc.edu/ARIC/search.php

   _X_ Yes   _______ No
There is no overlap between this proposal and current active proposals/published manuscripts. There are no metabolomics manuscript proposals examining the gene by sex interactions in metabolite levels.

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)?

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* __ AS#2014.20 ___)
   ___ 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/

12a. 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.
   Agree.

12b. The NIH instituted a Public Access Policy in April, 2008 which ensures that the public has access to the published results of NIH funded research. It is your responsibility to upload manuscripts to PubMed Central whenever the journal does not and be in compliance with this policy. Four files about the public access policy from http://publicaccess.nih.gov/ are posted in http://www.cscc.unc.edu/aric/index.php, under Publications, Policies & Forms. http://publicaccess.nih.gov/submit_process_journals.htm shows you which journals automatically upload articles to PubMed central.
   Yes, the lead author is aware of the policy.

13. Per Data Use Agreement Addendum, approved manuscripts using CMS data shall be submitted by the Coordinating Center to CMS for informational purposes prior to publication. Approved manuscripts should be sent to Pingping Wu at CC, at pingping_wu@unc.edu. I will be using CMS data in my manuscript ____ Yes __X__ No.

Reference: