ARIC Manuscript Proposal #2001

PC Reviewed: 9/11/12	**Status: A**	**Priority: 2**
SC Reviewed: _________	**Status: _____**	**Priority: ____**

1.a. Full Title: Genome-wide study of copy number variation (CNV) and serum urate

b. Abbreviated Title (Length 26 characters): **CNV and urate**

2. Writing Group:

Writing group members: Lynn Mireles, Linda Kao, Eric Boerwinkle, Adrienne Tin, Robert Scharpf. Other ARIC study members are welcome to participate.

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

F**irst author**: **Lynn Mireles**
Address: 615 N. Wolfe Street, W6513, Baltimore MD 21205
Phone: 410-614-0945
Fax:
E-mail: lmireles@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: **Linda Kao**
Address: 615 N. Wolfe St.
Room W6513
Baltimore, MD 21205

3. Timeline: September 2012, analysis completion
November 2012, first draft of manuscript for circulation

4. Rationale:

Hyperuricemia is associated with multiple diseases, including gout, cardiovascular disease, and renal disease. Serum uric acid concentrations are highly heritable suggesting a strong genetic component, yet genome wide association studies of single nucleotide polymorphism (SNP) and serum uric acid concentrations explain only a small fraction of the heritability. Therefore, it is hypothesized that some of the missing heritability might be attributed to copy number variants. The association of copy number variants (CNV) and risk of hyperuricemia has not been reported.

5. Main Hypothesis/Study Questions:

We will test the hypothesis that copy number variants are associated with serum urate.

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:

We will perform CNV discovery association analysis on participants who consented for genetic studies and were genotyped on the Affymetrix 6.0 platform. Exclusions include: individuals who did not consent to genetic research; individuals whose self-reported race is not “White”; samples missing outcomes or covariates; first-degree relatives; and individuals with excess autosomal heterozygosity, a mismatch between genotypic and phenotypic gender, or that are genetic outliers.

Methods for CNV discovery will include the hidden Markov model (HMM) implemented in PennCNV (Wang et al., 2007) and a HMM implemented in the R package VanillaICE (VI) (Scharpf et al., 2008). DNA sample quality will be ascertained by the variance of low-level summaries of copy number, as well as the number of CNVs called by PennCNV/VI.

Linear regression models will be used to test the association of serum uric acid (measured at the first visit) and CNV while adjusting for age, sex, and study site. CNV will be modeled as a three-category variable (deletion, normal, gain). A 2-degree of freedom test for significance will be used. In addition, we will assess and adjust for technical factors known to impact copy number estimates in genotyping arrays as needed. If an association of CNV and serum uric acid is found within a megabase (Mb) of a Genome-Wide Association study index SNP, the level of linkage disequilibrium between the index SNP and the CNV will be calculated and the models will include a term for genotypes.

There are several methodological challenges to CNV discovery and association analyses. First, low-level summaries of relative copy number, such as log R ratios, are noisy and highly susceptible to technological source of variation such as batch effects (Leek et al., 2010). We will leverage the genotypes at polymorphic markers on the array, which are more robust to batch effects, to guide the estimation of the low level summaries (Scharpf et al., 2011). Secondly, biases in PCR efficiency can occur as a result of differences in the GC content of probes. These biases can vary in magnitude and direction between samples, and manifest as waves in plots of normalized intensities versus physical position. Without correction, CNV calls may contain a number of false positives as peaks and valleys in the waves are confused with copy number alterations. We will explore previously published tools for wave correction (e.g, Diskin et al., 2008) and extend such methods for the analysis of ARIC samples as needed. Finally, there will be a number of samples for which the DNA quality is too poor or for which the above preprocessing steps do not adequately remove technical artifacts. Such samples will be excluded prior to assessing the association of copy number and serum uric acid.

References:

SJ Diskin, M Li, C Hou, S Yang, J Glessner, H Hakonarson, M Bucan, JM Maris, and K Wang. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. Nucleic Acids Res, 2008 ; 36(19):e126.

JT Leek, RB Scharpf, HC Bravo, D Simcha, B Langmead, WE Johnson, D Geman, K Baggerly, and RA Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet, 2010; 11(10): 733-739.

RB Scharpf, G Parmigiani, J Pevsner, and I Ruczinski. Hidden Markov models for the assessment of chromosomal alterations using high-throughput SNP arrays. Annals of Applied Statistics, 2008; 2(2):687-713.

RB Scharpf, I Ruczinski, B Carvalho, B Doan, A Chakravarti, RA Irizarry. A multilevel model to address batch effects in copy number estimation using SNP arrays. Biostatistics, 2011; 12(1): 33-50.

K Wang, M Li, D Hadley, R Liu, J Glessner, SFA Grant, H Hakonarson, and M Bucan. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res, 2007; 17(11):1665-1674.

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

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)? None found

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.**