ARIC Manuscript Proposal #2008

PC Reviewed: 10/9/12	**Status: A**	**Priority: 2**
SC Reviewed: _____	**Status: _____**	**Priority: ____**

1.a. Full Title: Best practices and joint calling of the HumanExome BeadChip: the CHARGE consortium

b. Abbreviated Title (Length 26 characters): CHARGE exome chip best practices

2. Writing Group:

Writing group members: Megan L. Grove, Bing Yu, Barbara J. Cochran, Talin Haritunians, Joshua C. Bis, Kent D. Taylor, Mark Hansen, Christopher J. O'Donnell, Jerome I. Rotter, and Eric Boerwinkle, on behalf of the CHARGE Exome Chip Genotyping Committee

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

First author: Megan L. Grove
Address: The University of Texas Health Science Center at Houston
Human Genetics Center
1200 Hermann Pressler, RAS W406A
Houston, TX 77030
Phone: 713-500-9833
Email: Megan.L.Grove@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: The University of Texas Health Science Center at Houston
Human Genetics Center
1200 Hermann Pressler, RAS E453
Houston, TX 77030
Phone: 713-500-9816
Email: Eric.Boerwinkle@uth.tmc.edu

3. Timeline:

We expect a manuscript to be ready for submission within 1 month.

4. Rationale:

Exome- and whole-genome sequencing is becoming increasingly affordable and allows for detection and genotyping of rare variants in the human genome. Yet, genotyping arrays remain a cost-effective approach when determining genetic polymorphisms previously identified in large populations. A limitation of using arrays to genotype rare variants is the ability of automated clustering algorithms to accurately detect and assign accurate genotype calls (Korn et al. 2008; Ritchie et al. 2011). Having a very large sample would increase the number of occurrences of rare variants and, therefore, facilitate automated clustering and genotyping. Participating studies in the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium (Psaty et al. 2009) consented to have their Illumina Infinium HumanExome BeadChip intensity data analyzed collectively (N=62,150) in order to increase the accuracy of rare variant genotype calls. The resulting cluster file is publically available and we show that its use increases genotype accuracy even in studies with small sample sizes.

5. Main Hypothesis/Study Questions:

To evaluate the performance of our rare variant calling approach of array data we plan to compare genotypes derived from the following three methods to available exome sequencing data of 530 ARIC individuals with sample call rates > 0.90. First, exome chip genotypes will be called with the Illumina issued cluster file (HumanExome-12v1.egt). Second, we will use zCall (Goldstein et al. 2012) to determine genotypes for the missing variant calls in Dataset 1 (Dataset 2). Third, we will use the CHARGE best practices and joint calling approach to ascertain exome chip genotypes (Dataset 3). Results will be reported as genotype concordance percentages. Additionally, we will compare the genotype concordance of singletons, doubletons and tripletons in the array data to exome sequencing data.

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).

Data from approximately 60,000 participants from the following eleven studies in the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium were included in this joint calling effort: Age, Gene/Environment, Susceptibility-Reykjavik (AGES) Study, Atherosclerosis Risk in Communities (ARIC) Study, Cardiac Arrest Blood Study (CABS), Cardiovascular Health Study (CHS), Coronary Artery Risk Development in Young Adults (CARDIA) , Erasmus Rotterdam Gezondheid Onderzoek (ERGO): the Rotterdam Study, Multi-Ethnic Study of Atherosclerosis (MESA), Family Heart Study (FamHS), Framingham Heart Study (FHS), Health, Aging, and Body Composition (HABC) Study, and Jackson Heart Study (JHS). In addition, 96 unrelated HapMap samples were genotyped along with each cohort to be utilized for quality control and determination of batch effects.

Study samples were processed on the HumanExome BeadChip (v1.0 bead pool, Illumina, Inc., San Diego, CA) querying 247,870 variable sites using standard protocols

suggested by the manufacturer at the following seven genotyping centers: Broad Institute (JHS), Cedars-Sinai Medical Center (CHS, FamHS and MESA), Erasmus Medical Center (ERGO), Illumina Fast Track Services (FHS), University of Texas Health Science Center at Houston (AGES, ARIC and CARDIA), University of Washington (CABS), and Wake Forest University (HABC). The two channel raw data files (.idat) for all samples were transferred to a central location at UT Houston and assembled into a single project for joint calling. Laboratory best practices were applied and genotype will be available for further cleaning and analyses in early October.

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? _X___ 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.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)? There are not related manuscript proposals.

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* _2009.12__)
___ 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.

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.