ARIC Manuscript Proposal # 1622

PC Reviewed: 3/9/10	**Status: A**	**Priority: 2**
SC Reviewed: _________	**Status: _____**	**Priority: ____**

1.a. Full Title: Consequences of excess rare variation in sequences of *HHEX* and *KCNJ11* from a large cohort study (Note: There are no phenotype data in this study.)

b. Abbreviated Title (Length 26 characters): Excess rare variation in sequences of *HHEX* and *KCNJ11*

2. Writing Group:

Writing group members: Alex Coventry, Lara M. Bull-Otterson, Xiaoming Liu, Andrew G. Clark, Taylor J. Maxwell, Jacy Crosby, James E. Hixson, Thomas J. Rea, Alan R. Templeton, Eric Boerwinkle, Richard Gibbs, Charles F. Sing

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

First author: Alex Coventry

Address: Dept of Biology and Genetics
Cornell University
Biotechnology Bldg, Room 101C
Ithaca, NY 14850

Phone: 607-255-6515 Fax: 607-255-6249
E-mail: Coventry@cornell.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
Address: Human Genetics Center
1200 Hermann Pressler, Suite E447
Houston, TX 77030

Phone: 713-500-9816 Fax: 713-500-0900
E-mail: eric.boerwinkle@uth.tmc.edu

3. Timeline: Immediate.

4. Rationale:

The next phase of research in the genetics of complex disease entails deep resequencing of large population-based samples of individuals (Collins 2010), enabling a first glimpse at the role that extremely rare variants play in human genetic and phenotypic diversity. Because they have not been subject to natural selection for more than a few hundred generations, loss-of-function alleles and other large biological effects will be present in much larger proportions among rare-variant sites (Relative Minor Allele Frequency (RMAF) < 0.01) than among the common SNPs (RMAF >= 0.01) on which most GWA studies have focused. Also, we will show that the super-exponential growth of the human population over the last few millenia means that the vast majority of human genetic variants are extremely rare. Sequencing two genes in a large cohort sample has allowed us to assess the potential impact of the full spectrum of variants on the genetics of medically relevant human phenotypes. We selected *KCNJ11* and *HHEX* for resequencing in ARIC, because they are good etiological candidates for type 2 diabetes risk (*KCNJ11* encodes a potassium channel in pancreatic cells, and *HHEX* is essential for ventral pancreatic tissue formation) and have been implicated in diabetes risk by multiple GWAS (*e.g.* Schwanstecher *et al.* 2002, Vliet-Ostaptchouk *et al.* 2008), including a recent meta-analysis including the ARIC study (Dupuis et al, 2010).

5. Main Hypothesis/Study Questions:

1. There is an excess of rare variants in these two diabetes-related genes in ARIC.
2. This excess can be accounted for by rapid population growth in recent human history.

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).

We applied Sanger sequencing of PCR products from genomic DNA to sequence 50 amplicons covering the genes, UTRs and introns in the 13,715 ARIC individuals, which includes 3,293 African-Americans and 10,422 individuals of European ancestry. We achieved excellent coverage of the sequencing loci, and found evidence for 52 insertions/deletions at 42 sites. We assigned probabilities to potential SNP variants on the basis of (1) independent phred scores for each peak at apparently heterozygous sites, (2) relative heights of peaks and risk of leakage from nearby peaks, and (3) calculated genotype frequencies at each site. The data reveal a vast number of rare variants: by averaging over many draws from the genotype probabilities, we compute the expected number of variant sites in this sample to be 742 in total, with 382 in HHEX and 360 in KCNJ11. By contrast, dbSNP reports just 28 and 68 SNPs in *HHEX* and *KCNJ11*. To confirm that these apparent variants were not sequencing errors, we validated 789 potential rare variants by barcoding and pooling the relevant PCR amplicons, then submitting them *en masse* to 454 Roche sequencing.

This vast excess of extremely rare variants sheds light on recent human demography. Most human population samples have shown that, consistent with population expansion, there is an excess of rare variation compared to the standard constant-population-size Wright-Fisher model (Schaffner *et al*. 2005; Nielsen *et al.* 2009; Guttenkunst *et al*. 2009). Because those inferences have been drawn from resequencing of modest samples (*n* ~ 100) and from HapMap SNPs ascertained with a bias toward common variation, they focused on common variants and have very little signal from the demography of the last few thousand years. Because mutations are uniformly distributed over these genealogies, a preponderance of mutations have fallen on those very recent branches, and thus most variants appear in very few contemporary individuals. To test this, we fitted a model of exponential growth to the site frequency spectrum of this European-American sample. The huge excess of rare variants in *HHEX* and *KCNJ11* fits well with this model, giving a mean posterior growth rate of 2.2%/generation. Despite relying on shallower regenotyping data, earlier models such as in (Guttenkunst *et al*. 2009) have also found a good fit to an exponential model, but with a substantially lower growth rate (Gutenkunst *et al.* found a modal value of 0.4%/generation). By comparing our results with Gutenkunst *et al.*, we conclude that Europe's population growth rate was over five times faster over the last 7,500 years than it had been prior to that time. These data are intriguing in light of the advent and expansion of agriculture, and the role of these genes in energy metabolism and diabetes.

7.a. Will the data be used for non-CVD analysis in this manuscript? _?___ Yes _?___ 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 (Just to be safe.)
(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)? There are no related manuscripts.

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* 2006-11_)

___ 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.

Agreed.