ARIC Manuscript Proposal # 1781

PC Reviewed: 5/10/11        Status: A        Priority: 2
SC Reviewed: __________    Status: ____       Priority: ____

1.a. Full Title: Next-generation sequencing for Atrial fibrillation: Results from the CHARGE-S Project

b. Abbreviated Title (Length 26 characters): Next-gen sequencing in AF

2. Writing Group: Atrial Fibrillation or CHARGE-S AF

Writing group members: Dan E. Arking, Alvaro Alonso, Alanna Morrison, Eric Boerwinkle (and/or other Houston personnel). Others ARIC authors welcome. Other authors from additional consortium cohorts will included, with a plan to maintain symmetry across cohorts.

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

First author:     Dan E. Arking
Address: Johns Hopkins University School of Medicine
               733 N. Broadway, BRB 453
               Baltimore, MD 21205
               Phone: 410-502-4867       Fax: 410-614-8600
               E-mail: arking@jhmi.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: Alvaro Alonso
Address: Div. Epidemiology & Community Health
         Sch Public Health, U of Minnesota
         1300 S 2nd St, Suite 300, Minneapolis, MN 55416
         Phone: 612-626-8597       Fax: 612-624-0315
         E-mail: alonso@umn.edu

4. **Rationale:**
Atrial fibrillation (AF) is the most common arrhythmia and is a major public health burden. It is associated with a five-fold increased stroke risk, doubling in dementia risk, tripling in heart failure risk, and nearly two-fold increase in mortality. Numerous risk factors predispose to AF, yet our understanding of the fundamental mechanisms for the arrhythmia remains limited. In recent years, the heritability of AF has been established and mutations have been described in rare families with AF. Genome wide association studies have identified genomic loci and common variants for AF; however, these loci explain only a fraction of the heritability of AF. One current concept in explaining the genetic background of AF is a combination of common risk alleles and rare variants. Whereas common variants are shared by a larger proportion of the population, rare variants can only be found in a small subset of patients.

5. **Main Hypothesis/Study Questions:**
Here, we propose to perform an analysis of targeted and exome sequencing data available from subjects in the CHARGE-S AF project. Three targeted sequencing regions were selected from among gene regions with genome-wide significant associations for AF previously identified via GWAS meta-analysis. Four additional targeted sequencing regions were selected due to association with PR interval. Additional AF-associated genes will be available for study from the CHARGE-S exome sequencing study, and will be analyzed in a similar manner as the targeted regions, described in detail here. We hypothesize that in the target regions, a combination of common and rare variants will add up to a more comprehensive explanation of the variability in AF risk. The target regions were selected based on prior knowledge from genome-wide association studies on AF and partly the PR interval as a proxy for AF.

**We will systematically evaluate genetic variants identified in the target regions by:**
1) Using bioinformatics analyses,
2) Screening for mutations and replication of identified rare variants in additional cohorts with AF.

**Functional analyses will vary depending upon the findings of Aims 1 & 2 and may include:**
3) Assessing the function of selected rare variants and mutations by gene knockdown in a zebrafish model,
4) Assessing the cellular electrophysiology of selected rare variants and mutations in cell lines.

Approaches 3) and 4) have to be considered optional at this point, however, we believe that they would add novelty to the analyses. We anticipate that such a multidisciplinary approach will enable the identification of rare genetic variants and mutations associated with AF that promote the clarification of the missing heritability for this arrhythmia.

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 samples

The CHARGE-S project design is a case cohort study, in which extremes chosen for specific traits are considered in conjunction with a cohort subsample selected randomly from the larger total possible sample. The cohort random sample consists of 1000 ARIC, 500 FHS, 500 CHS study participants, in approximate proportion to the total number of GWA study participants in each of these studies. Biologically unrelated individuals were chosen for this sample. A stratified random sample has been selected from each study, stratifying on men and women. For targeted sequencing, case groups of 200 unrelated study participants each from the extremes of 11 specific traits have been selected. These cases groups generally mimic the same study proportions as the cohort random sample with 100 ARIC, 50 CHS and 50 FHS study participants for each trait balanced by sex.

For the AF subproject, approximately 200 case subjects with early-onset AF were chosen from MGH. Case subjects with AF were derived from a selection of AF patients from MGH who developed AF at relatively early age (≤ 66 years, no structural heart disease at diagnosis). It has been demonstrated that subjects with early-onset AF have a more pronounced genetic background. Additional AF cases will be identified from the random cohort (n=303) and the other trait groups (n = 409), for a total of 921 AF cases. A distinction will be made between random sample AF cases, depending on whether they developed the arrhythmia as early-onset AF or as typical AF (> 65 years, structural heart disease possible). Referents will be derived from the CHARGE-S samples, if the subjects are free of AF (n=3321 across all studies). AF-free individuals from FHS will be used as controls for FHS AF patients as well as the AF cases from MGH. In the past, a similar combination of MGH cases and FHS controls has been successfully utilized, since both MGH patients and FHS participants are derived from the same geographic background. For ARIC and CHS, respectively, AF cases and controls will be compared within the study and the results will then be shared as outlined below.

Additional case groups were selected for exome sequencing, with the same proportions of ARIC, CHS, and FHS study participants, and will include 259 AF cases from across all 3 cohorts. Half of the cohort random sample (N=1000) will also be exome sequenced (AF cases=171), yielding a total of 430 AF cases, and controls will be identified as described above from both the random sample and case groups (n=1901).

Main Statistical analysis plans & methods

Prior to statistical analyses, rigorous quality control measures will be applied to the data. These measures will follow closely the analysis plan as proposed by the CHARGE-S analysis committee.

Target regions: A total of 59 kb genomic data will be analyzed, which are derived from the following candidate genes. The genes were selected according to their involvement in AF: CAV1 / CAV2, PRRX1, ZFHX3, or the PR interval: C12orf67, SCN5A, SCN10A, SOX5.

Exome sequencing: The sequenced regions of the genes that have genome-wide significant associations with AF and / or the PR interval will be analyzed. A list of genes
and respective regions will be available once the currently ongoing genome-wide studies have finalized their analyses.

**Bioinformatics prediction of the identified variants:** Genetic variants identified by sequencing will be categorized. One categorization will be based on the detected allele frequency, where common variants occur in ≥1% of cases, and rare variants occur in few subjects (<1%) only. A second categorization will be based on the predicted impact on the encoded proteins, such as synonymous, missense, nonsense, frameshift or splice-site alleles. PolyPhen-2 and SIFT will be used to predict potential deleterious effects of nonsynonymous mutations on gene function. Deleterious mutations will be further examined in terms of protein functional domains and 3D structures. Evolutionary conservation at variant loci will be studied by multiple sequence alignment of vertebrate genomes. Gene function and pathway information will be retrieved from Gene Ontology database and KEGG pathway database.

**Analysis:** According to their classification, identified variants will be treated differentially. For common variants, we will use a logistic regression model to assess associations with AF. Unweighted linear or logistic regression followed by a weighted adjustment to obtain unbiased population odds ratio estimates will be performed for each trait. The rationale for this strategy is the following: Unweighted linear regression was found to have the best power in simulations among three strategies considered. The unweighted approach has valid type 1 error, but biased estimates of the regression parameters. The other two approaches were a weighted least squares approach that weights each study participant by the probability of sample selection. Study participants in the extreme (all extremes regardless of trait) each have a weight of 1; study participants in the random cohort have weights that are the inverse of the probability of being selected. The weighted least squares approach was found to have inflated type 1 error. A second weighted approach includes a two-phase sampling adjustment for finite sample selection and provides a Horvitz-Thompson-like estimator. This approach has valid type 1 error and unbiased estimates of the population regression parameters, but it has lower power than the unweighted approach. Hence, we will use the unweighted approach to identify associations, and will compute population-level effect estimates using a weighted approach. This approach will provide the best power with unbiased estimates of the effect estimates in the population. Findings will be compared with previously identified GWAS findings. All analyses will be adjusted for age, sex and hypertension and an additive model of inheritance will be assumed.

For rare variants (≤1% minor allele frequency (MAF)), a single marker based association test has lower power, however, larger effect sizes of rare variants might be expected. We will assess and compare different approaches including a pooled association test or bioinformatics weighting of variants. Power for pooled tests of multiple rare variants with a certain combined frequency will be similar to the power for the single SNP tests with the same allele frequency, assuming all of the combined variants are truly associated and all increase or decrease risk a similar amount.
Very rare variants that occur in only few cases are considered mutations. A formal test of association will most likely not be sufficiently powered. We will therefore descriptively focus on variants with predicted deleterious effects on the encoded gene product.

**Replication:** Samples for targeted replication of identified variants are readily available. Confirmatory, replication genotyping will be possible in other samples including MGH / FHS, HVH, and AFNET / KORA.

**Sources of Data to be used**

The general goal is to retain as much information as possible without sharing individual level sequence data. Each study will conduct analyses for the data in their study as agreed upon by the working group and then summary statistics from each study will be combined into a single score test across studies to provide a global test of association. For the first pass of analyses using the unweighted approach, the weighted Z approach for combining p values and signs of regression coefficients will be used. MGH cases will be combined with FHS cases, and then be analyzed using FHS controls. Case-controls analyses in ARIC and CHS will be performed within each respective study.

As needed, GWAS data will be used internally in each cohort as a quality control measure for exome sequencing.

Additional analyses will be undertaken to evaluate whether sequence data explain the GWA signal that was the impetus for examining the target. These analyses will condition on the SNP that was the primary signal for the target in the GWA results as an additional covariate in the regression analyses described above.

**The following phenotype data will be required at each study site:** self-reported race, sex, age, PR interval, RR interval (inverse of heart rate), BMI, height, prevalent heart disease (myocardial infarction or heart failure), Wolff-Parkinson-White syndrome, bundle branch block, class I, II, III, or IV antiarrhythmic medications, beta-blocking medication, calcium channel blocking medication, digoxin, study site, prior pacemaker, second or third degree heart block, blood pressure, atrial fibrillation, duration of atrial fibrillation, onset of atrial fibrillation, date of blood draw, HTN, or treatment for HTN.

**7.a. Will the data be used for non-CVD analysis in this manuscript?**  
_**X**_ 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**_

(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)? MS1396, MS1397, MS1398 which are GWAS for AF, lone AF, and PR, which we lead.

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* ___ 2009.14 ______)  
  ___ 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.

References