ARIC Manuscript Proposal # 1763

PC Reviewed: 3/8/11	**Status: A**	**Priority: 2**
SC Reviewed: _________	**Status: _____**	**Priority: ____**

1.a. Full Title: CHARGE for BP (SBP, DBP, MAP, PP, HTN): CHARGE-S sequencing

b. Abbreviated Title (Length 26 characters): CHARGE-S BP

2. Writing Group:
Aravinda Chakravarti for CHARGE-BP, Eric Boerwinkle

I, the first author, confirm that all the coauthors have given their approval for this manuscript proposal:

First author: Aravinda Chakravarti
Address: Johns Hopkins University School of Medicine
733 N. Broadway, BRB 453
e-mail: georg@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: Eric Boerwinkle
Address: University of Texas Health Science Center at Houston School of Public Health
1200 Herman Pressler Dr.
Houston, Texas 77030
United States
Phone: (713) 500-9800
Eric.Boerwinkle@uth.tmc.edu

3. Timeline: spring 2012

4. Rationale: Persistent elevated blood pressure (BP), diagnosed as hypertension (HTN), is quantitatively the major cardiovascular risk factor with a population prevalence of ~30%. Pathogenic pathways that lead to HTN remain poorly understood. A distinct fraction of the hypertension risk is genetic and this opens the possibility for genetic investigations to contribute to a better understanding of this trait and possible identification of new molecular targets for drug therapy.

ARIC has published a first genome-wide association study on SBP and DBP within the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium. That experiment only explains few percent of the BP variability and our hypothesis is that variants within the rare spectrum of allele frequency will explain a sizable fraction of the heritability.

Here we propose to analyze the sequencing data made available by CHARGE-S (targeted and exome-wide data) to analyze BP traits. It is our intention to meta-analyze our results with other CHARGE studies and non-CHARGE studies that have similar results.

5. Main Hypothesis/Study Questions:

1. To identify common variation that may account for the observed GWAS associations.
2. To identify low frequency alleles in the same genes that may account for additional genetic variability.

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).

Overview

The fundamental data are the demographic data, the blood pressure (BP) values (across multiple visits), other CVD relevant phenotypes, and DNA sequence information for 5 BP association loci (*ATP2B1, SH2B3, PLEKHA7, CACNB2, CYP17A1*) in a set of 2,000 random cohort controls and a case group of 2,200 individuals sampled from the cohorts ARIC, FHS and CHS in the approximate proportions 50%:25%:25%, respectively. On a subset of the individuals whole exome sequence data will be available.

Study Design

The original BP proposal included only 200 individuals sampled from the extremes of the BP distribution (we used the age, age^2, gender and BMI corrected residuals of both SBP and DBP). Fortunately, since we are using a quantitative phenotype, we can have access to BP measurements from all case samples. Consequently, our analysis will utilize 2,200 cases (of which 200 are extreme BP values) and 2,000 random controls. The analysis plan will be the same for both targeted sequencing data and exome sequence data, but the interpretations may be different.

Genetic analyses

We are assuming that we will obtain for each individual a base location, the number of reads overlapping that base (including counts of forward a reverse reads) and the numbers of reads specifying the bases A, C, G and T, respectively. All sequence calls will be with respect to a fixed human genome assembly and assumed to be on the + strand (forward) for reference. We are also assuming that we will be provided with quality values for each base. In a first analysis these quality measures are sufficient, but subsequently more detailed information on read quality may be necessary. (For example, the nature and frequency of the second most popular base is often a good indicator of false-positive status.)

The specific analyses to be conducted are:

1) *Quality Analysis & Annotation:* Using the raw sequence data, we will threshold all reads and calls by some quality standard. Although absolute standards may be used, we wish to use the distribution of read and base quality to pick some value whereby 95% or more of the reads pass the threshold. This analysis will further look at the distribution of forward and reverse reads, and the four base calls to further trim the data, such as by imposing a threshold of F/R read ratio between 25%-75% and the two rarest bases being cumulatively under 5%. Once again the observed data will be used to set the specific threshold.

The above analysis will be conducted for both SNPs and CNVs.

We will also separate variants by coding and noncoding variants (the latter being further subdivided by ENCODE regulatory sites versus not), by frequency and by conservation. For coding variants, we will also classify them by their predicted effect (damaging or not). We suggest a frequency binning into four classes: (1) those >10% in the random cohort; (2) those less than 10% but having 10 or more observed variant homozygotes; (3) all other sites but for singletons; (4) singleton sites only. I suggest a rough binning for conservation of (1) below and (2) above the genome-wide average for non-coding variants.

2) *Basic Population Genetic Analysis:* The quality controlled variant data will be analyzed by variant frequency spectrum and fit to the neutral frequency model (to demonstrate excess as frequency decreases) by quality scores. We will also compare to existing 1000 Genomes and HapMap data to classify sites by type, frequency, previously observed or not, observed in other populations or not, frequency by quality. The main aim here is to identify coding or regulatory rare variants (classes 3 and 4 above) of high quality.

Next, we will convert all data to genotypes and analyze, in both case and control groups, allele frequency, linkage disequilibrium and fit to Hardy-Weinberg proportions.

Subsequently, we will use the 1000 Genomes data to impute additional associated variants and genotypes and threshold by imputation quality score.

3) *Association Analysis:* We will perform association analyses in the total data set for SBP, DBP, MAP, PP for each visit and LTA of the 4 BP traits across all visits using their age, age^2, gender, BMI-adjusted residuals.
(a) Single SNP analyses: For frequent sites (classes 1 and 2 above) this will be a standard association study with the aim of identifying all sites that show association of $r^2>0.9$ with the peak SNP. The aims here are to identify biologically annotated SNPs that may explain the observed GWAS associations and be the subject of future functional experiments. We will also perform conditional association analyses to discover additional independent association signals.

(b) Rare sites: For rare sites, we will test whether they individually or cumulatively show a difference in cases versus controls, in other words, a burden test. The detailed analytical approaches to the association and burden tests will depend on many aspects of the data and will be defined later. Significance thresholds for association testing will be determined by adjustment for the number of tests carried out.

We also wish to identify variants that associate with the increase / decrease of BP over time (BP changes across 4 visits), in addition to the BP traits per each visit.

We wish to obtain also the BAM files of the sequencing data in order to mine the dataset for a) the full mutation spectrum and coverage at all variable positions b) to explore the impact of other calling methods (e.g. SAM tools vs. GATK) on the BP associations with SNPs and CNVs. Details of the transfer of these BAM files will need to be worked out with Drs. Gibbs and Boerwinkle, since they are very large.

For coding and regulatory variants, we will use the Human Gene Mutation Database to review the potential damage that is caused by known mutations.

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

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?

NA

(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?

Yes

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"?

Yes

8.c. If yes, is the author aware that the participants with RES_DNA = 'not for profit' restriction must be excluded if the data are used by a for profit group?

Yes

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

Yes (based on list circulated in July 2008)

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)?
This manuscript proposal is an extension of proposal "CHARGE GWAS for BP (SBP and DBP) at first visit" and has largely similar authors.

11. a. Is this manuscript proposal associated with any ARIC ancillary studies or use any ancillary study data?
Yes

11.b. If yes, is the proposal
___ A. primarily the result of an ancillary study (list number* _________)
_x__ B. primarily based on ARIC data with ancillary data playing a minor role (usually control variables; list number(s)* 2006.03
*ancillary studies are listed by number at http://www.cscc.unc.edu/aric/forms/

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.