ARIC Manuscript Proposal #2096

1. Full Title: Genetic analyses of obstructive sleep apnea and related traits in The Atherosclerosis Risk in Communities Study (ARIC)

b. Abbreviated Title (Length 26 characters): ARIC Sleep Apnea GWAS

2. Writing Group:
Writing group members:
Eric Boerwinkle (UT), Xiaofeng Zhu (Case Western Reserve University); Brian Cade, Daniel Gottlieb, Sanjay Patel, Susan Redline, Shamil Sunyaev (Brigham and Women's Hospital); Xihong Lin (Harvard School of Public Health); Naresh Punjabi (Johns Hopkins University) and other nominated ARIC investigators. This proposal is part of a larger meta-analysis with sub-analyses and/or methods development performed by all of our proposed coauthors. We will also invite interested investigators from all meta-analysis cohorts to participate in regular analysis teleconferences and ultimately as coauthors.

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

First author: Brian Cade
Address: 221 Longwood Avenue BLI 228, Boston MA 02115
Phone: 857-307-0353  Fax: 617-278-6946
E-mail: bcade@rics.bwh.harvard.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: Naresh Punjabi
Address: Johns Hopkins Bloomberg School of Public Health
          Department of Epidemiology
          615 North Wolfe Street, W6508
          Baltimore, MD 21205
Phone: 410-550-5405  Fax: 410-550-2612
E-mail: npunjabi@jhmi.edu
3. **Timeline**: To meet the goals of the “Life After Linkage” NHLBI initiative, we hope to begin the ARIC-cohort specific analysis as soon as possible. We anticipate ARIC-specific analyses to be complete within three months of data receipt, and meta-analyses including ARIC to be complete within the next 18 months.

4. **Rationale**: Obstructive Sleep Apnea (OSA) is a common disease characterized by recurrent episodes of pharyngeal obstruction during sleep, often associated with profound hypoxemia, sleep fragmentation, and sympathetic nervous system activation, exposing the affected individual to severe nightly cardiovascular stresses. OSA affects more than 10% of the population. Prospective studies have established that OSA increases the incidence of stroke; heart failure and coronary artery disease; as well as mortality. Causal associations of OSA with cardiometabolic disease are supported by studies showing improved blood pressure, endothelial function, and insulin resistance after OSA treatment [1-3]. OSA may therefore contribute to cardiovascular health disparities among minority populations, thus supporting the need to identify the genetic bases for this common disorder.

Through the Cleveland Family Study (CFS), a genetic epidemiological study of rigorously phenotyped families ascertained through probands with OSA, we have established that OSA has a strong genetic basis ($h^2 \sim 0.30$ [4]). Variance component modeling of OSA and BMI indicates that ~35% of the genetic variance in OSA is shared with BMI, suggesting that a majority portion of the genetic basis of OSA is obesity-independent [5]. Our family studies also have identified promising areas of linkage to inform genetic association analysis and sequencing efforts. Through NHLBI’s RFA “Life After Linkage” we are funded to conduct a series of genetic association analyses using the CFS data, combined with sleep data from major NHLBI cohort studies that had sleep phenotyping performed by our group: Sleep Heart Health Study cohorts within ARIC, CHS and FHS; MESA; MrOS-Sleep; SOF-Sleep; Starr County Health Study and Hispanic Community Health Study (HCHS). Thus, this project will provide an opportunity to combine several sources of standardized sleep and genotype data to discover and replicate genetic variants for OSA in multiple populations. **The data requested of ARIC are the sleep data collected as part of the Sleep Heart Health Study (all relevant phenotype data reside at the BWH Sleep ReadingCenter); genotype data available via CARe (IBC chip data already analyzed at BWH) plus available Affy and exome chip data available via ARIC.)** The primary outcome, the apnea hypopnea index (AHI) was collected and analyzed under the direction of the senior author (S Redline).

We propose to use complementary strategies designed to identify common as well as low frequency and rare variants associated with OSA. We will: 1) perform a GWAS (analyzing each ethnic group within individual cohorts) of common variants, using individual SNP analysis weighted by linkage analysis LOD scores derived from our CFS study; and 2) examine rare-variant exon data using SKAT and SKAT-O [6], an advanced gene-based statistic developed by our group, using exomic and/or exome chip data to detect regions of
functional association, and weight the evidence by LOD scores and Polyphen scores from
gene-based and pathway analyses mindful of the effects of private mutations, inter-
population allele frequency differences, and negative selection effects on individual
burdensome alleles. This will be operationalized in two components. We will first perform
ARIC-only analyses of common variants based on the Affymetrix 6.0 and IBC chips. A
second phase will incorporate ARIC data within a larger meta-analysis by imputing ARIC
genotypes based on 1000 Genomes Project data. Imputed genotypes will be additionally
weighted by PolyPhen2 scores (developed by our group to examine non-synonymous
mutation severity) and overlapping ENCODE cell line evidence (e.g. transcription factor
binding sites) to incorporate available biological information. As available, we will
incorporate exome chip data.

5. **Main Hypothesis/Study Questions:**
Covariates that will be adjusted for will include age, age^2, sex, age x sex, and 10
population principal components. A second model will further adjust for BMI, BMI^2, and
sex x BMI to explore whether associations are through obesity dependent vs obesity
independent pathways.

   - **Primary Hypothesis:** Novel genes are associated with obstructive sleep apnea
   (OSA) as measured by the primary phenotype (log of
   apnea-hypopnea index [AHI]) through BMI- dependent and
   independent pathways (using models including and
   excluding BMI and BMI^2 as covariates).

   - **Secondary Hypotheses:**
     1. Additional genetic associations exist with the secondary
     OSA measures of oxygen saturation percentage; AHI
     within REM and non-REM sleep; self-reported snoring
     plus sleepiness; arousal index; and sleep efficiency.

     2. OSA genetic associations are consistent with the primary analysis when further
     controlling for additional sleep phenotypes such as sleep
     duration and when stratifying by obesity (BMI>30).

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**

**Subjects:** All ARIC cohort members who underwent overnight polysomnography in the
Sleep Heart Health Study I ancillary study with available genotypes through the prior
Affymetrix 6.0 and IBC assays (n ~ 1,673).
Genotyping: All Affymetrix and IBC chip data, plus 1000 Genomes Project-based imputed genotypes. We will return copies of the imputed genotypes to ARIC. When available, we would like to incorporate the exome chip data.

Main Outcome Variables: Age and sex adjusted AHI (RDI3p) with and without BMI adjustment.

Secondary Outcome Variables: NREM RDI3p and average SaO2; REM RDI3p and average SaO2, arousal index; sleep efficiency, habitual snoring and excessive sleepiness (questionnaire report of frequency of excessive daytime sleepiness and/or Epworth Sleepiness Score); all with and without BMI adjustment.

Covariates: Age, age², sex, age x sex, 10 population principal components, BMI, BMI² and sex x BMI. Meta-analysis will be performed independently on each population group, then on a combined sample to interrogate the transferability of findings.

Methods/Analysis Plan

Genetic imputation will be based on 1000 Genomes Project Phase 1 Release 3 data (or a later version if available) and follow GIANT guidelines (http://genome.sph.umich.edu/wiki/Minimac:_1000_Genomes_Imputation_Cookbook). Imputed SNPs not meeting a minimal allele frequency-based rsq threshold [7] will be omitted. For common alleles, we will perform individual SNP analysis as well as gene- and pathway-level analysis. For individual SNP analysis, individual cohort results will be calculated using PLINK or GWAF (Framingham R routine) for family data. Both programs accommodate imputed dosage files and covariates. For gene- and pathway-level analysis, we will use the SKAT package.

For rare variants, gene-level SKAT [13] and SKAT-O [6] analysis will examine SNPs within 100 bp of Ensembl-defined exons. 100 bp of untranslated region sequence flank will also be included, along with DNase 1 hypersensitivity sites within 1000 bp of Ensembl-defined exon coordinates (if available). Analysis will be performed by restricting all variants to MAF<5% (apart from Ensembl-defined missense, nonsense, or splice status; or location within a hypersensitivity site) or using all variants by up-weighting rare variants. SKAT-O will be used for the analysis, which adaptively best chooses between burden and SKAT to maximize the power when the effects of variants in a gene are in the same direction or different directions. Meta-analysis of SKAT using techniques developed by our group will be performed by combining results across studies. As SKAT identifies regions of association, this meta-analysis can combine results across populations and cohorts which likely contain an excess of singleton and doubleton deleterious SNP results not observed in other cohorts. The meta-analysis can also combine results across different ethnicities, accounting for heterogeneous genetic effects across cohorts.
Weighting of p value: CFS LOD score of linkage will be calculated using SOLAR based on LD-thinned SNPs, with tests for heterogeneity. LOD scores for individual SNPs will be extrapolated based on 1000 Genomes Project genetic distance maps. Polyphen data will be downloaded from Ensembl Biomart. RegulomeDB will be used to integrate ENCODE regulatory data [8]. A weighted false discovery rate approach will be used to prioritize biologically plausible SNPs [9]. Weights will be equally distributed between 1) CFS LOD score; and 2) Polyphen and ENCODE evidence. For common variants, meta-analysis will examine all post-weighted p-values using METAL with genomic control such that biologically plausible SNPs are up-weighted and SNPs without RegulomeDB, Polyphen, or linkage evidence are down-weighted. SKAT significance will be assessed following Bonferroni correction of the number of Ensembl-defined protein-coding genes tested.

**Discussion of Power**

**Common-variant based tests:** We estimated the statistical power of the gene-based association test for GWA, performing power calculations using SKAT (Table 1). The haplotype information was constructed randomly by the calibrated coalescent model and assumed a gene size of 30 kb and 50 kb. Power was estimated by calculating average power of a large number of simulated genes for a continuous outcome (e.g., AHI). To mimic GWA for common variant effects, only SNPs with MAF > 0.05 were used for the gene based test. 5% of the common SNPs were selected as causal variants. All causal variants were assumed to have the same small absolute effect sizes, with 80% of βs being positive and 20% being negative (+/-=80/20). The power was estimated for 8,000 subjects with α= $1.25 \times 10^{-5}$ which corresponds to selecting the top 5 genes among 20,000 genes at the overall type I error rate 0.05 using the Bonferroni correction. The power was estimated accounting for the family design in CFS. We have at least 80% power to detect genes explaining phenotype variance $r^2$ as small as 1.6 % (both 30 kb and 50 kb genes).

Table 1. Estimated power for detecting a disease-associated gene in the top 5 genes for common variants effects using n=8000 subjects accounting for multiple comparisons (α= $1.25 \times 10^{-5}$) assuming 5% of common variants in the gene are causal, where β is the effect size of each causal variant and $r^2$ represents proportion of phenotypic variance due to genetic effects.

<table>
<thead>
<tr>
<th>Gene size</th>
<th>$\beta = 0.11$</th>
<th>$\beta = 0.13$</th>
<th>$\beta = 0.15$</th>
<th>$\beta = 0.17$</th>
<th>$\beta = 0.20$</th>
</tr>
</thead>
<tbody>
<tr>
<td>30kb</td>
<td>0.75 ($r^2=0.013$)</td>
<td>0.85 ($r^2=0.018$)</td>
<td>0.91 ($r^2=0.024$)</td>
<td>0.94 ($r^2=0.033$)</td>
<td>0.96 ($r^2=0.041$)</td>
</tr>
<tr>
<td>50kb</td>
<td>0.86 ($r^2=0.019$)</td>
<td>0.92 ($r^2=0.027$)</td>
<td>0.95 ($r^2=0.035$)</td>
<td>0.96 ($r^2=0.045$)</td>
<td>0.97 ($r^2=0.060$)</td>
</tr>
</tbody>
</table>
Our study sample could increase to 25,000, based on inclusion of CHS, CFS, FHS, MESA, MrOS, SOF, Starr County and Hispanic Community Health Study. We anticipate adding additional cohorts, including Jackson Heart Study, which will have sleep apnea measures available in several years and may serve as a replication cohort. We will also perform multiple phenotype analysis using the SMAT method that was developed by our group to study pleiotropy and increase analysis power [11].

Summary

OSA is a very common disorder, with burdens contributing to multiple diseases of interest to the ARIC study. The genetic mechanisms contributing to OSA are largely unknown due to a lack of power in previous studies. This project will identify genes that increase susceptibility to OSA in a multi-ethnic sample, thus potentially revolutionizing the scientific understanding of the molecular pathways leading to it and its co-morbidities, such as heart disease and diabetes.

Reference List


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 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.csc.unc.edu/ARIC/search.php

__X__ Yes    _______ No (This meta-analysis will include many more individuals and novel methodologies relative to past CARe-based sleep apnea proposals).

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)?

We are unaware of any sleep apnea genetics proposals incorporating imputation, bioinformatics, and/or gene-based tests. The meta-analysis will involve additional cohorts relative to past CARe analyses, which have been published on (Patel SR et al, 2012) and involved several of the current proposal authors.

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* 1995.12__)

____ 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.csc.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.csc.unc.edu/aric/index.php, under
Publications, Policies & Forms. [http://publicaccess.nih.gov/submit_process_journals.htm](http://publicaccess.nih.gov/submit_process_journals.htm) shows you which journals automatically upload articles to Pubmed central.