ARIC Manuscript Proposal # 1360

PC Reviewed: 4/8/08  Status: A  Priority: 2
SC Reviewed: _________  Status: _____  Priority: ____

1a. Full Title: Genome-Wide Association Study of Chronic Obstructive Pulmonary Disease (COPD) Phenotypes and Lung Function Parameters: The Atherosclerosis Risk in Communities Study

1b. Abbreviated Title (Length 26 characters): GWAS, COPD, and Lung Function

2. Writing Group: Matthew B. Schabath, Rubina Inamdar, Stephanie London, Kari North, Jack Follis, Eric Boerwinkle (Plus other interested ARIC investigators)

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

First author: Matthew B. Schabath
Address: Division of Epidemiology
UTHSCH School of Public Health
1200 Herman Pressler Suite #617
Houston, Texas 77030
Phone: 713-500-9216 Fax: 713-500-6264
E-mail: Matthew.B.Schabath@uth.tmc.edu

Corresponding/senior author (if different from first author correspondence will be sent to both the first author & the corresponding author):
Name: Matthew B. Schabath
Address: Division of Epidemiology
UTHSCH School of Public Health
1200 Herman Pressler Suite #617
Houston, Texas 77030
Phone: 713-500-9216 Fax: 713-500-6264
E-mail: Matthew.B.Schabath@uth.tmc.edu

3. Timeline:
   Statistical Analysis: March 2008 to July 2008
   Manuscript Revision: Dec 2008
   Manuscript Submission: Jan 2009
4. **Rationale:**

Non-malignant respiratory diseases have considerable impact on public health. In the US, chronic obstructive pulmonary disease (COPD) is the fourth leading cause of death and nearly 11% of the population is estimated to have some form of COPD (1). Spirometrically measured lung function (LF) parameters in conjunction with clinically-relevant symptoms are the most important phenotypes of COPD. As pulmonary function deteriorates in susceptible individuals, LF data distinguish patients with minimal pathologic changes from those with normal airways (2-4). Genetic and environmental factors, especially smoking, and their interactions influence COPD phenotypes and LF parameters. Although smoking is the major environmental risk factor for COPD, only a fraction of smokers develop COPD or experience any impact on LF (5,6), indicating that other factors are involved. Familial, genetic, and association studies provide evidence that these factors are genetic. The evidence from these studies suggest: 1) susceptibility factors have a substantial influence on COPD phenotypes and LF parameters and 2) that COPD is a complex genetic disease. Hence, multiple genetic factors are likely involved in the pathophysiology of COPD and LF.

At the molecular and cellular level, COPD is a complex disease that involves a variety of cell types, mediators, and enzymes, but at present their relative importance is still poorly understood and under studied. Risk of COPD is likely a complex interplay between multiple gene and gene pathways and environmental exposures. For a better understanding for the role of genetics on COPD and LF, it is critical to determine how genetic variation influences COPD phenotypes and LF parameters. Presently, the majority of COPD and LF associated studies have focused on investigating genetic variation within genes involved in inflammation, metabolism-oxidation, and proteolysis (i.e. proteases/anti-proteases). Recent advances in genetic polymorphism discovery, population genetics, and genotyping technologies have yielded a genome-wide collection of single nucleotide polymorphisms (SNPs) that span the human genome and predict (or “tag”) other unmeasured SNPs because of linkage disequilibrium (7). Combined with appropriate population-based samples, good phenotyping, details on relevant exposures and a robust analytical plan, it is now possible to use a genome-wide association (GWA) approach to identify genetic factors that influence COPD phenotypes and LF parameters.

The goal of this proposed research is to identify genetic factors (i.e., SNPs and haplotypes) that influence COPD phenotypes and LF parameters using a GWA analysis of ~1 million SNPs in the ARIC cohort. To date, this application is proposing the largest GWA analysis of COPD phenotypes and LF parameters.

5. **Main Hypothesis/Study Questions:**

1. Utilizing the ~1,000,000 SNP GWAS genotype data available for the entire ARIC cohort, apply analytical approaches to assess individuals SNPs and inferred and sliding window haplotypes to investigate the influence of genetics on COPD phenotypes and LF function parameters.

2. Further and detailed analyses will also be performed including stratified analyses to identify highly susceptible subgroups (e.g. smoking, age, gender, ethnicity, BMI, etc) and statistical interactions.
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 Design and Outcomes:**
The GWA analysis will consider two types of outcomes, lung function parameters (i.e., quantitative traits) and COPD phenotypes (i.e., qualitative traits). The proposed analyses include a cross-sectional analysis of the COPD phenotypes, performed at the baseline visit, and age-normalized mean of the lung function parameters from visits 1 and 2 as performed in the Framingham GWAS study. By taking the mean value of the LF parameters from two subsequent and close follow-up visits, we will be able to reduce standard error because we are essentially taking multiple measures within a time range where there will be no or limited longitudinal changes. As reported by Vollmer (8), given the measurement error in PFTs there is too much noise to detect longitudinal changes with two measurements < 4 years apart. Therefore, we will have limited ability, if any at all, to detect longitudinal changes within these two time points. However, we can take advantage and maximize the available data and study design by the approach described above.

The main outcomes for these analyses will be COPD phenotypes and lung function parameters. The primary LF parameters are FEV₁, FVC, FEV₁ % Predicted, FEV1/FVC, and FEF25-75 among the 30 different LF parameters (Table 1) available on all study subjects. However, as performed in the Framingham GWAS, all LF parameters will be analyzed. Analyses will attempt to remove the known sources of variation in LF parameters in order to detect the associations with genetics. These include age, gender, race, and height which are the factors that go into calculation of percent predicted pulmonary function, where appropriate. Smoking history influencing lung function and will be controlled for and stratified analyses will be assessed. Other factors that may influence pulmonary function that will be considered to be controlled for include adiposity, BMI, and environmental tobacco smoke exposure.

We also propose case-comparison analyses of the COPD phenotypes at the baseline visit. The proposed COPD phenotypes are described in Table 2. Although the standard definition of COPD uses post-bronchodilator PFTs, in large epidemiologic studies, pre-bronchodilator LF parameters are acceptable COPD outcome and COPD defined in the manner has been associated with mortality in the ARIC cohort (Mannino et al.). Because COPD is a heterogeneous disease consisting largely of two different and distinct pathologies, we will also examine emphysema (EMPH) and chronic bronchitis (CBSR) (Table 2). To eliminate the possibility of reversible concomitant disease (i.e., asthma), we will exclude individuals with a self-reported history of asthma. COPD will be defined using PFT data and modified GOLD criteria i.e. FEV1/FVC <0.70 & FEV1 < 80% predicted. We will utilize clinically relevant physician diagnosed self-report data to discern between emphysema and chronic bronchitis. Chronic bronchitis by physician diagnosed self-report (i.e., CBSR) will be defined as individuals who self-report a physician diagnosis of chronic bronchitis or who self-reported a history of chronic cough or phlegm production for 3 or more months for 2 or more years. The CBSR phenotype will be particularly useful since these criteria are quite similar to the clinical criteria used to diagnosis chronic bronchitis. Emphysema (i.e., EMPH) will be defined as individuals who self-report a physician diagnosis of emphysema. We found that 77% of individuals who had a physician diagnosed history of emphysema had a FEV₁/FVC of < 0.70 which suggests that physician diagnosed emphysema has a relatively high concordance with the LF criteria. And it is established that individuals can have emphysema without obstruction. So, the individuals who self-report a physician diagnosis of emphysema should be a reliable measure of this phenotype. Normal (i.e., NORM) will be defined as individuals who do not meet any of the phenotype criteria and have a FEV₁/FVC ≥ 0.70 and FEV₁ ≥ 80% predicted.

Although not a primary endpoint of these analyses, we will have the ability to explore additional chronic bronchitis phenotypes. Not all individuals who have chronic bronchitis will have a FEV₁/FVC < 0.70, so this
phenotype will not exclude individuals who have chronic bronchitis but a preserved FEV$_1$/FVC. Classification in this manner will also allow us to analyze two subgroups of CBSR, i.e., those with and without a preserved FEV$_1$/FVC. Overall, by using these phenotype criteria to discriminate between emphysema and chronic bronchitis with and without “normal” PFTs, we will have the potential to explore these disease processes separately and determine if specific diseases segregate for a given genetic association.

Table 2. Criteria to Define the COPD and “Normal” Phenotypes in ARIC

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Abbreviation</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Major COPD Phenotypes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>COPD defined by LF</td>
<td>COPD</td>
<td>FEV$_1$/FVC $&lt;$ 0.70 and FEV$_1$ $&lt;$ 80% predicted</td>
</tr>
<tr>
<td>Physician diagnosed emphysema</td>
<td>EMPH</td>
<td>Self-report physician diagnosed emphysema</td>
</tr>
<tr>
<td>Chronic bronchitis CB by self-report</td>
<td>CBSR</td>
<td>Self-report physician diagnosed CB or self-reported history of chronic cough or phlegm production for 3 or more months for 2 or more years</td>
</tr>
<tr>
<td>COPD by any of the above definitions</td>
<td>ANY</td>
<td>An individual is classified in one or more of the above COPD phenotypes</td>
</tr>
<tr>
<td>Chronic bronchitis subtypes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physician diagnosed chronic bronchitis with “non-normal” LF</td>
<td>CBPFT-$-$</td>
<td>CBSR individuals with a FEV$_1$/FVC $&lt;$ 0.70</td>
</tr>
<tr>
<td>Chronic bronchitis by self-report with “normal” PFT</td>
<td>CBPFT+$+$</td>
<td>CBSR individuals with a FEV$_1$/FVC $\geq$ 0.70</td>
</tr>
<tr>
<td>Normal Phenotype</td>
<td>NORM</td>
<td>Not classified above and FEV$_1$/FVC $\geq$ 0.70 and FEV$_1$ $\geq$ 80% predicted</td>
</tr>
</tbody>
</table>

To further explore the genetic influence of COPD and LF, we also plan to examine sliding window- and inferred haplotypes:

**Sliding-Window Haplotype Analysis:** To characterize nonrandom associations between loci (linkage disequilibrium [LD]), multilocus tests will be performed to measure the total pairwise LD between any two loci. Haplotypes represent multivariate information about LD in and around a gene and therefore capture greater genetic variability than single polymorphic sites alone. Thus, SNP haplotypes may be more informative than individual SNPs for detecting causative alleles (10), especially causative alleles of low frequency. Employing a sliding-window haplotype analyses will allow us to “extract” more information from the ~1,000,000 SNP dataset than a simple SNP-by-SNP analysis (11).

In this method, SNPs are arranged by chromosome position from p- to q-terminal. For a 3-SNP haplotype sliding-window analysis, SNPs 1 through 3 from the p-terminal are identified, haplotypes are inferred for each individual in the sample, and the appropriate association analysis performed. Then SNPs 2 through 4 are identified, haplotypes are inferred, and the appropriate association analysis performed. This process is repeated, sliding one SNP toward the q-terminal at each step, until the last SNP is analyzed. The resulting P-values, either raw or adjusted for multiple-testing, may be plotted by the position of the first SNP in each haplotype.

**Haplotype Inference:** The sliding-window haplotype method discussed above requires a very fast and efficient method for inferring haplotype information. Haplotype frequencies and haplotype probabilities for each individual will be inferred from un-phased genotype data using the expectation maximization (EM) algorithm (14). The EM provides maximum-likelihood estimates of haplotype frequencies under the assumption of Hardy-Weinberg Equilibrium. An iteration of the EM algorithm involves two steps: 1) expectation, in which probabilities for proposed multi-locus genotypes are computed based on the probabilities of their corresponding haplotypes, and 2) maximization, in which haplotype probabilities are re-estimated based on the ratios of the computed genotype probabilities compared to the actual ones. The EM is run until frequencies at expectation and maximization converge. The accuracy of EM-based estimation is good, even when some alleles are not in HWE, in moderate to large sample sizes (13).

**Inclusion/exclusion:** The usual DNA and non-CVD restrictions, ethnic group and missing data exclusion criteria will be used.

**Other variables of interest:** Smoking, SES, occupation, physical activity, diet, second hand smoke, prior medical history (e.g., history of heart disease, stroke, etc), BMI and/or weight, and covariate data will be required.
Summary of data analysis:

**Overview:** If required we will perform any necessary data merging, validation, and query resolution including excess missing data, low allele frequency (< 0.05), excessive homozygosity, departures from Hardy-Weinberg Equilibrium, etc. Multivariable linear regression models will be used to assess association between each genetic variant and the lung function parameters, while multivariate logistic regression will be used to assess the association diary and the COPD phenotypes. The P-value from the linear regression model and the Wald Statistic will be used to determine significance, respectively. Variables that will be adjusted for, but not limited to, include: age, gender, age, SES, (i.e. education/income), ethnicity, smoking status (never, former, current), and pack-years smoked. Other variables that will be considered include western diet (factor 1), prudent diet (factor 2), total calories/day, physical activity, occupation, second hand smoke exposure, pre-existing medical conditions (e.g. history of CHD), and any other statistically significant or biologically/clinically relevant variables. For each LF parameter and COPD phenotype, the P-values, following correction for multiple comparisons (see below for description), will be ranked from most significant to least significant (i.e. 0 to 1). As in a manner similar to the Framingham GWAS analyses of LF and COPD, the significant associations will be reported and published. Unless noted otherwise, all statistical analysis will be performed using the Intercooled Stata 10.0 statistical software package (Stata Corp., College Station, TX).

**LF parameters (i.e. quantitative traits):** The association between LF parameters and the genetic factors will be analyzed via multivariable linear regression models. As mentioned above, the following approaches will be applied: 1) a cross-sectional analysis of LF parameters at the baseline visit, and 2) the age-normalized mean of two measurements taken at visits 1 and 2 (these are the only two visits where pulmonary function testing was performed). By taking the mean value of the LF parameters from two subsequent and close follow-up visits, we should be able to reduce standard error because we are essentially taking multiple measures within a time range where there will no or limited longitudinal changes. This approach has been employed previously by the FHS GWA results for pulmonary function. We will also examine whether the findings for visit one are replicated in visit two.

**COPD Phenotypes (i.e. qualitative traits):** Multivariable logistic regression analyses will largely be used to assess the association between SNPs/haplotypes and the COPD phenotypes. Any given COPD phenotype will be coded as “1” and the NORM phenotype will be coded as “0”. Genotypes will be coded, by a *priori* information, and analyzed using both the co-dominant model (i.e., wildtype = 0, heterozygote = 1 and variant = 2) and the mutation-dominant model (i.e., wildtype = 0, heterozygote & variant = 1). Additionally, sex-, age-, and smoking-specific stratified analyses will be assessed for interaction and highly susceptible subgroups (i.e. Aim 2). Multivariable analysis will be conducted to assess if each genotype variable is independently associated with risk of the COPD phenotypes adjusting for potential confounders (as noted above). Because COPD is a multi-causal disease whereby many factors may contribute to and modify risk, careful consideration will be followed to assess the impact of existing medical conditions, such as obesity, cardiac disease, stroke, etc. Using stratified analyses we will determine if these factors are potential effect modifiers. If any factor is considered to be an effect modifier, we will decide, based on sample size considerations whether to exclude individuals with the specific factor or report strata-specific effects. Additionally, formal analytical methods will be applied to determine if any of these factors are a confounder. If any factor is deemed a confounder, we will control for the confounding the multivariable analysis.

**Adjustment for Multiple Comparisons:** The relative ease of genotyping a very large number of SNPs/Haplotypes to test for associations with multiple endpoints is offset by the difficulty in taking into account the number of statistical tests that will be conducted. To avoid the possibility of avoid false positives, the statistical significance of individual results in these analyses must be evaluated in the context of all of the hypothesis tests carried out. With a fixed sample size, there is a trade-off between power of the study and conservative-nature of the results deemed “statistically significant”. Specifically, there is a trade-off between type II and type I errors. So, we plan to employ the False Discovery Rate (FDR) to account for multiple comparisons. The FDR provides an analytical step to account for multiple comparisons (14). FDR controls the expected proportion of incorrectly rejected null hypotheses (i.e., type I errors) in a list of rejected hypotheses.
and is a less conservative comparison procedure with greater power than the Family-wise error rate. The FDR is the expected proportion of false discoveries among the actual discoveries, and controlling the FDR increases error from multiplicity while losing less in the ability to discover real differences.

**Hidden Population Substructure:** One potential confounding problem with large-scale genetic association studies is hidden population substructure and genetic homogeneity of the study sample. Previously, we have tested the entire ARIC cohort for evidence of population substructure. In an ethnic-specific manner, we utilized the program STRUCTURE to assess evidence of population substructure using targeted genetic polymorphisms and subsets of the ARIC participants for a larger number of SNPs (100’s). **There was no evidence of hidden population substructure in any analysis or subset of the cohort.** The analysis to compute the posterior probability of K (the number of populations) was run 2,000 times.

It is possible that these negative results are due to the relatively small number of markers and/or the nature of the markers. Therefore, we will continue to analyze the ARIC cohort for evidence of hidden population substructure. For these analyses, we will use appropriate subsamples (No.~8,000) of SNPs that are known to show allele frequency differences among populations. To test for statistical significance, we will compare the results from the selected SNPs to the result obtained from repeated random sampling (No. = 1,000) of the same number of SNPs. If population substructure is found, appropriate adjustment will be made using latent variables and a structured association method (15).

**Any anticipated methodological limitations or challenges:** At this point we do not anticipate any substantial methodological challenges. However, it should be noted that the cross-sectional design of this study is a limitation since we cannot assess temporal changes of lung function or incident COPD. Given the measurement error in lung function testing, there is too much noise to detect longitudinal changes with two measurements < 4 years apart. Therefore, we will have limited ability, if any at all, to detect longitudinal changes if we also applied visit 2 to these analyses. However, we can take advantage and maximize the available data and study design by the approach described above.

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 ICTDER02 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 ICTDER02 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 ICTDER02 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)? None at this time.

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 A. primarily the result of an ancillary study (list number* ___________)

B. primarily based on ARIC data with ancillary data playing a minor role (usually control variables; list number(s))* GxE ancillary study (#1995.07)

*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.

13. Literature Cited


