1.a. Full Title: Peroxisome Proliferator-Activated Receptor Polymorphisms Associated with Lung Function and COPD: Atherosclerosis Risk in Communities (ARIC) cohort study.

b. Abbreviated Title (Length 26 characters): PPAR genotypes and Lung Function

2. Writing Group: Matthew B. Schabath, Kelly Volcik, Jack Follis, Eric Boerwinkle, and any other ARIC member(s) as deemed by the coordinating center.

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

Address: 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:  
Manuscript Prep: May 2008 to July 2008  
Manuscript Revisions: Aug 2008  
Manuscript Submission: Sept 2008
4. **Rationale:**

Chronic Obstructive Pulmonary Disease (COPD) is a treatable and largely preventable disease state characterized by airflow limitation, but not fully reversible. Although smoking is a major risk factor for the development of COPD, only a relatively small proportion (~15%) of smokers will develop clinically relevant COPD. Familial, genetic, and association studies have shown that genetic factors contribute to susceptibility to lung function impairment and risk for COPD (1). Because COPD has a considerable impact on public health, affecting nearly 11% of the US population, it is imperative to identify those genetic factors that contribute to disease risk.

Peroxisome proliferator-activated receptors (PPARs) are ligand-activated transcriptional factors that are members of the nuclear hormone receptor superfamily, of which three major types have been identified: PPAR-α, PPAR-γ and PPAR-β/δ (2-4). Originally identified for their role in lipid homeostasis, all three have similar structure and function with PPAR-α and PPAR-γ playing roles in the intensity, duration and consequences of inflammatory events (3,5-8). PPAR-α and PPAR-γ have been found expressed in monocytes/macrophages, eosinophils and neutrophils (9-11) and have been found on human airway smooth muscle cells and peripheral human lung tissue (6,7). The proportion of PPAR-α positive cells in the alveolar wall and alveolar macrophages have been found to be higher in COPD patients than non-COPD patients (7). Human airway epithelial cells express PPAR-γ, which blocked cytokine induced iNOS expression and down-regulated IL-8 secretion, two possible contributors to the pathogenesis of COPD (12,13). PPAR-α and PPAR-γ have also been found to inhibit expression of MMP-9, a possible contributor to COPD (14-17). Murine studies demonstrate that PPAR-α activation may be beneficial in airway inflammatory diseases involving neutrophil and monocyte recruitment (18), that PPAR-γ regulates lung inflammation (19) and epithelial cell PPAR-γ contributes to normal lung maturation (4). Even though there is evidence for a biological role of PPARs in the pathogenesis of COPD, at present there have been no published association studies for the role of PPAR genotypes and risk of impaired lung function and COPD.

The SNPs chosen for the proposed analysis have been genotyped on the entire ARIC cohort (Table 1).

<table>
<thead>
<tr>
<th>Gene</th>
<th>Variant Site</th>
<th>dsSNP ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPAR-γ</td>
<td>Pro12Ala</td>
<td>rs1801282</td>
</tr>
<tr>
<td></td>
<td>C &gt; G</td>
<td>rs1805192</td>
</tr>
<tr>
<td></td>
<td>C161T</td>
<td>rs3856806</td>
</tr>
<tr>
<td>PPAR-α</td>
<td></td>
<td>rs3892755</td>
</tr>
<tr>
<td></td>
<td>A401G</td>
<td>rs6008259</td>
</tr>
<tr>
<td></td>
<td>A101G</td>
<td>rs9615784</td>
</tr>
<tr>
<td></td>
<td>T1019C</td>
<td>rs1042311</td>
</tr>
<tr>
<td></td>
<td>A1454T</td>
<td>rs9615759</td>
</tr>
</tbody>
</table>
5. **Main Hypothesis/Study Questions:**

1. Determine the relationship between 8 single nucleotide polymorphisms (SNP) within the candidate genes PPAR-α and -γ. The association between genotypes and lung function (LF) parameters will be assessed as will the association between genotypes and COPD risk. The primary measures of lung function include: Percent predicted forced expiratory volume at 1 sec (FEV1), forced vital capacity (FVC) and the FEV1 to FVC ratio. **We hypothesize that individuals with the risk genotypes/alleles will be associated with statistically significant lower lung function values and increased risk for COPD.** Cross-sectional analyses at baseline will be performed. All analyses will be adjusted for, where appropriate, relevant confounding factors such as age, gender, field center, smoking status, pack-years smoked, and prior history of lung disease. Additionally, gender-specific and smoke status-specific analyses will be performed to identify possible susceptible subgroups.

2. By ethnic-specific groups and using standard and novel analytical approaches, evaluate gene-gene (GxG) interactions of the SNPs listed in Specific Aim 1 on the influence of lung function and COPD. In addition to traditional modeling approaches, Bayesian Networks (BN) and Recursive Partitioning (RP) will be used to explore complex interactions and to explore the data for patterns and clusters of genotypes and haplotypes, epidemiologic covariates, and LF parameters and COPD. **We hypothesize that traditional model building and novel statistical techniques, such as BN and RP, will allow for the identification of SNP combinations and haplotypes that influence LF and/or risk of COPD.**

3. By ethnic-specific groups, evaluate the extent of gene-environment (GxE) interaction between the genes listed in Specific Aim 1 and smoking (status, intensity, cessation) and occupational exposures (based on job coding) as they combine to influence LF and COPD. We will test two-way interactions and use novel analytical approaches as proposed in Specific Aim 2. **We hypothesize that gene-environmental statistical interactions will allow for the identification of highly susceptible subgroups and improve the understanding of the multiple factors that influence LF and risk of COPD.** Smoking specific analyses also will be performed to examine the effects of never, former, and current smokers. Moreover, the effects of amount smoked will also be explored by stratifying the data on pack-years smoked.

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:** The proposed analysis will be a cross-sectional analysis utilizing the clinical, epidemiologic, and lung function data from visit 1 (i.e. baseline).

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

**Outcome:** The main outcome for this analysis will be lung function and COPD as categorized by lung function parameters. The main lung function (i.e. pulmonary function test) measurements that will be used for this analysis will be the Forced Expiratory Volume at 1 s (FEV1), which is the volume of gas exhaled in the first second of expiration, FVC (i.e., the total volume of gas exhaled), and the ratio of FEV1:FVC. The lung function variables will be analyzed
as a categorical predictor variable by using a modified Global Initiative on Obstructive Lung Disease (GOLD) classification originally described by Mannino et al. (Am J Resp Crit Care Med, 2006). The modified GOLD classification uses clinically relevant cutpoints that will allow us to capture at-risk individuals ranging from modest or subclinical disease to advanced lung function impairment. Only individuals with the appropriate acceptability code regarding the pulmonary function test will be included in the analysis.

**Other variables of interest:** Genotypes, environmental exposures (smoking, SES, occupation, physical activity, diet, etc), prior medical history (e.g., prior respiratory disease), and covariate data will be required.

**Summary of data analysis:** Where appropriate, parametric and non-parametric analyses will be used to describe and characterize the study population, the genotypes, covariates, and to assess the association between the lung function data and the genotypes. COPD will be defined using a modified GOLD classification. For example, using the GOLD criteria a “case” can be defined as an individual with the following lung function criteria: FEV1/FVC <0.70 and FEV1 < 80% predicted. A control can be defined as an individual who does not fall into any of the GOLD criteria and have lung function criteria of: FEV1/FVC ≥ 0.70 and FEV1 ≥ 80% predicted. Thus, we can estimate risk via odds ratios (ORs) and 95% confidence intervals (CI) using unconditional logistic regression analytical methods. Multivariate logistic (or ordinal) regression analyses can also be explored if we decide to use more than one category to define COPD. Regression models will be constructed to include statistically significant and biologically relevant variables. From previous experience with the ARIC cohort, genotype distributions are markedly different between ethnic groups, so all analyses will be stratified by ethnicity. Additionally, sex-, age-, and smoking specific stratified analyses will be explored. The association between lung function and the genotype(s) will be analyzed by main effects, joint effects, haplotypes, and gene-gene (GxG) and gene-environmental (GxE) interactions. The false discovery rate (FDR) will be used to help account for multiple comparisons. 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. Interaction and trend tests will be performed where applicable.

Exploratory “data mining” approaches will also be employed to detect higher order GxE and GxG interactions. It is likely that a single genotype will only have a modest effect, if any at all, on risk of impaired lung function, yet multiple genes in relevant pathways may reveal a more accurate represent of risk and reveal susceptible subgroups. So, we will evaluate higher-order effects using novel statistical techniques such as:

i) **Bayesian Networks (BN):** a graphical representation of a joint probability distribution, representing dependence and conditional independence relationships. The primary goal of BN in the proposed study is descriptive modeling to determine interactions and patterns in the data by employing reverse-engineering and visualization of the biological relationships between SNPs and other factors (environmental variables, covariates, phenotypes, etc). Therefore, BN modeling is a prototypical “systems biology” method. Applying BN modeling in molecular epidemiologic research has only recently become feasible due to the computational complexity of the model selection process. BN modeling has been deployed successfully in the research of medical diagnoses, gene expression data, and genetic network reconstruction.
ii) Recursive Partitioning (RP) Classification Methods: RP methods are capable of addressing non-additive variable interactions because many possible variable combinations are encountered repeatedly during the classifier's construction. Similarly, RP methods can account for genetic heterogeneity because, in addition to classification, these methods effectively perform “clustering” of the observations into smaller subsets with shared input variable profiles. There are three types of RP methods: 1) single decision trees (e.g., CART); 2) randomized decision tree ensembles (e.g., Random Forests); and 3) adaptive decision tree ensembles. Single-tree methods are moderately expressive and easy to interpret, but they are not robust. Randomized ensembles are exceedingly robust, expressive and highly scalable, but the results are difficult to interpret. Therefore, in this study, we will focus on the adaptive decision tree ensembles, specifically “boosted” classifiers.

Any anticipated methodological limitations or challenges: At this point we do not anticipate any substantial methodological imitation or challenges.

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 website at: http://www.csc.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))**

*ancillary studies are listed by number at [http://www.cscc.unc.edu/aric/forms/](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


