1.a. Full Title: Matrix Metalloproteinase Polymorphisms Associated with Impaired Lung Function and COPD

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

2. Writing Group: Matthew B. Schabath, Kelly Volcik, Rubina Inamdar, Eric Boerwinkle.

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

Manuscript Revision: Dec 2007
Manuscript Submission: Jan 2008
4. Rationale:

Respiratory diseases have considerable impact on public health. In the US, non-malignant pulmonary disease is the fourth leading cause of death and nearly 11% of the population is estimated to have some form of chronic obstructive pulmonary disease (COPD). Lung function measures are the most important phenotypes of COPD. Although smoking is the major environmental risk factor for lung function impairment and COPD, only a fraction of smokers experience accelerated lung function impairment and actually develops COPD (1,2), indicating that other factors are involved. Familial, genetic, and association studies provide evidence that these factors are genetic. The genes encoding the family of matrix metalloproteinases (MMPs) are plausible candidate genes because of their protease-antiprotease activities associated with the pathogenesis of COPD (3-9).

MMPs are zinc-dependent endopeptidases, known for their ability to cleave one or several constituents of the extracellular matrix (ECM). They contain a large family of proteases that share common structural and functional elements and are produced by different genes (10). There is substantial evidence for the role of MMPs in lung function and COPD. Neutrophils have been implicated in causing tissue damage in COPD through the release of a number of mediators, including proteases such as neutrophil elastases and MMPs. A primary role for macrophages is also proposed because of their capacity to produce several metalloproteinases, including MMPs such as MMP-1, -9 and -12 (11). Expression and production of MMP-1 and MMP-9 mRNA is enhanced in macrophages from patients with COPD (12). Inhaled cigarette smoke may induce alveolar macrophages to produce MMP-12, which subsequently induces chemotactic fragments that attract monocytes to the lung parenchyma (13). There is an increase in BAL concentration and macrophage expression of MMP-1 and MMP-9 in patients with emphysema (14,15). There is an increase in MMP-9 activity in the lung parenchyma of patients with emphysema (16), and this is correlated with FEV$_1$ (17). Alveolar macrophages from normal smokers express more MMP-9 than those from normal subjects (18), and there is an ever greater increase in cells in patients with COPD (19), which greatly enhances elastolytic activity (20). MMPs also account for most of the elastolytic activity released by alveolar macrophages of COPD patients over prolonged periods of time (20). MMP-9 and the MMP-9/TIMP-1 ratio are increased in induced sputum of patients with COPD (21). At present there is limited but supportive evidence from association studies for the role of MMP genotypes and risk of impaired lung function and COPD (22-25).

The SNPs chosen for the proposed analysis have been genotyped on the entire ARIC Cohort (Table 1). The SNPs were selected based on known functional effects or prior association study results. Although previous studies have explored the relationship between MMP genotypes and lung function, this will be the largest analysis to explore the MMP family genes for risk of impaired lung function and lung function.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Variant Site</th>
<th>dsSNP ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMP-2</td>
<td>G1575A</td>
<td>rs243866</td>
</tr>
<tr>
<td></td>
<td>C1306T</td>
<td>rs243865</td>
</tr>
<tr>
<td></td>
<td>T790G</td>
<td>rs243864</td>
</tr>
<tr>
<td>MMP-3</td>
<td>5A/6A</td>
<td>rs35068180</td>
</tr>
<tr>
<td></td>
<td>Lys45Glu</td>
<td>rs679620</td>
</tr>
<tr>
<td>MMP-7</td>
<td>A181G</td>
<td>rs11568818</td>
</tr>
<tr>
<td></td>
<td>C153T</td>
<td>rs11568819</td>
</tr>
<tr>
<td>MMP-9</td>
<td>Arg279Gln</td>
<td>rs17576</td>
</tr>
<tr>
<td>MMP-12</td>
<td>A82G</td>
<td>rs2276109</td>
</tr>
<tr>
<td>MMP-13</td>
<td>A77G</td>
<td>rs2252070</td>
</tr>
</tbody>
</table>
5. Main Hypothesis/Study Questions:

1. Determine the relationship between genetic variation of 10 SNPs in 6 biologic candidate genes and multiple measures of lung function in the ARIC cohort study. The candidate genes include: MMP-2, -3, -7, -9, -12, -13. The primary measures of lung function include: Percent predicted forced expiratory volume at 1 sec (FEV1), forced vital capacity (FVC) and the ratio of FEV1 to FVC. **We hypothesize that individuals with the risk genotype/allele of the MMP genes will be associated with impaired lung function.** Baseline cross-sectional and exploratory longitudinal analyses will be performed.

2. By ethnic-specific groups, evaluate the extent of gene-gene interactions, of the SNPs from the 6 MMP genes, on the influence of lung function. Exploratory “data mining” approaches will also be employed including Bayesian Networks (BN), Multifactorial Dimensionality Reduction (MDR), and Classification and Regression Tree (CART) analyses. **We hypothesize that model building and statistical techniques, such as BN, CART and MDR, will allow identification of SNP combinations and haplotypes associated with impaired lung function and clinical outcomes.**

3. By ethnic-specific groups, evaluate the extent of gene-environment interaction between the MMP genes and smoking (status, intensity, ETS, and cessation) as they combine to influence the multiple measures of lung function. **We hypothesize that gene-environmental statistical interaction will allow for the identification of highly susceptible subgroups and improve the understanding of lung function.** Certainly, smoking is the most important environmental factor that will be explored; other exposures that may be considered include diet, occupation, and physical activity. Again, all analyses will 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. The smoking specific analyses will 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; however prevalent COPD will also be explored. 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 either dichotomizing the measurement (e.g., by percentile or a clinically important cutpoint [i.e. 80%]) or by generating several levels (e.g., percentile or several clinically important cutpoints). Additionally, a modified Global Initiative on Obstructive Lung Disease (GOLD) criteria originally described by Mannino et al (Am J Resp Crit Care Med, 2006) will be utilized. 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, and covariates. Impaired lung function will be defined using a modified GOLD classification and several methodological approaches will be employed to define impaired lung function. For example, individuals in the GOLD classifications and the restricted group can be combined to define impaired lung function (i.e. “case subjects”) and compared to individuals classified as “normal” (i.e. “controls subjects”). Additional analysis could focus on specific classifications (e.g. “normal” versus “GOLD 3 or 4”) or ordinal logistic regression can be employed since the dependent variable can be ranked, which is preferred to multinomial logistic regression. Multivariate logistic (or ordinal) regression analyses will be performed to assess risk of impaired lung function by the inflammation-related genotype(s). 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.

ii) Multifactorial Dimensionality Reduction (MDR): a method for detecting and characterizing high-order gene-gene and gene-environment interactions by using data-reduction and combinatorial-partitioning methods for the exploratory analysis of quantitative traits. With MDR, multilocus genotypes are pooled into high-risk and low-risk groups, effectively reducing the genotype predictors from $n$ dimensions to one dimension.

iii) Classification and Regression Tree (CART): a graphical method for predicting continuous dependent variables (regression) and categorical predictor variables (classification) by using decision trees. Decision trees contain a binary question (yes/no answer) about some feature at each node in the tree.

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 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* __________)

__X__ 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.
References for rationale


