1.a. Full Title: Genome-wide gene-by-smoking interaction analysis of pulmonary function

b. Abbreviated Title (Length 26 characters): Lung GxSmoking Analysis

2. Writing Group:
   Writing group members: Alanna Morrison, Peng Wei, Tianzhong Yang, Stephanie London, and members of the CHARGE Pulmonary Function Working Group

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

First author: Tianzhong Yang
   Address: University of Texas Health Science Center at Houston
   School of Public Health
   1200 Pressler St., Houston, TX 77030
   Phone: 617-817-6793   Fax:
   E-mail: Tianzhong.Yang@uth.tmc.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: Alanna Morrison, PhD
   Address: University of Texas Health Science Center at Houston
   School of Public Health
   1200 Pressler St., Houston, TX 77030
   Phone: 713.500.9913   Fax: 713.500.0900
   E-mail: alanna.c.morrison@uth.tmc.edu

3. Timeline:
   Studies that are a part of the CHARGE Pulmonary Function Working Group will provide results from the project analysis plan by July 2017.
   A manuscript for the first project aim will be ready for submission by August 2017.
   A manuscript for the second project aim will be ready for submission by January 2018.
4. **Rationale:**
The architecture of complex traits involves the interplay of genetic and environmental factors. For example, lung function and its decline in older adulthood is the result of genetic and environmental influences. Cigarette smokers have an accelerated decline in lung function and are at increased risk of chronic obstructive pulmonary disease (COPD) and death. Although the environmental context of cigarette smoking is a key risk factor for loss of lung function over time, genetic variation also plays an important role. Family and twin studies investigating longitudinal change in lung function report heritability estimates between 10 and 39%. Genome-wide association studies (GWAS) have identified 26 genetic loci associated with cross-sectional spirometric measures of lung function. Recent GWAS of the longitudinal change in lung function have identified additional novel loci. To date, there is only one published genome-wide study of gene-by-environment (GxE) interaction on lung function that considers smoking as the environment of interest. Hancock et al. performed a genome-wide joint analysis of single nucleotide polymorphism (SNP) and SNP-by-smoking interaction using cross-sectional information on lung function and smoking, and identified three novel loci not previously associated with lung function. These studies have made important contributions to understanding the etiology of lung function; however, it’s clear that additional investigation is needed to further understand how smoking interacts with genetic factors to influence lung function.

The previous GWAS and genome-wide gene-by-smoking study for lung function were facilitated by the organizational structure and support of the Cohorts for Heart and Aging in Genomic Epidemiology (CHARGE) consortium. The CHARGE collaborative infrastructure also forms the basis for this proposed research. Ongoing work within the CHARGE Pulmonary Function Working Group includes analysis of data from the Illumina HumanExome BeadChip genotyping array (the “exome chip”) for individuals of European ancestry with spirometric measures of lung function, all of whom also have longitudinal measures of smoking history and lung function. Additional individuals of African ancestry have measures of lung function, smoking history, and exome chip data, and also have longitudinal measures. Spirometric measures include forced expiratory volume in one second (FEV1), forced vital capacity (FVC), and their ratio (FEV1/FVC). These measures of lung function are important clinical tools for diagnosing pulmonary disease, classifying its severity, and evaluating its progression over time.

5. **Main Hypothesis/Study Questions:**
The objective of this proposal is to elucidate the complex interplay of genes and environment underlying lung function using state-of-the-art statistical methods and analysis strategies that leverage available data resources. This application has two novel aspects: (1) investigation of rare variation and environmental interactions, and (2) investigation of longitudinal measures of environmental factors. The exome chip, which includes ~240,000 rare and low frequency exonic variants and GWAS-identified loci, provides an opportunity to investigate the interplay of rare and low frequency genetic variation and environmental factors influencing lung function.

**Aim 1.** Identify rare genetic variation and smoking interactions influencing lung function using novel gene-based statistical methods. We will employ the rareGE method, a recently proposed powerful statistical test, to identify rare variants aggregated in genes and biological pathways.
that interact with smoking history to influence measures of lung function (FEV1, FVC, and their ratio).

**Aim 2.** Identify genes and pathways modifying the association between longitudinal smoking history and lung function using a generalized estimating equations (GEE)-based approach. As a complementary approach to Aim 1, where smoking is a single cross sectional measure, this aim incorporates time-varying smoking information.

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 Phenotypes for Aim 1**

Table 1 below describes the data available for the 8 CHARGE Consortium studies and 5 SpiroMeta Consortium studies that will participate in Aim 1. The total sample size for Aim 1 is 50,512 individuals (44,234 EA and 6,117 individuals of African ancestry [AA]). Additionally, we received commitment from the UK Biobank/UK Believe with a total sample size of 105,000 EA individuals for replication.

| Table 1. Studies with measures of lung function, smoking history, and exome chip data |
|-----------------------------------------------|-----------------|-----------------|-----------------|
| Cohort Name                                 | Cohort Abbreviation | N               |
| **European ancestry (EA)**                  |                  |                 |
| Atherosclerosis Risk in Communities Study   | ARIC             | 10,679          |
| Cardiovascular Health Study                 | CHS              | 3,496           |
| Framingham Heart Study                      | FHS              | 6,172           |
| Rotterdam Study                             | RS               | 567             |
| Multi-Ethnic Study of Atherosclerosis       | MESA             | 1,298           |
| The Netherlands Epidemiology of Obesity Study | NEO              | 6,203           |
| Health ABC                                  | HABC             | 1,454           |
| Age, Gene/Environment Susceptibility        | AGES             | 1,459           |
| 1958 British Birth Cohort                   | B58C             | 4,594           |
| Lothian Birth Cohort 1936                   | LBC1936          | 970             |
| Northern Finland Birth Cohort 1966          | NFBC66           | 1,431           |
| Study of Health in Pomerania                | SHIP             | 4,694           |
| Cooperative Health Research in the Region of Augsburg | KORA         | 1,217           |
| **Total EA**                                |                  | **44,234**      |

| **African ancestry (AA)**                   |                  |                 |
| Atherosclerosis Risk in Communities Study   | ARIC             | 3,695           |
| Cardiovascular Health Study                 | CHS              | 465             |
| Multi-Ethnic Study of Atherosclerosis       | MESA             | 852             |
| Health ABC                                  | HABC             | 1,105           |
| **Total AA**                                |                  | **6,117**       |

All of the studies in Table 1 have measures of lung function (FEV1, FVC, and their ratio) at their baseline exam. Smoking exposure at baseline is the environmental variable of interest and this is modeled in two ways as (1) current vs. not current smoker, and (2) ever vs. never smoker.
Methods for Aim 1
All smoking variables will be included in the GxE regression model as covariates while one of them is modeled as the primary exposure variable of interest and is tested for interaction with “G”. Following Hancock et al., we model the main effects as well as interaction effects such that we can avoid global inflation of GxE test p-values. Additional covariates are age, gender, height, and the first 10 principal components (PCs) are included to account for any underlying population substructure. Rare variants with minor allele frequency (MAF) <5% on the exome chip are included in the gene-based and pathway-based analyses.

We will use the method developed by Chen et al. to test GxE for rare genetic variants. This method has been implemented in a user-friendly R package(https://cran.r-project.org/web/packages/rareGE/index.html). A fixed-effect meta-analysis model will be used to combine p-values among all the studies.

In addition to inspection of summary statistics, we will visualize results using QQ plots. Significant findings will be identified based on an a priori threshold using Bonferroni correction (e.g., p< 3.4e-06 based on 14,591 genes evaluated).

Study design and Phenotypes for Aim 2
The goal of this aim is to identify genes and pathways modifying the association between longitudinal smoking history and lung function using a generalized estimating equations (GEE)-based approach. For this aim, we will use 1000 Genomes (1000G) imputed genotype data and longitudinal measures of lung function (FEV1, FVC, and their ratio) and smoking (e.g., current vs. not current smoker and ever vs. never smoker) from the ARIC study. Replication will be sought from studies shown in Table 1 that have longitudinal data for lung function and smoking status.

Methods for Aim 2
In addition to single SNP-based Gene x Smoking interaction analysis of longitudinal lung function, we will utilize a newly developed SNP-set-based Gene x longitudinal exposure interaction test in our analytic pipeline. The new method, called LGEWIS (Tests for Genetic Association/Gene-Environment Interaction in Longitudinal Gene-Environment-Wide Interaction Studies) has been implemented in a user-friendly R package (https://cran.rproject.org/web/packages/LGEWIS/index.html). In our extensive evaluation using simulations, we found that the LGEWIS method is computationally fast, robust to nonlinear main effects of the exposure variable, and is powerful across a wide range of genetic architecture.

We will also use the Wald test to identify genome-wide SNP-environment interactions influencing longitudinal lung function.
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.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)?

The most closely related manuscript proposal is #2131, titled “Meta-analysis of exome chip variants and pulmonary function in the CHARGE consortium.” The proposed work scope in this manuscript proposal is an extension to rare variant tests for GxE and incorporation of longitudinal measures.

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

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.csec.unc.edu/aric/index.php, under Publications, Policies & Forms. http://publicaccess.nih.gov/submit_process_journals.htm shows you which journals automatically upload articles to PubMed central.

13. Per Data Use Agreement Addendum, approved manuscripts using CMS data shall be submitted by the Coordinating Center to CMS for informational purposes prior to publication. Approved manuscripts should be sent to Pingping Wu at CC, at pingping_wu@unc.edu. I will be using CMS data in my manuscript ____ Yes _X_ No.

References


