ARIC Manuscript Proposal # 3120

1.a. Full Title: Meta-analysis of the relation of DNA methylation patterns, lung function, and chronic obstructive pulmonary disease

b. Abbreviated Title: DNA methylation in relation to PFTs and COPD

2. Writing Group:

ARIC Authors: Kari North, Geetha Chittoor, Anne Justice, Myriam Fornage, Jan Bressler, Weihua Guan, other interested ARIC Authors are welcome.

CHARGE Research Team: Stephanie London, Mi Keyong Lee, Lies Lahousse, Natalie Terzikhan, Sinjini Sikdar

I, the first author, confirm that all the coauthors have given their approval for this manuscript proposal. KEN

Corresponding/senior author (if different from first author correspondence will be sent to both the first author & the corresponding author):

Kari North
Associate Professor
Department of Epidemiology
Carolina Center for Genome Sciences
UNC Gillings School of Global Public Health
Chapel Hill, NC
919-966-2148 (voice)
919-966-9800 (Fax)

3. Timeline (The following are estimates. Please see project phases and timeline below for more details):

   Individual cohort statistical analyses: Spring 2018
   Consortium meta-analyses: Fall 2019
   Manuscript preparation: Spring 2019
   Manuscript submission: Fall 2020

4. Rationale: Pulmonary function test traits (PFTs) form the basis for the diagnosis of chronic obstructive pulmonary disease (COPD), as well as monitoring of the severity of various lung diseases including COPD and asthma [1]. The commonly measured traits of pulmonary function in clinical practice are the forced expiratory volume in the first second (FEV₁), the forced vital capacity (FVC) and their ratio (FEV₁/FVC). In addition to their utility in the diagnosis and
monitoring of lung diseases, these traits are independent predictors of morbidity and mortality in the general population, even within the normal range [2]. These traits are influenced by environmental factors, most notably smoking, but also air pollution and other exposures [3, 4]. They also appear to be under a high degree of genetic control with over 100 loci contributing to COPD risk as identified in genome-wide association studies (GWAS) [3, 5-8]. Most of the loci related to COPD were previously identified as related to the lung function. It is possible that these associations between genetic variants and COPD are mediated through DNA methylation, especially those that interact with smoking, as smoking is known to influence methylation [9, 10]. Although, there is reason to believe that DNA methylation might be related to pulmonary function, independently of smoking or genetic risk factors, and thus also to COPD.

5. Main Hypotheses/Study Questions: The aim of this project is to perform a meta-analysis of the association between DNA methylation and lung function and COPD. Because of the strong influence of smoking on both of these outcomes as well as on DNA methylation we will also stratify by smoking status.

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).

Outcomes – Quantitative PFTs and COPD

A. Outcome: Quantitative measures from pulmonary function tests (PFTs).
As PFT was measured at visit 2, when the majority of ARIC epi-type occurred, we will have concurrently measured methylation and PFT. However, for the subset of participants where methylation and PFT data are not from the same time point, we will calculate the elapsed time (ideally in years with 2+ digits after the decimal, provide units) between PFTs and methylation and provide descriptive statistics: mean, SD, min, max and the following percentiles – 5th, 25th, 50th, 75th, 95th. This subset will be analysed separately, provided the sample size is well-powered. We will remove outliers based on > 5*SD from the mean after adjustment for covariates.

The specific pre-bronchodilator PFTs parameters we will analyze are:

Forced Expiratory Volume in one second – FEV1 – in units of millilitres (ml)
Forced Expiratory Volume – FVC – in units of millilitres (ml)

Ratio (=FEV1/FVC) – as a proportion ranging from 0 to 1 - to four (or more) decimal places

B. Outcome: COPD – Chronic Obstructive Pulmonary Disease

Cases: FEV1 < 80% predicted with FEV1/FVC < 0.7000 (GOLD criteria - [12])
Controls: FEV1 >= 80% predicted and FEV1/FVC >= 0.7000
To calculate predicted values, we will either use Hankinson et al. [13] reference equations, the newer GLI criteria [14]. Both variables needed to perform calculations are available in the ARIC derived data sets for visit 1 and 2 (PFTB31- FEV1/FVC and PFTB26 – FEV1).

Exposure – Methylation
We will analyze the BMIQ normalized methylation beta values. These are the exposures (predictors) in our models.

a. Quality control (QC) for methylation data
Sample level QC:
- Exclude samples with detection p-value > 1.0*10^{-10} in more than 5% of the probes

Probe level QC:
- Exclude probes with detection p-value > 1.0*10^{-10} in more than 5% of the samples

Identification of extreme outlier methylation values using “gap hunting”. This is a method to identify probes that have gaps in their distribution – these could be due to SNPs, technical errors, or other reasons (Andrews et al. [15]). We are using this for a slightly different purpose than the paper - to identify the CpGs where there are 1 or very small number of individuals with values at the other end of the distribution that can be very influential even in meta-analysis (for example, a CpG with mean beta=0.15, SD=0.05 with one individual with beta=0.99. Depending on this person’s phenotype value this single data point can generate a genome wide significant result in the meta-analysis). We will filter these values before analyzing, using the software is the gaphunter() function in the minfi package (Minfi URL: http://bioconductor.org/packages/release/bioc/html/minfi.html). The resulting data matrix with the extreme outliers removed will also be useful for other EWAS analyses and will be shared with other members of the ARIC methylation working group.

Covariates
Please use covariates assessed at the same time as the PFT phenotype (cross-sectional). If a covariate was not assessed at the same time as the phenotype, use the closest visit in time but please contact us first to discuss.

- **Sex** (0 for male, 1 for female)
- **Age** (in years)
- **Age^2** (in years^2).
- **Height** (in cm)
- **Height^2** (in cm^2)
- **Weight** (in kg) – Include weight in FVC models
- **Smoking** status – As two dummy variables, one for current smoker (CS), one for past smoker (PS) (both relative to 0 for never smokers)
- **Packyears** of smoking (in years)
- **Study center**
- **White blood cell proportions** estimated using the Houseman method on the Reinius reference data. (these are the varnames in the minfi estimated cell counts function): CD8T, CD4T, NK, Bcell, Mono, Neu and Eos. Calculate the “lymphocytes” by summing
together CD8T, CD4T, NK and Bcell. We will use the following four estimated cell counts for adjustment: lymphocytes, Mono, Neu and Eos.

- **Batch** effects (methylation technical covariates) – chip and row
- **Ancestry principal components (PCs)** to control for population stratification.
- **Family structure** if relevant for your cohort using cohort-specific methods. Specify method in README file.
- **If you have more than one major ethnic group** (for example, European, African, Asian, Hispanic) please provide results for each group analyzed separately. Provide descriptive statistics by each ethnic group in the Excel sheet.

**Inclusion criteria**

- For all analyses, include only individuals ≥ 40 years old
- For analyses of COPD
  - if <15 cases were available overall or in any subgroup (defined by ethnic group and smoking), that analysis will not be performed.

**Models and analysis**

EWAS will be carried out using R and stratified by ARIC European- and African- American participants. For each CpG site, regression modelling will assess the impact of DNA methylation (predictor) on PFTs or COPD (outcomes), with adjustment for covariates.

Each analysis of four outcomes (three PFTs and COPD) will be performed separately in three groups: All subjects, Never smokers only, Ever smokers only.

- **All subjects** (current, past, and never smokers all together)
  1. \( \text{FEV1} \sim \text{CpGs (Beta value)} + \text{Sex} + \text{Age} + \text{Age}^2 + \text{Height} + \text{Height}^2 + \text{CS} + \text{PS} + \text{Packyears} + \text{Cell proportions} + \text{Batch (+ Ancestry PCs + Study center + Selection factor)} \)
  2. \( \text{FVC} \sim \text{CpGs (Beta value)} + \text{Sex} + \text{Age} + \text{Age}^2 + \text{Height} + \text{Height}^2 + \text{Weight(kg)} + \text{CS} + \text{PS} + \text{Packyears} + \text{Cell proportions} + \text{Batch (+ Ancestry PCs + Study center + Selection factor)} \)
  3. \( \text{Ratio} \sim \text{CpGs (Beta value)} + \text{Sex} + \text{Age} + \text{Age}^2 + \text{Height} + \text{Height}^2 + \text{CS} + \text{PS} + \text{Packyears} + \text{Cell proportions} + \text{Batch (+ Ancestry PCs + Study center + Selection factor)} \)
  4. \( \text{COPD} \sim \text{CpGs (Beta value)} + \text{Sex} + \text{Age} + \text{Age}^2 + \text{Height} + \text{Height}^2 + \text{CS} + \text{PS} + \text{Packyears} + \text{Cell proportions} + \text{Batch (+ Ancestry PCs + Study center + Selection factor)} \)

- **Never smokers only** (NS)
  5. \( \text{FEV1} \sim \text{CpGs (Beta value)} + \text{Sex} + \text{Age} + \text{Age}^2 + \text{Height} + \text{Height}^2 + \text{Cell proportions} + \text{Batch (+ Ancestry PCs + Study center + Selection factor)} \)
  6. \( \text{FVC} \sim \text{CpGs (Beta value)} + \text{Sex} + \text{Age} + \text{Age}^2 + \text{Height} + \text{Height}^2 + \text{Weight(kg)} + \text{Cell proportions} + \text{Batch (+ Ancestry PCs + Study center + Selection factor)} \)
  7. \( \text{Ratio} \sim \text{CpGs (Beta value)} + \text{Sex} + \text{Age} + \text{Age}^2 + \text{Height} + \text{Height}^2 + \text{Cell proportions} + \text{Batch (+ Ancestry PCs + Study center + Selection factor)} \)
  8. \( \text{COPD} \sim \text{CpGs (Beta value)} + \text{Sex} + \text{Age} + \text{Age}^2 + \text{Height} + \text{Height}^2 + \text{Cell proportions} + \text{Batch (+ Ancestry PCs + Study center + Selection factor)} \)

- **Ever smokers only** (ES; past and current)
9. \( \text{FEV1} \sim \text{CpGs (Beta value)} + \text{Sex} + \text{Age} + \text{Age}^2 + \text{Height} + \text{Height}^2 + \text{Smoking (CS=1 versus PS=0)} + \text{Packyears} + \text{Cell proportions} + \text{Batch (+ Ancestry PCs + Study center + Selection factor)} \)

10. \( \text{FVC} \sim \text{CpGs (Beta value)} + \text{Sex} + \text{Age} + \text{Age}^2 + \text{Height} + \text{Height}^2 + \text{Weight(kg)} + \text{Smoking (CS=1 versus PS=0)} + \text{Packyears} + \text{Cell proportions} + \text{Batch (+ Ancestry PCs + Study center + Selection factor)} \)

11. \( \text{Ratio} \sim \text{CpGs (Beta value)} + \text{Sex} + \text{Age} + \text{Age}^2 + \text{Height} + \text{Height}^2 + \text{Smoking (CS=1 versus PS=0)} + \text{Packyears} + \text{Cell proportions} + \text{Batch (+ Ancestry PCs + Study center + Selection factor)} \)

12. \( \text{COPD} \sim \text{CpGs (Beta value)} + \text{Sex} + \text{Age} + \text{Age}^2 + \text{Height} + \text{Height}^2 + \text{Smoking (CS=1 versus PS=0)} + \text{Packyears} + \text{Cell proportions} + \text{Batch (+ Ancestry PCs + Study center + Selection factor)} \)

References:


7.a. Will the data be used for non-CVD analysis in this manuscript?
8. a. Will the DNA data be used in this manuscript?
   __x__ Yes
   ____ No

   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.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)?

    GWAS of SNP-by-Serum Vitamin D Interaction on Pulmonary Function. Geetha Chittoor and
    Kari North were authors and members of the CHARGE pulmonary function group.

    GWAS of pulmonary function. Kari North was an author on this work and is a member of the
    CHARGE pulmonary function group.

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 (AS #2006.03 & 2007.02_)
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.cscc.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.