1. **Full Title:** Meta-analysis of exome chip variants and pulmonary function in the CHARGE consortium

b. **Abbreviated Title (Length 26 characters):** Exome chip and PFTs: CHARGE

2. **Writing Group:**
   Writing group members: Stephanie London, Alanna Morrison, Kari North, Laura Loehr, David Couper, Nora Franceschini, Bonnie Joubert. Consortium paper – number of authors per cohort subject to change.

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

   **First author:** Stephanie London – subject to change. I am organizing this and Alanna is doing the analysis
   Address: NIEHS, PO Box 12233, MD A3-05. RTP NC 27709

   Phone: 919-541-5772     Fax: 
   E-mail: london2@niehs.nih.gov

   **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).
   Address: The University of Texas at Houston, School of Public Health, Division of Epidemiology, Human Genetics and Environmental Science
   1200 Herman Pressler; Suite 453E, Houston, TX 77030

   Phone: 713-500-9913     Fax: 713-500-0900
   E-mail: Alanna.C.Morrison@uth.tmc.edu


4. **Rationale:**
Spirometric measures of pulmonary function, such as forced expiratory volume in one second (FEV1), forced vital capacity (FVC) and their ratio (FEV1/FVC) are heritable traits influenced by both environment and genes. Recent genome-wide association studies (GWAS) in subjects of European ethnicity have identified at least 27 loci associated with FEV1 and/or FEV1/FVC. In ongoing analyses we are finding independent signals for FVC. Unfortunately, these signals are responsible for a minimal amount of the trait variation: for FEV1, the GWAS loci explain ~1.5% of inter-individual variance (adjusting for age, gender and height) and for FEV1/FVC the implicated loci explain 3.2% of the variation in this trait. The extent to which rare variants contribute to variation in these traits within the population is unknown. We seek to test the hypothesis that rare coding variation in conjunction with common variants contributes to inter-individual variability in these three spirometric phenotypes (FEV1, FVC, and FEV1/FVC).

Recently, a novel genotyping array (“exome chip”) was created to comprehensively evaluate rare coding variants. The “exome chip” contains ~250,000 coding variants discovered through exome sequencing in ~12,000 individuals and, in addition, includes all common GWAS variants previously associated with pulmonary function. Collectively, the array represents nearly all non-synonymous coding and splice-site variation with a >1:1000 allele frequency in the European population.

Here we propose analyses of ARIC samples with both exome chip and spirometry phenotype data (FEV1, FVC, and FEV1/FVC) with the goal of discovering new genes harboring rare coding variants associated with pulmonary function. The ARIC results will be combined with other studies participating in the CHARGE Pulmonary Working Group using meta-analyses.

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

1. Do rare variants or functional variants not covered on GWAS platforms explain the GWAS signals that we have already identified for these phenotypes.
2. Conditioning on our previous GWAS hits, are exome chip variants independently associated with these phenotypes?
3. Are there variants in new genes harboring variants associated with pulmonary function?

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).**

Detailed CHARGE Pulmonary Working Group Exome chip analysis plan is attached at the end of this document.

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?  ____ Yes  _x_ 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: \( \text{http://www.cscce.unc.edu/ARIC/search.php} \)
   \( _x \) 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)?
   Our analysis plan is modeled on the exome chip CHARGE adiposity group analysis plan which one of our co-authors, Kari North, kindly provided to us. In addition Alanna Morrison, is involved in exome chip analysis of lipid traits and other phenotypes.

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 \( \text{http://www.cscce.unc.edu/aric/forms/} \)

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 \( \text{http://publicaccess.nih.gov/} \) are posted in \( \text{http://www.cscce.unc.edu/aric/index.php} \), under Publications, Policies & Forms. \( \text{http://publicaccess.nih.gov/submit_process_journals.htm} \) shows you which journals automatically upload articles to Pubmed central.
Background and Rationale:
Spirometric measures of pulmonary function, such as forced expiratory volume in one second (FEV1), forced vital capacity (FVC) and their ratio (FEV1/FVC) are heritable traits influenced by both environment and genes. Recent genome-wide association studies (GWAS) in subjects of European ethnicity have identified at least 27 loci associated with FEV1 and/or FEV1/FVC. In ongoing analyses we are finding independent signals for FVC. Unfortunately, these signals are responsible for a minimal amount of the trait variation: for FEV1, the GWAS loci explain ~1.5% of inter-individual variance (adjusting for age, gender and height) and for FEV1/FVC the implicated loci explain 3.2% of the variation in this trait. The extent to which rare variants contribute to variation in these traits within the population is unknown. We seek to test the hypothesis that rare coding variation in conjunction with common variants contributes to inter-individual variability in these three spirometric phenotypes (FEV1, FVC, and FEV1/FVC).

Recently, a novel genotyping array (“exome chip”) was created to comprehensively evaluate rare coding variants. The “exome chip” contains ~250,000 coding variants discovered through exome sequencing in ~12,000 individuals and, in addition, includes all common GWAS variants previously associated with adiposity. Collectively, the array represents nearly all non-synonymous coding and splice-site variation with a >1:1000 allele frequency in the European population.

Using this exome chip platform for the study of insulin traits, rare variants underlying GWAS signals have been identified and novel loci have been identified. Here we propose analyses of ARIC subjects with spirometry phenotypes (FEV1, FVC, and FEV1/FVC) with the goal of discovering new genes harboring rare coding variants associated with pulmonary function. The ARIC analysis will be meta-analyzed with other CHARGE cohorts. We may also combine data with the SpiroMeta cohorts although their exome chip data are not yet ready so we will start with CHARGE analyses.

Methods:
The analysis will be done within each cohort and only summary results will be shared. The CHARGE Pulmonary WG has discussed QC procedures and we are using comparable criteria across cohorts—at this time point QC procedures within each cohort have already been refined based on working groups that began this process earlier. When possible across cohort summary statistics may be used to examine quality of variants across studies. Summary results will be deposited to a password-protected UW sharepoint that has been set up specifically for the Pulmonary WG exome chip analysts and key investigators and meta-analysis will be performed. We plan for one investigator to take the lead on the analysis and for a second to repeat the analysis for verification.
I. Each individual study
   a. Quality control: each cohort will use the data that has already gone through the central quality control for your cohort. The recommended steps are listed below so please indicate where your cohort has deviated from any of these criteria.
      1. Excluding those missing > 5% genotypes
      2. Population clustering to identify outliers
      3. High inbreeding coefficient or heterozygote rate far from median (to detect possible contamination) based on distribution observed in the data
      4. Gender mismatch
      5. One from duplicate pairs, keeping the individual with less missing data
      6. Unexpectedly high proportion IBD sharing, with consideration for family studies, based on high quality variants
      7. Unexpected relatives, with consideration of family structures, based on high quality variants
   iii. Variant QC for single variant analysis
      1. Exclusion of those strongly associated with plate assignment
   iv. Additional variant QC for single variant analysis during meta-analysis
      1. Remove monomorphic SNPs
      2. Remove variants with missing rate > 5%
      3. Remove variants with HWE p-value < 1E-06 (only for autosomal variants with MAF > 5%)
   v. Variant QC for burden analysis
      1. Remove monomorphic SNPs
      2. Remove variants with missing rate > 5%
   vi. The GWAS PCs, if available will be used
      a. If GWAS PCs are not available, then one can use the ~3,000 AIMs on the exome chip to calculate PCs subject to maf > 1%, optionally adding exonic variants with appropriate maf.

For ARIC, subjects who overlap with JHS will be removed because JHS is also planning to participate in the CHARGE Pulmonary WG exome chip analysis. JHS should do likewise – remove ARIC overlap subjects.

b. Phenotype Model

Outcome FEV1 – units in milliliters (please check units – saves much work later on when we discover that some were analyzed in liters)
Covariates: former smoking, current smoking and pack-years of smoking, age, age^2, sex, height (cm), height^2, center/cohoot, principle components

Outcome-FVC - in milliliters- difference from above is adding weight to the model
Covariates: former smoking, current smoking and pack-years of smoking, age, age$^2$, sex, height, height$^2$, weight (kg), center/cohoot, principle components.

Outcome FEV1/FVC – please use percentage (%), not proportion. Again, please check this because units matter in the meta-analysis.

Covariates: former smoking, current smoking and pack-years of smoking, age, age$^2$, sex, height, height$^2$, center/cohoot, principle components

1. **Single Variant Analyses**: For each trait, please run the association analysis in the combined sample (see below if your study has nonwhites), but also in never and ever smoker subgroups. If your cohort has non-whites, the combined sample means separately by major ethnic group (European-Americans, African-Americans, etc). If you have family data, use a linear mixed effects model to account for non-independence of family members.

   - Phenotypes – check for outliers – do this in the population as whole, not stratified by anything. The reason for this is that rare variant analyses are more sensitive to outlier than our previous GWAS analyses.
     - Each cohort should deal with outliers in a rational manner so they do not unduly affect the association results. Please indicate number of subjects deleted as outliers.
     - A method for outlier evaluation was discussed on the CHARGE PFTs call on March 12, 2013. Using the covariates in our phenotype models above, examine the residuals. We suggest using residuals > 5 SD from the mean to evaluate for exclusion. Once you have identified potential outliers, a quick way to eyeball them is to list out the percent predicteds for them because these take age, height, race and gender into account.
     - Will evaluate genomic lambda values to determine if there appears to be elevated type 1 error and if so, will follow-up with analysis of rank-normalized phenotypes.

   - Limit your dataset to individuals with valid data for both FEV1 and FVC so that the Ns will be the same across the three outcomes (FEV1, FVC and FEV1/FVC).

After removing any outliers, generate numbers that will be needed for a descriptive table. Save it as an Excel document “Exome.descriptive.STUDYNAME.xls” and put on the Sharepoint. We need the following fields and information:

Study name
N – total
Females N, %
Males, N, %
Age – min, max, median, mean, SD
Height (cm) – min, max, median, mean, SD
Weight (kg) – min, max, median, mean, SD
FEV1 (ml) – min, max, median, mean, SD
FVC (ml) – min, max, median, mean, SD
FEV1/FVC – min, max, median, mean, SD
Never smokers, N%
Past smokers, N%
Current smokers, n%
Pack-years among ever smokers only (ever=past+current) – min, max, median, mean, SD
Visit from which spirometry values were taken

If you find any min or max values that are implausible for either outcomes or covariates, please check and revisit exclusions.

- Genotypes – use cleaned / filtered genotypes (those passing your study QC procedures)

1. Variant-wise analysis
   - For variant-wise tests, do all SNPs – filters for low maf or mac will be applied centrally on the basis of maf and mac determined in the common calling.

   a. FEV1 (in ml)
      i. By linear regression in the subgroups separately:
      \[
      \text{FEV1 (ml)} = \text{SNP} + \text{former smoking, current smoking and pack-years of smoking, age, age}^2, \text{sex, height, height}^2, \text{center/cohort, principle components}
      \]
      ii. report beta coefficient for SNP, SE, EA (effect allele), EAF (effect allele frequency)

   b. FVC (ml)
      \[
      \text{FVC (ml)} = \text{SNP} + \text{former smoking, current smoking and pack-years of smoking, age, age}^2, \text{sex, height, height}^2, \text{weight, center/cohort, principle components}
      \]
      ii. report beta coefficient for SNP, SE, EA (effect allele), EAF (effect allele frequency)

   c. FEV1/FVC (%)

2. Rare variant tests (T1, SKAT)
• T1
  o Include all variants (no lower limit on frequency, but maf < 1% in common calling) –
    ▪ Sites to include defined by the CHARGE dbNSFP annotation/SNPINFO file:
      o Variants with MAF <= 1% (for EA/AA depending on study being analyzed)
      o Variants labeled as missense, nonsense, splice
      o Count of # alleles in a specified region
      o “Region” will be defined as “gene” here.
      o Gene will be the unit of inference.
      o This is the more powerful burden test when all alleles have effects in the same direction
      o FEV1 = allele count + covariates + PC
      o FVC = allele count + covariates + PC
      o FEV1/FVC = allele count + covariates + PC

• SKAT
  o Include all variants (no lower or upper limit on frequency)
  o Use script off the wiki (package develop by the CHARGE Analysis group (available on the CHARGE wiki)
    ▪ Wu weights of α = 1 and β = 25)
  o Method is more powerful if alleles have effects in different directions.
  o Same phenotype model as other analyses
  o Missing data are handled by imputation in skatMETA package

• Conditional analyses to determine if low frequency variants in a region are independent of GWAS hits
  o Implement by adding top identified GWA SNPs as covariates in analyses (single variant and burden tests)
  o Use imputed data from GWAS for covariate SNPs. (If SNP is available on chip could use actual genotypes)

We plan to run all analyses (single SNP, T1 burden, SKAT) using the skatMETA package on the CHARGE wiki.

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
a. Files to be shared
i. Descriptive Phenotypic information (STUDY.PHENO.DATE.xls)
   1. Sample size, mean/min/max/sd of each spirometry phenotype, age, sex
ii. Results from linear regression (STUDY.TRAIT.LR.DATE.txt)
1. Sample size, Variant identifier, effect allele, other allele, beta, se, p-value, HWE p-value, callrate, EAF, count of number of individuals with minor allele
   iii. Results from Burden tests (STUDY.TRAIT.SKAT.DATE.txt, STUDY.TRAIT.GENDER.VT.DATE.txt)
   1. Gene, number of variants in test, test statistic, p-value
   iv. Fields should be tab delimited
   v. skatMETA R object from running skatCohort or skatfamCohort (for SKAT results)

II. Meta-analysis
   b. Inverse variance weighted fixed effects
   c. QQ plots will be generated
   d. Significance will be assessed at $2.0 \times 10^{-7}$ for single variant tests (chip-wide significance threshold; 0.05/250,000 variants); $2.5 \times 10^{-6}$ for burden tests (0.05/20,000 genes)

File naming convention
Please use the following naming scheme:
STUDY.PHENOTYPE.METHOD.DATE.txt

STUDY is a short (14 characters or less) identifier for the population studied. If you have a case-control study and will be providing data separately for cases and controls, please use the suffix “.CASE” or “.CONTROL” after the study identifier.

PHENOTYPE options:
“FEV1”
“FVC”
“FEV1/FVC”

METHOD options:
“LR” for linear regression
“SKAT” for SKAT burden test
“T1” for T1 burden test

DATE is the date on which the file was prepared, in the format “YYYYMMDD”

Example: using Family Heart Study (Family data) and FEV1 phenotype
FamHS.FEV1.COMBINED.LR.20121010.txt
   i. GWAS results for linear regression of standardized residuals adjusting for age, age^2, sex, study specific field center, and PCs (when covariates cannot be adjusted for in the model) among all participants
FamHS.FEV1.SMOKERS.LR.20121010.txt
i. GWAS results for linear regression of standardized residuals adjusting for age, age², study specific field center, and PCs (when covariates cannot be adjusted for in the model) among men
FamHS.FEV1.NEVERSMOKERS.LR.20121010.txt
   i. GWAS results for linear regression of standardized residuals adjusting for age, age², study specific field center, and PCs (when covariates cannot be adjusted for in the model) among women
FamHS.FEV1.COMBINED.SKAT.20121010.Rdata
   ii. Results for SKAT burden test of standardized residuals adjusting for age, age², study specific field center, and PCs (when covariates cannot be adjusted for in the model) among all participants
FamHS.FEV1.SMOCKERS.SKAT.20121010.Rdata
   ii. Results for SKAT burden test of standardized residuals adjusting for age, age², study specific field center, and PCs (when covariates cannot be adjusted for in the model) among men
FamHS.FEV1.NEVERSMOKERS.SKAT.20121010.Rdata
   ii. Results for SKAT burden test of standardized residuals adjusting for age, age², study specific field center, and PCs (when covariates cannot be adjusted for in the model) among women
FamHS.FEV1.COMBINED.T1.20121010.txt
   iii. Results for T1 burden test of standardized residuals adjusting for age, age², sex, study specific field center, and PCs (when covariates cannot be adjusted for in the model) among all participants
FamHS.FEV1.SMOCKERS.T1.20121010.txt
   iii. Results for T1 burden test of standardized residuals adjusting for age, age², study specific field center, and PCs (when covariates cannot be adjusted for in the model) among men
FamHS.FEV1.NEVERSMOKERS.T1.20121010.txt
   iii. Results for T1 burden test of standardized residuals adjusting for age, age², study specific field center, and PCs (when covariates cannot be adjusted for in the model) among women

References

