ARIC Manuscript Proposal #2157

PC Reviewed: 6/11/13	**Status: A**	**Priority: 2**
SC Reviewed: _____	**Status: _____**	**Priority: ____**

1.a. Full Title: Pathway analysis based on meta-analysis of genome wide association studies of FEV_1 and FEV_1/FVC

b. Abbreviated Title (Length 26 characters): **Pathway analysis: GWAS of PFTs**

2. Writing Group:

Writing group members: Stephanie London, Bonnie Joubert, Kari North. It is possible that other ARIC authors will be included. This analysis uses only GWAS results files used for the meta-analysis that we published in 2011 (2). The current proposal is for 3 authors per CHARGE cohort. This may change in negotiation in either direction. Sina Gharib from CHS is the first author.

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

F**irst author**: **Stephanie London**
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).
Name: Kari North
Address: 137 E. Franklin St., Suite 306 , Campus Box 7435
Chapel Hill, NC

Phone: (919) 966-2148 Fax: (919) 966-9800
E-mail: kari_north@unc.edu

3. Timeline: Manuscript submitted by October 30, 2013.

4. Rationale: text is modified from the CHS manuscript proposal submitted by Sina Gharib, the first author.

ARIC participated in the CHARGE Consortium meta-analysis of genome wide association studies (GWAS) of lung function[1]. More recently, the CHARGE and SpiroMeta consortia have published large-scale GWAS meta-analyses of lung function[2] and airflow obstruction[3]. We propose to use these datasets to test our hypothesis that a pathway-based approach can increase the number of biologically plausible candidate SNPs for these clinically relevant phenotypes.

Genome wide association studies have the potential to elucidate the genetic underpinnings of complex phenotypes in a comprehensive and unbiased manner[4]. Although a powerful approach, GWAS suffer from a number of limitations[4]. A key issue is identification of relatively limited SNPs that associate with the phenotype of interest because of rigorous statistical thresholds. These strict criteria are implemented, in part, to account for the large number of SNPs that are assessed simultaneously in genome-wide studies.

We propose to circumvent some of these limitations by exploiting known biological functions and relationships of gene products to expand the pool of putative candidates. The rationale for this approach is straightforward: since complex diseases are the result of perturbations in biological pathways, genetic variation in more than one member of a given process will increase the likelihood that the pathway is causally linked to the disease. This hypothesis implies that less strict statistical cutoffs can be used if SNPs in multiple gene members of a biological process associate with the phenotype of interest.

We would like to acknowledge that this approach has several limitations. First of all, our analysis must be limited to SNPs that map to loci of known genes (or within a defined region around them). Furthermore, polymorphisms in a given gene may not alter its expression level (at the mRNA or protein level). Finally, our assumption that multiple members of a biological module can have genetic variations may occur infrequently, thereby limiting the utility of our methods.

However, there are a number of attractive features in a "pathway-centric" approach to GWAS. Incorporating known relationships among genes that also fulfill statistical criteria for association with a disease brings substantial biological relevance to an otherwise purely statistical decision on selecting putative candidates. Importantly, a pathway-based analysis may provide a much more meaningful benchmark for comparing multiple GWAS by focusing on "enriched" biological modules common to related phenotypes or across different populations.

Ultimately, associating biological pathways with disease phenotypes has the potential to expand the pool of candidate genes used for biomarker analysis, provide mechanistic insights into the pathophysiology of complex disorders, and identify logical targets for therapeutic interventions.

5. Main Hypothesis/Study Questions:
Integrating biological knowledge with genome-wide association studies expands the number of lung function-associated candidate SNPs and may identify mechanistic pathways.

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).

The inputs for the pathway analysis are the meta-analyzed GWAS results for FEV1 and FEV1/FVC across the CHARGE cohorts that participated in the combined meta-analysis with the SpiroMeta consortium (2). Although that paper meta-analyzed across all CHARGE and SpiroMeta cohorts in this paper we propose a replication approach. We will meta-analyze the CHARGE cohorts and will perform pathway analysis on them. We will then repeat in the SpiroMeta cohorts. We will then look at top pathways from the overlap. We will use airflow obstruction jointly meta-analyzed across both consortia as a secondary analysis. We provide more detail below.

We will begin by simplifying the complex structure of this dataset. As part of the completed lung function meta-analyses, the CHS dataset has already undergone standard statistical analyses including phenotype modeling, genotype analysis (e.g., determination of Hardy-Weinberg equilibrium, imputation), and phenotype-genotype association (e.g., modeling of residuals to determine effect estimates, standard errors, *P*-values). Therefore, no additional analyses at the genotype-phenotype level will be needed.

The procedure described below will be applied separately to the following pulmonary phenotypes:

a. FEV_1.
b. FEV_1/FVC.
c. Airflow obstruction, defined as FEV_1 and FEV_1/FVC below the lower limit of normal[7]; this analysis will be performed in all participants and in subsets of subjects based on smoking status.

We will begin by simplifying the complex structure of this dataset. Genotyped and imputed SNVs from lung function GWAS ($n \approx 2.5\times10^6$) will be mapped to genes if within a 100 kb distance (upstream or downstream). For a given SNV, if multiple genes are located within this range, the closest gene will be selected and assigned the association *P*-value. Since multiple SNVs can map to the same gene, a SNV label permutation will be used to reduce biases caused by larger loci having disproportionately higher number of SNVs. Log-transformed association *P*-values ($-\log_{10}[P]$) will be used to rank order the resulting gene list (~17,700 genes) and calculate gene set enrichment

scores (ES). Approximately 2,000 gene sets will be obtained from the Molecular Signatures Database (http://www.broadinstitute.org/gsea/msigdb) (8,9). The gene sets will be limited to curated pathways derived from multiple resources such as KEGG, BioCarta, REACTOME, and functional annotations extracted from the Gene Ontology database. A modified version of GSEA procedure will be performed using SNV label permutation analysis to generate a distribution of ES and adjust for multiple testing using false discovery rate (FDR). Significant enrichment of gene sets will be determined at FDR < 5%.

Step-wise validation of functional enrichment analysis. A two-step approach will be taken to independently validate enriched pathways associated with lung function. The i-GSEA4GWAS algorithm will be initially applied to the CHARGE pulmonary function GWAS and enriched gene sets will be identified if they meet an FDR < 5% for either FEV_1 or FEV_1/FVC. Next, the same procedure will beimplemented in the SpiroMeta consortium GWAS for FEV_1 and FEV_1/FVC. We will restrict further analysis to those enriched pathways in CHARGE (FDR < 5%) that are also significantly enriched in SpiroMeta (FDR < 5%).

Cluster analysis. Two-way unsupervised hierarchical clustering will be performed on enriched pathways based on the membership profile of gene sets and their associated genes' log-transformed *P*-values ($-\log_{10}[P]$) using Pearson's correlation metric (10).

Literature mining. We will use PubMatrix (11) an online multiplex comparison tool for querying "search" and "modifier" terms within NCBI's PubMed database, to index published literature on the role of enriched gene sets and their associated gene members in influencing lung function. The search terms are either pathway-associated gene symbols (n = 3878) or pathway names (n = 131), and the modifier term is "pulmonary function".

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

But only GWAS results files are being used.

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

But only GWAS results files are being used.

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)? I am submitting another MS proposal that has similarities with this. It is "GWAS follow-up for lung function using expression quantitative trait locus analysis". The current MS would be submitted first. The second proposal would be submitted later – it includes a pathway analysis but that one just uses eQTL SNPs and the pathway analysis is not the main element of the paper.

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.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.

References:

1. Hancock DB, Eijgelsheim M, Wilk JB, Gharib SA, Loehr LR, Marciante KD, Franceschini N, van Durme YM, Chen TH, Barr RG, Schabath MB, Couper DJ, Brusselle GG, Psaty BM, van Duijn CM, Rotter JI, Uitterlinden AG, Hofman A, Punjabi NM, Rivadeneira F, Morrison AC, Enright PL, North KE, Heckbert SR, Lumley T, Stricker BH, O'Connor GT, London SJ. Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. *Nat Genet*. 2010;42:45-52

2. Soler Artigas M, Loth DW, Wain LV, Gharib SA, Obeidat M, Tang W, Zhai G, Zhao JH, Smith AV, Huffman JE, Albrecht E, Jackson CM, Evans DM, Cadby G, Fornage M, Manichaikul A, Lopez LM, Johnson T, Aldrich MC, Aspelund T, Barroso I, Campbell H, Cassano PA, Couper DJ, Eiriksdottir G, Franceschini N, Garcia M, Gieger C, Gislason GK, Grkovic I, Hammond CJ, Hancock DB, Harris TB, Ramasamy A, Heckbert SR, Heliovaara M, Homuth G, Hysi PG, James AL, Jankovic S, Joubert BR, Karrasch S, Klopp N, Koch B, Kritchevsky SB, Launer LJ, Liu Y, Loehr LR, Lohman K, Loos RJ, Lumley T, Al Balushi KA, Ang WQ, Barr RG, Beilby J, Blakey JD, Boban M, Boraska V, Brisman J, Britton JR, Brusselle GG, Cooper C, Curjuric I, Dahgam S, Deary IJ, Ebrahim S, Eijgelsheim M, Francks C, Gaysina D, Granell R, Gu X, Hankinson JL, Hardy R, Harris SE, Henderson J, Henry A, Hingorani AD, Hofman A, Holt PG, Hui J, Hunter ML, Imboden M, Jameson KA, Kerr SM, Kolcic I, Kronenberg F, Liu JZ, Marchini J, McKeever T, Morris AD, Olin AC, Porteous DJ, Postma DS, Rich SS, Ring SM, Rivadeneira F, Rochat T, Sayer AA, Sayers I, Sly PD, Smith GD, Sood A, Starr JM, Uitterlinden AG, Vonk JM, Wannamethee SG, Whincup PH, Wijmenga C, Williams OD, Wong A, Mangino M, Marciante KD, McArdle WL, Meibohm B, Morrison AC, North KE, Omenaas E, Palmer LJ, Pietilainen KH, Pin I, Pola Sbreve Ek O, Pouta A, Psaty BM, Hartikainen AL, Rantanen T, Ripatti S, Rotter JI, Rudan I, Rudnicka AR, Schulz H, Shin SY, Spector TD, Surakka I, Vitart V, Volzke H, Wareham NJ, Warrington NM, Wichmann HE, Wild SH, Wilk JB, Wjst M, Wright AF, Zgaga L, Zemunik T, Pennell CE, Nyberg F, Kuh D, Holloway JW, Boezen HM, Lawlor DA, Morris RW, Probst-Hensch N, Kaprio J, Wilson JF, Hayward C, Kahonen M, Heinrich J, Musk AW, Jarvis DL, Glaser S, Jarvelin MR, Ch Stricker BH, Elliott P, O'Connor GT, Strachan DP, London SJ, Hall IP, Gudnason V, Tobin MD. Genome-wide association and large-scale

follow up identifies 16 new loci influencing lung function. *Nat Genet*. 2011;43:1082-1090
3. Wilk JB, Shrine NR, Loehr LR, Zhao JH, Manichaikul A, Lopez LM, Smith AV, Heckbert SR, Smolonska J, Tang W, Loth DW, Curjuric I, Hui J, Cho MH, Latourelle JC, Henry AP, Aldrich M, Bakke P, Beaty TH, Bentley AR, Borecki IB, Brusselle GG, Burkart KM, Chen TH, Couper D, Crapo JD, Davies G, Dupuis J, Franceschini N, Gulsvik A, Hancock DB, Harris TB, Hofman A, Imboden M, James AL, Khaw KT, Lahousse L, Launer LJ, Litonjua A, Liu Y, Lohman KK, Lomas DA, Lumley T, Marciante KD, McArdle WL, Meibohm B, Morrison AC, Musk AW, Myers RH, North KE, Postma DS, Psaty BM, Rich SS, Rivadeneira F, Rochat T, Rotter JI, Soler Artigas M, Starr JM, Uitterlinden AG, Wareham NJ, Wijmenga C, Zanen P, Province MA, Silverman EK, Deary IJ, Palmer LJ, Cassano PA, Gudnason V, Barr RG, Loos RJ, Strachan DP, London SJ, Boezen HM, Probst-Hensch N, Gharib SA, Hall IP, O'Connor GT, Tobin MD, Stricker BH. Genome wide association studies identify chrna5/3 and htr4 in the development of airflow obstruction. *Am J Respir Crit Care Med*. 2012
4. Kruglyak L. The road to genome-wide association studies. *Nat Rev Genet*. 2008;9:314-318
5. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102:15545-15550
6. Zhang K, Cui S, Chang S, Zhang L, Wang J. I-gsea4gwas: A web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study. *Nucleic Acids Res*. 2010;38:W90-95
7. Swanney MP, Ruppel G, Enright PL, Pedersen OF, Crapo RO, Miller MR, Jensen RL, Falaschetti E, Schouten JP, Hankinson JL, Stocks J, Quanjer PH. Using the lower limit of normal for the fev1/fvc ratio reduces the misclassification of airway obstruction. *Thorax*. 2008;63:1046-1051
8. Subramanian, A., Kuehn, H., Gould, J., Tamayo, P. & Mesirov, J.P. GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics* **23**, 3251-3 (2007).
9. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-50 (2005).
10. Saeed, A.I. *et al.* TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* **34**, 374-8 (2003).
11. Becker, K.G. *et al.* PubMatrix: a tool for multiplex literature mining. *BMC Bioinformatics* **4**, 61 (2003).