ARIC Manuscript Proposal #4125

PC Reviewed: 9/13/22 **Status: _____** **Priority: 2**
SC Reviewed: _____ **Status: _____** **Priority: ____**

1.a. Full Title: Novel Privacy Considerations for Large Scale Proteomics
b. Abbreviated Title (Length 26 characters): pQTLs prediction

2. Writing Group:
Writing group members:
Russell Bowler, Betty Gorbet, Josef Coresh and Eric Boerwinkle.
Other are welcome.

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

First author: **Russell Bowler**
Address: Division of Pulmonary Medicine, Department of Medicine
National Jewish Health, Denver, Colorado
1400 Jackson Street, Room K715a
Denver, Colorado 80206
Phone: 303-398-1639 Fax: 303-270-2249
E-mail: BowlerR@NJHealth.org

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: **Bing Yu**
Address: Human Genetics Center
1200 Pressler Street, Suite E-407
Houston, TX 77030

Phone: 713-500-9285 Fax: 713-500-9000
E-mail: bing.yu@uth.tmc.edu

3. Timeline:
The discovery analyses in COPDGene is completed. The manuscript is to be completed as soon as the replication results from ARIC is available. We expect that the manuscript will be prepared within three months from approval of the analysis plan.

4. Rationale:
Identifying individuals by genomics is a rising concern in research because advances in genotyping and sequencing have resulted in large genetic databases (dbGaP; GEO; EMBL-EBI) for both research and commercial use. The existence of newer genotyping technologies and large genomic databases has created concerns among policy makers regarding discrimination in health

insurance and employment and resulted in new laws that address genetic information (e.g., the Genetic Information Non-discrimination Act of 2008) as well as privacy protection efforts such as the Global Alliance for Genomics and Health, which has created frameworks to ensure responsible and secure sharing of genomic and health-related data. A key feature of these policies in the United States is that they explicitly addressed genomic data only. Despite these policies, there have been multiple instances of "deidentified" personal information linked back to individual genetic profiles (1), including well publicized individuals such as Henrietta Lacks (2). There have also been methods proposed which can link expression data to genotype through eQTLs (3). Concurrently there are studies which demonstrate that many proteins (4, 5) have genetic quantitative trait loci (QTLs), but current practice is to consider these datasets as deidentified data. In COPDGene, we have shown that even limited proteome profiles without peptide sequencing can be linked to specific individuals by using prior independent knowledge of these QTLs and we provide a bioinformatic solution which obfuscates reidentification, yet still preserves at least some biomarker-phenotype relationships. Here we seek replication of our findings in the ARIC study.

5. Main Hypothesis/Study Questions:

1. To predict genotypes in ARIC using 250 pre-trained cis-pQTLs in COPDGene.

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).

This is a cross-sectional study that consists of ~11,000 ARIC participants at visit 2 with plasma proteomes and imputed genotypes using TOPMed panel. Proteomes were measured by SOMAscan assay (Somalogic Inc., Boulder, CO) and were log2 transformed prior to the analyses. 250 cis-pQTLs were pre-trained in COPDgene. The corresponding protein levels and variant gentoptyes were pulled out from ARIC whites and blacks to replicate the findings from COPDGene.

For predicting the probability of a genome matching, a Naïve Bayesian method is used which estimates the probability of observing genotype vector *g* using the genotype specific mean (μ) and standard deviation (s) estimated from the training data. This is similar to an approach used in genotype estimation from eQTLs (*3*). Under the naïve Bayes framework, we will estimate the probability of the subject possessing each of the three genotype classes, given an observed protein level. By repeating this process for each of the *N* protein/SNP pairs, we will obtain the probability of each genotype class for the top 250 SNPs. We will calculate the odds of each genotype being the true genotype, and then using the known genotype values $g_1 \ldots g_N$ for each subject, we can compute the odds of observing the correct or "true" genotype vector $\boldsymbol{g^{true}}$ for a subject as the product of the odds of observing the individual true genotype values.

$$Odds(\boldsymbol{g^{true}}|\boldsymbol{x^{true}}) = \prod_{i=1}^{N} \frac{P(g_i^{true}|x_i^{true})}{1 - P(g_i^{true}|x_i^{true})}$$

For each subject with proteome data, we will calculate the odds of the genotype vector of every genotyped subject in the dataset. Assuming one of the genotyped subjects within the dataset is the true identity S_{true} with observed protein levels $\boldsymbol{x_{true}}$ we will take the genotype with the highest odds given the observed protein values as the "match" for this subject. If the genotype with the highest odds of match (top 1) belongs to the subject whose protein levels were observed, we consider this a match. We will also test whether the true match is among the three highest odds (top 3) and 1% highest odds (in top 1%).

7.a. Will the data be used for non-CVD analysis in this manuscript? __ Yes _X_ 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 ____ 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)?

MS#3324 Yu B, et al. Whole Genome Sequence and Proteomics for Gene Discovery in the Atherosclerosis Risk in Communities (ARIC) Study

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
_X_ A. primarily the result of an ancillary study (list number* _AS2017.27)
___ 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.

Yes, the lead author is aware that manuscript preparation is expected to be completed in 1-3 years, and if this expectation is not met, 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.

Yes, the lead author is aware of the policy.

References:

1. L. Sweeney, A. Abu, J. Winn, Identifying Participants in the Personal Genome Project by Name *CoRR*, (2013).
2. L. M. Beskow, Lessons from HeLa Cells: The Ethics and Policy of Biospecimens. *Annu Rev Genomics Hum Genet* **17**, 395-417 (2016).
3. E. E. Schadt, S. Woo, K. Hao, Bayesian method to predict individual SNP genotypes from gene expression data. *Nat Genet* **44**, 603-608 (2012).
4. W. Sun *et al.*, Common Genetic Polymorphisms Influence Blood Biomarker Measurements in COPD. *PLoS Genet* **12**, e1006011 (2016).
5. B. B. Sun *et al.*, Genomic atlas of the human plasma proteome. *Nature* **558**, 73-79 (2018).