ARIC Manuscript Proposal #3972

PC Reviewed: 12/14/21	**Status: _____**	**Priority: 2**
SC Reviewed: _____	**Status: _____**	**Priority: ____**

1.a. Full Title:
Associations between polygenic risk scores for common cancers and the plasma proteome in ARIC

b. Abbreviated Title (Length 26 characters):
Proteins and cancer risk scores

2. Writing Group:
Writing group members:
Jingning Zhang, Joe Coresh, Montserrat Garcia-Closas, Bing Yu, Eric Boerwinkle, Elizabeth Platz, Nilanjan Chatterjee. Others are welcome.

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

First author: **Jingning Zhang**
Address: Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
615 N. Wolfe Street, E3527
Baltimore, MD 21205

Phone: 443-739-3979
E-mail: jzhan218@jhu.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: **Nilanjan Chatterjee**
Address: Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
615 N. Wolfe Street, E3527
Baltimore, MD 21205

Phone: 410-955-3067
E-mail: nchatte2@jhu.edu

3. Timeline:

Analysis will begin immediately using visit 3 plasma protein data and focusing on identifying associations between plasma protein measurements and cancer polygenic risk scores in ARIC individuals (expected to be completed within 2 months). We will then explore direct association between identified proteins and risk of selected common cancers within ARIC (expected to be completed within another 2 months). The manuscript is expected to be completed within 6 months to 1 year.

4. Rationale:

Genomewide association studies for cancers to date have identified many common susceptibility variants associated with risks. While each individual SNP contributes only small amount of risk, polygenic risk score (PRS), which captures genetic burden of a disease associated with multiple risk variants, can explain significant heritability for many diseases including cancers. There is currently great interest in using PRS to improve disease risk prediction. Further, it has been proposed that PRS for a disease can be used identify mediating biomarkers of genetic association in a more powerful manner than that is possible based on individual weakly associated SNPs [1-3]. Recently, high-throughput technology, such as SOMAscan, has provided an opportunity to explore the role of proteins at large in mediating complex trait genetic association[4]. We propose to utilize recently available data on ~5000 SOMAscan from the large and diverse ARIC sample to explore relationship between cancer PRS and plasma proteome. Further, for identified protein-PRS associations, we will explore potential causal basis of them by using Mendelian Randomization approach where the causal effects of the PRS associated proteins on the risk of the underlying cancers will be tested using cis-pQTLs that we have recently characterized using the ARIC data itself [5]. Further, for the identified proteins, we propose to investigate direct association between the level of proteins and prospective risks within the ARIC study for common cancers that have sufficient number of cases (>100 cases). Different approaches have complementary strengths and thus we expect they together can help to triangulate potential causal proteins that mediates cancer genetic associations.

5. Main Hypothesis/Study Questions:
Identification proteins that may causally mediate associations between polygenic risk scores and cancers.

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).

Construction of PRS and Selection of Cancers
We will use data reported on PGS catalog [6] to construct polygenic risk scores across cancers. We will restrict analyses to cancers which has been associated with at least 10 independent common SNPs based on recent genome-wide association studies. Preliminary exploration indicates according to this criterion, we can include a total of 16 cancers in our PRS analysis. The list include basal cell carcinoma, glioma, melanoma, liver cirrhosis, chronic lymphoid leukemia and cancers of bladder, breast, cervix, colorectum, kidney cancer, lung, ovary, pancreas, prostate, and thyroid..

PRS by Proteome Analysis in ARIC

We will construct PRS for each cancer for individuals in ARIC using prespecified weights available from the PGS catalog curated from the latest literatures. We have previously developed a QC pipeline for cleaning, transforming and adjustment for known and hidden confounder for association analysis of the ARIC proteome data in a linear regression framework. We will use a similar pipeline to perform linear regression analysis of each individual protein on each individual cancer PRS. For each cancer, we will also adjust for additional known cancer-specific risk factors. For PRS corresponding to gender specific cancers, we will restrict the analysis to individuals of the same gender. For cancers that affect both sexes, we will pool data and use self-reported sex as a covariate. We will use the ARIC European American (EA) samples as the discovery set for conducting PRS by proteome association analysis. For any given cancer PRS and the proteins which reach proteome-wide significance (p-value < $3.7x10^{-5}$), we will replicate the association in the ARIC African American (AA) samples. Following are our inclusion/exclusion criterion and required ARIC variables for our primary analysis.

Inclusion/exclusion criteria:
ARIC participants with:

1) SomaLogic proteomic data at visit 3
2) Genotype imputed using Affy 6.0.
3) Exclude individuals with cancer diagnosis prior to visit 3

Data requirement

1) SomaLogic Proteomic data (visit 3)
2) Imputed genotype data
3) Prevalent and incident cancer outcomes
4) covariate data: age, sex, race, study site, smoking history, body mass index, other known cancer risk factors

Mendelian Randomization Analysis

A PRS by protein association may not necessarily indicate a causal role of the underlying protein on the risk of the cancer. In fact, coincidental associations may arise due to hidden confounders that may influence both cancer risk and the level of the proteins. Thus, for identified proteins in our primary analysis, we propose to further conduct Mendelian Randomization (MR) analysis to investigate whether these proteins can be causally linked to risk of the underlying cancers. In this step, we will leverage results from our own recent study [5] that has mapped cis-pQTLs for ~2000 genes based on comprehensive analysis of the ARIC data itself. Cis-pQTLs, which are less likely to be affected by horizontal pleiotropy, provide excellent instruments for robust MR analysis.

Direct Association Analysis of Protein and Cancer Risk within ARIC

For the identified proteins, we will also investigate direct association between protein levels and prospective cancer risk within the ARIC study. We will use the Cox proportional hazard regression model to relate visit 3 protein levels with post-visit 3 cancer risk after eliminating all follow-up times prior to visit 3. We will adjust for cancer specific risk factors as covariates. We will pool data across EA and AA populations with suitable covariate adjustment for race. As ARIC is a relatively small cohort study, the number of cancer events accrued to date is low for many

individual cancers and thus to ensure sufficient power we will restrict this analysis to only those cancers which has at least 100 cases (post visit-3).

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/mantrack/maintain/search/dtSearch.html

___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)?

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 https://www2.cscc.unc.edu/aric/approved-ancillary-studies

12a. Manuscript preparation is expected to be completed within one year. 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.

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

1. Choi, S. W., Mak, T. S. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols* **15**, 2759-2772 (2020).

2. Mooney, M. A. *et al*. Large epigenome-wide association study of childhood ADHD identifies peripheral DNA methylation associated with disease and polygenic risk burden. *Translational psychiatry* **10**, 1-12 (2020).

3. Fromer, M. *et al*. Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci.* **19**, 1442-1453 (2016).

4. Ritchie, S. C. *et al*. Integrative analysis of the plasma proteome and polygenic risk of cardiometabolic diseases. *BioRxiv* (2019).

5. Zhang, J. *et al*. Large Bi-Ethnic Study of Plasma Proteome Leads to Comprehensive Mapping of cis-pQTL and Models for Proteome-wide Association Studies. *bioRxiv* (2021).

6. Lambert, S. A. *et al*. The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. *Nat. Genet.* **53**, 420-425 (2021).