ARIC Manuscript Proposal #3754

PC Reviewed: 12/8/20	**Status: _____**	**Priority: 2**
SC Reviewed: _____	**Status: _____**	**Priority: ____**

1.a. Full Title:
Novel sparse canonical correlation analysis to identify trans-regulated protein networks associated with known complex trait variants.

b. Abbreviated Title (Length 26 characters):
Trans-regulated protein sets.

2. Writing Group:
Writing group members:
Diptavo Dutta, Jingning Zhang, Josef Coresh, Morgan Grams, Eric Boerwinkle, Alexis Battle, Nilanjan Chatterjee and others are welcome

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

First author: Diptavo Dutta
Address: Department of Biostatistics
E3001, 615 N Wolfe Street
Baltimore, MD 21205
Phone: 734-223-1009 Fax:
E-mail: diptavo21@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: Josef Coresh
Address: Department of Epidemiology, Biostatistics and School of Medicine
2024 E Monument Street, Room 2-635, Suite 2-600
Baltimore, MD 21287
Phone: 410-955-0495 Fax:
E-mail: coresh@jhu.edu

3. Timeline:
Analysis will begin immediately using protein abundance level data from visit 3 and genetic data. The manuscript is expected to be completed in 8-10 months.

4. Rationale:
Regulation of molecular phenotypes like gene expression or protein levels by genetic variants related to traits and diseases can illuminate the intermediate biological mechanisms and causal processes[1,2] underlying phenotypic change. Extensive research has been done to understand the

regulatory effects of local (cis) variation on protein levels but identifying distal (trans) regulation remains challenging due to statistical power[3]. In particular, detecting distal targets of known trait related variants can highlight their complex downstream effects and affected pathways for the trait[2,4,5]. **Power to detect such weaker distal effects can be increased by aggregating associations across groups of proteins and groups of known trait-related genetic variants in contrast to standard single variant-single protein analysis**. Our previous work in this regard has shown that such aggregative strategies can effectively identify trait-specific gene-networks whose expression levels are trans-regulated by known trait-related variants[6]. In this project, we will apply sophisticated statistical methods to effectively identify groups of target proteins that are trans-regulated by groups of known trait-related variants, representing trait-specific protein networks that mediate the effects of the trait-related genetic variants. From public repositories, we will curate list of known genetic variants that have been reported to be associated to different traits and diseases and identify the networks of proteins being trans-regulated by these variants using sparse canonical correlation approach. The identified protein networks will provide novel insights into the key target proteins and pathways which propagate the effects of variants related to different traits and diseases.

5. Main Hypothesis/Study Questions:

To identify trans-regulated protein sets or networks that mediate the effects of known genetic variants related to a trait. The identified protein networks can uncover trait-specific affected pathways and biological mechanisms.

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).

Study Design: Cross sectional analysis.
Inclusion criteria: Participants with protein abundance level data from SOMAscan and genetic data.
Outcome and Variables of interest: Human plasma protein concentrations measured by SOMAscan assay (Somalogic Inc., Boulder, CO). We will investigate an appropriate exclusion threshold for proteins that were flagged in at least one plate, and the appropriate transformation of the relative abundance of protein levels, such as logarithmic and/or inverse normal transformation.
We will curate a list of variants of interest, which will include variants that have been shown to be associated to various traits in GWAS[7] including but not limited to height, cholesterol levels, cardiometabolic traits, psychiatric traits, cancers, neurological traits, endocrinological traits, autoimmune diseases. We will use the genotype-call/dosages of genetic variants associated to different traits as the covariate of interest. We expect the number of variants of interest to be around 20,000.
Other Covariates: Known and available covariates such as age, sex, race, center, body mass index (BMI), current drinking, blood pressure and others. We shall further estimate other sources of variation and use them as covariates[8].
Broad analysis plan:

1. We will first identify genetic variants that have been associated with different traits using GWAS catalog and other publicly available resources. (Source: External GWAS repositories).
2. Next, we will extract the standard trans-summary statistics for these genetic variants across the available plasma proteome by using the standard linear regression model adjusting for covariates like age, sex, BMI, genetic principal components and other sources of variation. (Source: ARIC Proteomic and genetic data).
3. Using these summary statistics, we will use sparse canonical correlation-based models to identify protein networks associated with the groups and subgroups of the genetic variants.
4. We will use several downstream analyses to quantify whether the identified trans-regulated protein networks have additional evidence to be associated to the trait, using enrichment analysis, heritability analysis and network construction.
5. We will also perform the above analysis separated by ethnic groups to identify protein networks associated to different traits across ancestry. We will check for cross-replicability of the protein networks identified for each trait/disease for (dis-) similarities in the networks topology and identified proteins.

Our analysis is separate from and will be complementary to any ongoing standard trans-pQTL (protein quantitative trait loci) analysis in the fact that we aim to select the proteins networks associated to groups of known trait-related variants as opposed to single protein-single variant association conducted using standard analysis. Our method uses sparse canonical correlation analysis which is a novel approach in this respect and has demonstrated promise previously in similar trans-regulation analysis of gene-expression data for known trait-related variants[6].

7.a. Will the data be used for non-ARIC analysis or by a for-profit organization in this manuscript? ____ Yes __X__ No

b. If Yes, is the author aware that the current derived consent file ICTDER05 must be used to exclude persons with a value RES_OTH and/or RES_DNA = "ARIC only" and/or "Not for Profit"? ____ Yes __X__ No

(The 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 current derived consent file ICTDER05 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 https://sites.cscc.unc.edu/aric/approved-ancillary-studies

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. Battle A, Khan Z, Wang SH, et al. Impact of regulatory variation from RNA to protein. *Science (80-)*. 2015. doi:10.1126/science.1260793
2. Umans BD, Battle A, Gilad Y. Where Are the Disease-Associated eQTLs? *Trends Genet*. September 2020:1-16. doi:10.1016/j.tig.2020.08.009
3. Sun BB, Maranville JC, Peters JE, et al. Genomic atlas of the human plasma proteome. *Nature*. 2018;558(7708):73-79. doi:10.1038/s41586-018-0175-2
4. Võsa U, Claringbould A, Westra H-J, et al. Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *bioRxiv*. 2018. doi:10.1101/447367
5. Aguet F, Barbeira AN, Bonazzola R, et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *bioRxiv*. 2019. doi:10.1101/787903
6. Dutta D, He Y, Saha A, Arvanitis M, Battle A, Chatterjee N. *Novel Aggregative Trans-EQTL Association Analysis of Known Genetic Variants Detect Trait-Specific Target Gene-Sets*.; 2020. https://doi.org/10.1101/2020.09.29.20204388.
7. Buniello A, MacArthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*. 2019;47(D1):D1005-D1012. doi:10.1093/nar/gky1120
8. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc*. 2012. doi:10.1038/nprot.2011.457