Louis I. and Thomas D. Dublin Award Research Statement

Jingning Zhang, Advisor: Nilanjan Chatterjee

I am conducting dissertation research using genetic and proteomic data to understand the complex mechanism of human traits to build models for disease risk prediction. Below I describe my ongoing research in two major areas of genetic epidemiology, and a statement of how my research can benefit epidemiology and biostatistics.

1. Plasma proteome analyses

Benefit from the availability of recent high-throughput technologies for measuring proteins, great opportunities have arisen to substantially increase our understanding of the causal role of proteins in complex traits. By utilizing data from Atherosclerosis Risk in Communities (ARIC) study, I led a research project for a comprehensive set of analyses of *cis*-genetic regulation of the plasma proteome for over 10,000 European and African Americans. In this work, we proposed genetic prediction models for plasma proteome. By integrating those models with the genome-wide association studies (GWAS) of complex traits and diseases, people are able to perform proteome-wide association analysis (PWAS) and thus identify associated plasma proteins that might be potential drug target. This paper was published in *Nature Genetics* [Link], and the accompanying website (http://nilanjanchatterjeelab.org/pwas) containing *cis*-GWAS, fine-mapping, and the proposed genetic prediction models for plasma proteins has become a valuable resource and been widely used for proteome-related biostatistics and epidemiology research.

Study of genetic architecture of plasma proteins. We carried out a set of association and fine-mapping analyses to identify common *cis*- protein quantitative trait locus (pQTL) and compare results across ancestries to explore shared and unique genetic architecture. We further showed the benefit of integrating multi-ancestry data to better identify causal variants. In addition, we performed colocalization analysis for identified *cis*-pQTL and curated *cis*- expression quantitative trait locus (eQTL), and observed large overlap which indicates the shared underlying causal variants of proteome and transcriptome.

Identification of diseases-associated plasma proteins. There are many large-sample size GWAS for complex disease. However, an ultimate translational goal is to identify causal genes and thus inform potential therapeutic targets for those diseases. For each population, we built models for genetically predicting levels of plasma proteins. By integrating these models to disease GWAS, we can then conduct PWAS to find diseases-associated plasma proteins, study their mediating role behind the genetic-disease associations, and thus identify potential therapeutic target proteins. In the application of PWAS to the complex disease gout, we identified an associated protein, IL1RN, which had been a drug target for treating another disease. Independent of our research, there was an ongoing trial of gout showing this drug to be as good as usual treatment, which suggested the promise of the drug repurposing potential of our proposed models and PWAS.

2. Polygenic risk prediction for complex human traits and diseases

The proposal of polygenic risk scores (PRS) provides a quantitative measurement of the total genetic risk assessed simply by the genotypes. It raises the hope to prevent and screen diseases by simple genetic testing, and promote precision medicine. My comprehensive research work on PRS includes both methodology and application.

Methodology -- multi-ancestry polygenic risk prediction. PRS are commonly built from GWAS summary statistics. However, existing large GWAS have been primarily conducted in European population (EUR), and studies have shown that the transferability of EUR PRS is poor to non-EUR populations, which exaggerates the health disparities in clinical applications. To address these issues, I proposed a novel method for generating multi-ancestry Polygenic Risk scOres based on enSemble of PEnalized Regression models (PROSPER) for predicting complex traits/disease risks by integrating

summary statistics from GWAS and individual-level reference data. Results in simulation and real data analyses show that PROSPER can substantially improve multi-ancestry polygenic prediction compared to alternative methods across a wide variety of genetic architectures. In real data analyses, for example, PROSPER increased out-of-sample prediction R^2 for a variety of continuous traits from the datasets of Global Lipids Genetics Consortium (GLGC) and All of Us (AoU) by an average of 81.5% compared to a state-of-the-art Bayesian method (PRS-CSx) in the African ancestry population. Further, PROSPER is computationally highly scalable for the analysis of large genetic variants and many diverse populations. The algorithm is wrapped in a command-line tool (https://github.com/Jingning-Zhang/PROSPER). This paper has won the *2023 JSM student paper award* in the Risk Analysis Section, and is ready to be submitted to *Nature Methods* [Slides].

Application -- proteomic mediation of genetic risks of cancers. 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 genetic variants. We proposed to use PRS in a triangular analysis to identify major cancer proteins through which the genetic risk of cancers may be mediated. We associated PRS to the proteome in ARIC cohort and partitioning the associations into effects from *cis* and *trans*, and further performed the analysis of Aggregative tRans assoCiation to detect pHenotype specIfic gEne-sets (ARCHIE). We performed the association analysis for more than 20 common cancers, and identified multiple potential trans-regulated cancer proteins. Examples include CD72 and CXCL2 for Hodgkin's lymphoma. This manuscript is under preparation.

3. Benefits for Epidemiology and Biostatistics

In the first work on plasma proteome analyses, I have made several innovative methodological contributions to biostatistics and epidemiology. I showed that probabilistic estimation of expression residuals (PEER) factors can be used in high throughput protein measurements to improve statistical power for pQTL identification. This can benefit future biostatistics research conducted using similar high throughput measuring technologies. I also developed a joint conditional regression analysis for the association of a trait with genetic scores associated with gene expression and protein levels using only GWAS summary statistics. The proposed PWAS framework can be used by epidemiologists to easily identify plasma proteins and study potential causal pathways associated with traits/disease risks. This helps the study of disease mechanism and novel drug targeting. The accompanying reference table for existing protein-targeted drugs further provides valuable resource for the identification of potential drug repurposing opportunities.

In the second work of the methodology development of multi-ancestry PRS, I have made significant contribution for improving the predictive accuracy of complex trait and diseases. I proposed a novel multi-ancestry PRS method, PROSPER, that uses a penalized regression to integrating data from multi-ancestry summary statistics. PROSPER has substantial improvement of predictive accuracy especially for minority populations, which could contribute to alleviating disparity in the performance across ancestry groups and hence help to bring more equity in genetic epidemiology research. There is especially significant benefit for the African ancestry population where risk prediction remains the most challenging.

In the third work, we proposed to use PRS as a genetic instrument to discover key proteins for more than twenty common cancers. Access to data is crucial for initiating research in biostatistics and epidemiology. However, obtaining individual-level genetic data for cancer research is often difficult due to privacy concerns. PRS is a valuable and user-friendly tool that can be applied to summary-level data, and provides a summary of genetic risk. We integrated the data of plasma proteome and PRS, and proposed a triangular analysis framework. By further combining other complementary analysis strategies, including ARCHIE, it is likely to identify causal proteins that mediate genetic associations from either *cis*- or *trans*-regulation for common cancers, which could provide a comprehensive framework and contribute to cancer epidemiology research.