Proteome-Wide Association Study for Cardiovascular Diseases and Related Traits in the Women's Health Initiative Study

Brian Chen[1*], Chanhwa Lee[1*], Amanda Tapia[1], Alexander P. Reiner[2], Hua Tang[3], Charles Kooperberg[4], Yun Li[1,5], Laura M. Raffield[5]

(1) Department of Biostatistics, at University of North Carolina, Chapel Hill (2) Department of Epidemiology, University of Washington (3) Department of Genetics, Stanford University School of Medicine (4) Division of Public Health Sciences, Fred Hutchinson Cancer Center (5) Department of Genetics, University of North Carolina, Chapel Hill, *contributed equally to this work

Introduction

In Proteome-Wide Association Studies (PWAS), genetic variants near the corresponding protein-coding gene (+/- 1 Mb of encoding region), also known as local SNPs, are used to predict the protein level[1]. These predicted protein levels are then associated with phenotypes of interest.

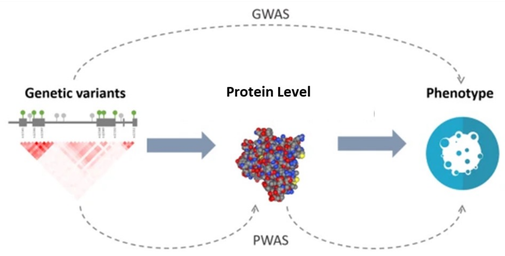


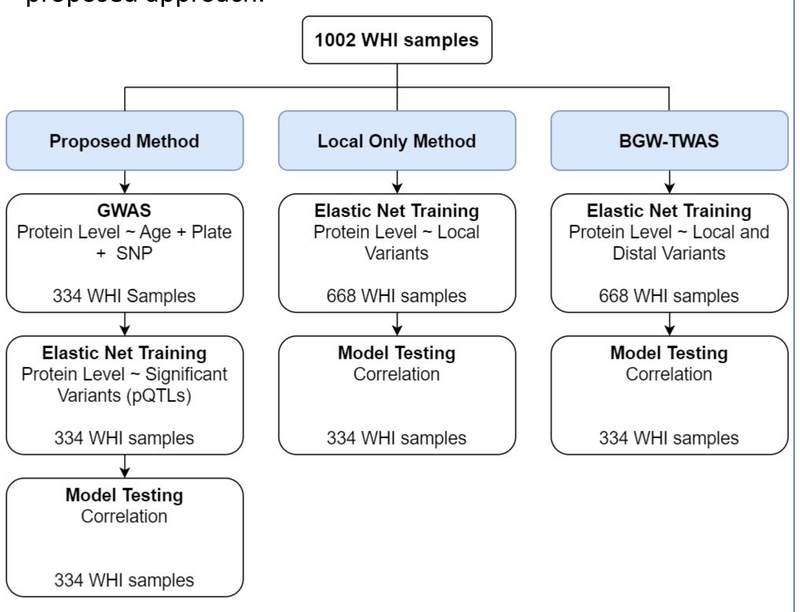
Figure from Brandes *et al.* [2]

However, proteins can be regulated through variants outside of the nearby region[3]. Thus, we propose an intermediate GWAS step to select for protein quantitative trait loci (pQTL), which can be from anywhere in the genome, to select variants for consideration in PWAS model training. This allows for the inclusion of SNPs outside of the protein coding gene region, called distal SNPs, when creating elastic net models for protein level prediction.

Methods

Using the Women's Health Initiative (WHI) proteomic data, the sample included 1002 unique participants with complete cases and passed quality control for 552 different protein level data.

- Used predetermined p-value thresholds to classify significant variants defined as pQTLs, and then used in Elastic net training
- Compared this to using an elastic net with only local SNPs and BGW-TWAS[4] (a TWAS method that handles distal and local SNPs) using the testing r^2 of predicted protein level using our proposed approach.



Protein levels were genetically predicted for 8131 WHI individuals without measured proteomics data using the prediction weights from proposed method

- Associated predicted proteins with different lipid traits controlling for top PCs and age
- Compared the association results from predicted protein levels with the association results measured protein levels

Discussion

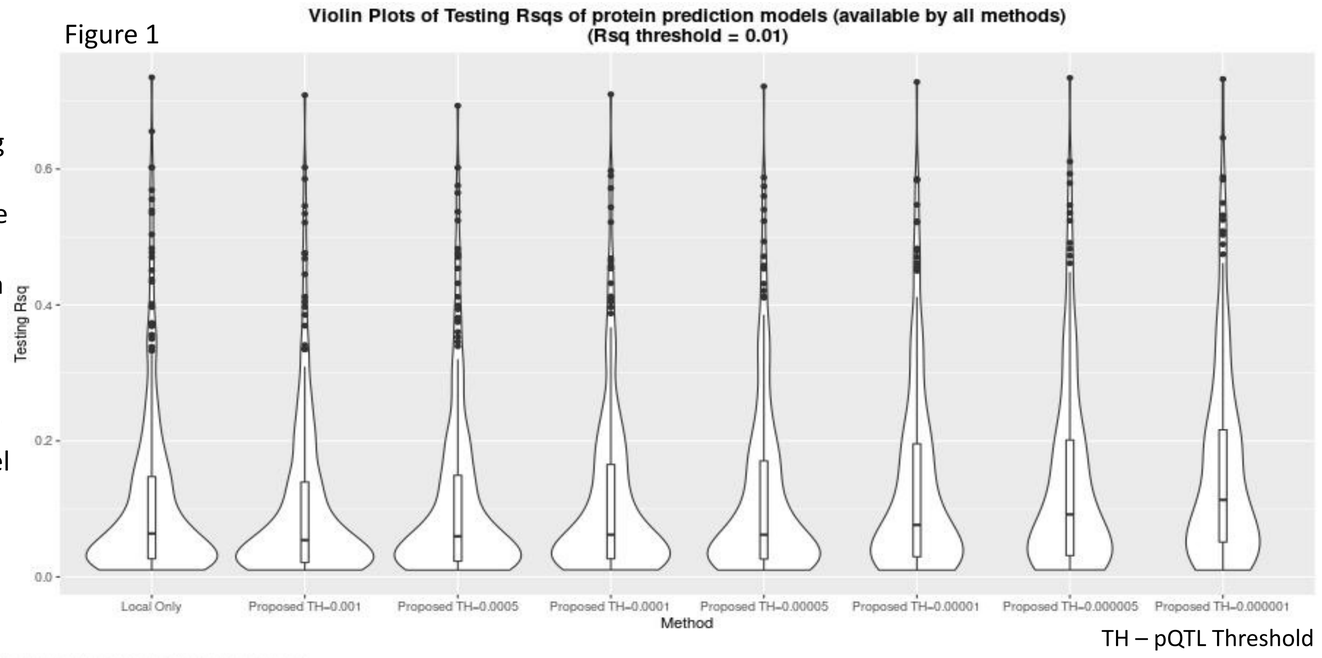
Overall, our method shows similar prediction results across different pQTL thresholds, and similar results when compared to local only method and BGW-TWAS. Moreover, we have a improvement in r^2 over the local-only method for a subset of proteins when the top GWAS pQTL was outside coding region, despite decrease in sample size for model training. Likely due to the fact that the genetic component only explains a portion of the protein level and a small sample size in WHI alone, we only have a select few predicted proteins that are significant when associated with lipid outcomes. However, with rapidly growing proteomic and genomic datasets, our results suggest inclusion of both local and distal SNPs will improve protein prediction.

Acknowledgements

The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts 75N92021D00001, 75N92021D00002, 75N92021D00003, 75N92021D00004, 75N92021D00005.

Results

Figure 1



TH – pQTL Threshold

Figure 1 shows violin plots the distribution of r^2's with a minimum of 0.01. Comparing the testing r^2 between different pQTL thresholds, we see a similar trend in distribution of r^2 values. As a result, we have decided to use a pQTL threshold of 0.0001 in our downstream analysis to maximize number of proteins with a valid model while minimizing computational costs for model fitting.

Figure 2

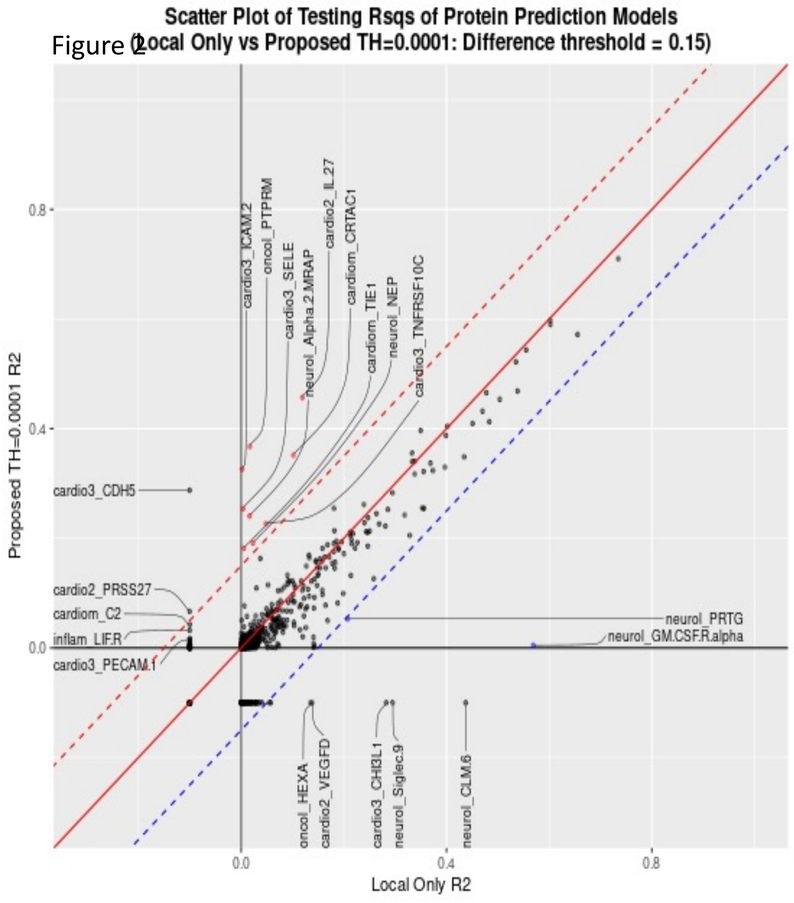


Figure 2 is a scatterplot plot of the local method r^2 against the proposed method with a pQTL threshold of 0.0001. A majority of proteins showed similar r^2 values between the two methods, with a r^2 of 0.9365 between the r^2 values. The dotted lines indicate a difference of 0.15 in the r^2 value. The negative values indicate that the model did not converge for that protein.

Table 1 shows a subset of proteins that were shown to have a higher correlation between predicted and measured protein level in our proposed method than the local only method and BGW-TWAS (with the exception of Tumor necrosis factor receptor superfamily member 10C).

Table 2 shows top associated predicted proteins (that had model $r^2 > 0.05$) with lipid outcomes. Our of 552 proteins, 96 were not imputed because of failed EN training or less than 2 SNPs included in EN training. Therefore, we have a Bonferroni corrected p-value of 0.05/456 = 1.096E-4.

Table 1

Protein	Proposed r^2†	Local-only r^2	BGW-TWAS r^2	Coding region	Top GWAS pQTL	Top GWAS pQTL Gene
Interleukin-27	0.456	0.119	0.108	16:28499362-28512051	19:4236999:G:A	*EBI3*
E-selectin	0.253	0.003	0.010	1:169722640-169764705	9:133273983:A:G	*ABO*
Tumor necrosis factor receptor superfamily member 10C	0.226	0.048	0.253	8:23102921-23117445	19:43648948:A:G	*PLAUR*
Cartilage acidic protein 1	0.351	0.101	0.165	10:97865000-98030828	3:186675758:T:C	*HRG, LOC105374258*
Alpha-2-macroglobulin receptor-associated protein	0.240	0.016	- ‡	4:3563612-3532446	6:31322175:G:A	Non-coding (HLA-B region)*

† pQTL threshold of 1E-4 was used ‡ BGW-TWAS method did not converge

Table 2

Outcome	Protein	Predicted Protein Level Estimate	Predicted Protein Level P-value	Model r^2	Measured Protein Level Estimate	Measured Protein Level P-Value
Cholesterol	Progranulin	11.520	1.60E-06	0.412	-2.803	0.017
Triglycerides	Angiopoietin-related protein 3	20.399	4.89E-05	0.201	8.832	1.53E-07
LDL	Neural cell adhesion molecule L1	-9.560	7.32E-05	0.184	0.690	0.507
LDL	E-selectin	-3.103	9.07E-05	0.462	-0.412	0.695
HDL	Phospholipid transfer protein	3.403	1.48E-4	0.161	6.016	4.21E-43
LDL	Galectin-4	-7.767	1.81E-4	0.120	-6.497	3.79E-10
Platelet	Granulocyte Colony-Stimulating Factor	22.302	2.43E-4	0.162	0.978	0.604
Hemoglobin	Matrix metalloproteinase-9	-0.624	2.75E-4	0.170	0.308	5.85E-15
Hematocrit	Fibroblast growth factor 5	0.363	4.07E-4	0.275	-0.038	0.747
LDL	C-X-C motif chemokine 1	-15.076	5.31E-4	0.176	1.001	0.338

References

[1] Schubert R., Geoffroy E., Gregga I., *et al*. Protein prediction for trait mapping in diverse populations. *PLoS One*. (2022).
[2] Brandes, N., Linial, N. & Linial, M. PWAS: proteome-wide association study—linking genes and phenotypes by functional variation in proteins. *Genome Biol* **21**, 173 (2020).
[3] Suhre, K., McCarthy, M.I. & Schwenk, J.M. Genetics meets proteomics: perspectives for large population-based studies. *Nat Rev Genet* **22**, 19–37 (2021).
[4] Luningham J.M., Chen J., Tang S. *et al*. Bayesian Genome-wide TWAS Method to Leverage both cis- and trans-eQTL Information through Summary Statistics. *Am J Hum Genet*. (2020).

Table 1

Protein	Proposed r^2†	Local-only r^2	BGW-TWAS r^2	Coding region	Top GWAS pQTL	Top GWAS pQTL Gene
Interleukin-27	0.456	0.119	0.108	16:28499362-28512051	19:4236999:G:A	*EBI3*
E-selectin	0.253	0.003	0.010	1:169722640-169764705	9:133273983:A:G	*ABO*
Tumor necrosis factor receptor superfamily member 10C	0.226	0.048	0.253	8:23102921-23117445	19:43648948:A:G	*PLAUR*
Cartilage acidic protein 1	0.351	0.101	0.165	10:97865000-98030828	3:186675758:T:C	*HRG, LOC105374258*
Alpha-2-macroglobulin receptor-associated protein	0.240	0.016	- ‡	4:3563612-3532446	6:31322175:G:A	Non-coding (HLA-B region)*

† pQTL threshold of 1E-4 was used ‡ BGW-TWAS method did not converge

Table 2

Outcome	Protein	Predicted Protein Level Estimate	Predicted Protein Level P-value	Model r^2	Measured Protein Level Estimate	Measured Protein Level P-Value
Cholesterol	Progranulin	11.520	1.60E-06	0.412	-2.803	0.017
Triglycerides	Angiopoietin-related protein 3	20.399	4.89E-05	0.201	8.832	1.53E-07
LDL	Neural cell adhesion molecule L1	-9.560	7.32E-05	0.184	0.690	0.507
LDL	E-selectin	-3.103	9.07E-05	0.462	-0.412	0.695
HDL	Phospholipid transfer protein	3.403	1.48E-4	0.161	6.016	4.21E-43
LDL	Galectin-4	-7.767	1.81E-4	0.120	-6.497	3.79E-10
Platelet	Granulocyte Colony-Stimulating Factor	22.302	2.43E-4	0.162	0.978	0.604
Hemoglobin	Matrix metalloproteinase-9	-0.624	2.75E-4	0.170	0.308	5.85E-15
Hematocrit	Fibroblast growth factor 5	0.363	4.07E-4	0.275	-0.038	0.747
LDL	C-X-C motif chemokine 1	-15.076	5.31E-4	0.176	1.001	0.338