Evaluation of PrediXcan for prioritizing GWAS associations and predicting gene expression[1*]

Binglan Li[1], Shefali S. Verma[1,2], Yogasudha C. Veturi[2], Anurag Verma[1,2], Yuki Bradford[2], David W. Haas[3,4] and Marylyn D. Ritchie[1,2]

[1]The Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA; [2]Biomedical and Translational Informatics Institute, Danville, PA; [3]Department of Medicine, Pharmacology, Pathology, Microbiology & Immunology, Vanderbilt University School of Medicine, Nashville, TN; [4]Department of Internal Medicine, Meharry Medical College, Nashville, TN, USA

Genome-wide association studies (GWAS) have been successful in facilitating the understanding of genetic architecture behind human diseases, but this approach faces many challenges. To identify disease-related loci with modest to weak effect size, GWAS requires very large sample sizes, which can be computational burdensome. In addition, the interpretation of discovered associations remains difficult. PrediXcan was developed to help address these issues. With built in SNP-expression models, PrediXcan is able to predict the expression of genes that are regulated by putative expression quantitative trait loci (eQTLs), and these predicted expression levels can then be used to perform gene-based association studies. This approach reduces the multiple testing burden from millions of variants down to several thousand genes. But most importantly, the identified associations can reveal the genes that are under regulation of eQTLs and consequently involved in disease pathogenesis. In this study, two of the most practical functions of PrediXcan were tested: 1) predicting gene expression, and 2) prioritizing GWAS results. We tested the prediction accuracy of PrediXcan by comparing the predicted and observed gene expression levels, and also looked into some potential influential factors and a filter criterion with the aim of improving PrediXcan performance. As for GWAS prioritization, predicted gene expression levels were used to obtain gene-trait associations, and background regions of significant associations were examined to decrease the likelihood of false positives. Our results showed that 1) PrediXcan predicted gene expression levels accurately for some but not all genes; 2) including more putative eQTLs into prediction did not improve the prediction accuracy; and 3) integrating predicted gene expression levels from the two PrediXcan whole blood models did not eliminate false positives. Still, PrediXcan was able to prioritize GWAS associations that were below the genome-wide significance threshold in GWAS, while retaining GWAS significant results. This study suggests several ways to consider PrediXcan's performance that will be of value to eQTL and complex human disease research.

Keywords: PrediXcan; GWAS; prioritization; prediction accuracy.

[1*] The project described was supported by Award Number U01AI068636 from the National Institute of Allergy and Infectious Diseases (NIAID) and supported by National Institute of Mental Health (NIMH), National Institute of Dental and Craniofacial Research (NIDCR). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Allergy and Infectious Diseases or the National Institutes of Health. This work was supported by the AIDS Clinical Trials Group funded by the National Institute of Allergy and Infectious Diseases (AI068636, AI038858, AI068634, AI038855). Additional grant support included AI077505, AI069439, TR000445, AI054999, and AI110527.

Clinical Research Sites that participated in ACTG protocol A5202, and collected DNA under protocol A5128, were supported by the following grants from NIAID: AI069477, AI027675, AI073961, AI069474, AI069432, AI069513, AI069423, AI050410, AI069452, AI069450, AI054907, AI069428, AI069439, AI069467, AI045008, AI069495, AI069415, AI069556, AI069484, AI069424, AI069532, AI069419, AI069471, AI025859, AI069418, AI050409, AI069423, AI069501, AI069502, AI069511, AI069481, AI069465, AI069494, AI069472, AI069470, AI046376, AI072626, AI027661, AI034853, AI069447, AI032782, AI027658, AI-27666, AI058740, and AI046370, and by the following grants from the National Center for Research Resources (NCRR): RR00051, RR00046, RR025747, RR025777, RR024160, RR024996, RR024156, RR024160, and RR024160. Study drugs were provided by Bristol-Myers Squibb Co., Gilead Sciences, and GlaxoSmithKline, Inc.

1. Introduction

Genome-wide association studies (GWAS) have successfully identified disease susceptibility loci for complex traits. Yet, disease related loci discovered to date explain a small portion of the variance in disease risk[1]. It is not known whether the missing heritability is predominantly driven by variants with small effect sizes or by causal factors beyond genic regions. As a consequence, GWAS have relied on increasing sample size which increases the power to find disease-related loci and provides opportunities for rare variant analysis. However, analysis based on larger datasets consume an excessive amount of computational resources, which may not be available to everyone. The excessive number of single nucleotide polymorphism (SNP) loci in comparison to sample size leads to "the curse of dimensionality"[2]. Moreover, loci in intergenic regions may be robustly associated with complex traits, but the mechanisms behind such associations are generally not apparent.

Researchers have been trying to integrate functional genomics into GWAS in the anticipation that mechanistic studies of complex diseases will be facilitated by better interpretation of identified associations[3-6]. Much attention has been paid to the study of regulatory elements that change genes' transcriptional activities and consequently alter phenotypes. Expression quantitative trait loci (eQTLs) are one important class of such regulatory elements[7]. The Genotype-Tissue Expression (GTEx) Project[8] was initiated to identify a comprehensive set of eQTLs from different human tissues and their relationship to gene expression.

PrediXcan[9] is a computational algorithm developed to exploit GTEx data, including eQTLs identification and their relationship to complex traits. PrediXcan evaluates the aggregate effects of cis-regulatory variants (within in 1MB upstream or downstream of genes of interest) on gene expression via an elastic net regression method, and consequently, PrediXcan may identify loci with modest to weak effect sizes that do not achieve significance in variant-based association studies. In theory, PrediXcan has a greatly reduced multiple testing burden as compared to single-variant-single-trait association tests. For example, given one trait and a genotypic dataset of 10 million SNPs, there are at most about 20,000 tests for PrediXcan (~20,000 genes), but 10 million tests for single-variant-single-trait association study. Putative eQTLs and their effect sizes on gene expression level in each GTEx tissue type are available online in PredictDB (http://predictdb.org/).

Several cases have been recently identified where eQTLs are likely to play a causal role in disease by regulating gene expression[26,27]. But while more eQTLs have been identified in recent years, it remains challenging to prioritize the 'true' causal variants. Thus, as PrediXcan is designed to predict gene expression levels and prioritize GWAS results, PrediXcan can also be of great use for mechanistic studies. Here, PrediXcan performance was examined by two datasets where the PrediXcan whole blood models, the most similar tissue type to the samples, were used. One is the genotypic and transcriptomic data of the Yoruba (YRI) cohort from the 1000 Genomes Project[10]. While perhaps not the optimal dataset, it is very accessible which makes it convenient for readers to replicate this study. The other is based on the AIDS Clinical Trials Group (ACTG) protocol A5202[11,12,24], which we refer to as the A5202 cohort hereafter. A5202 cohort has a large enough sample size for evaluating the association tests (see methods) and has underwent a thorough variant-based association study[24] to compare with. To test prediction accuracy, PrediXcan's predicted gene expression levels were compared to the actual gene expression levels measured in the YRI cohort.

We also investigated possible influential factors and filter criterion to increase the possibility of identifying true predictions. As for GWAS prioritization, we carried out a transcriptome-wide association study (TWAS) based on PrediXcan predictions to obtain gene-trait associations and evaluate whether these associations prioritized the GWAS results. Our study provides insight into PrediXcan's capabilities and more importantly eQTL relationships to molecular phenotypes and disease traits, which is of great value in studying transcriptional regulation and disease pathogenesis.

2. Methods & materials

2.1. *Data preparation*

The YRI cohort from the 1000 Genomes Project was used to evaluate the prediction accuracy of PrediXcan for gene expression levels. The YRI cohort comprises 75 individuals. All specimens and 4,395,198 variants passed genotype quality control (based on Hardy-Weinberg Equilibrium ($P > 0.05$) and minor allele frequency (MAF) > 5%). From these 75 individuals, gene expression levels of 23,723 genes in RPKM (Reads Per Kilobase of transcript per Million mapped reads) were provided by the 1000 Genomes Project.

Another 1000 Genomes Project cohort, the Northern Europeans from Utah (CEU) cohort, was also included in this experiment to perform some components of the prediction accuracy test. The CEU cohort comprised 72 individuals and 3,660,275 variants after quality control (Hardy-Weinberg Equilibrium ($P > 0.05$) and MAF > 5%). But since the CEU cohort is part of the Depression Genes and Networks (DGN) cohort that was used to construct the DGN whole blood model by PrediXcan, we did not apply the DGN model to predict expression for the CEU cohort. This is the primary rationale for selecting the YRI cohort for our analyses.

Genotypic and phenotypic data from the A5202 cohort (data based on ACTG protocol A5202[11,12,24]) were used to evaluate PrediXcan's ability to prioritize GWAS results. The A5202 cohort comprises 47% European, 26% African, and 25% Hispanic Americans according to self-reported race or ethnicities. A5202 genotype and imputed data have been previously studied and reported[24]. Imputed genotypic data was quality checked using PLINK and non-ambiguous-stranded variants with imputation score > 0.7, MAF > 1%, and in Hardy-Weinberg Equilibrium ($P > 0.05$) were retained, resulting in 1221 individuals and 5,091,820 variants. Phenotypic data contained 690 continuous traits, which were based on laboratory assay results from HIV-infected patients before and after initiating antiretroviral therapy. The 690 traits were derived from plasma atazanavir pharmacokinetics, plasma efavirenz pharmacokinetics, change in CD4+ T-cell count, fasting low-density lipoprotein (LDL)-cholesterol, and fasting triglyceride data. Details about population structures, phenotypes, genotypes, and GWAS strategy are described elsewhere[24].

2.2. *Heritability Estimation*

To obtain the upper bound of how well a gene expression level can be predicted using genotypic data, we estimated the narrow-sense heritability between SNP variants and gene expression levels. Restricted maximum likelihood (REML) analysis was performed using GCTA[13] for each gene that

is included in both the PrediXcan models and the YRI cohort's gene expression data. Variant-gene relationships were retrieved from the weights table in the PrediXcan models so as to use the same exact set of variants for heritability and prediction accuracy estimations.

2.3. *Performance of gene expression prediction*

PrediXcan provides tissue-specific genotype-expression models, including 44 tissues from GTEx and 1 tissue (whole blood) from DGN[14]. As the 1000 Genomes project uses cultured cell lines derived from blood for genotypic and transcriptomic data, GTEx whole blood and DGN whole blood models were analyzed with the genotypic data from the YRI cohort to predict gene expression levels. The square of Pearson correlation (R^2) between predicted and observed gene expression levels was calculated to measure prediction accuracy. To assess directionality, the Pearson Correlation Coefficient (PCC) between predicted and actual gene expression levels was calculated and is called directionality estimates in the following context. For example, PCC is positive when predicted and observed gene expression levels both increase or decrease at the same time; PCC is negative when the predicted and observed directions are discordant. Of note, some genes had flat predicted gene expression levels across individuals whose genotypes differed. Standard deviations for these predicted gene expression levels were 0, which forced these genes to be dropped from the prediction estimation using R^2 or PCC.

To test which factors influence PrediXcan's prediction accuracy, we examined relationships between a few different model characteristics and accuracy estimates (R^2). For each predicted gene expression level, we evaluated whether the prediction accuracy is influenced by the following model characteristics: 1) the number of variants, 2) the number of variants adjusted by gene length, 3) the percentage of variants over the number of all variants in a PrediXcan model used, and 4) choice of PrediXcan models (tissue specific models). Gene length was annotated using Biofilter[25].

2.4. *Filtering for possibly more accurately predicted genes*

In most experimental data analyses, we have either genotypic data or transcriptomic data, but not both, to perform GWAS or TWAS (see method 2.5. for details). Thus, it is unlikely that we can estimate prediction accuracy or genotype-expression heritability and accordingly select more accurately predicted genes for downstream analyses. To address this issue, we explored whether it is possible to filter the gene list for a subset of more accurately predicted genes without prior knowledge of actual gene expression levels. The filter criterion we tested was based on the similarity of the predicted gene expression levels from the two whole blood models, GTEx and DGN, as predictions from different models will be the easiest to obtain for every PrediXcan users. PCC was used to measure the similarity between prediction results.

2.5. *GWAS Prioritization*

In addition to predicting gene expression levels for individuals who have SNP data but no gene expression data, we also tested PrediXcan's ability to prioritize GWAS results. Some SNP loci may be omitted from mechanistic studies because they only have modest to weak impact on traits and

thus the association signals are not strong enough to pass the multiple testing thresholds set by GWAS or phenome-wide association studies (PheWAS[15]). We were interested in whether PrediXcan could prioritize such association signals. Thus, we carried out PrediXcan followed by TWAS and compared the association hits to PheWAS (since we had multiple phenotypes). To obtain gene-trait association p-values, PrediXcan GTEx whole blood model was applied to the genotypic data from ACTG A5202 to predict gene expression levels. Then predicted gene expression levels and 690 traits were used to perform phenome-wide TWAS via PLATO[16]. Sex, age, and the first three principal components were used as covariates to adjust for sampling biases and underlying population structure. As for variant-trait association studies, to reduce computation time and burden, we only explored the variants within and close to (1MB upstream or downstream) the PrediXcan-TWAS significant genes (Bonferroni-corrected $P < 0.05$). Filtering of variants was done using Biofilter[17]. The criterion of vicinity was in accordance with the region window used by PrediXcan for expression prediction. We then carried out PheWAS using PLATO on the PrediXcan significant traits and the variants nearby PrediXcan significant genes. The association p-values of PrediXcan-TWAS and PheWAS were visualized using *ggplot2*[18] in R.

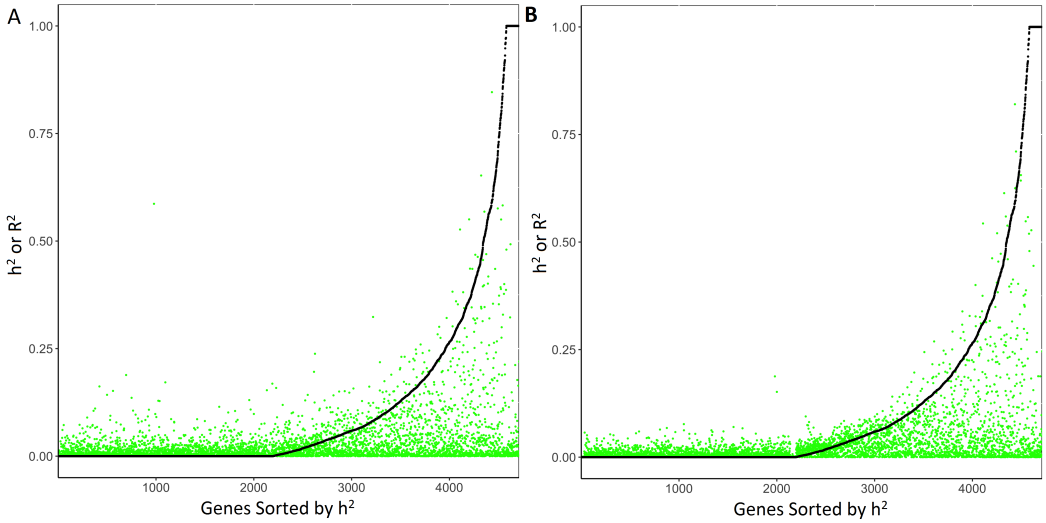

Fig. 1. Prediction performance of DGN (**A**) and GTEx (**B**) whole blood tissue model on the YRI cohort. DGN and GTEx whole blood tissue models were applied to the genotypic data from the YRI cohort. Prediction accuracy (R^2 of predicted versus observed gene expression levels; green) was compared to the narrow-sense heritability (h^2) estimates (black).

3. Results

3.1. *Prediction accuracy*

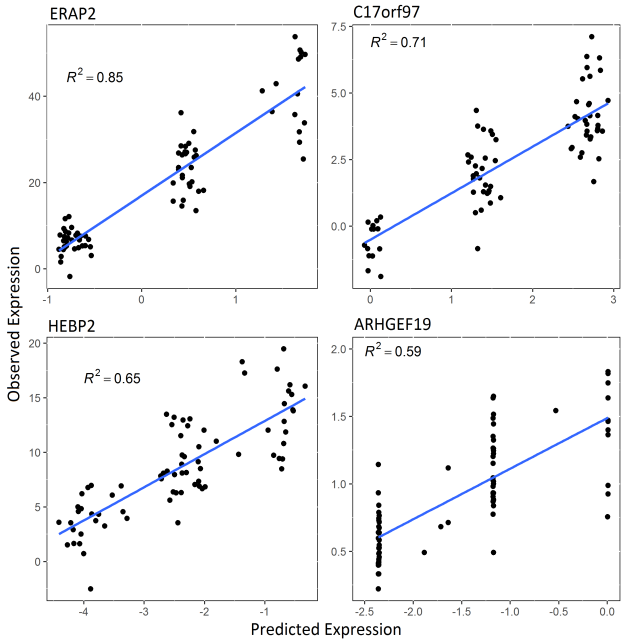

Fig. 2. Examples of well-predicted genes. These plots show the top four performing genes based on PrediXcan's prediction accuracy. Predicted gene expression levels were generated using the DGN whole blood model. Observed expression levels (in RPKM) for the YRI cohort were provided the 1000 Genome Project.

Using the genotypic data of the YRI cohort, the PrediXcan DGN and GTEx tissue models predicted expression of 11,538 and 6,695 genes, respectively. Prediction performance was evaluated using PCC and R^2 for 10,387 DGN genes and 6,127 GTEx genes, respectively (see method 2.3 for why some genes did not have estimates and the justification of using PCC and R^2). Due to the finite number of genes that were common to both models and transcriptomic data, heritability estimation was limited to 4,711 genes.

We first evaluated how well PrediXcan predictions capture the regulatory effects of variants on gene expression levels (**Fig. 1**). We found that genes with higher expression heritability were more likely to have higher R^2 values than genes with lower expression heritability. These results are consistent with what has been published in initial PrediXcan paper[9]. In theory, the better PrediXcan performs at capturing additive regulatory effects imposed by variants, the closer h^2 estimates (black line) and R^2 (green dots) should be, which was what we observed for the genes whose expression levels were influenced by genetic factors ($h^2 > 0$). These results (**Fig. 1**) suggest that PrediXcan predictions were able to capture the transcriptome/gene expression level variability.

We next sought out to evaluate PrediXcan's prediction accuracy. We found that PrediXcan's DGN and GTEx model had similar performance in predicting of gene expression. As indicated in the initial PrediXcan paper[9], PrediXcan precisely predicted gene expressions for some genes (DGN results shown in **Fig. 2**, GTEx results in supplementary figure 2), but prediction accuracy was overall unsatisfactory as most genes had accuracy estimates near 0 (**Fig. 1**). For the two whole blood models, the directionality estimates centered on zero with a small standard deviation, which suggested that most predicted gene expression levels did not correlate with the observed gene expression levels (**Fig. 3**). The GTEx model on the CEU cohort from 1000 Genomes Project performed similarly, with mean of -0.067 and variance of 0.03 (supplementary figure 3). In addition, for all three tests, about one-half of all predictions had negative correlation between predictions and observed values, which made interpretation difficult. In short, based on our evaluation, PrediXcan did not predict gene expression well when DGN and GTEx models were used as training sets to predict gene expression levels in YRI and CEU cohorts. While this finding may not be surprising, many researchers have assumed that PrediXcan could be used for this purpose. Thus, this examination was worthwhile.

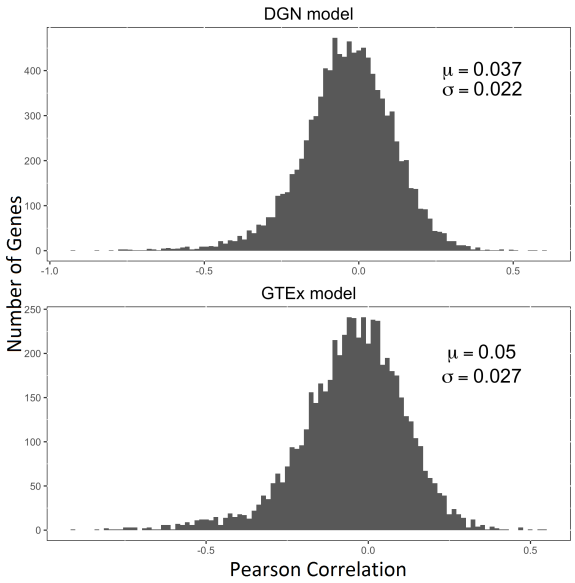

Fig. 3. Performance of prediction directionality of PrediXcan models, DGN (top) and GTEx (bottom), on the YRI cohort. Directionality was computed between predicted and observed gene expression levels.

Next, we examined factors that were responsible for predicting gene expression and more importantly which factors could improve the prediction performance of PrediXcan. We first evaluated whether prediction performance was dependent on specific model properties. For example, would prediction accuracy for a certain gene improve if more variants were included in the input genotypic data for expression prediction? To address this possibility, we explored the relationships between the prediction accuracy and three model properties: 1) the number of model variants used for prediction (**Fig. 4A**); 2) the percentage of the model variants used for prediction (**Fig. 4B**); and 3) the number of model variants used with adjustment for gene length (**Fig. 4C**). A slight improvement in prediction accuracy was apparent in these scatterplots when more variants were taken into account to predict gene expression levels. However, relationships were so weak that these model properties could not be used to favorably assess or improve PrediXcan's prediction performance.

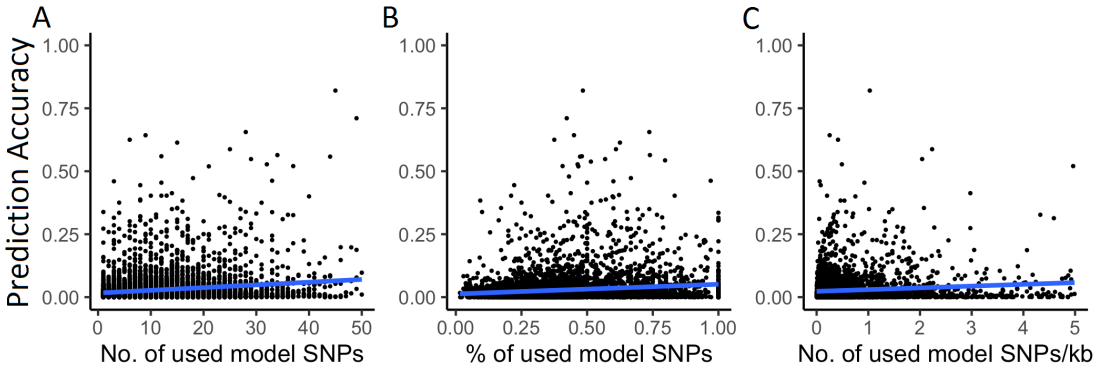

Fig. 4. Prediction accuracy has weak relationship to the model properties. R^2 was computed between observed and GTEx whole blood model predicted expressions. A few genotype-expression model properties were explored, including the number **(A)** and the percentage **(B)** of model variants used for prediction, and the number of used model variants adjusted to gene length **(C)**. But neither of them explained the unsatisfactory prediction, nor could be used as a filtering criterion.

Another potential filtering criterion, the similarity of predicted gene expression levels in the two whole blood models, was also explored. Blood is the most accessible tissue, which makes whole blood models of great practical value and their prediction accuracy critical. The fact that PrediXcan provided two whole blood tissue models offered the opportunity to examine the prediction results based on the two distinct model cohorts. If gene expression was truly regulated by genetic factors, then genotype-expression relationship would be captured regardless of the cohort, and predicted values should be the same given the same genotype data. With this assumption, we hypothesized that the predicted expression for a given gene would likely be more reliable and accurate if the predictions were similar in both whole blood models. As shown in **Fig. 5A**, we selected three sets of genes whose correlations between predicted expression were low, median, or high between the two models. If our hypothesis was correct, we would observe an increase of prediction accuracy from genes with low similarity to those with high similarity, which was indeed what we observed in **Fig. 5B**. The average of prediction accuracy increased from 0.023 to 0.084 for the DGN model and from 0.02 to 0.083 for the GTEx model. In effect, genes whose predicted expressions were more similar between models showed higher prediction accuracy in either PrediXcan whole blood model. However, the filtered results still contained genes whose predicted gene expression levels were directionally different from actual gene expression levels (figures not shown in this paper). In summary, similarity between models was a useful but not ideal filter criterion to improve prediction performance. However, the test of prediction similarity between models can be expanded to using models of different tissue types or using samples from different populations. It may also be worthwhile to investigate genes whose predictions are accurate and similar across models, which could be a good resource or reference set for future investigation of prediction accuracy. In short,

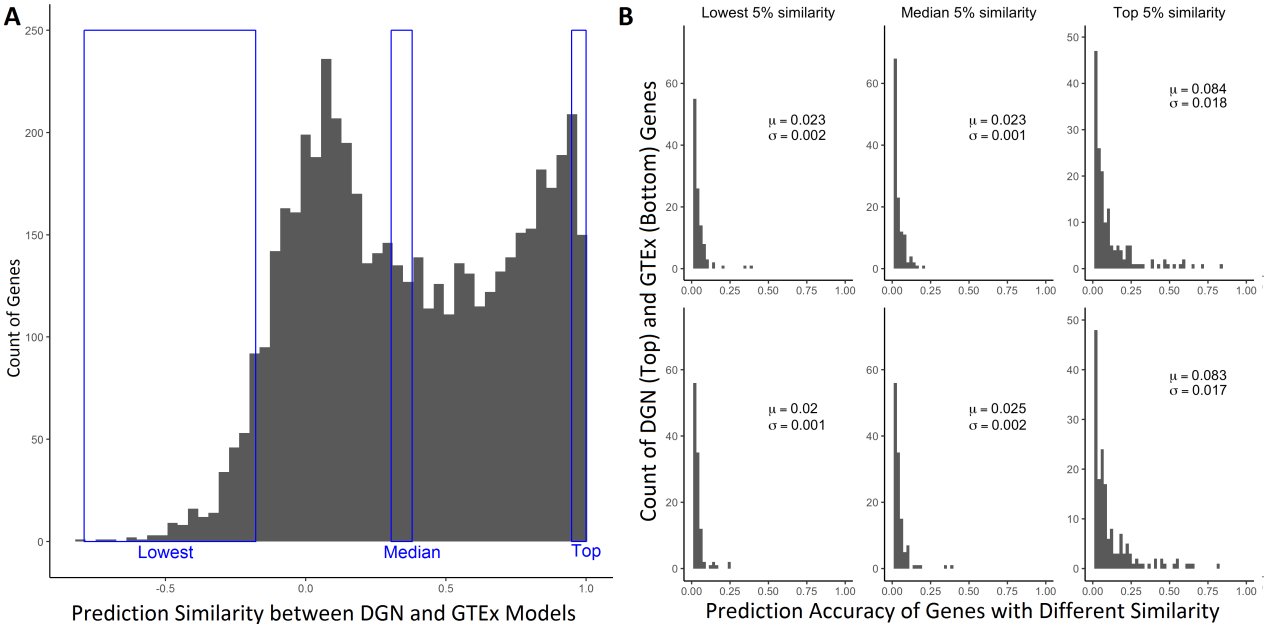

Fig. 5. Prediction similarity between two models has weak, if any, indication on prediction accuracy. Prediction similarity was measured by the Pearson correlation of predicted expressions between the DGN and the GTEx model. **(A)** Distribution of prediction similarity. **(B)** Indication of prediction similarity on prediction accuracy.

many more evaluations could be done with the PrediXcan models or the underlying GTEx data to better understand the SNP-expression relationships in different populations, different tissues, and different genes.

3.2. *Prioritizing GWAS results*

We were also interested in evaluating another use of PrediXcan – prioritization of GWAS results. We wanted to determine whether PrediXcan-TWAS could prioritize important genetic associations that could not be identified by PheWAS due to biological or statistical limitations. To address this question, variant-trait associations that were located within 1MB upstream or downstream of genes were compared to the gene-trait associations identified by PrediXcan-TWAS, using data from the A5202 cohort (**Fig. 6**). Nineteen significant genes identified by PrediXcan-TWAS ($P < 10^{-5}$) were all associated with triglyceride change from baseline to 24 or 48 weeks on treatment. For example, "tgch24_42" in **Fig. 6A** indicates the change in triglyceride from baseline (before starting HIV therapy) to week 24, and was the 42nd phenotype collected. **Fig. 6A** showed that if there were significant variant-trait associations, PrediXcan-TWAS was able to retain the significant signals ($P < 10^{-5}$). This included 3 genes, *DLEU7, DDX1*, and *NARF*. On the other hand, PrediXcan-TWAS prioritized PheWAS associations that almost reached certain significance thresholds ($P = 10^{-5}$; **Fig. 6B**). This highlighted 9 genes – *GPN3, RAP1A, TTC8, SLC5A6, ELOVL7, SUMO1, BAIAP2, OCM,* and *SPRYD4.* The remaining 7 genes had no GWAS association signals in the vicinity regions and thus were likely false positives. Loci within *DLEU7*, *DDX1*, *RAP1A*, *TTC8*, *SLC5A6*, *SUMO1*, and *SPRYD4* were related to triglycerides in previous studies[19,20] according to GRASP[28]. *DDX1* was reported to play a role in HIV-1 infection[21]. More studies are needed to see whether these genes are involved in changes in triglyceride levels on HIV therapy. Other identified genes did not have apparent connections with viral infections or triglycerides, but they could be disease related genes or simply genes that could help to fine-map causal genetic factors. In summary, we demonstrated the ability of PrediXcan to prioritize GWAS results, but the identified gene-trait associations warrant further investigation.

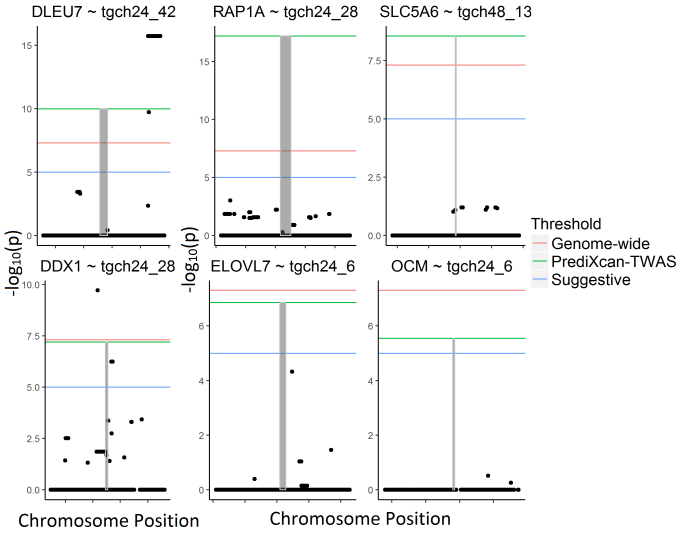

Fig. 6. PrediXcan is able to prioritize GWAS associations. ACTG A5202 imputed genotypic data after quality control was used as input for PrediXcan using GTEx whole blood model and followed by phenome-wide TWAS. Variants within 1MB upstream or downstream of PrediXcan-TWAS significant genes were used to carry out PheWAS. The figures showed the comparison of p-values between PrediXcan-TWAS associations (green line; grey shaded areas represent the size of genes) and PheWAS associations (black dots; blue and red lines denote the suggestive and genome-wide significant p-value, respectively). (A) PrediXcan-TWAS was able to replicate PheWAS results. (B) PrediXcan was able to prioritize non-significant PheWAS results.

4. Discussion

In this study, we carried out a preliminary investigation of the PrediXcan capabilities to predict gene expression levels and to prioritize GWAS signals. If PrediXcan accurately predicts gene expression from SNP data, there could be many potential uses of the algorithm such as imputation of missing transcriptomic data and exploring the biological mechanisms that link genotype to phenotype. But these future analyses are all contingent on the assumption that PrediXcan can accurately predict both the direction of a variants' effect and levels of gene expression. We tested the prediction accuracy of the two PrediXcan whole blood models, DGN and GTEx. PrediXcan was able to accurately predict gene expression for some but not all genes. The slopes of correlation between predicted and actual gene expression levels were negative for almost one-half of genes. This limited the utility of PrediXcan as a transcriptomic data imputation/prediction tool. Several model properties that we explored failed to explain the suboptimal predictions. Dr. Im and her colleagues examined tissues from GTEx and DGN and the results suggested that the local architecture of gene expression traits is simple rather than polygenic[22]. In effect, gene expression is genetically regulated by few rather than multiple eQTLs. This simple local genetic architecture of gene expression might explain why including more putative eQTLs did not improve prediction accuracy in our study. Using prediction similarity between the two whole blood models as a filter improved prediction accuracy somewhat, but did not avoid the negative linear correlation between some predicted and observed gene expression levels. When it came to prioritizing genetic association study results, PrediXcan was able to identify genes that were not significant in GWAS, and also retained significant variant-trait associations. These results were reassuring of the utility of PrediXcan. PrediXcan possessed promising features to reduce research burden by focusing on genes instead of SNPs, and map regulatory effects of distant SNPs onto responding genes, which are overlooked by most studies where only genes adjacent to SNPs are investigated.

Overall, the present study found that PrediXcan performed differently when evaluated for different functions. There are limitations to our study and PrediXcan models. First, whole blood itself is a heterogeneous tissue. And we applied the PrediXcan whole blood model to the YRI cohort whose transcriptomic data actually comes from immortalized blood cell lines. Second is the sample size and population specificity of the test cohort. The YRI cohort (75 individuals) was the most accessible cohort with both genotypic and transcriptomic data, but has a different population structure than the model cohorts from PrediXcan, either DGN or GTEx. While the GTEx cohort includes African Americans, the GTEx model did not yield better expression predictions. To better investigate the influence of population structure and sample size, we would need genotypic and transcriptomic data from multiple populations and of much larger sample sizes. If available, these datasets of different population background will also allow us to explore allelic heterogeneity and population-specific eQTLs. Third, we only evaluated the whole blood models. However, the trait of interest may be regulated by other tissue(s). For example, change of triglyceride in blood may be regulated by metabolism in liver. Thus, it is of biological interest and necessity to explore other tissue models to better understand the tissue specific SNP-expression-trait relationships in the future. Last but not least, PrediXcan is based on two assumptions, 1) loci are equivalent in their functional roles as potential eQTLs, despite the fact that loci at different functional regions may influence gene

expression via different biological mechanisms; and 2) different alleles have the same effect on gene expression. Our study did not specifically evaluate these assumptions. Investigating the relationship of locus functional regions and their roles as eQTLs depends on more detailed annotation and categorization of different types of eQTLs. On the other hand, researchers have looked into allelic expression, which could be a future development for PrediXcan's SNP-expression model design[23].

Although there are challenges, PrediXcan has illuminated a new path for GWAS – incorporating functional genomics and providing mechanistic insights for derived genetic associations. PrediXcan-TWAS results indicated that behind the association, a group of cis-eQTLs regulated gene expression and consequently affected the phenotype. More study is needed to assess PrediXcan's ability to predict gene expression levels and prioritize GWAS results, which will hopefully further our understanding of relationships between eQTLs, gene expression levels, and phenotypes or disease traits.

5. Supplementary

At http://ritchielab.psu.edu/files/PrediXcan_PSB_2018_Binglan_Supplementary_Figures.pdf can supplementary material be found.

6. Acknowledge

The authors are grateful to the many persons with HIV infection who volunteered for A5202 and A5128. In addition, they acknowledge the contributions of study teams and site staff for these protocols. We thank Paul J. McLaren, PhD (Public Health Agency of Canada, Winnipeg, Canada) for prior involvement and collaborations that used these genome-wide genotype data.

References

1. Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., ... & Cho, J. H. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747-753.
2. Van Steen, K. (2011). Travelling the world of gene–gene interactions. *Briefings in bioinformatics*, *13*(1), 1-19.
3. Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., ... & Cherry, J. M. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome research*, *22*(9), 1790-1797.
4. Portela, A., & Esteller, M. (2010). Epigenetic modifications and human disease. *Nature biotechnology*, *28*(10), 1057-1068.
5. Battle, A., Khan, Z., Wang, S. H., Mitrano, A., Ford, M. J., Pritchard, J. K., & Gilad, Y. (2015). Impact of regulatory variation from RNA to protein. *Science*, *347*(6222), 664-667.
6. Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., ... & Bitton, A. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature genetics*, *40*(8), 955-962.
7. Albert, F. W., & Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, *16*(4), 197-212.
8. GTEx Consortium. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, *348*(6235), 648-660.
9. Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., ... & Im, H. K. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics*, *47*(9), 1091-1098.

10. 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68-74.
11. Sax PE, Tierney C, Collier AC, et al. Abacavir-lamivudine versus tenofovir-emtricitabine for initial HIV-1 therapy. *N Engl J Med* 2009; 361:2230-40.
12. Daar ES, Tierney C, Fischl MA, et al. Atazanavir plus ritonavir or efavirenz as part of a 3-drug regimen for initial treatment of HIV-1. *Ann Intern Med* 2011; 154:445-56.
13. Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, *88*(1), 76-82.
14. Battle, A., Mostafavi, S., Zhu, X., Potash, J. B., ... & Urban, A. E. (2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome research*, *24*(1), 14-24.
15. Pendergrass, S. A., Brown-Gentry, K., Dudek, S., Frase, A., Torstenson, E. S., Goodloe, R., ... & Deelman, E. (2013). Phenome-wide association study (PheWAS) for detection of pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. *PLoS genetics*, *9*(1), e1003087.
16. Grady, B. J., Torstenson, E., Dudek, S. M., Giles, J., Sexton, D., & Ritchie, M. D. (2010). Finding unique filter sets in plato: a precursor to efficient interaction analysis in gwas data. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (p. 315). NIH Public Access.
17. Bush, W. S., Dudek, S. M., & Ritchie, M. D. (2009). Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (p. 368). NIH Public Access.
18. Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Springer.
19. Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., ... & Johansen, C. T. (2010). Biological, clinical, and population relevance of 95 loci for blood lipids. Nature, 466(7307), 707.
20. Kathiresan, S., Willer, C. J., Peloso, G. M., Demissie, S., Musunuru, K., Schadt, E. E., ... & Voight, B. F. (2009). Common variants at 30 loci contribute to polygenic dyslipidemia. Nature genetics, 41(1), 56-65.
21. Fang, J., Acheampong, E., Dave, R., Wang, F., Mukhtar, M., & Pomerantz, R. J. (2005). The RNA helicase DDX1 is involved in restricted HIV-1 Rev function in human astrocytes. *Virology*, *336*(2), 299-307.
22. Wheeler, H. E., Shah, K. P., Brenner, J., Garcia, T., ... & GTEx Consortium. (2016). Survey of the heritability and sparse architecture of gene expression traits across human tissues. PLoS genetics, 12(11), e1006423.
23. Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E., & Lappalainen, T. (2015). Tools and best practices for data processing in allelic expression analysis. *Genome biology*, *16*(1), 195.
24. Verma A, Bradford Y, et al. Multiphenotype association study of patients randomized to initiate antiretroviral regimens in AIDS Clinical Trials Group protocol A5202. *Pharmacogenet Genomics* 2017; 27:101-11.
25. Bush, W. S., Dudek, S. M., & Ritchie, M. D. (2009). Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. In Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing (p. 368). NIH Public Access.
26. Göring, H. H., Curran, J. E., Johnson, M. P., Dyer, T. D., ... & Mahaney, M. C. (2007). Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. Nature genetics, 39(10), 1208.
27. Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N. E., Ahfeldt, T., Sachs, K. V., ... & Pirruccello, J. J. (2010). From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature, 466(7307), 714.
28. Leslie, R., O'Donnell, C. J., & Johnson, A. D. (2014). GRASP: analysis of genotype–phenotype results from 1390 genome-wide association studies and corresponding open access database. Bioinformatics, 30(12), i185-i194.