Prediction Using Models Informed by Chromatin conformations, Epigenomics and Summary Statistics

Lida Wang[1] , Chachrit Khunsriraksakul[2,3], Havell Markus[2,3], Laura Carre[4], Dajiang J. Liu[1,2,4]

1 Department of Public Health Sciences; Pennsylvania State University College of Medicine; Hershey, Pennsylvania, 17033; USA.
2 Bioinformatics and Genomics PhD Program; Pennsylvania State University College of Medicine; Hershey, Pennsylvania, 17033; USA.
3 Institute for Personalized Medicine; Pennsylvania State University College of Medicine; Hershey, Pennsylvania, 17033; USA.
4 Department of Biochemistry and Molecular Biology; Pennsylvania State University College of Medicine; Hershey, Pennsylvania, 17033; USA

Background

Genome-wide association studies of human diseases have led to numerous variants associated with many human diseases and traits. Most identified associations are non-coding and influence disease risk via their regulatory effects on gene expressions. Interpreting the functional consequence of non-coding variants is challenging and requires integrating functional genomic data from disease relevant cell types and tissues.

Recently, transcriptome-wide association studies (TWAS) have become a popular gene-based association analysis method for understanding non-coding variants. Many studies applied TWAS to identify risk genes for complex human diseases. Briefly, TWAS first derives the gene expression prediction models from datasets with matched genotypes and gene expression data. Via these models, it imputes gene expression levels in a GWAS dataset and tests for associations with complex traits to identify significant gene-trait associations.

The power of TWAS critically depends on the accuracy of the gene expression prediction models. Existing TWAS methods focus on datasets with individual level genotype and phenotype information, which often have limited sample sizes and are often more difficult to obtain. On the other hand, many large consortia efforts to aggregate large eQTL datasets only release summary association statistics. It is of great interest to extend TWAS to exploit large consortium datasets such as eQTLGen, which can greatly improve the prediction accuracy of gene expression levels and improve the power of association analysis.

Method

In **EXPRESSO (EXpression PREdiction with Summary Statistics Only)**, we use epigenetic and 3D genome to prioritize genomic regions containing causal eQTL variants. We define essential variants as the ones that overlap four broadly available epigenomic annotation from ENCODE database. We denote the genotypes of essential variants as $\boldsymbol{X^e}$. We denote variants that do not overlap the above epigenetic annotation as non-essential variants, i.e., $\boldsymbol{X^{ne}}$. Gene expression prediction is based on a linear model,

$$\boldsymbol{y = X\beta + \epsilon = X^{ne}\beta^{ne} + X^{e}\beta^{e} + \epsilon}$$

We fit the model using elastic net penalty, i.e.,

$$L(\boldsymbol{\beta}\,;\lambda,\phi,w) = ||\boldsymbol{y - X^{ne}\beta^{ne} - X^{e}\beta^{e}}\,||_2^2 + \frac{1}{2}\times\frac{\lambda}{2}(\phi\|\boldsymbol{\beta_e}\|_2^2 + \|\boldsymbol{\beta_{ne}}\|_2^2) + \frac{\lambda}{2}(\phi\|\boldsymbol{\beta_e}\|_1^1 + \|\boldsymbol{\beta_{ne}}\|_1^1)$$

$$= \boldsymbol{y^T y + \beta^T X^T X\,\beta} - 2\boldsymbol{\beta^T X^T y} + \frac{1}{2}\times\frac{\lambda}{2}(\phi\|\boldsymbol{\beta_e}\|_2^2 + \|\boldsymbol{\beta_{ne}}\|_2^2) + \frac{\lambda}{2}(\phi\|\boldsymbol{\beta_e}\|_1^1 + \|\boldsymbol{\beta_{ne}}\|_1^1)$$

where $\|\quad\|_1$ and $\|\quad\|_2$ denote the L_1 and L_2 norms.

We can calculate the objective function $L(\boldsymbol{\beta}\,;\lambda,\phi,w)$ with only summary association statistics. For example, we can estimate $\boldsymbol{X^TX}$ from the LD matrix using a reference panel of matched ancestries, and $\boldsymbol{X^Ty}$ is proportional of the marginal eQTL effect size estimates.

The PVS (Pseudo Variable Selection) method generates a set of pseudo variables $\boldsymbol{X^\pi}$ that have the same covariance structure as the observed set of predictors but are not associated with the phenotypes of interest. Specifically, we introduce an auxiliary loss function that includes both the measured predictors as well as pseudo variables, i.e.,

$$L^*(\boldsymbol{\beta},\boldsymbol{\beta_\pi};\lambda,w,\phi)$$

$$= \|\boldsymbol{Y - X\beta - X^\pi\beta_\pi}\|_2^2 + + \frac{1}{2}\times\frac{\lambda}{2}(\phi\|\boldsymbol{\beta_e}\|_2^2 + \|\boldsymbol{\beta_{ne}}\|_2^2 + \|\boldsymbol{\beta_\pi}\|_2^2)$$

$$+ \frac{\lambda}{2}(\phi\|\boldsymbol{\beta_e}\|_1^1 + \|\boldsymbol{\beta_{ne}}\|_1^1 + \|\boldsymbol{\beta_\pi}\|_1^1)$$

Bigger values of λ impose strong penalty on the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\beta_\pi}$, and lead to stronger shrinkage. Since the control values are not associated with the outcomes, we expect to choose a large enough shrinkage parameter λ to ensure that the model eliminate all pseudo variables. Based on this intuition, for each pair of window size and mitigation factor values, we gradually increase the magnitude of the tuning parameter λ until all coefficients of the pseudo variables become zero.

When individual data is available, the most straightforward approach to generate pseudo variable is permutation. In the absence of individual level data, we can generate summary association statistics of pseudo variables by Monte Carlo simulation. We noted that the covariance between measured variables, pseudo variables, and the gene expression satisfy

$$\boldsymbol{X^{\pi T}X^\pi = X^TX};\ E[\boldsymbol{X,X^\pi}]^T[\boldsymbol{X,X^\pi}] = \begin{bmatrix}\boldsymbol{\Sigma} & 0\\ 0 & \boldsymbol{\Sigma}\end{bmatrix};\ E\left(\boldsymbol{X^{\pi T}Y}\right) = 0$$

We can simulate the summary statistics for pseudo variables as

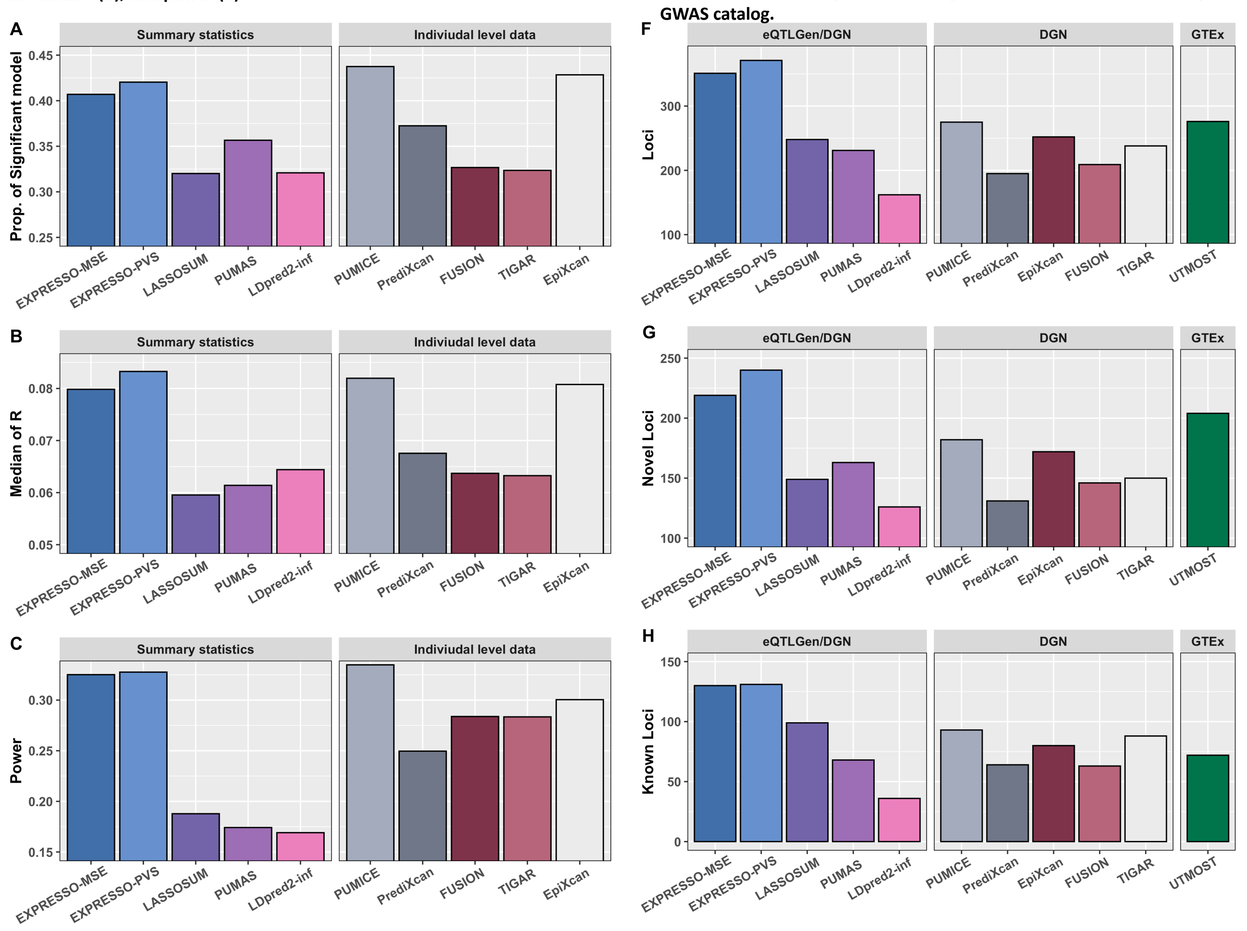
$$\boldsymbol{X^{\pi T}Y} \sim N(\boldsymbol{0, X^TX}\times var(Y))$$

We simulate ten sets of pseudo values. For each set of simulated pseudo variables, we estimate tuning parameters $\lambda(\phi, w)$, which we then average across ten simulated datasets to get final estimates to improve the stability of results.

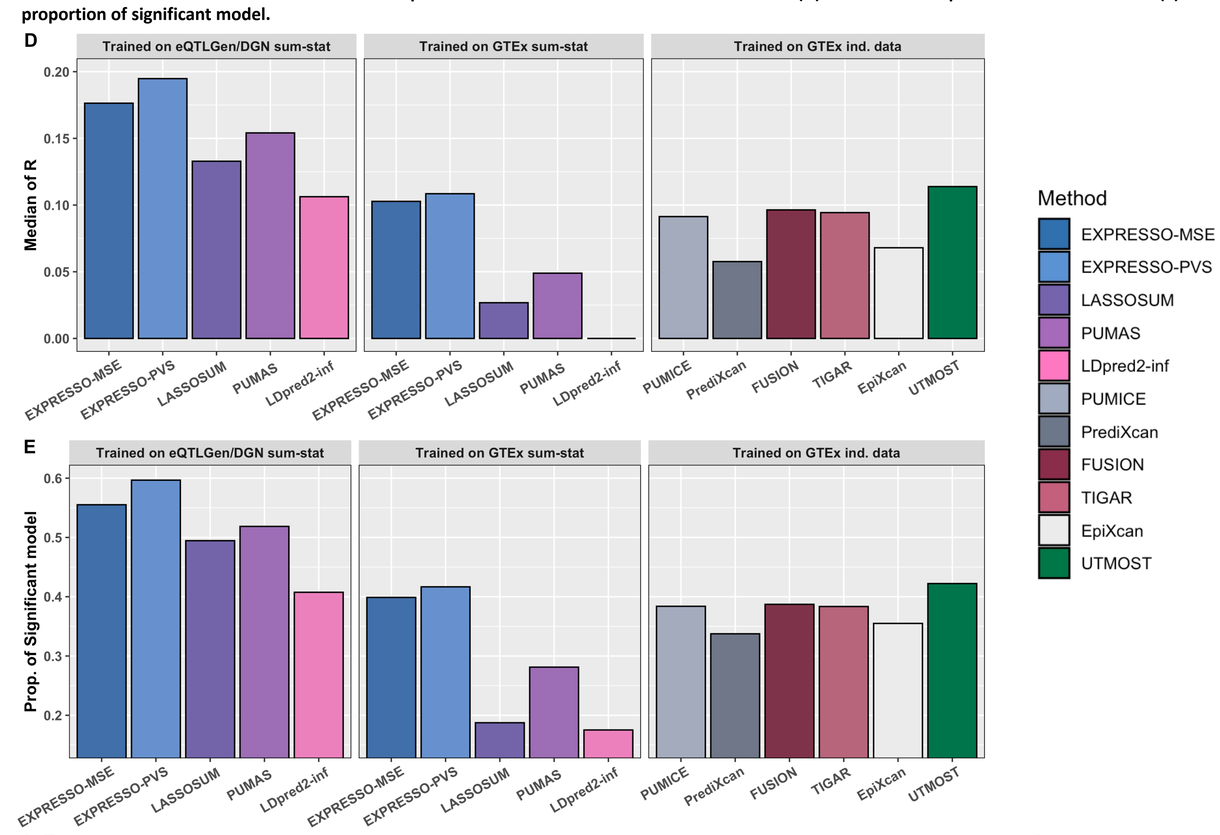
Results

Simulation results: Panel A-C illustrates the comparison of EXPRESSO to other TWAS methods for proportion of significant model (A), median of Spearman's correlation (B), and power (C).

Identified loci: Panel F-H illustrates the comparison of EXPRESSO to other TWAS methods for (F) the number of total loci, (J) the number of novel loci that do not overlap GWAS catalog and (H) known loci that overlap GWAS catalog.



Real data validation: Panel D-E illustrates the comparison of EXPRESSO to other TWAS methods for (D) the median of Spearman's correlation and (E) the proportion of significant model.



1) The number of significant models: the predicted and measured gene expression are significantly correlated with Spearman's correlation coefficient > 0.1 and p-value < 0.05.
2) R: the Spearman correlation between predicted and measured gene expression levels,
3) The power for TWAS: we define as the fraction of genes with significant gene expression prediction model and significant TWAS p-value,
4) We define loci outside 1 million base pairs window of GWAS catalog hits as novel loci.

We compared EXPRESSO against existing TWAS methods that rely on individual level genotype and phenotype data, and polygenic risk score methods adapted to analyze eQTL datasets. We demonstrated substantial improvement in the power of TWAS in both simulation and applied data analysis. We applied the new methods to eQTLGen summary statistics and 14 autoimmune diseases to discover novel gene-level association, performed cell type enrichment analysis, and identify drugs that we may repurpose to treat these disorders.

Conclusion

In conclusion, we presented an integrative framework to perform gene-based association analysis. EXPRESSO is built on publicly available eQTL summary data of bulk-RNASeq, allowing the framework to identify prediction models with higher imputation accuracy and discover cell type-specific risk genes. As the research community continues to generate and assemble large eQTL datasets, EXPRESSO and its future extensions will be a valuable tool for integrative analysis and play an important role for understanding the phenotypic impact of regulatory variants.