Application of Transcriptome-Wide Association Studies to Identifying Genes Associated with Inflammatory Bowel Disease

Esha Desai, Jacqueline Lee, Moksha Poladi, Samuel Zhou

Abstract

Understanding how genetic variation impacts gene expression can help us identify gene-based mechanisms of disease risk. For the last two decades, genome-wide association studies (GWAS) have been utilized to identify disease-associated genetic variants. However, these associated variants often do not lie in gene exons, creating uncertainty as to which genes are associated with disease. Our project aims to fill this gap by leveraging a technique known as transcriptome-wide association studies (TWAS). TWAS is a powerful strategy that can detect gene–trait associations if variation in the expression of a gene colocalizes with phenotypic variation. It combines expression quantitative trait locus (eQTL) data with GWAS summary statistics to identify disease-associated genes. We leverage gene expression and genotype data from the 1000 Genomes project and GWAS from UK Biobank. Here, we focused our analysis on Inflammatory Bowel Disease. Ultimately, the identification of disease-associated genes will accelerate the development of therapeutics and treatment options for patients.

Introduction

Identifying the gene-based mechanism of disease risk is crucial to facilitate early diagnosis and treatment of individuals. Previous approaches attempted to understand this mechanism by detecting key variants associated with gene expression, which in turn is responsible for regulating the production of certain key proteins that can lead to the presence of certain diseases[1]. For example, GWAS finds thousands of trait-associated variants, however, 93% of disease and trait-associated variants emerging from these studies lie within noncoding sequences of the DNA[2]. This makes it difficult to understand the functionality of these variants and their association with disease risk. In addition, it is infeasible to measure gene expression in as many people that are in a GWAS cohort. In order to address these concerns, a new technique called eQTL colocalization was developed. It leverages transcriptomic data in order to inform gene discovery by connecting non-coding disease-associated variants to changes in transcript levels[3]. Colocalization determines whether a single variant is responsible for both GWAS and eQTL signals in a locus, however this approach is often underpowered[4]. For our project, we decided to use a technique called TWAS that aggregates variant effects on gene expression and estimates a gene level association[5].

TWAS aims to identify genes that lead to manifestation of complex human traits due to genetically regulated transcriptional activity[6]. It integrates genome-wide association studies (GWAS) and gene expression datasets to identify gene–trait associations . TWAS leverages eQTL cohorts with expression and genotype data to discover gene–trait associations from GWAS

summary statistics. The eQTL cohort is then used to find predictive models of gene expression by using allele counts of genetic variants in the gene's vicinity[7]. The model is then used to impute the genetic component of gene expression in a large sample of people with genotyping results (ex. a GWAS cohort). Finally, TWAS correlates the disease phenotype and predicts gene expression to find disease associated genes[8]. Ultimately TWAS is powerful in identifying the disease associated gene by aggregating the effects of multiple variants into a single testing unit[5]. Identification of a gene, rather than just a variant, can enable scientists to create drugs targeting a specific gene that can potentially offset or mitigate the effects of the associated disease.

Inflammatory Bowel Disease (IBD) is a highly heritable disease, which causes a chronic inflammation of tissues in an individual's digestive tract[9]. This disease affects 3.1 million adults in the United States and adversely impacts their quality of life[10]. Even though GWAS has identified hundreds of variants associated with Inflammatory Bowel Disease (IBD), there are few known associated genes. For our project we decided to leverage TWAS to identify the genes associated with IBD. To this end, we utilized gene expression data from whole blood samples because IBD is an immune disease and blood will contain the relevant cell types. We applied TWAS to this data and found 7 genes associated with Inflammatory Bowel Disease. Our analysis not only found genes that were previously known to be associated with IBD but it found more as well. This can help scientists create drugs in the future that can target a specific gene, potentially offsetting the effects of IBD.

Methods

Data:

The genotype data that we worked with was collected from Phase 1 of the 1000 Genomes Project, Release Version 3 for chromosome 22. This dataset contains data on millions of SNPs for hundreds of individuals, which we attempt to connect to our gene expression dataset to identify potential relationships. We combined this single nucleotide polymorphisms (SNP) data with gene expression data from RNA-sequencing on LCL samples from the Geuvadis RNA-sequencing project, retrieved from EBI ArrayExpress. Together, these data sources contain genetic data for 344 individuals across four populations: CEPH (CEU), Finns (FIN), British (GBR), and Toscani (YRI). Combining these data sources, we are able to identify relevant SNPs to each gene. We identified cis/local-SNPs as those found within 1Mb of the transcription start site (TSS) of each gene from the Geuvadis project. We considered SNPs that had a minor allele frequency greater than 0.05 to be common, and obtained variant information for these SNPs across all individuals. To perform TWAS, we also required the summary statistics of our chosen disease, IBD. The GWAS summary statistics for IBD were obtained from the analysis carried out in the paper Finucane 2015, Nature Genetics.

Methods and Process Flow:

In order to perform our analysis, we downloaded the genotype and gene expression data from 1000 Genomes project, phase 1 release 3. We filtered the gene expression data to focus our analysis on chromosome 22. We also retrieved additional information about the individual's population groups and merged it with our gene expression data for these individuals. To work with the genotype data, we used plink to convert the original VCF files into the bed, bim and fam files containing the genotype information about the individuals. When extracting the genotype information using plink we removed variants that were not biallelic and had an allele frequency less than .05.

Once we identified the significantly heritable genes and their corresponding cis-SNPs using the GCTA script, for each gene we created a linear model to predict gene expression, weighted by the SNPs. We assume that SNPs additively contribute to a phenotype. For each gene, this follows the following linear model:

$$y_i = \sum_j X_{ij}\beta_j$$

where y_i represents the gene expression for an individual i, X_{ij} represents the estimated minor allele count of SNP j for the individual i, and β_j is an unknown weight on cis-SNP j. To estimate the set of $\widehat{\beta_j}$ for each gene, we used three different modeling techniques - lasso regression, elastic net regression, and only using the single-best eQTL. Both lasso regression and elastic net are regularization techniques that aim to reduce overfitting of our linear regression model. Lasso regularization aims to do so by minimizing the following, using an L1 penalty:

$$||y - \sum_j X_{ij}\widehat{\beta_j}||^2 + \lambda||\widehat{\beta}||$$

Elastic net regression minimizes the following error, using both L1 and L2 penalties:

$$||y - \sum_j X_{ij}\widehat{\beta_j}||^2 + \lambda_1||\widehat{\beta}|| + \lambda_2||\widehat{\beta}||^2$$

These two methods incorporate information from all of the cis-SNPs of a gene. With the single-best eQTL method, the estimated minor allele count for only the most significantly associated SNP was used to model a gene's expression. For this method, the rest of the SNP weights are set to zero.

After fitting these three models for each gene, we performed the summary-based imputation TWAS analysis. To perform the analysis, for each gene we use each set of the model weights obtained above along with GWAS summary statistics for IBD for the corresponding cis-SNPs. Let Z be the standardized GWAS effect sizes of the corresponding SNPs on the trait. We impute the effect size of the expression-trait association using a linear combination of the estimated weights $\widehat{\beta}$ with Z. To test for significance, our null hypothesis is that there is no association and that Z follows a multivariate normal distribution $Z \sim N(0, \Sigma)$, where Σ is the covariance/LD matrix among all SNPs. Under these assumptions, $\widehat{\beta}^T Z$ has a variance of $\widehat{\beta}^T \Sigma \widehat{\beta}$, and the standardized imputed Z-score of the association is as follows:

$$\widehat{\beta}^T Z / (\widehat{\beta}^T \Sigma \widehat{\beta})^{1/2}$$

We then compute imputed Z-scores for every gene with relation to IBD using each of the different regression methods and identify the best-performing model for each gene. We can then use these standardized Z-scores to calculate p-values and identify significant associations between genes and the presence of IBD. This methodology is particularly useful as it allows us to impute gene expression into the GWAS summary-based data and exploit the GWAS data's large sample size to draw significant associations.

Discussion

From our initial sample of 633 genes, we identified 100 that were significantly heritable. Of these 100 genes, 7 were found to have a significant association with IBD. These 7 significant genes are illustrated in Fig. 1.

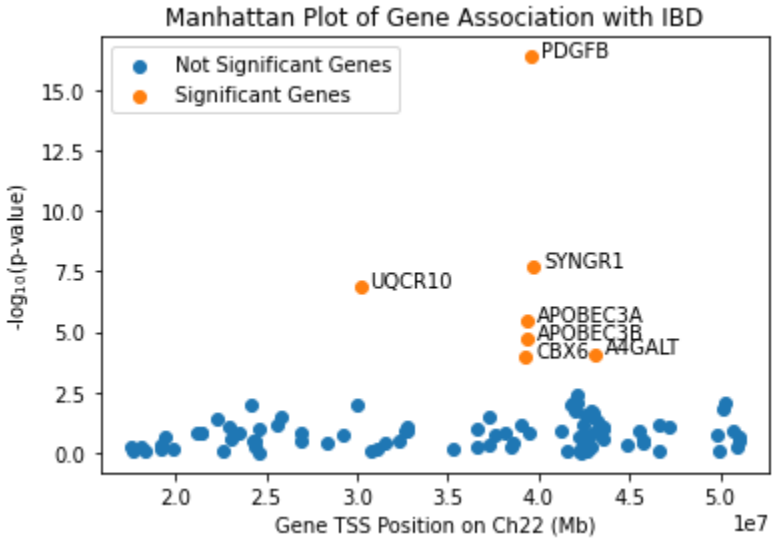

Fig. 1

Five of the seven significant genes are very close together and share the same best GWAS SNP, indicating that they might contribute to the same signal. This highlights TWAS's ability to identify the actual genes associated with a particular disease, rather than just the GWAS SNPs. From the figure above, we find that PDGFB had the most significant association with a p-value of 4.17×10^{-17}.

For each gene, we were able to identify which modeling technique performed the best. Fig. 2 depicts this distribution.

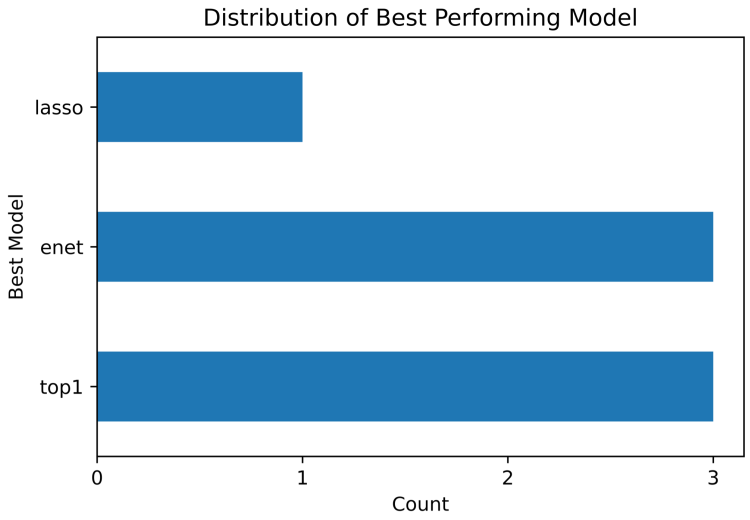

Fig 2.

For significant genes, elastic net regression and using single-best eQTL were the best performing models for three genes each, and lasso regression was the best model for only one.

In addition, we found that the R^2 for each gene was reasonably bounded by heritability. This is to be expected and hence serves as a sanity check for our findings.

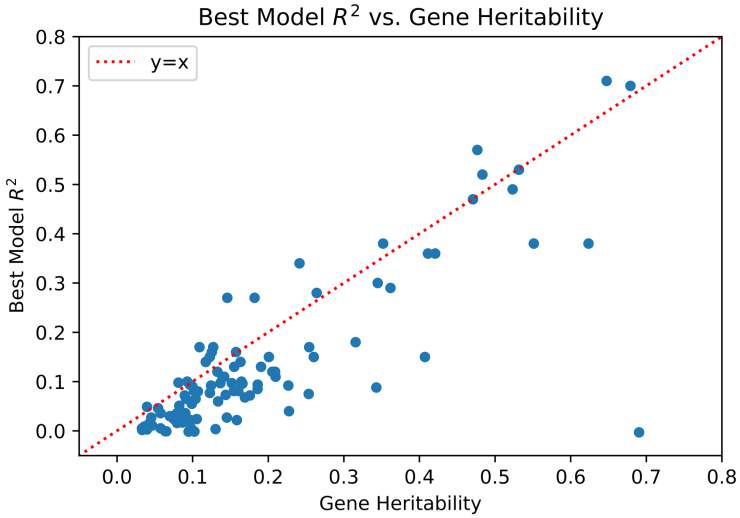

Fig 3.

Conclusion

In conclusion, TWAS was ultimately able to identify 7 genes that are associated with IBD One of the genes, A4GALT, did not have any genome-wide significant GWAS SNPs nearby. This highlights that TWAS is able to find relevant genes even if that locus is not genome-wide significant in GWAS. It was also interesting to see that our findings could be verified by other sources. For example, Marigorta et. al (2017) showed that there was an association between SYNGR1 and IBD, which our analysis was also able to identify. Thus, the use of a technique like TWAS for the identification of these genes can be crucial for disease prevention and early detection.

Acknowledgements

M.P., E.D., J.L., S.Z. researched relevant papers and methodologies. J.L., S.Z. downloaded FUSION packages and other relevant data sets. M.P., E.D. researched relevant genes of interest. J.L., S.Z. conducted statistical and genetic analysis to find association statistics. M.P., E.D consolidated the findings and wrote the final manuscript. All co-authors contributed to the final poster and website of the project.

References

1. “Gene Expression.” *Nature News*, Nature Publishing Group, https://www.nature.com/scitable/topicpage/gene-expression-14121669/.
2. Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun,

S., Johnson, A. K., Canfield, T. K., Giste, E., Diegel, M., … Stamatoyannopoulos, J. A. (2012, September 7). *Systematic localization of common disease-associated variation in regulatory DNA*. Science (New York, N.Y.). Retrieved February 8, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3771521/

3. Al-Barghouthi, B. M., Rosenow, W. T., Du, K.-P., Heo, J., Maynard, R., Mesner, L., Calabrese, G., Nakasone, A., Senwar, B., Gerstenfeld, L., Larner, J., Ferguson, V., Ackert-Bicknell, C., Morgan, E., Brautigan, D., & Farber, C. R. (2022, November 23). *Transcriptome-wide association study and EQTL colocalization identify potentially causal genes responsible for human bone mineral density Gwas Associations*. eLife. Retrieved February 8, 2023, from https://elifesciences.org/articles/77285
4. Hormozdiari, F., van de Bunt, M., Segrè, A. V., Li, X., Joo, J. W. J., Bilow, M., Sul, J. H., Sankararaman, S., Pasaniuc, B., & Eskin, E. (2016, December 1). *Colocalization of GWAS and EQTL signals detects target genes*. American journal of human genetics. Retrieved February 8, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5142122/
5. Went, M., Kinnersley, B., Sud, A., Johnson, D. C., Weinhold, N., Försti, A., van Duin, M., Orlando, G., Mitchell, J. S., Kuiper, R., Walker, B. A., Gregory, W. M., Hoffmann, P., Jackson, G. H., Nöthen, M. M., da Silva Filho, M. I., Thomsen, H., Broyl, A., Davies, F. E., … Houlston, R. S. (2019, August 20). *Transcriptome-wide association study of multiple myeloma identifies candidate susceptibility genes - human genomics*. BioMed Central. Retrieved February 8, 2023, from https://humgenomics.biomedcentral.com/articles/10.1186/s40246-019-0231-5
6. Li, B., & Ritchie, M. D. (2021, September 30). *From GWAS to gene: Transcriptome-wide association studies and other methods to functionally understand GWAS discoveries*. Frontiers in genetics. Retrieved February 8, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8515949/
7. Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A. N., Knowles, D. A., Golan, D., Ermel, R., Ruusalepp, A., Quertermous, T., Hao, K., Björkegren, J. L. M., Im, H. K., Pasaniuc, B., Rivas, M. A., & Kundaje, A. (2019, March 29). *Opportunities and challenges for transcriptome-wide association studies*. Nature News. Retrieved February 8, 2023, from https://www.nature.com/articles/s41588-019-0385-z
8. *Transcriptome-wide association study*. Bioinformatics Analysis – CD Genomics. (n.d.). Retrieved February 8, 2023, from https://bioinfo.cd-genomics.com/transcriptome-wide-association-study.html
9. Mayo Foundation for Medical Education and Research. (2022, September 3). *Inflammatory bowel disease (IBD)*. Mayo Clinic. Retrieved February 8, 2023, from https://www.mayoclinic.org/diseases-conditions/inflammatory-bowel-disease/symptoms-causes/syc-20353315
10. Centers for Disease Control and Prevention. (2022, April 15). *People with IBD have more chronic diseases*. Centers for Disease Control and Prevention. Retrieved February 8, 2023, from https://www.cdc.gov/ibd/features/IBD-more-chronic-diseases.html