Determining the Genetic Component of Protein Levels Using TWAS

Henry Wittich and Heather Wheeler

Introduction

Human genetics research has demonstrated that all human beings share greater than 99.9% of genetic material, meaning that a very small fraction of the human genome contributes to the wealth of phenotypic diversity displayed in the human population.[1] The central dogma of biology describes how an individual's genome contributes to their unique phenotypes: genes encoded in DNA get transcribed into mRNA molecules, which get translated into proteins. Nevertheless, the path from genotype to phenotype is not straightforward. Unraveling the molecular mechanisms that take place at every step of the process, especially for complex traits that are controlled by a network of genes on top of environmental factors, is a challenging task.

At the genomic level, variation is measured in single nucleotide polymorphisms (SNPs), a point in the DNA sequence at which two individuals have a different nucleotide base-pair. Genome-wide association studies (GWAS) are a statistical test designed to compare the pattern of SNPs across a large population of individuals and calculate associations between the millions of SNPs and a specific complex trait.

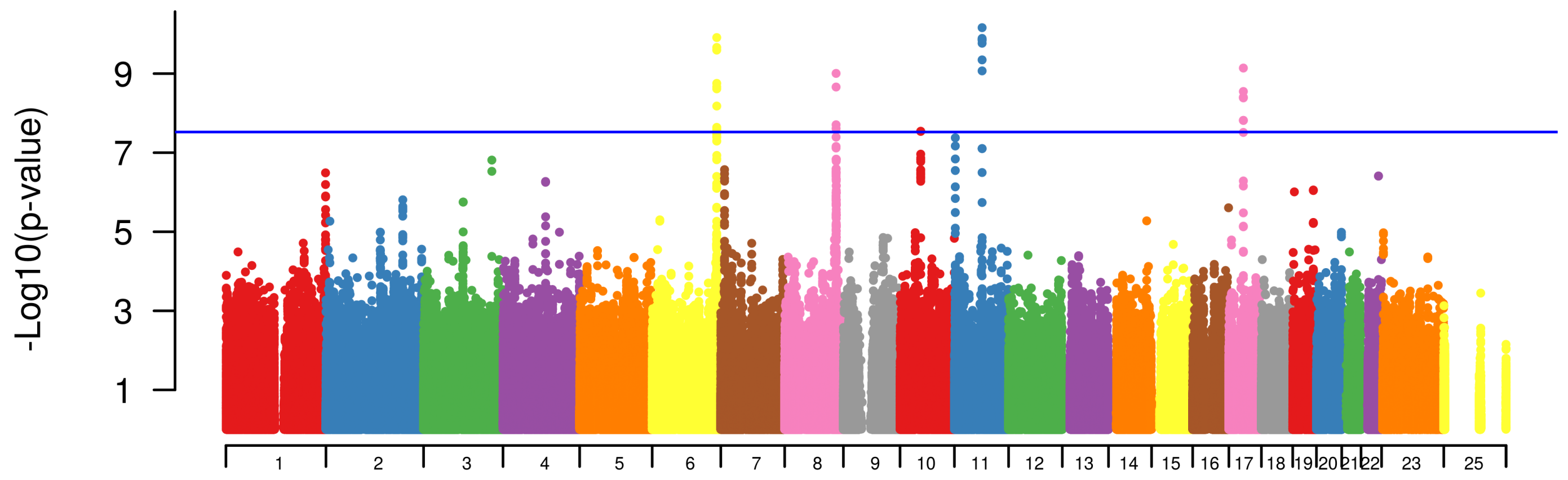


Fig. 1. A Manhattan plot, showing the output of a GWAS. Each point represents a SNP, with its chromosomal position on the x-axis and the p-value of its association with a particular trait on the y-axis.

With the advent of assays designed to collect RNA molecules, transcriptome-wide association studies (TWAS) have been performed to associate transcripts with phenotypes. Furthermore, GWAS have been performed to associate SNPs with transcript levels, identifying expression quantitative trait loci (eQTLs).

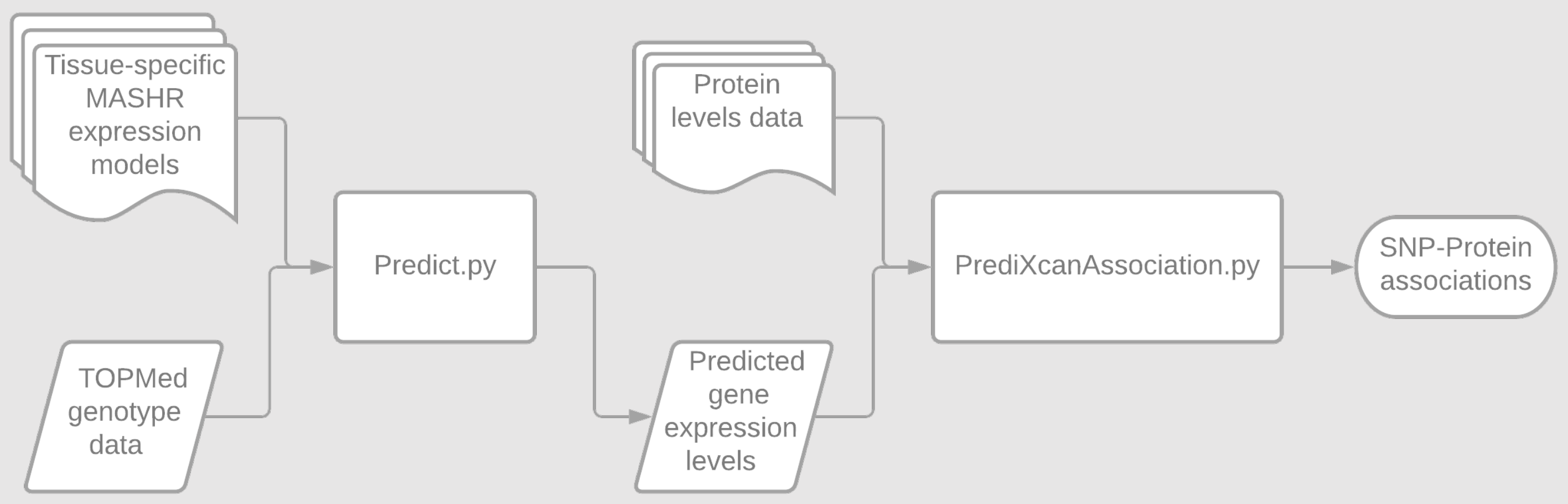
The rise of omics technologies, high-throughput tools for identifying and measuring molecules from a biological sample, has enabled researchers to begin exploring the way in which an individual's genome, transcriptome, proteome, metabolome, and more interact to result in their phenotype. This multi-omics approach has large implications on the possibility of precision medicine, which involves treatment that is personalized to an individual's genotype.

Background

Nevertheless, the effectiveness of tools like GWAS and TWAS is dependent on the input data they are provided. Different ancestral populations have different sets of SNPs. Thus, genetic studies that aren't conducted on diverse populations won't produce results applicable to the entire human population. As of 2017, it was estimated that 87.96% of individuals used in GWAS self-identified as European, highlighting a need to diversify the populations utilized in genetic studies.[2]

Methodology

For this study, I am interested in exploring the genetic component of protein levels by performing a TWAS to associate gene transcripts with proteins. Below I present my bioinformatics pipeline designed to conduct this analysis.


Tissue-specific MASHR expression models

TOPMed genotype data

Predict.py

Protein levels data

Predicted gene expression levels

PrediXcanAssociation.py

SNP-Protein associations

Fig. 2. TWAS for proteins pipeline.

This pipeline utilizes genotype and protein levels data from around 1000 individuals in the Trans-Omics for Precision Medicine (TOPMed) cohort. This project aims to study heart, lung, blood, and sleep disorders in diverse populations with a multi-omics approach.

Due to a lack of expression data from the individuals in the study, this pipeline implements PrediXcan, a tool that imputes gene expression levels from an individual's genotype.[3] PrediXcan relies on models that have been trained off eQTL data to predict gene expression. In this case, we are using expression models for 54 different tissues developed by the Genotype-Tissue Expression (GTEx) Project.[4]

Literature Cited

1. 1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, 526(7571), 68–74.

2. Mills, M.C., Rahal, C. A scientometric review of genome-wide association studies. *Commun Biol* 2, 9 (2019).

3. Gamazon, E., Wheeler, H., Shah, K. et al. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet 47, 1091–1098 (2015).

4. Barbeira, AN, Melia, OJ, Liang, Y, et al. Fine-mapping and QTL tissue-sharing information improves the reliability of causal gene identification. *Genetic Epidemiology*. 2020; 44: 854– 867.

Results

Thus far, PrediXcan has been applied to the genotypes and measured protein levels of the 971 individuals from the TOPMed ALL population (every ancestral population). Expression levels were predicted using each of the 54 GTEx tissue models and associated with the levels of the 1,355 proteins measured in the study.

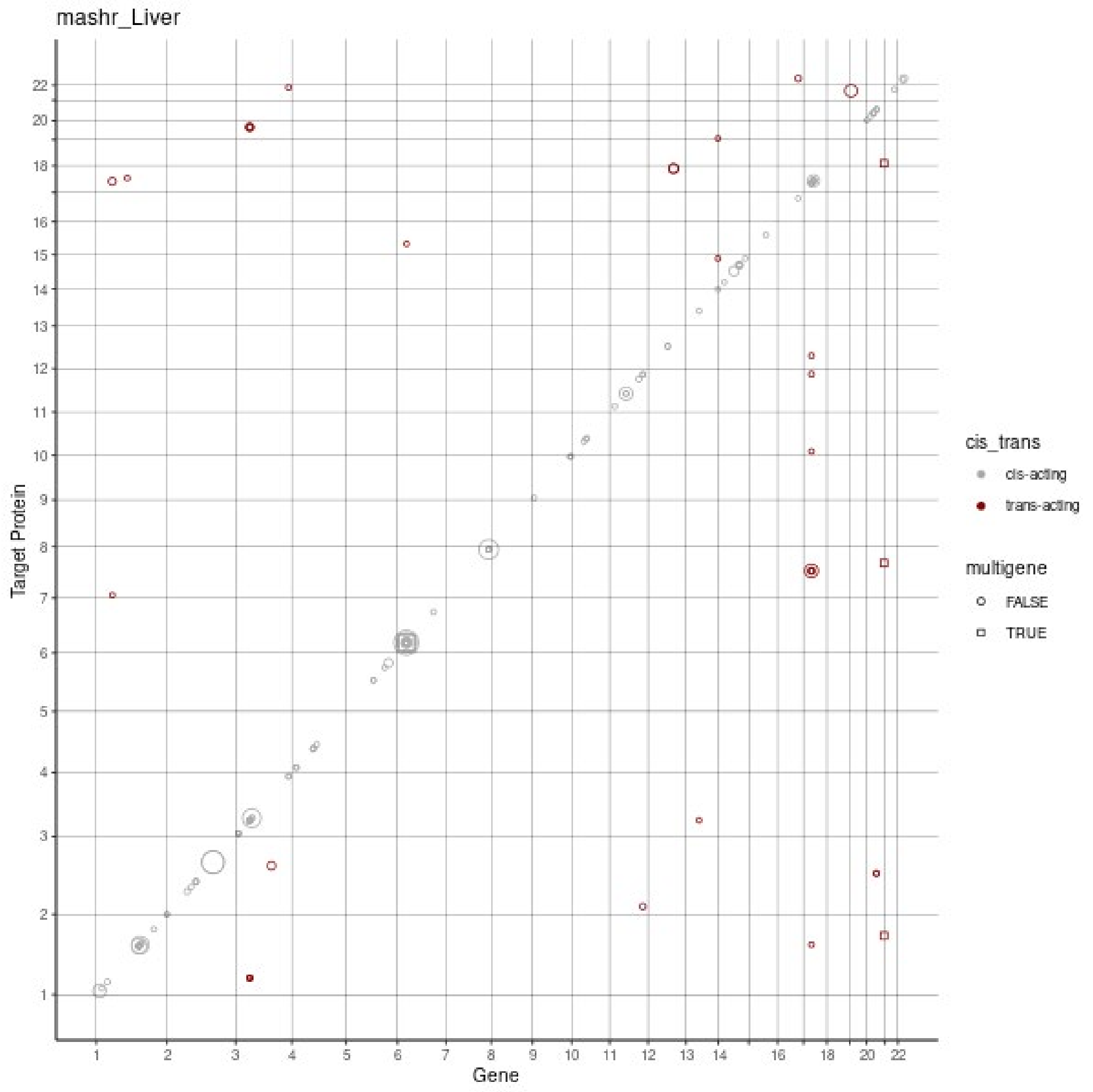


Fig. 3. Chromosomal positions of significantly associated transcript-protein pairs in the liver tissue. Cis-acting transcripts are for genes found within 1Mb of the associated protein, while trans-acting transcripts lie outside of this range. Multigene proteins are those that are picked up by the same aptamer in the assay used to measure protein levels.

Discussion

With these early results, it is clear that the pipeline is effective in finding transcript-protein associations, identifying hundreds of significant pairs across every tissue. Many of these pairs involve cis-acting transcripts, which makes biological sense; a transcript should associate with the protein it encodes.

Interestingly, this analysis revealed a number of trans-acting transcripts, that is, transcripts that are significantly associated with proteins in distant chromosomal locations that they don't code for. These associations might indicate relationships between these genes in a gene network, validating the utility of this analysis in clarifying the biological mechanisms by which an individual's genotype influences their phenotype.

Ultimately, further analysis into the function of these trans-acting genes is necessary to determine what type of relationship they may have with their associated protein. Furthermore, comparing the associations across tissues might reveal tissue-dependent expression patterns with phenotypic importance.

Acknowledgements

- Dr. Heather Wheeler, my research mentor
- The TOPMed Consortium
- Members of the Wheeler Lab:
  Ryan Schubert, Elyse Geoffrey, Ashley Mulford, Isabella Gregga, Daniel Araújo, Chris Nguyen, and Rhea Prag

Contact

hwittich@luc.edu