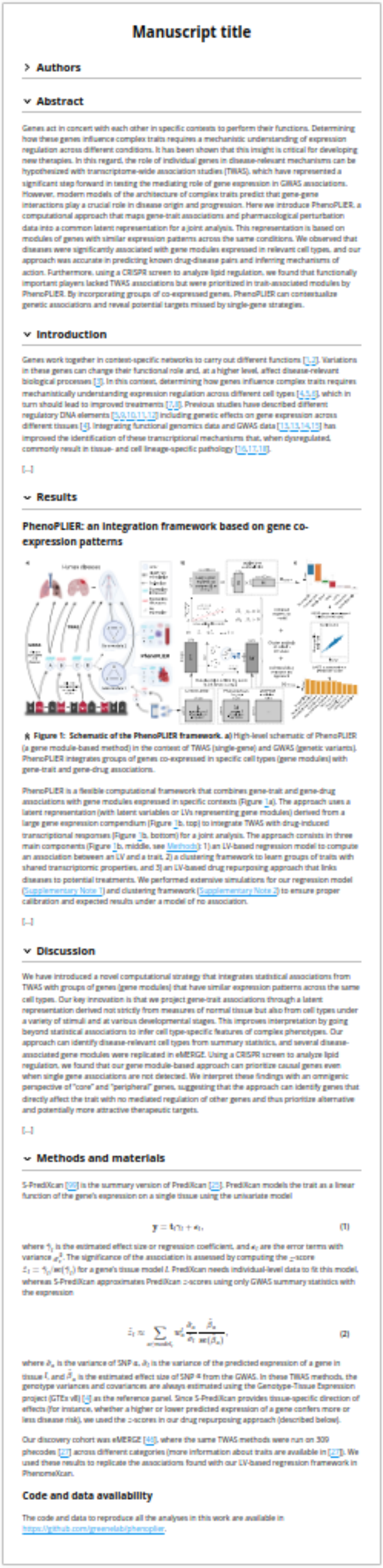
a)



Manuscript title

> Authors

Abstract

Genes act in concert with each other in specific contexts to perform their functions. Determining how these genes influence complex traits requires a mechanistic understanding of expression regulation across different conditions. It has been shown that this insight is critical for developing new therapies. In this regard, the role of individual genes in disease-relevant mechanisms can be hypothesized with transcriptome-wide association studies (TWAS), which have represented a significant step forward in testing the mediating role of gene expression in GWAS associations. However, modern models of the architecture of complex traits predict that gene-gene interactions play a crucial role in disease origin and progression. Here we introduce PhenoPLIER, a computational approach that maps gene-trait associations and pharmacological perturbation data into a common latent representation for a joint analysis. This representation is based on modules of genes with similar expression patterns across the same conditions. We observed that diseases were significantly associated with gene modules expressed in relevant cell types, and our approach was accurate in predicting known drug-disease pairs and inferring mechanisms of action. Furthermore, using a CRISPR screen to analyze lipid regulation, we found that functionally important players lacked TWAS associations but were prioritized in trait-associated modules by PhenoPLIER. By incorporating groups of co-expressed genes, PhenoPLIER can contextualize genetic associations and reveal potential targets missed by single-gene strategies.

Introduction

Genes work together in context-specific networks to carry out different functions [1,2]. Variations in these genes can change their functional role and, at a higher level, affect disease-relevant biological processes [3]. In this context, determining how genes influence complex traits requires mechanistically understanding expression regulation across different cell types [4,5,6], which in turn should lead to improved treatments [7,8]. Previous studies have described different regulatory DNA elements [5,9,10,11,12] including genetic effects on gene expression across different tissues [4]. Integrating functional genomics data and GWAS data [13,13,14,15] has improved the identification of these transcriptional mechanisms that, when dysregulated, commonly result in tissue- and cell lineage-specific pathology [16,17,18].

[...]

Results

PhenoPLIER: an integration framework based on gene co-expression patterns

Figure 1: Schematic of the PhenoPLIER framework. a) High-level schematic of PhenoPLIER (a gene module-based method) in the context of TWAS (single-gene) and GWAS (genetic variants). PhenoPLIER integrates groups of genes co-expressed in specific cell types (gene modules) with gene-trait and gene-drug associations.

PhenoPLIER is a flexible computational framework that combines gene-trait and gene-drug associations with gene modules expressed in specific contexts (Figure 1a). The approach uses a latent representation (with latent variables or LVs representing gene modules) derived from a large gene expression compendium (Figure 1b, top) to integrate TWAS with drug-induced transcriptional responses (Figure 1b, bottom) for a joint analysis. The approach consists in three main components (Figure 1b, middle, see Methods): 1) an LV-based regression model to compute an association between an LV and a trait, 2) a clustering framework to learn groups of traits with shared transcriptomic properties, and 3) an LV-based drug repurposing approach that links diseases to potential treatments. We performed extensive simulations for our regression model (Supplementary Note 1) and clustering framework (Supplementary Note 2) to ensure proper calibration and expected results under a model of no association.

[...]

Discussion

We have introduced a novel computational strategy that integrates statistical associations from TWAS with groups of genes (gene modules) that have similar expression patterns across the same cell types. Our key innovation is that we project gene-trait associations through a latent representation derived not strictly from measures of normal tissue but also from cell types under a variety of stimuli and at various developmental stages. This improves interpretation by going beyond statistical associations to infer cell type-specific features of complex phenotypes. Our approach can identify disease-relevant cell types from summary statistics, and several disease-associated gene modules were replicated in eMERGE. Using a CRISPR screen to analyze lipid regulation, we found that our gene module-based approach can prioritize causal genes even when single gene associations are not detected. We interpret these findings with an omnigenic perspective of "core" and "peripheral" genes, suggesting that the approach can identify genes that directly affect the trait with no mediated regulation of other genes and thus prioritize alternative and potentially more attractive therapeutic targets.

[...]

Methods and materials

S-PrediXcan [26] is the summary version of PrediXcan [25]. PrediXcan models the trait as a linear function of the gene's expression on a single tissue using the univariate model

$$\mathbf{y} = \mathbf{t}_l \gamma_l + \epsilon_l, \qquad (1)$$

where $\hat{\gamma}_l$ is the estimated effect size or regression coefficient, and ϵ_l are the error terms with variance σ_ϵ^2. The significance of the association is assessed by computing the z-score $\hat{z}_l = \hat{\gamma}_l / \mathrm{se}(\hat{\gamma}_l)$ for a gene's tissue model l. PrediXcan needs individual-level data to fit this model, whereas S-PrediXcan approximates PrediXcan z-scores using only GWAS summary statistics with the expression

$$\hat{z}_l \approx \sum_{a \in model_l} w_a^l \frac{\hat{\sigma}_a}{\hat{\sigma}_l} \frac{\hat{\beta}_a}{\mathrm{se}(\hat{\beta}_a)}, \qquad (2)$$

where $\hat{\sigma}_a$ is the variance of SNP a, $\hat{\sigma}_l$ is the variance of the predicted expression of a gene in tissue l, and $\hat{\beta}_a$ is the estimated effect size of SNP a from the GWAS. In these TWAS methods, the genotype variances and covariances are always estimated using the Genotype-Tissue Expression project (GTEx v8) [4] as the reference panel. Since S-PrediXcan provides tissue-specific direction of effects (for instance, whether a higher or lower predicted expression of a gene confers more or less disease risk), we used the z-scores in our drug repurposing approach (described below).

Our discovery cohort was eMERGE [46], where the same TWAS methods were run on 309 phecodes [27] across different categories (more information about traits are available in [27]). We used these results to replicate the associations found with our LV-based regression framework in PhenomeXcan.

Code and data availability

The code and data to reproduce all the analyses in this work are available in https://github.com/greenelab/phenoplier.

b) Section-specific prompt generator:

Revise the following paragraph from the abstract of an academic paper (with the title 'Manuscript title' and keywords 'keyword1, keyword2, ...') so the research problem/question is clear, the solution proposed is clear, the text grammar is correct, spelling errors are fixed, and the text is in active voice and has a clear sentence structure.

Revise the following paragraph from the Introduction... so most of the citations to other academic papers are kept, the text minimizes the use of jargon, ...

Revise the following paragraph from the Results... so most references to figures and tables are kept, the details are enough to clearly explain the outcomes, sentences are concise and to the point, ...

Revise the following paragraph from the Discussion... so most of the citations to other academic papers are kept, the text minimizes the use of jargon, ...

Revise the following paragraph from the Methods... so most of the citations to other academic papers are kept, most of the technical details are kept, most references to equations (such as “Equation (@id)”) are kept, all equations definitions (such as “$$ ... $$ {#id}”) are included, the most important symbols in equations are defined, ...

...