MULTI-OMIC CAUSAL DISCOVERY USING GENOTYPES AND GENE EXPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal inference in genomics is inherently challenging due to high dimensionality, hidden confounders, and the intricate task of distinguishing direct regulatory interactions from indirect ones. We introduce GENESIS (Gene Network Inference from Expression Signals and SNPs), a novel algorithm that fuses genotype (SNP) data with transcriptomic profiles to reconstruct directed gene regulatory networks. Our algorithm adopts a robust two-stage hypothesis testing strategy: (1)it conducts marginal testing of SNP-gene associations within defined genomic windows; (2)it applies conditional independence tests to eliminate indirect regulatory effects. We run our method on a real-world dataset and compare the results. This parametrically rigorous framework offers a powerful avenue for uncovering causal pathways in complex traits, with promising applications in functional genomics, therapeutic interventions, and precision medicine.

1 INTRODUCTION

In recent years, high-throughput sequencing and advanced microarray technologies have generated vast volumes of multi-omic data across different layers such as transcriptomics, proteomics, metabolomics, and genomics Mohammadi-Shemirani et al. (2023); Qiao et al. (2024); Reuter et al. (2015). While this wealth of information holds immense potential for advancing our understanding of biological systems, it simultaneously poses formidable analytical challenges, particularly in the realm of causal inference from complex, high-dimensional data. Although considerable progress has been made in applying causal discovery methods within individual omics layers, particularly genomics these approaches often fall short of capturing the intricate interplay between the different molecular layers. Integrating information across multiple omics layers promises to reveal more detailed explanations that would remain obscured if we solely rely on single-omics analyses. Adopting a multi-omic causal discovery approach is very essential for meaningful biological interpretation, clinical translations, and accurate therapeutic interventions.

At the heart of modern genomics is the quest to unravel gene regulatory networks (GRNs) that dictate cellular behaviours. A clear and precise understanding of these networks can illuminate the pathways that lead to complex traits and diseases. However, the complexities of the underlying data, characterized by thousands of gene (nodes) and millions of single nucleotide polymorphism (SNPs) demands innovative methods that are robust against the pitfalls of high-dimensionality, unobserved confounders, and the statistical challenges posed by multiple testing. One promising strategy, which we employ, is the usage of the analysis of cis-expression quantitative trait loci (cis-eQTL) which leverages the spatial proximity between SNPs and their target genes. By confining the analysis to a defined genomic window, we concentrate on genetic variants most likely to exert a direct regulatory influence. This targeted approach boosts statistical power by reducing the search space.

In this context, we introduce GENESIS (Gene Network Inference from Expression Signals and SNPs), a novel algorithm to integrate genotype and transcriptomic data to reconstruct directed gene regulatory networks. It harnesses the natural variability of SNPs to distinguish between direct and indirect gene regulatory effects by following the inference rules proposed by Magliacane et al. (2016). We implement a two-step hypothesis testing framework to identify marginal SNP-gene associations within cis-windows and then conditional independence tests to filter out indirect effects. Our algorithm is uniquely positioned to deliver a statistically robust and biologically interpretable network model.

2 DEFINITIONS AND NOTATION

We use upper case italics X to denote random variables and lower case x to denote their values. Calligraphic $\mathcal{X}$ denotes the support of the random variable X. Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ be two random variables with joint distribution $\mathbb{P}_{X,Y}$. Consider a graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ with nodes $\mathcal{V}$ that represent variables and edges $\mathcal{E} : \mathcal{V} \times \mathcal{V}$ that denote the relationships between the variables. If all edges do not encode an ordering on nodes, we say the graph is *undirected*; otherwise, the graph is *directed*. In the context of this study, the nodes represent our genes. At the same time, the edges may represent SNP-gene or gene-gene interactions. The acyclic directed mixed graph (ADMG) shown in figure (1) tells us the encoded family relationship of our nodes. We use kinship terms to describe the relationship between nodes. We say $\{X, Z\}$ are the *parents* of Y (denoted $\{X, Z\} \in Pa(Y)$) and Y is a *child* of its parents. We also call $\{X, Z, W\}$ as the ancestral set of Y ($Anc(Y)$). Y has no descendants ($Des(Y) = \{\}$) Spirtes et al. (2001); Pearl (2009). The dotted curved bidirectional edge between Z and X represent the presence of an unknown common cause or confounder. Ancestry graphs have the following properties as proposed by Claassen & Heskes (2012). (1) *irreflexive*: A node can not be the ancestor of itself (2) *acyclic*: No cycles are permitted amongst variables and (3) *transitive*: If X is an ancestor of Y, and Y is an ancestor of Z, then X must also be an ancestor of Z.

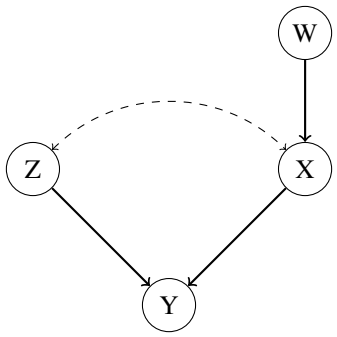

Figure 1: A mixed graph containing directed and bidirected edges.

We say that X and Y are *independent* if and only if their joint distribution factorizes as follows: $\forall x, y : P(x, y) = P(x)P(y)$. We denote this via the shorthand $X \perp\!\!\!\perp Y$. We can informally describe this as a case where X contains no information about the value of Y and vice versa. Let $Z \in \mathcal{Z}$ be a third random variable. We say that X and Y are *conditionally independent* given Z if and only if the joint distribution factorizes as follows: $\forall x, y, z : P(x, y, z) = P(x \mid z)P(y \mid z)$. We denote this via the shorthand $X \perp\!\!\!\perp Y \mid Z$. Informally, this describes a case where X contains no information about the value of Y (and vice versa) once we have processed the information in Z. Following the definitions of minimal (de)activators as stated by Claassen & Heskes (2012), we define the following.

Definition 1: A variable W is called a *minimal deactivator* of the relationship between X and Y given Z if $X \perp\!\!\!\perp Y \mid Z \cup W$ (i.e., X and Y become independent when conditioning on Z and W), and $X \not\perp\!\!\!\perp Y \mid Z$ (i.e., X and Y are dependent when conditioning only on Z). In this case, we write $X \perp\!\!\!\perp Y \mid Z \cup [W]$.

Definition 2: A variable W is called a *minimal activator* of the relationship between X and Y given Z if $X \not\perp\!\!\!\perp Y \mid Z \cup W$ (i.e., X and Y are dependent when conditioning on Z and W), and $X \perp\!\!\!\perp Y \mid Z$ (i.e., X and Y are independent when conditioning only on Z). In this case, we write $X \not\perp\!\!\!\perp Y \mid Z \cup [W]$.

Definition 3: Two variables X and Y are a *causally unrelated pair* given a set of variables Z if their statistical dependence can be fully explained by Z. That is, X and Y are conditionally independent given Z, meaning: $X \perp\!\!\!\perp Y \mid Z$. This implies that no additional variable W outside of Z alters the conditional dependence between X and Y.

3 ALGORITHM

Our algorithm as shown in Algorithm 1 infers gene regulatory networks from SNP genotype and gene expression data by employing a combination of marginal and conditional statistical tests. It operates under the assumption that SNPs can influence the expression levels of nearby genes (cis-eQTLs). For each SNP, the algorithm first identifies candidate target genes within a defined genomic window. Then it performs marginal linear regression tests to assess the association between each SNP and the expression level of each candidate gene. Conditional linear regression tests are conducted to investigate the dependence of a gene's expression on a given SNP while accounting for the expression levels of other genes within the same genomic window. This was based on the inferences from Magliacane et al. (2016). This conditional testing aims to disentangle direct regulatory effects from indirect associations mediated by other genes. Additionally, the algorithm evaluates the correlation between gene pairs. Finally, based on the p-values from these tests and pre-defined significance thresholds, it constructs a binary adjacency matrix representing the inferred gene regulatory network. Specifically, an edge is drawn from a SNP to a gene if the marginal test is significant and the conditional test (conditioning on other genes) is not, suggesting a direct regulatory effect. Similarly, an edge is drawn between two genes if the conditional test (conditioning on the SNP) is significant, indicating that their expression relationship is not fully explained by the SNP. Correlated gene pairs, above our defined significance threshold, are excluded from the inferred network to potentially remove spurious edges due to co-expression. This multi-stage statistical approach allows to identify both direct and indirect regulatory relationships while attempting to control for confounding effects.

Algorithm 1 Multiomic Causal Discovery

Inputs: Z: Genotype matrix, X: Expression matrix
$\alpha \leftarrow 0.05$: Significance threshold, $b_p \leftarrow 5000$: Genomic window size
Output: A: Binary adjacency matrix

Align matrix dimensions: $n_{\mathbf{Z}} \leftarrow n_{\mathbf{X}}$
for each SNP $z_i \in \mathbf{Z}$ **do**
    $\text{chr}, \text{pos} \leftarrow \mathbf{Z}_{pos}[z_i]$
    $\mathcal{G}_i \leftarrow \{g_j \in \mathbf{X} \mid \mathbf{X}_{pos}[g_j].\text{chr} = \text{chr} \wedge \text{pos} \in [\text{lo} - b_p, \text{hi} + b_p]\}$
end for
for each SNP $z_i \in \mathbf{Z}$ **do**
    for each gene $g_j \in \mathcal{G}_i$ **do**
        $\beta^m_{ij}, p^m_{ij} \leftarrow \text{lm}(\mathbf{X}[g_j] \sim \mathbf{Z}[z_i])$
    end for
end for
for each SNP $z_i \in \mathbf{Z}$ **do**
    for each pair $(g_j, g_k) \in \mathcal{G}_i^2,\ j \neq k$ **do**
        $\beta^c_{ijk}, p^c_{ijk} \leftarrow \text{lm}(\mathbf{X}[g_j] \sim \mathbf{Z}[z_i] + \mathbf{X}[g_k])$
        $p^{r3a}_{jk} \leftarrow \text{cor.test}(\mathbf{X}[g_j], \mathbf{X}[g_k])$
    end for
end for
$\mathcal{R}_1 \leftarrow \{(z_i, g_j) \mid p^m_{ij} \leq \alpha \wedge p^c_{ijk} \geq \alpha\}$
$\mathcal{R}_2 \leftarrow \{(g_k, g_j) \mid p^m_{ik} \geq \alpha \wedge p^c_{ijk} \leq \alpha\}$
$\mathcal{R}_3 \leftarrow \{(g_j, g_k) \mid p^{r3a}_{jk} \geq \alpha \vee p^{r3b}_{jk} \geq \alpha\}$
$\mathbf{A} \leftarrow \mathbf{0}^{n \times n}$
for $(u, v) \in \mathcal{R}_1 \cup \mathcal{R}_2 \setminus \mathcal{R}_3$ **do**
    $\mathbf{A}[u, v] \leftarrow 1$
end for
return $\mathbf{A}$

4 RESULTS

Our analysis of the yeast dataset using TRIGGER by Chen et al. (2007) reveals distinct p-value distributions corresponding to the three defined relationship types. It contained 3244 SNPs and 6216 genes. In Figure 4, the histogram for minimal deactivation (Definition 1) shows that when an additional variable W is conditioned upon, the relationship between X and Y becomes independent (i.e $X \perp\!\!\!\perp Y \mid Z \cup [W]$). Our method returned 235 genes that show this relationship. Conversely, the histogram for minimal activation (Definition 2) displays a broader distribution of p-values, indicating that introducing W induces a dependency between X and Y. 97 genes showed this relationship. Finally, the histogram for causally unrelated pairs (Definition 3)—where conditioning on Z completely accounts for the statistical association between X and Y exhibit a distribution consistent with the absence of additional causal influence from any external variable

In a separate multivariate simulation comparison (Figure 2), we evaluated the performance of GENESIS, GES, and RFCI on simulated data. Our results indicate that while GENESIS outperforms the other methods in terms of average accuracy, it also exhibits large variability in its estimates, suggesting that its performance could benefit from further refinement. In contrast, RFCI not only required substantially longer computation times but also produced less reliable estimates on our simulated dataset.

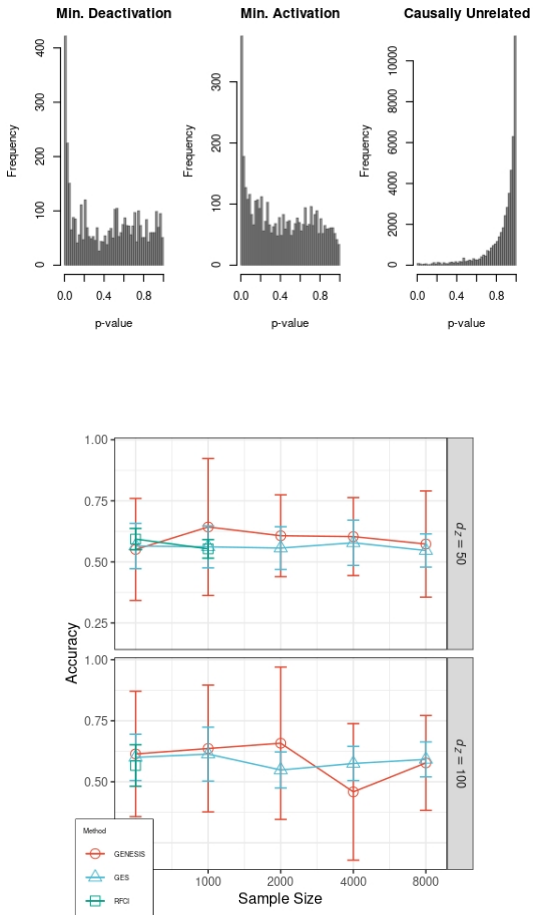

Figure 2: Multivariate comparison of network inference methods on simulated data. The figure compares GENESIS, GES, and RFCI in terms of accuracy across different simulation parameters. GES shows superior average performance, though with large variability (whiskers), while RFCI required substantially more time and produced less reliable estimates.

REFERENCES

Lin S Chen, Frank Emmert-Streib, and John D Storey. Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome biology*, 8:1–13, 2007.

Tom Claassen and Tom Heskes. A logical characterization of constraint-based causal discovery. *arXiv preprint arXiv:1202.3711*, 2012.

Sara Magliacane, Tom Claassen, and Joris M Mooij. Ancestral causal inference. *Advances in Neural Information Processing Systems*, 29, 2016.

Pedrum Mohammadi-Shemirani, Tushar Sood, and Guillaume Paré. From 'omics to multi-omics technologies: the discovery of novel causal mediators. *Current atherosclerosis reports*, 25(2): 55–65, 2023.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Yi Qiao, Tianguang Cheng, Zikun Miao, Yue Cui, and Jing Tu. Recent innovations and technical advances in high-throughput parallel single-cell whole-genome sequencing methods. *Small Methods*, pp. 2400789, 2024.

Jason A Reuter, Damek V Spacek, and Michael P Snyder. High-throughput sequencing technologies. *Molecular cell*, 58(4):586–597, 2015.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2001.