Modified CStaR using LV-IDA to identify downstream targets of MAP3K5-AS2 gene in childhood-onset asthma

Zakari Billo
P8124: Graphical Models for Complex Health Data

December 17, 2025

Abstract

Using B lymphocyte (B-cell) gene expression in 395 children from families recruited through a proband with asthma (Liang et al., 2013), we developed a modified Causal Stability Ranking (CStaR, Stekhoven et al., 2012) pipeline: (1) neighborhood selection via Lasso stability selection, (2) constraint-based PAG (Partial Ancestral Graph) learning with the FCI (Fast Causal Inference) and RFCI (Really Fast Causal Inference) algorithms, and (3) bounding causal effects in the presence of latent confounders using Latent Variable Intervention Calculus when the Directed Acyclic Graph (DAG) is Absent (LV-IDA) (Malinsky and Spirtes, 2016). We compare the causal networks of asthmatic and non-asthmatic subjects to identify causally downstream therapeutic targets of MAP3K5-AS2 (MAP3K5 Antisense RNA 2) for further investigation in COA (childhood-onset asthma).

1 Introduction

Allergen-induced bronchial asthma is a pressing global health challenge requiring novel therapeutic targets, particularly for childhood-onset asthma (COA). While Genome-Wide Association Studies (GWAS) have identified numerous susceptibility loci, moving from association to causation in high-dimensional regulatory networks remains a significant hurdle. The mechanisms by which specific regulatory genes influence the broader transcriptome often involve unmeasured common causes.

In this study, we focus on *MAP3K5-AS2* (MAP3K5 Antisense RNA 2), a long non-coding RNA implicated in the regulation of the mitogen-activated protein kinase (MAPK) signaling pathway. Given the ubiquitous role of MAPK pathways in inflammation and oxidative stress response, the *MAP3K5-AS2* gene represents a plausible, yet unexplored, driver of COA (Khorasanizadeh et al., 2017).

To estimate the effects of *MAP3K5-AS2* expression on the transcriptome, we employ a modified CStaR approach (Stekhoven et al., 2012). We first reduce the dimensionality of the search space using stability selection, then learn the causal structure allowing for latent variables with FCI and RFCI using the `pcalg` package in R, and finally apply LV-IDA method (Malinsky and Spirtes, 2016). This approach allows us to estimate the bounds of top causal effects even when the underlying causal graph is only recoverable up to a Markov equivalence class (a PAG).

2 Data and Preprocessing

2.1 Study Population

We analyzed gene expression data profiling human lymphoblastoid cell lines derived from 400 children recruited through a proband with asthma in the MRCA family panel (Liang et al., 2013). The dataset was obtained from the ArrayExpress archive (Accession E-MTAB-1425). The samples include both asthmatic ($n = 258$) and non-asthmatic ($n = 134$) siblings, providing a common asthma-predisposed genetic background for comparison.

2.2 Preprocessing Pipeline

To ensure the validity of causal discovery algorithms, which rely on the faithfulness of the statistical distribution to the underlying graph, pre-processing was performed on the omics data following the protocol outlined by Wang et al. (2016).

The obtained omics data was already RMA-normalized (Liang et al., 2013) to correct for technical variations. We removed non-informative probes by filtering the bottom 40% of probes based on mean expression. From the remaining set, we selected the top 1,000 probes with the highest variance.

Confounding by demographic variables was addressed with univariate regression on the expression levels of the top 1,000 probes against sex and age. Genes significantly associated with sex or age (Benjamini-Hochberg adjusted $q < 0.01$) were removed. Probes were mapped to gene symbols using the `hgu133plus2` (Affymetrix Human Genome U133 Plus 2.0 Array) annotation package in R. Immunoglobulin genes (IGH, IGK, and IGL) inherent to lymphoblastoid cell lines were excluded. The final dataset consisted of 642 genes across 395 samples.

3 Methods

3.1 Neighborhood Selection via Stability Selection

Causal discovery algorithms that account for latent variables, like FCI, are computationally intensive for high-dimensional networks. To address this, feature selection was performed using a stability selection framework with Lasso regression to identify a close approximation for the Markov neighborhood of the target gene *MAP3K5-AS2*.

To determine the optimal level of sparsity, we performed 10-fold cross-validation. We selected the optimal λ value within one standard error of the minimum mean cross-validated error (λ_{1se}) to prioritize a parsimonious model. Stability scores were calculated as the selection probability of each gene across 10,000 $n/2$ subsampled iterations (without replacement) at the optimal cross-validated λ. We kept the top 25 genes that were most frequently selected as predictors of *MAP3K5-AS2* across the bootstrap replicates and used it as the input data for the causal structure learning.

3.2 Causal Structure Learning (FCI and RFCI)

Standard structure learning algorithms like the PC algorithm assume Causal Sufficiency (no unmeasured confounders). Given the complexity of gene regulatory networks, weaker assumptions should be made to ensure more informative graphs are discovered. Therefore, we used the FCI (Fast Causal Inference) and RFCI (Really Fast Causal Inference) algorithms to learn PAGs.

We performed a sensitivity analysis by varying the significance level (α) for the gaussian conditional independence tests ($\alpha = \{0.01, 0.05, 0.10\}$) to obtain reliable graph sparsity and edge discovery. The maximum size of the conditioning sets was limited to 4 for computational efficiency. RFCI was utilized for comparative validation due to its lower sparsity and speed. The optimal network sparsity was identified at $\alpha = 0.01$ and we proceeded with the RFCI-generated PAGs.

3.3 Causal Stability Ranking with LV-IDA

To quantify the impact of *MAP3K5-AS2* on its Markov neighborhood, we applied the Local LV-IDA (Latent Variable - Intervention-calculus when the DAG is Absent) algorithm (Malinsky and Spirtes, 2016). Unlike standard IDA, which assumes causal sufficiency (a DAG), LV-IDA operates on the equivalence class of PAGs, which allows for the inclusion of unmeasured confounders or selection variables in the resulting graph. LV-IDA estimates the multi-set of possible causal effects (not necessarily unique) for each target node by considering all valid directed graphs within the equivalence class represented by the PAG.

We computed the minimum absolute effect size for each target across 100 bootstrap replicates. A target was considered stable if the lower bound of its estimated causal effect was consistently non-zero.

3.4 Network Comparison

To investigate disease-specific mechanisms, we stratified the data into Asthma" and Normal subsets. We independently learned the PAG structures and estimated causal effects for both groups.

4 Results

4.1 Stability Selection and Neighborhood Discovery

The Lasso stability selection successfully reduced the dimensionality from 642 genes to a tractable neighborhood of the 25 genes most strongly associated with *MAP3K5-AS2*. The resulting gene set includes known inflammatory mediators, validating the biological relevance of the selection process.

4.2 Sensitivity of Graph Structure

Using the FCI algorithm, we constructed PAGs for the neighborhood of *MAP3K5-AS2*. As expected, higher α levels resulted in denser graphs with more edges. We selected $\alpha = 0.01$ for downstream LV-IDA analysis with RFCI as it provided a balance between power (detecting true edges) and parsimony (avoiding false positives).

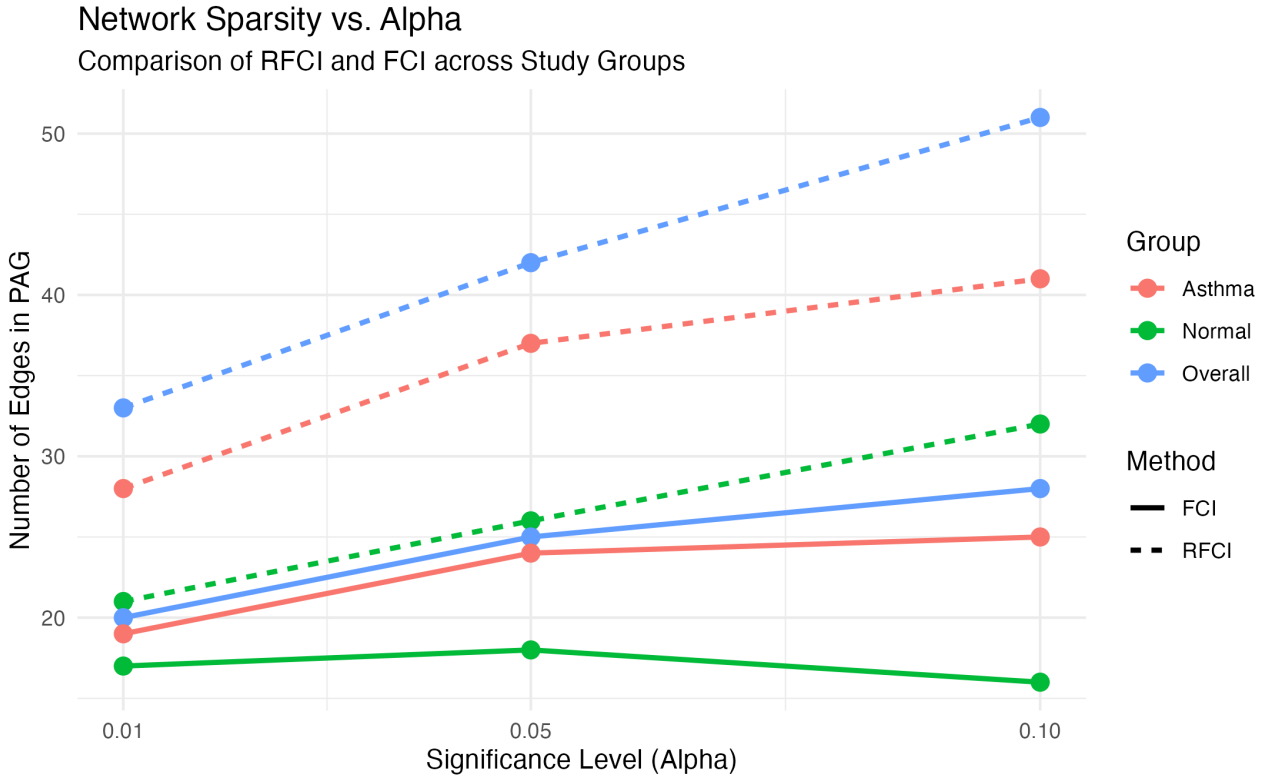

Figure 1: Sensitivity Analysis of PAG Sparsity. The number of edges discovered by FCI is plotted against the significance level (α), demonstrating the impact of the threshold on graph complexity.

4.3 Causal Effect Estimation (LV-IDA)

The LV-IDA algorithm identified several stable downstream targets of *MAP3K5-AS2*. The table below displays the distribution of estimated causal effects. Genes with effect intervals strictly excluding zero are strong candidates for direct regulation. Notably, we did not observe significant differences in the magnitude of effects between the combined, asthmatic, and normal datasets, suggesting *MAP3K5-AS2* does not induce context-dependent regulation.

LV-IDA Results

Comparison of Median Effects between Asthma and Normal Groups

Target Gene	Asthma Group[1]	Normal Group[1]
IL18R1	0.497 (0.289, 0.818)	0.975 (0.276, 1.079)
LAMP5	0.617 (0.512, 0.821)	0.655 (0.501, 0.923)
STAG3	-0.596 (-0.778, -0.461)	-0.598 (-0.904, -0.366)
G0S2	-0.695 (-0.854, -0.415)	-0.711 (-1.009, -0.512)
FLT1	0.357 (0.050, 0.798)	0.629 (0.629, 0.629)
CR2	0.003 (-0.076, 0.296)	0.871 (0.871, 0.871)
PRDM1	0.838 (0.718, 0.975)	0.756 (0.398, 0.860)
BFSP2	0.577 (0.433, 0.691)	0.743 (0.606, 1.006)
H3C10	0.558 (0.365, 0.776)	0.650 (0.540, 0.854)
JCHAIN	0.618 (0.324, 0.823)	0.430 (0.103, 1.003)
BLOC1S5-TXNDC5.1	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)
IRF5	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)
BLOC1S5-TXNDC5	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)
DDX6	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)
SMIM14	0.407 (0.205, 0.596)	0.000 (0.000, 0.000)
SGO2	0.687 (0.303, 0.822)	0.567 (0.333, 0.761)
SKAP2	0.690 (0.588, 0.795)	0.293 (0.131, 0.591)
TSC22D3	-0.604 (-0.911, -0.412)	-0.498 (-0.509, -0.486)
FCRL5	0.559 (0.441, 0.833)	0.539 (0.243, 0.665)
C11orf96	-0.607 (-0.777, -0.219)	-0.176 (-0.395, 0.019)
BHLHE40	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)
LAPTM4B	0.578 (0.426, 0.750)	0.667 (0.667, 0.667)
ARHGAP6	-0.487 (-0.551, -0.104)	-0.106 (-0.106, -0.106)
MIR29B2CHG	-0.531 (-0.904, -0.332)	-0.394 (-0.491, -0.268)

[1] Values displayed as Median (95% Confidence Interval).

Figure 2: Summary of stable causal targets identified by LV-IDA across the different cohorts.

5 Discussion

In this paper, we modified the CStaR approach to identify candidate regulatory targets of *MAP3K5-AS2*, an unexamined target of novel therapeutics for COA. By accounting for latent variables using PAGs and LV-IDA, we significantly reduced the risk of false positives compared to the original CStaR approach. The failure of standard IDA to return valid DAGs underscores the necessity of using methods like LV-IDA that accommodate relaxed causal sufficiency assumptions in complex biological systems. The identification of stable targets suggests that the graph returned reasonable regulatory network targets; however, the lack of significant differences between the cohorts suggests that the causal impact of *MAP3K5-AS2* on its neighborhood may be a fundamental regulatory backbone in B-cells that is not overtly dysregulated in childhood asthma within this sample.

A limitation of this study is the sample size of the stratified groups, which reduces the power of conditional independence tests. Furthermore, the assumption of linearity in the Gaussian conditional independence tests may not capture non-linear gene interactions. Future work should incorporate non-parametric independence tests such as the Generalized Covariance Measure (GCM) test and validate the top causal targets in an independent cohort.

References

Khorasanizadeh, M., Eskian, M., Gelfand, E. W., & Rezaei, N. (2017). Mitogen-activated protein kinases as therapeutic targets for asthma. *Pharmacology & therapeutics*, 174, 112–126.

Liang, L., Morar, N., Dixon, A. L., Lathrop, G. M., Abecasis, G. R., Moffatt, M. F., & Cookson, W. O. (2013). A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. *Genome research*, 23(4), 716–726.

Wang, T., Ren, Z., Ding, Y., Fang, Z., Sun, Z., MacDonald, M. L., Sweet, R. A., Wang, J., & Chen, W. (2016). FastGGM: An Efficient Algorithm for the Inference of Gaussian Graphical Model in Biological Networks. *PLoS computational biology*, 12(2), e1004755.

Malinsky, D., & Spirtes, P. (2016). Estimating Causal Effects with Ancestral Graph Markov Models. *JMLR workshop and conference proceedings*, 52, 299–309.

Stekhoven, D. J., Moraes, I., Sveinbjörnsson, G., Hennig, L., Maathuis, M. H., & Bühlmann, P. (2012). Causal stability ranking. *Bioinformatics*, 28(21), 2819–2823.