RESEARCH ARTICLE

Learning gene networks underlying clinical phenotypes using SNP perturbation

Calvin McCarter[1], **Judie Howrylak**[2], **Seyoung Kim**[3]*

1 Machine Learning Department, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, **2** Pulmonary, Allergy and Critical Care Division, Penn State Milton S. Hershey Medical Center, Hershey, Pennsylvania, USA, **3** Computational Biology Department, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

* sssykim@cs.cmu.edu

OPEN ACCESS

Citation: McCarter C, Howrylak J, Kim S (2020) Learning gene networks underlying clinical phenotypes using SNP perturbation. PLoS Comput Biol 16(10): e1007940. https://doi.org/10.1371/journal.pcbi.1007940

Editor: Christina S. Leslie, Memorial Sloan-Kettering Cancer Center, UNITED STATES

Received: September 9, 2019

Accepted: May 11, 2020

Published: October 23, 2020

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1007940

Copyright: © 2020 McCarter et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The PerturbNet software is available at https://github.com/SeyoungKimLab/PerturbNet.

Funding: The study was funded by NSF MCB-1149885, PA CURE, and NIH 1R21HG011116. The

Abstract

Availability of genome sequence, molecular, and clinical phenotype data for large patient cohorts generated by recent technological advances provides an opportunity to dissect the genetic architecture of complex diseases at system level. However, previous analyses of such data have largely focused on the co-localization of SNPs associated with clinical and expression traits, each identified from genome-wide association studies and expression quantitative trait locus mapping. Thus, their description of the molecular mechanisms behind the SNPs influencing clinical phenotypes was limited to the single gene linked to the co-localized SNP. Here we introduce PerturbNet, a statistical framework for learning gene networks that modulate the influence of genetic variants on phenotypes, using genetic variants as naturally occurring perturbation of a biological system. PerturbNet uses a probabilistic graphical model to directly model the cascade of perturbation from genetic variants to the gene network to the phenotype network along with the networks at each layer of the biological system. PerturbNet learns the entire model by solving a single optimization problem with an efficient algorithm that can analyze human genome-wide data within a few hours. PerturbNet inference procedures extract a detailed description of how the gene network modulates the genetic effects on phenotypes. Using simulated and asthma data, we demonstrate that PerturbNet improves statistical power for detecting disease-linked SNPs and identifies gene networks and network modules mediating the SNP effects on traits, providing deeper insights into the underlying molecular mechanisms.

Author summary

We describe PerturbNet, a statistical framework for learning a gene network that modulates the influence of genetic variants on phenotypes, using genetic variants as naturally occurring perturbation of a biological system. PerturbNet directly models the cascade of perturbation from genetic variants to the gene network to the phenotype network, thus integrating the existing computational tools for eQTL mapping, GWAS, co-localization analysis of eQTL and GWAS variants, and gene network discovery under SNP perturbation within a single statistical framework. We demonstrate that PerturbNet improves

funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

statistical power for detecting disease-linked SNPs and uncovers gene networks mediating the SNP effects on traits, with computational efficiency that allows for human data analysis within several hours.

Introduction

One of the key questions in biology is how genetic variants perturb gene regulatory systems to influence disease susceptibility in population. Leveraging the naturally-occurring perturbation of gene expression by genetic variants to study gene regulatory systems has many advantages over experimental perturbation methods such as gene knockdown [1] and genome editing techniques [2], as it is more cost effective, is more easily applicable to humans, and can lead to more meaningful discoveries via subtle perturbations that occur in nature [3]. Several computational methods have been proposed to learn gene networks perturbed by single nucleotide polymorphisms (SNPs) given expression quantitative trait locus (eQTL) data [4–9]. Further combining eQTL data with clinical phenotype data to study how this gene network perturbed by SNPs in turn influences disease phenotypes has the potential to reveal the complex molecular mechanisms that explain the link between genetic variants and diseases.

However, the existing computational methods for integrating eQTL and clinical phenotype data lack the ability to learn a gene network as a mediator that receives SNP perturbation and then passes the perturbation effects onto clinical traits. Many of these methods have disregarded gene networks entirely, focusing on the co-localization of an eQTL and a trait-associated SNP [10–13], each of which was identified in a separate eQTL mapping [7, 14–16] and a genome-wide association study (GWAS) [17, 18]. Thus, their description of the regulatory role of the trait-associated SNPs was limited to the single gene whose eQTL co-localized with the GWAS SNP. Other works have used a known gene network to evaluate GWAS SNPs [19–21] or have looked for correlated associations to a trait module in a post processing step only after identifying the links between individual SNPs and traits [22, 23], thus missing out on the opportunity to discover gene networks via SNP perturbation. Bayesian networks have been used to learn gene regulatory networks and to form a predictive model for diseases with this network [24]; however, this approach relied on an elaborate pipeline of data analysis to identify disease-related gene modules and genetic variants, leading to loss of statistical power.

Here, we introduce PerturbNet, a computational framework for learning a gene network that underlies clinical traits, using genetic variants as a source of perturbation. PerturbNet unifies eQTL mapping, GWAS, co-localization analysis of eQTLs and GWAS SNPs, and gene network recovery within a single statistical framework.

PerturbNet consists of three components: a probabilistic graphical model, learning algorithms, and inference methods. The PerturbNet model represents the cascade of perturbation from SNPs to gene network and from gene network to phenotype network by stacking two sparse conditional Gaussian graphical models (sCGGMs) [4, 5], previously developed for learning gene networks perturbed by SNPs (Fig 1A). PerturbNet learning algorithm estimates the entire model from data by solving a single optimization problem that is convex with globally optimal solution, achieving high accuracy and minimal loss of statistical power. We introduce Mega-sCGGM, an sCGGM learning algorithm that is orders-of-magnitude faster with no memory restriction compared to the previous Newton coordinate descent (NCD) [25], and use it as a key module of the PerturbNet learning algorithm to allow for large human data analysis. The probabilistic graphical model framework [26] of PerturbNet naturally leads to a set of inference methods for revealing perturbations and regulatory relationships that are not

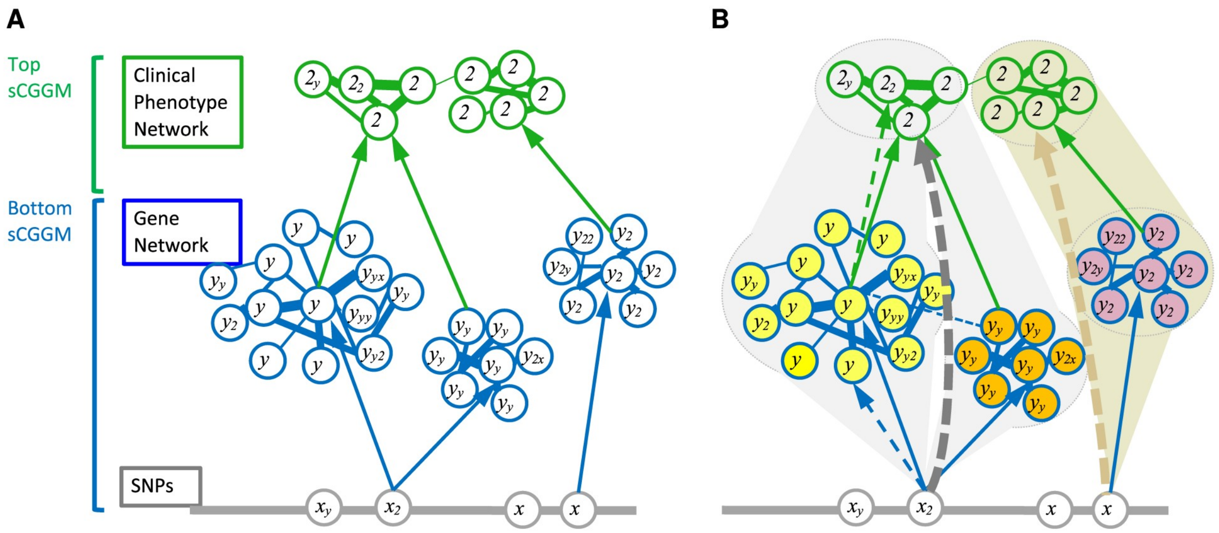

Fig 1. Overview of PerturbNet. (A) PerturbNet model as a cascade of two sCGGMs. The blue sCGGM for $p(\mathbf{y}|\mathbf{x})$ models a gene network perturbed by SNPs and the green sCGGM for $p(\mathbf{z}|\mathbf{y})$ models a clinical trait network perturbed by gene expression levels. (B) PerturbNet inference methods. Given the PerturbNet model in (A), the PerturbNet inference methods infer SNP effects on phenotypes mediated by the gene network (gray and tan dashed arrows), a decomposition of SNP effects on phenotypes into component SNP effects mediated by each gene module (yellow and orange modules for gray dashed arrow and pink module for tan dashed arrow), and the posterior gene network with additional edges between the gene modules influencing the same phenotypes (blue dashed edge). PerturbNet also inherits the inference methods of sCGGMs for inferring indirect perturbation effects resulting from the direct perturbation effects propagating through the networks (blue and green dashed arrows).

https://doi.org/10.1371/journal.pcbi.1007940.g001

explicitly represented as edges but only implicitly present in the model (Fig 1B). Given the model estimated from data, PerturbNet uses these inference methods to characterize how the gene network and modules mediate the SNP effects on clinical traits, uncovering the molecular mechanisms that underlie the link between SNPs and traits.

We demonstrate PerturbNet using simulated and patient cohort data from the Childhood Asthma Management Program (CAMP) [27–29]. Compared to the existing methods that combine eQTL and GWAS data [12, 13], we show PerturbNet provides higher sensitivities for detecting trait-perturbing variants and for identifying genes that mediate the genetic effects on traits, as it learns and leverages gene networks rather than individual genes as mediators. We show the PerturbNet inference methods extract rich information on the role of the gene networks as mediators.

Results

Methods overview

We briefly describe the model, learning algorithms, and inference methods for PerturbNet (see Methods for detail). The PerturbNet model represents the cascaded perturbation from SNPs to gene network to trait network as a Gaussian chain graph model obtained by threading two sCGGMs, each modeling a network and perturbation of this network at the given layer (Fig 1A). The PerturbNet learning algorithm obtains a sparse estimate of the model with few edges in the network by minimizing the negative log-likelihood of data with L_1-regularization. Since the PerturbNet learning algorithm uses an sCGGM learning method as a key module, we introduce Mega-sCGGM, an efficient sCGGM learning algorithms for reducing computation time and for removing memory constraint of the previous algorithm.

PerturbNet provides three inference methods for extracting detailed information on the cascaded network perturbation from the estimated model: INF-I for inferring SNP effects on traits mediated by gene network (gray and tan dashed arrows in Fig 1B), which in turn can be used to reveal the co-localization of the inferred trait-associated SNPs and the eQTLs; INF-II for decomposing SNP effects on traits into components mediated by each gene module (yellow, orange, and pink modules in Fig 1B) within the network, uncovering the mediator genes and modules for SNP effects on traits and revealing the pathways underlying these SNP-trait links; and INF-III for obtaining the posterior gene network after accounting for correlated clinical traits (blue dashed edge in Fig 1B). Additionally, PerturbNet inherits the inference methods of sCGGMs [5] for inferring indirect perturbation effects that arise from another direct perturbation exerting its influence through the network (blue and green dashed edges in Fig 1B).

Using simulated data and asthmatic patient cohort data from the CAMP [29], we evaluate PerturbNet against eCAVIAR [13], PrediXcan [12], and two-layer Lasso (S1 Text). All of these previous methods combined eQTL and GWAS data to identify genes underlying GWAS SNPs, but none considered a gene network as a mediator of SNP effects on traits.

Simulation experiments

Using data simulated from real genotypes in the CAMP study and the known ground-truth models, we evaluated PerturbNet and other methods on the accuracy of detecting trait-perturbing SNPs and mediator genes, and on the accuracy of gene network recovery (see Methods). We simulated datasets with 5, 000 genes and 100 clinical traits, given 10, 000 SNPs on chromosome 1 of 540 non-Hispanic Caucasians from the CAMP study.

For the tasks of identifying trait-perturbing SNPs and mediator genes, PerturbNet had the highest sensitivities across all false discovery rates (FDRs), regardless of the model types used to simulate data, and regardless of whether a gene is a mediator acting independently or cooperating with other genes in the network Figs (2A–2C). Unlike other methods, PerturbNet has the ability to distinguish between direct and indirect perturbation of a network. This enables a categorization of the role of the mediator gene into four possible combinations, based on whether the perturbation effect the mediator gene receives from SNPs and passes onto clinical trait network is direct or indirect. PerturbNet identified these categories for each mediating gene with high accuracy (Fig 2B).

On gene network recovery, PerturbNet had higher accuracy on data simulated with SNP perturbation, compared to sparse Gaussian graphical models (GGMs) [30, 31], a popular statistical method for learning gene networks (Fig 2D). PerturbNet's accuracy did not suffer on data simulated without SNP perturbation. PerturbNet achieved higher accuracy by estimating the entire model of cascaded network perturbation in a single statistical analysis.

Analysis of CAMP data

PerturbNet is scalable for human data analysis. We compared the computation time of PerturbNet against eCAVIAR, PrediXcan, and two-layer Lasso. We also compared the scalability of Mega-sCGGM, the key subroutine of PerturbNet, against Lasso and NCD [25], the previous state-of-the-art optimization method for learning an sCGGM. We used all CAMP data for 11,598 genes and 35 phenotypes, while varying the number of SNPs from 40,056 SNPs in chromosome 1 to 212,757 SNPs in chromosomes 1-6 and to 495,597 SNPs in all autosomal chromosomes.

PerturbNet analyzed all CAMP data in less than four hours and scaled similarly to other methods that are not concerned with gene networks (Fig 3A). On expression and SNP data, Mega-sCGGM scaled similarly to Lasso [32], a computationally efficient but less powerful method, as it only learns eQTLs but not the gene network (Fig 3B). Mega-sCGGM was also

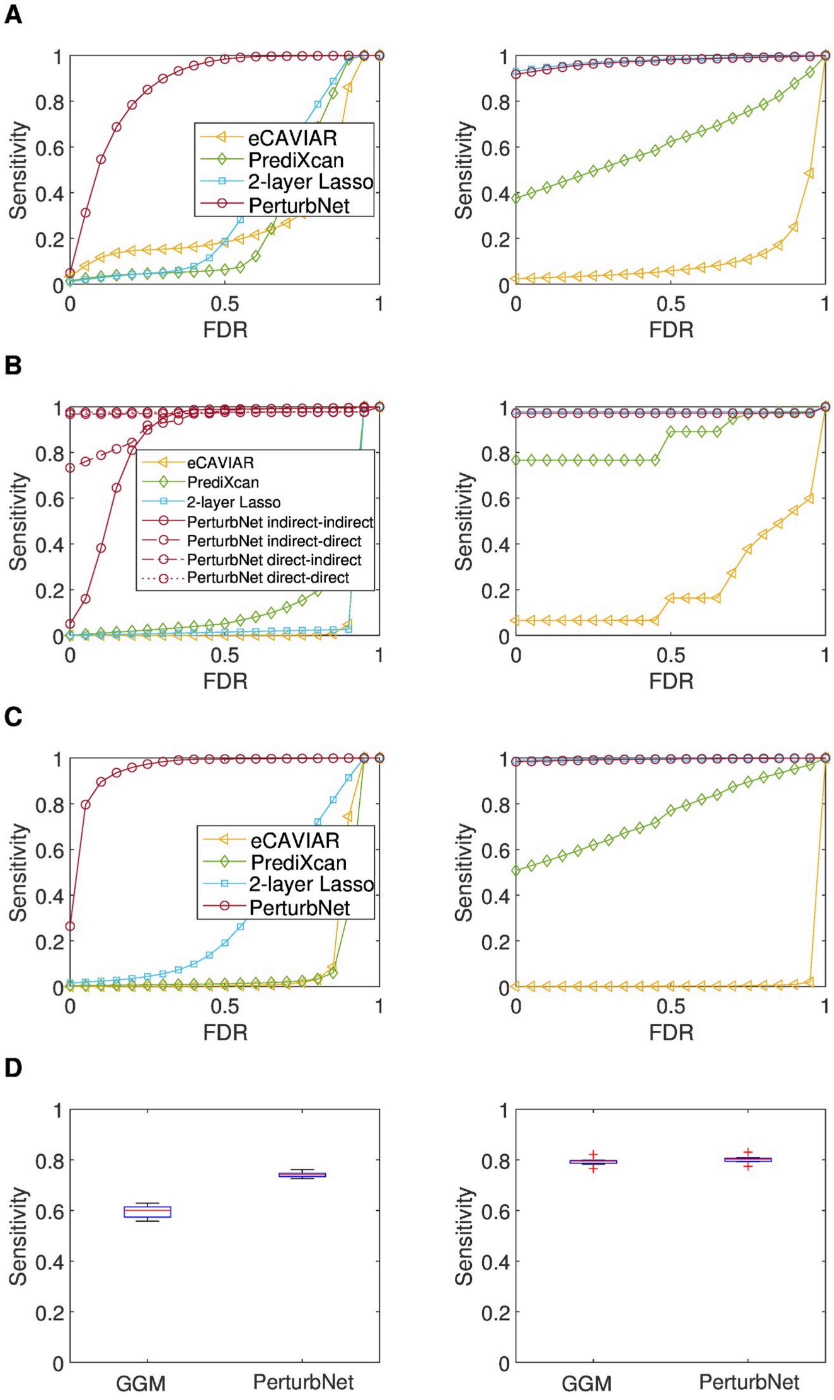

Fig 2. Results on simulated data. The accuracy of PerturbNet, eCAVIAR, PrediXcan, and two-layer Lasso is shown for (A) the recovery of SNPs perturbing traits, (B) the recovery of genes mediating the effect of each SNP on each trait (for PerturbNet, categories for the role of each mediator gene are shown for four possible combinations of direct vs. indirect perturbations that the mediator gene receives from SNPs and passes onto clinical trait network), and (C) the recovery of genes mediating the overall SNP effects on each trait. Ground-truth models with two-layer sCGGMs (left) and two-layer linear regression models (right) were used to simulate data. (D) The accuracy of PerturbNet and GGM on gene network recovery. Networks with SNP perturbation (left) and without SNP perturbation (right) were used as the ground-truth models. Sensitivities at FDR = 0.05 are shown.

https://doi.org/10.1371/journal.pcbi.1007940.g002

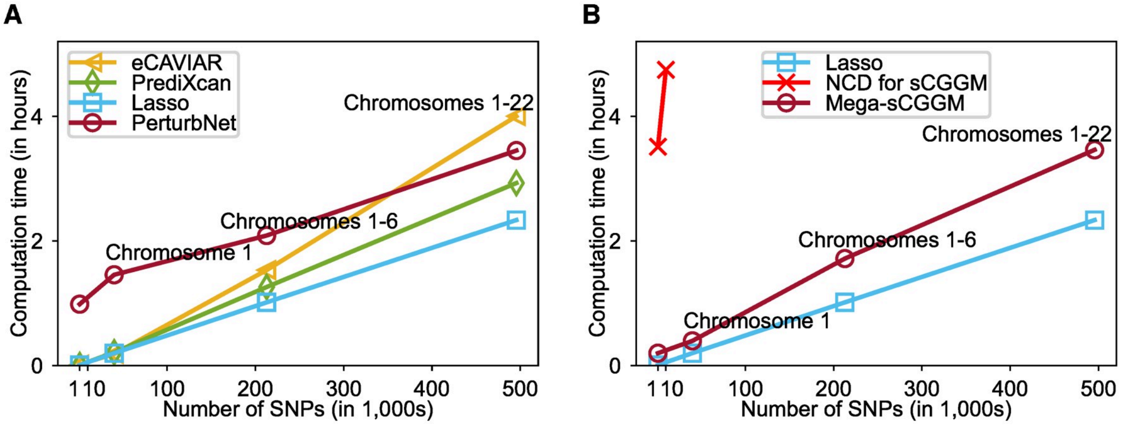

Fig 3. Comparison of computation time. (A) Computation time for analysis of SNP, expression, and clinical trait data from the CAMP study. (B) Computation time for analysis of eQTL data from the CAMP study. NCD, the previous algorithm for sCGGMs, ran out of memory at chromosome 1. The expression levels of 11,598 genes, 35 traits, and varying numbers of SNPs were used.

https://doi.org/10.1371/journal.pcbi.1007940.g003

significantly more efficient than NCD in terms of both memory requirement and computation time: it analyzed all 495, 597 SNPs within three and a half hours, whereas NCD spent the same amount of time on 1,000 SNPs and ran out of memory on SNPs from chromosome 1. When an estimate from NCD could be obtained on the dataset with 1, 000 SNPs, NCD and Mega-sCGGM obtained an identical estimate, as was expected since they solve the same convex optimization problem with a single global optimum using an exact method without approximation.

PerturbNet asthma gene network and gene modules. PerturbNet has the unique ability to determine the gene network perturbed by SNPs underlying clinical traits. To demonstrate this, we first examine the PerturbNet model estimated from the CAMP data (S1 Fig), focusing on the asthma gene network. In the gene network, out of 11,598 genes, 6,102 genes had at least one neighbor in the network, while the rest of the genes were singleton nodes in the network. To characterize the module structure in this network for further functional analysis, we post-processed the estimated gene network by excluding the singleton nodes and clustering the part of the network corresponding to the 6,102 genes into 20 modules, using the *k*-way network clustering algorithm METIS [33]. We tried a variety of settings for the number of clusters *k* and chose *k* = 20 modules as it produced modules that map to dense subnetworks within the full network from visual inspection (S1(B) Table).

A close examination of the 20 modules in the PerturbNet gene network revealed that a subset of those modules, modules 13-20, are likely to be involved in asthma. To determine the functional role of the 20 gene modules, we performed GO gene set enrichment analysis [34]. For each module, we performed a Fisher's exact test to find significantly enriched GO categories in biological processes (p-value < 0.05 after Bonferroni correction for multiple testing), using the GO database with annotations for 21,002 genes. Modules 13-15 had statistically significant enrichment of GO terms related to immune system function (S1 Table), suggesting these modules are likely to play a role in asthma, since asthma is an immune disorder. Even though modules 16-20 did not have any significant enrichment of the same GO categories, the set of 374 genes in these modules that are connected to genes in modules 13-15 in the posterior gene network from the PerturbNet inference method INF-III (S2 Fig) were significantly enriched for several GO categories related to immune system processes (p-value < 0.05 after Bonferroni correction; S2 Table). These genes with immune system GO annotations formed sub-clusters within each of modules 13-20. Overall, the enrichment of immune-related genes

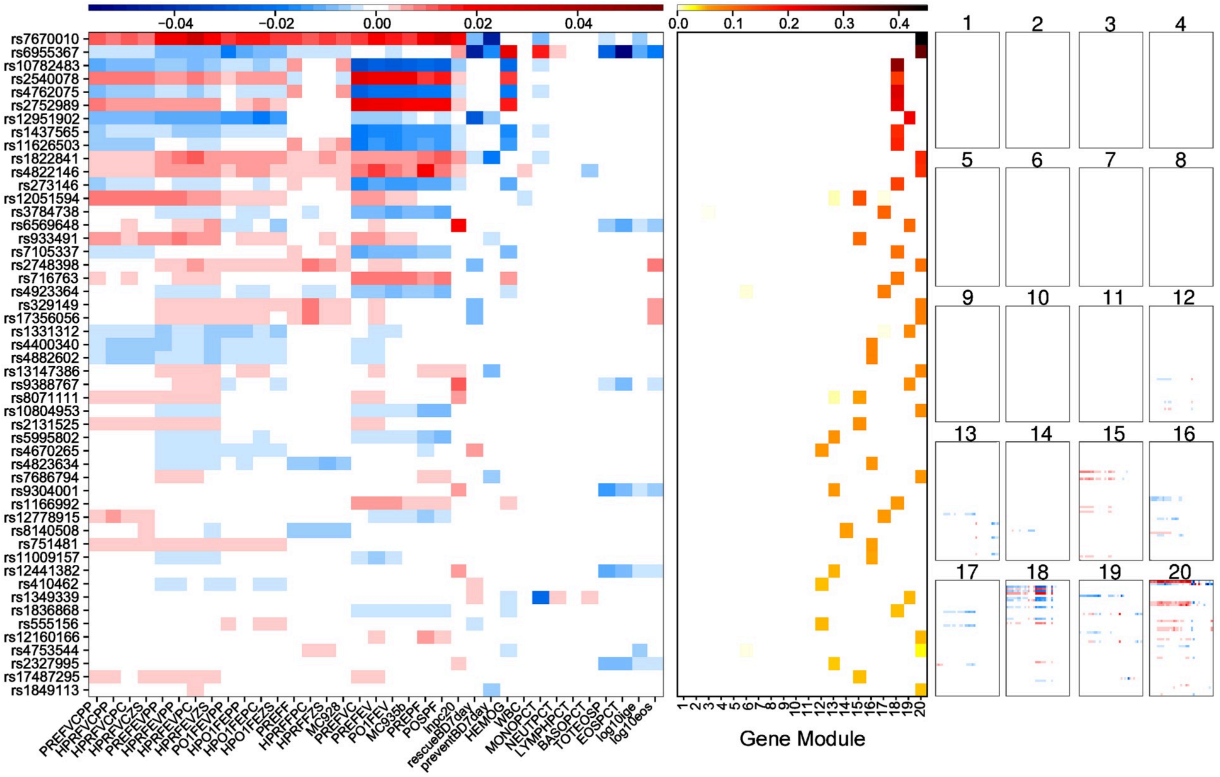

Fig 4. SNP effects on asthma traits mediated by gene network and gene modules from PerturbNet. SNP effects on traits for top 50 SNPs from the PerturbNet inference method INF-I (left). The decomposition of these SNP effects into 20 component effects mediated by each gene module from the PerturbNet inference method INF-II (right). Summary of these component effects by summing over traits (center).

https://doi.org/10.1371/journal.pcbi.1007940.g004

in modules 13-20 suggests that these modules are likely the network components involved in the disease process of asthma.

Furthermore, in the overall PerturbNet model, while all gene modules were perturbed by SNPs in the SNP-to-gene sCGGM (S3(A) Fig), modules 13-20 had substantially larger effects on the asthma phenotypes in the gene-to-trait sCGGM (S3(B) Fig), providing additional evidence that modules 13-20 are likely to be implicated in asthma.

PerturbNet finds gene modules 13-20 mediate most of the genetic effects on asthma traits. Another unique feature of PerturbNet is its inference methods for uncovering how SNP effects on traits are mediated by the gene network and modules. Given the estimated PerturbNet model, we first applied the inference method INF-I to infer the genetic effects on traits mediated by the network, identifying top 500 SNPs with the largest overall effects on traits (S4 Fig and Fig 4 left).

Then we applied INF-II to infer the decomposition of these SNP effects from INF-I into component effects mediated by each of the 20 gene modules (Fig 4 right). For 99 out of top 500 SNPs, the primary mediators were modules rather than single isolated genes in the network. Moreover, for all of those 99 SNPs, modules 13-20 mediated nearly all of those genetic effects (Fig 4 right), providing evidence that the modules enriched with genes involved in immume system function above (S1 and S2 Tables, S3 Fig) also explain the molecular mechanisms behind the top asthma SNPs. Additionally, the effect of each SNP on traits was mediated by only one of modules 13-20 (Fig 4 center), except for two SNPs with two mediator modules, suggesting a SNP influences asthma traits via localized perturbation of the gene network. Our results demonstrate that PerturbNet inference methods can identify candidate gene modules that are mediators of SNP effects on traits. Unlike eCAVIAR, PerturbNet explains the molecular mechanisms behind trait-associated SNPs not simply through an individual gene, whose eQTLs co-localize with GWAS SNPs, but through gene modules mediating SNP effects on traits.

Most of these asthma SNPs from the inference co-localized with eQTLs for modules 13-20 in the estimated SNP-to-gene sCGGM component of the PerturbNet model (S5 Fig). Although a small subset of these trait-perturbing SNPs co-localized with eQTLs for modules 1-12, the genetic effects mediated by these modules were weak, suggesting these modules do not propagate the SNP perturbations to the traits and are not likely to play a central role in the development of the disease.

PerturbNet SNPs are most strongly supported by external annotations. For the top asthma-associated SNPs identified by PerturbNet, eCAVIAR, PrediXcan, and two-layer Lasso, we examined the functional annotations in three external resources (see Methods): the Encyclopedia of DNA Elements (ENCODE) DNase hypersensitivity sites (GSM1008572) in GM12878 B-lymphoblastoid cell line [35], a blood cell type as in CD4+ T lymphocytes in the CAMP study; RegulomeDB [36] integrating ENCODE and other sources of annotations for diverse cell types; and SNPs in the National Human Genome Research Institute (NHGRI) GWAS catalog.

PerturbNet SNPs were most strongly supported as functional by all of the three external resources. PerturbNet SNPs had the largest overlap with the DNase-seq hypersensitivity sites from the ENCODE B-lymphoblastoid cell-line [35] with statistical significance (Fig 5A), and also with functionally annotated SNPs in RegulomeDB [36], especially with SNPs in the RegulomeDB category 1, which corresponds to the strongest evidence of being located in the functional region and includes SNPs previously identified as eQTLs (Fig 5B). Furthermore, PerturbNet had the largest numbers of SNPs located within 10, 20, 50, and 100kb of 557 asthma SNPs in the NHGRI GWAS catalog [37] (Fig 5C and S7 Fig). Applying the PerturbNet inference method to the PerturbNet asthma SNPs within 20kb of the NHGRI GWAS SNPs revealed for all except for one SNP the SNP effects on traits were mediated by modules 13-20 (Fig 5D), showing they may be the key modules underlying these SNPs from the GWAS catalog.

Module 13 mediates the genetic effects of SNP rs12441382 on asthma traits. We examined module 13 mediating the perturbation of traits by SNP rs12441382, one of the top PerturbNet SNPs (Fig 6). This SNP is located 17kb from rs1841128 in the NHGRI catalog [37] for a lung function trait, forced vital capacity [38]. These two SNPs have normalized linkage disequilibrium coefficient $D' = 1.0$ in the 1000 Genomes Project CEU population [39]. PerturbNet found two genes, *ATF3* and *EGR2*, that are directly perturbed by this SNP with the strongest effect sizes. Both genes have been previously linked to allergic asthma. *ATF3* is a known negative regulator of allergic asthma and was recently proposed to be a hub of the cellular adaptive-response network, playing a key role in immune diseases [40]. The perturbation of *ATF3* by this SNP is also reflected in statistically significant association in univariate regression analysis (FDR $q = 6.851 \times 10^{-5}$). *EGR2* has been linked to migration of CD4+ T cells to lung and to blood eosinophil levels in asthma [41–43]. Indirectly perturbed genes near *ATF3* and *EGR2* in the network include *C2* [44–46], *SERPING1* [47], *ZG16B* [48], and *CEACAM3* [49], all of which have been linked to immune response, asthma, or auto-immune diseases. PerturbNet provided new insights into the molecular characterization based on *ATF3*-centered pathway for the known GWAS SNP whose functional role in asthma has not been fully elucidated.

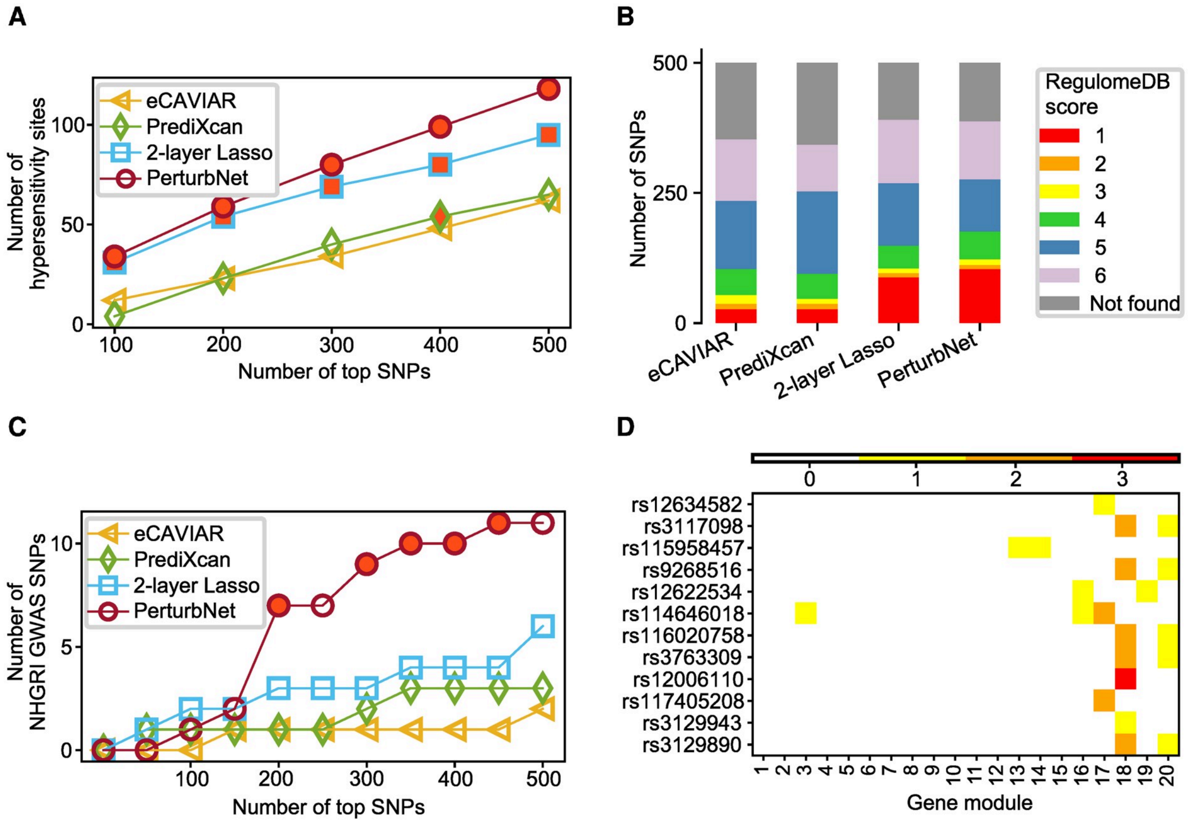

Fig 5. Comparison of top asthma SNPs from different methods. Top asthma SNPs identified by PerturbNet, eCAVIAR, PrediXcan, and two-layer Lasso are compared against external annotations. (A) Overlaps between top asthma SNPs from each method and the ENCODE DNase hypersensitivity sites for B-lymphoblastoid cell line. Statistically significant enrichments (p-value < 0.05) are highlighted with red. (B) Overlaps between top 500 asthma SNPs from each method and functionally annotated SNPs in RegulomeDB. Lower scores indicate more likely to be functional. (C) Overlaps with asthma SNPs in the NHGRI GWAS catalog. An overlap is defined as less than 20kb between two SNPs. Statistically significant enrichments (permutation test p-value < 0.05) are highlighted with red. (D) PerturbNet gene modules mediating the effects of NHGRI GWAS SNPs that overlap with PerturbNet asthma SNPs in (C). The colorbar indicates the number of PerturbNet asthma SNPs overlapping with each NHGRI SNP.

https://doi.org/10.1371/journal.pcbi.1007940.g005

PerturbNet robustness analysis on asthma data. We evaluated the robustness of PerturbNet by fitting the model to perturbed CAMP datasets with either fewer samples or samples with added noise. To generate a modified CAMP dataset with fewer samples, given a desired fraction of samples, we took a random subset of samples without replacement of the fully-observed samples and of the partially-observed samples. To generate CAMP datasets with added noise, given a desired noise standard deviation, we added independent Gaussian noise to both gene expression and phenotype data. For each of the modified datasets, we fit a PerturbNet model. Then, we evaluated the nonzero elements of the estimated parameters in terms of the area under curve (AUC) of receiver operating characteristic (ROC) curves, assuming the parameters obtained from the original CAMP dataset is the ground-truth. We found that the PerturbNet estimates tend to be robust across different fractions of samples and noise levels and do not show significant changes with small perturbation in data (Fig 7).

Discussion

PerturbNet combined genetic variant, expression, and trait data in a single statistical analysis to model the cascade of influence from SNPs to gene network to phenotype network in a

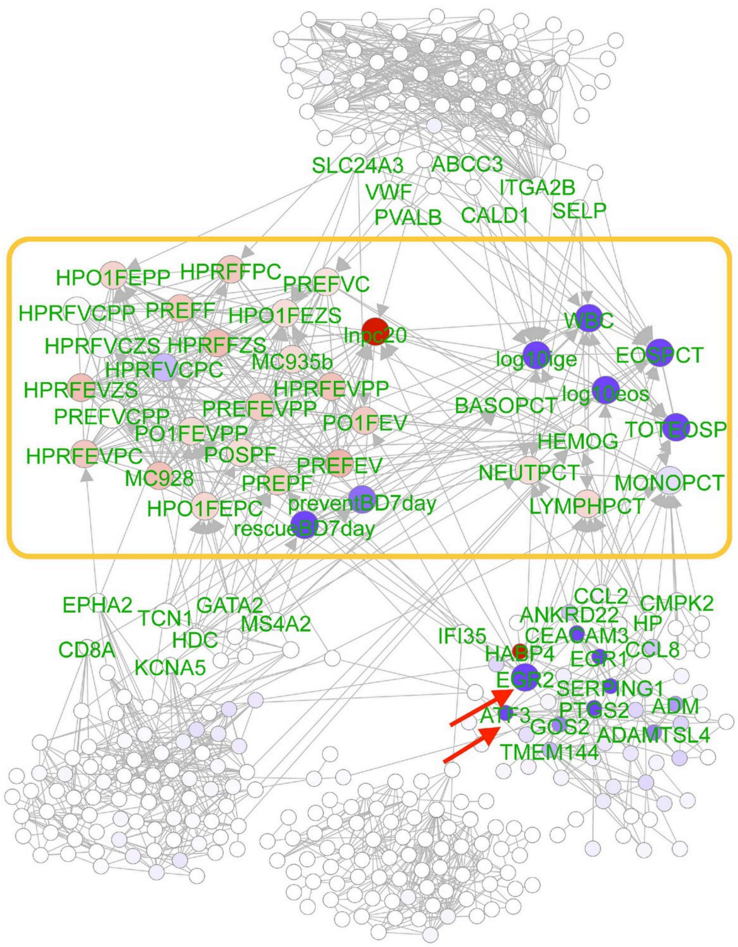

Fig 6. PerturbNet module 13 mediating the effects of SNP rs12441382 on asthma traits. The PerturbNet cascade from SNP rs12441382 to module 13 (outside of yellow box) to trait network (inside of yellow box). Direct SNP perturbations (red arrows), indirect SNP perturbations of expression levels and phenotypes (node colors), and the component of the SNP effects mediated by each gene, summed across all traits (gene node sizes) are shown.

https://doi.org/10.1371/journal.pcbi.1007940.g006

multi-layered biological system. Our results demonstrated several key advantages of our approach. First, PerturbNet detected SNPs perturbing traits through gene expression with higher sensitivity by leveraging gene networks, compared to the existing methods that narrowly focused on detecting the co-localization of eQTLs and GWAS variants with multi-stage analysis [10–13]. Second, unlike other methods, PerturbNet, using inference methods, could infer how different parts of the gene network act as mediators of the genetic influence on phenotypes, providing insights into the gene regulatory mechanisms underlying the genetic effects on the phenotypes. Finally, PerturbNet accomplished these without sacrificing the computation time, allowing for human data analysis within a few hours.

PerturbNet is a flexible tool that can be easily extended to model a more complex biological system under genetic influence. Because PerturbNet uses sCGGMs as building blocks, sCGGMs could be threaded in different ways to integrate different data types. To model tissue-specific gene networks underlying genetic effects on clinical traits, a PerturbNet model could be set up with multiple SNP-to-gene sCGGMs, one for each tissue type, all of which are linked to the same gene-to-trait sCGGM. Such a model could explain the phenotypic variability by the gene expression of the relevant tissue types and enable a discovery of SNPs that affect

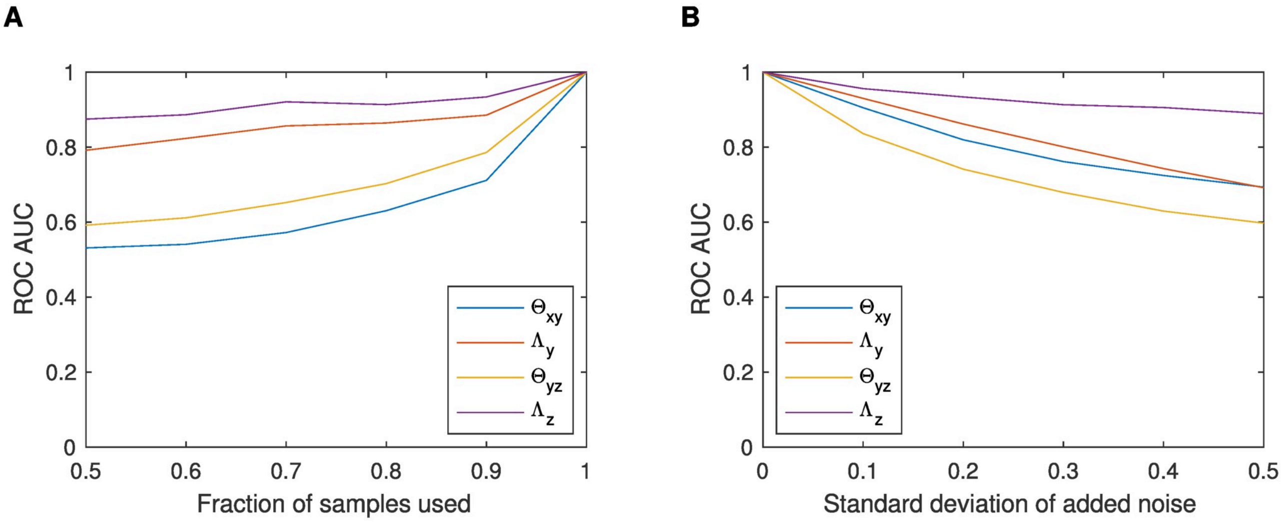

Fig 7. Robustness of the PerturbNet estimates on perturbed CAMP datasets. ROC AUCs are shown for PerturbNet models fit to the CAMP datasets modified with (A) fractions of samples and (B) noise added to the expression and clinical trait data. The non-zero elements in the PerturbNet parameters obtained from the modified datasets were compared against those from the original CAMP data.

https://doi.org/10.1371/journal.pcbi.1007940.g007

disease phenotypes through specific tissue types. To model genetic perturbation of a biological system with more than two layers that includes microRNAs, proteomes, and metabolomes, multiple sCGGMs could be combined in a PerturbNet model by assigning an sCGGM to each level of the cascaded perturbation. While in this work, we only considered continuous-valued traits, discrete-valued traits could be modeled with conditional random fields (CRFs) [26, 50], a discrete counterpart of sCGGMs, where efficient learning and inference algorithms are available from the huge literature on applications of CRFs to image and text data analysis.

Our study considered genetic variant, expression, and clinical data collected from the same patient cohort. One future direction is to apply PerturbNet to datasets, where different data types were collected from different cohorts. Each component sCGGM in the model could be estimated using data collected from different cohorts, if not all data types are available from the same cohort.

Another future direction is to extend PerturbNet to account for the characteristics of non-independent samples such as population structure. Linear mixed effects models (LMMs) have been widely used in genetics to model non-homogenous samples with population stratification. Multi-trait mixed effects models have been proposed to extend an LMM to model multivariate traits through matrix normal distribution [51]. A similar strategy could be used to combine LMMs with sCGGMs in a PerturbNet model.

Finally, PerturbNet could be extended to model non-linear relationships among expression/clinical traits and between traits genetic variants. The sCGGM component of PerturbNet assumed that multivariate expression traits are Gaussian distributed. This is essentially equivalent to assuming that the expression of each gene is a linear function of genetic variants and the expression of other neighboring genes in the gene network. Gaussian graphical models have been extended to non-parametric models to model non-linear dependencies among multiple random variables in high-dimensional setting [52]. Our future work includes employing a similar strategy to extend sCGGMs of PerturbNet to model non-linear dependencies.

The PerturbNet software is available at https://github.com/SeyoungKimLab/PerturbNet.

Methods

We describe the probabilistic graphical model, learning algorithms, and inference methods of our PerturbNet framework.

PerturbNet model

We introduce the probabilistic graphical model of the PerturbNet framework. Let $\mathbf{x}$ denote genotypes at p loci for an individual, given as the number of minor alleles at the loci taking values from {0, 1, 2}. Let $\mathbf{y} \in \mathbb{R}^q$ denote expression levels for q genes and $\mathbf{z} \in \mathbb{R}^r$ measurements for r phenotypes for the same individual. Then, PerturbNet models the cascade of influence from SNPs to a gene network to a phenotype network as a Gaussian chain graph model (Fig 1A), a factorized conditional probability distribution defined as follows:

$$p(\mathbf{y}, \mathbf{z}|\mathbf{x}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{z}|\mathbf{y}). \quad (1)$$

Each probability factor above is modeled as an sCGGM [4, 5, 25]:

$$p(\mathbf{y}|\mathbf{x}) = \exp\left(-\frac{1}{2}\mathbf{y}^T\mathbf{\Lambda}_\mathbf{y}\mathbf{y} - \mathbf{x}^T\mathbf{\Theta}_{\mathbf{xy}}\mathbf{y}\right)/Z_1(\mathbf{x}), \quad (2)$$

$$p(\mathbf{z}|\mathbf{y}) = \exp\left(-\frac{1}{2}\mathbf{z}^T\mathbf{\Lambda}_\mathbf{z}\mathbf{z} - \mathbf{y}^T\mathbf{\Theta}_{\mathbf{yz}}\mathbf{z}\right)/Z_2(\mathbf{y}). \quad (3)$$

The first probability factor in Eq (2) models the gene network perturbed by SNPs, representing the gene network as a $q \times q$ positive definite matrix $\mathbf{\Lambda}_\mathbf{y}$ and the SNP perturbation of this network as $\mathbf{\Theta}_{\mathbf{xy}} \in \mathbb{R}^{p\times q}$. The second probability factor in Eq (3) models the phenotype network perturbed by gene expression levels, where the phenotype network $\mathbf{\Lambda}_\mathbf{z}$ is an $r \times r$ positive definite matrix and the perturbation of this network by gene expression levels is modeled as $\mathbf{\Theta}_{\mathbf{yz}} \in \mathbb{R}^{q\times r}$. $Z_1(\mathbf{x})$ and $Z_2(\mathbf{y})$ in Eqs (2) and (3) are constants for ensuring that each sCGGM is a proper probability distribution that integrates to one. Our model in Eq (1) defines a probability distribution over the graph shown in Fig 1A. A non-zero value in the (i, j)th element of the network parameters, $[\mathbf{\Lambda}_\mathbf{y}]_{i,j}$ of $\mathbf{\Lambda}_\mathbf{y}$ and $[\mathbf{\Lambda}_\mathbf{z}]_{i,j}$ of $\mathbf{\Lambda}_\mathbf{z}$, corresponds to presence of an edge between the ith and jth nodes of the corresponding network. Similarly, a non-zero value in the (i, j)th element of the perturbation parameters, $[\mathbf{\Theta}_{\mathbf{xy}}]_{i,j}$ of $[\mathbf{\Theta}_{\mathbf{xy}}]$ and $[\mathbf{\Theta}_{\mathbf{yz}}]_{i,j}$ of $[\mathbf{\Theta}_{\mathbf{yz}}]$, indicates an edge between the ith perturbant and the jth node in the network.

This Gaussian chain graph model corresponds to the continuous counterpart of the chain graph model obtained by threading CRFs for discrete random variables [50]. CRFs and the chain graph models built from CRFs have been hugely popular in other application areas of statistical machine learning such as text modeling and image analysis for modeling multiple correlated output features influenced by input features [26, 50, 53]. Here, we explore the use of a chain graph model constructed with sCGGMs. We develop an efficient learning algorithm that runs on human data within a few hours and a set of inference algorithms for dissecting the gene regulatory mechanisms that govern the influence of SNPs on phenotypes.

PerturbNet inference methods

The probabilistic graphical model for PerturbNet naturally leads to a set of inference methods for revealing dependencies that are not explicitly represented as edges in the model. In general, inference methods in probabilistic graphical models are computationally expensive [26].

However, because the PerturbNet model is built upon sCGGMs, a form of Gaussian distributions, it is possible to obtain efficient inference methods that involve simple matrix operations. Below, we describe two inference methods that PerturbNet inherits from sCGGM and introduce three inference methods for the Gaussian chain graph model of PerturbNet.

PerturbNet inherits the following two inference methods directly from the inference method for an sCGGM [4, 5]. They are used to infer the indirect perturbation effects that arise from the direct perturbation effects propagating to other parts of the network.

- **Indirect SNP perturbation effects on gene expression levels:** $\mathbf{B}_{\mathbf{xy}} = -\boldsymbol{\Theta}_{\mathbf{xy}}\boldsymbol{\Lambda}_{\mathbf{y}}^{-1}$, where $[\mathbf{B}_{\mathbf{xy}}]_{k,i}$ represents the indirect perturbation effect of SNP k on the expression level of gene i (Fig 1B, blue dashed arrow). This can be seen by re-writing the sCGGM component $p(\mathbf{y}|\mathbf{x})$ of the PerturbNet model as follows:

$$p(\mathbf{y}|\mathbf{x}) = N(\mathbf{B}_{\mathbf{xy}}^T\mathbf{x}, \boldsymbol{\Lambda}_{\mathbf{y}}^{-1}), \tag{4}$$

from which the marginal distribution for the expression level $[\mathbf{y}]_i$ of gene i can be obtained as $p([\mathbf{y}]_i|\mathbf{x}) = N([[\mathbf{B}_{\mathbf{xy}}]_{:,i}]^T\mathbf{x}, [\boldsymbol{\Lambda}_{\mathbf{y}}^{-1}]_{i,i})$, where $[\mathbf{B}_{\mathbf{xy}}]_{:,i}$ represents the ith column of $\mathbf{B}_{\mathbf{xy}}$. While $[\boldsymbol{\Theta}_{\mathbf{xy}}]_{k,i}$ models the direct perturbation effect of SNP k on the expression of gene i, $[\mathbf{B}_{\mathbf{xy}}]_{k,i}$ corresponds to the overall perturbation effect that aggregates all indirect influence of SNP k on gene i through other genes. When SNP k does not influence the expression of gene i directly but exerts influence on gene i through other genes connected to gene i in the network $\boldsymbol{\Lambda}_{\mathbf{y}}$, we have $[\boldsymbol{\Theta}_{\mathbf{xy}}]_{k,i} = 0$ but $[\mathbf{B}_{\mathbf{xy}}]_{k,i} \neq 0$.

- **Indirect effects of gene expression levels on clinical phenotypes:** $\mathbf{B}_{\mathbf{yz}} = -\boldsymbol{\Theta}_{\mathbf{yz}}\boldsymbol{\Lambda}_{\mathbf{z}}^{-1}$, where $[\mathbf{B}_{\mathbf{yz}}]_{k,i}$ represents the indirect influence of the expression level of gene k on phenotype i (Fig 1B, green dashed arrow). Similarly as above, this can be seen by deriving the marginal distribution from the sCGGM component $p(\mathbf{z}|\mathbf{y})$ of the PerturbNet model as follows:

$$p(\mathbf{z}|\mathbf{y}) = N(\mathbf{B}_{\mathbf{yz}}^T\mathbf{y}, \boldsymbol{\Lambda}_{\mathbf{z}}^{-1}).$$

Then, the marginal distribution for $[\mathbf{z}]_i$ of phenotype i can be obtained as $p([\mathbf{z}]_i|\mathbf{y}) = N([[\mathbf{B}_{\mathbf{yz}}]_{:,i}]^T\mathbf{y}, [\boldsymbol{\Lambda}_{\mathbf{z}}^{-1}]_{i,i})$.

PerturbNet builds upon the sCGGM inference methods above and provides the following three inference methods on the sparse Gaussian chain graph model for characterizing the role of the gene network as a mediator of SNP effects on clinical traits.

- **INF-I for inferring SNP effects on clinical phenotypes mediated by gene network:** INF-I computes $\mathbf{B}_{\mathbf{xz}} = \mathbf{B}_{\mathbf{xy}}\,\mathbf{B}_{\mathbf{yz}}$, where $[\mathbf{B}_{\mathbf{xz}}]_{k,i}$ represents the overall influence of SNP k on phenotype i mediated by gene network $\boldsymbol{\Lambda}_{\mathbf{y}}$ (Fig 1B, gray and tan dashed arrows). Such SNP effects on phenotypes are not directly modeled in the PerturbNet model but can be inferred by deriving the marginal distribution $p(\mathbf{z}|\mathbf{x})$ as follows:

$$p(\mathbf{z}|\mathbf{x}) = N(\mathbf{B}_{\mathbf{xz}}^T\mathbf{x},\ \boldsymbol{\Lambda}_{\mathbf{z}}^{-1} + \boldsymbol{\Lambda}_{\mathbf{z}}^{-1}\boldsymbol{\Theta}_{\mathbf{yz}}^T\boldsymbol{\Lambda}_{\mathbf{y}}^{-1}\boldsymbol{\Theta}_{\mathbf{yz}}\boldsymbol{\Lambda}_{\mathbf{z}}^{-1}). \tag{5}$$

From this, the marginal distribution of $[\mathbf{z}]_i$ for phenotype i given $\mathbf{x}$ can be obtained as $p([\mathbf{z}]_i|\mathbf{x}) = N([[\mathbf{B}_{\mathbf{xz}}]_{:,i}]^T\mathbf{x}, [\boldsymbol{\Lambda}_{\mathbf{z}}^{-1} + \boldsymbol{\Lambda}_{\mathbf{z}}^{-1}\boldsymbol{\Theta}_{\mathbf{yz}}^T\boldsymbol{\Lambda}_{\mathbf{y}}^{-1}\boldsymbol{\Theta}_{\mathbf{yz}}\boldsymbol{\Lambda}_{\mathbf{z}}^{-1}]_{i,i})$. Eq (5) can also be obtained from the marginalization of $\mathbf{y}$ in the joint distribution $p(\mathbf{z}, \mathbf{y}|\mathbf{x})$ derived from Eq (1):

$$p(\mathbf{z}, \mathbf{y}|\mathbf{x}) = N(-\boldsymbol{\Lambda}_{\mathbf{zy}}{}^{-1}\boldsymbol{\Theta}_{\mathbf{zy},\mathbf{x}}^T\mathbf{x}, \boldsymbol{\Lambda}_{\mathbf{zy}}{}^{-1}), \tag{6}$$

where $\mathbf{\Theta_{zy,x}} = (\mathbf{0}_{p\times r}, \mathbf{\Theta_{xy}})$ with $\mathbf{0}_{p\times r}$ for $p \times r$ matrix of 0's and $\mathbf{\Lambda_{zy}} = \begin{pmatrix} \mathbf{\Lambda_z} & \mathbf{\Theta}_{\mathbf{yz}}^T \\ \mathbf{\Theta_{yz}} & \mathbf{\Lambda_y} + \mathbf{\Theta_{yz}}\mathbf{\Lambda_z}^{-1}\mathbf{\Theta}_{\mathbf{yz}}^T \end{pmatrix}$

This joint distribution is an alternative representation of the PerturbNet model in Eq (1) and corresponds to another sCGGM over **y** and **z** conditional on **x** with network $\mathbf{\Lambda_{zy}}$ and perturbation $\mathbf{\Theta_{zy,x}}$.

- **INF-II for inferring SNP effects on clinical phenotypes mediated by a gene module:** Let $M_1, \ldots, M_s$ be disjoint subsets of q genes, where $\cup_{m=1,\ldots,s} M_m$ is the full set of q genes (Fig 1B, gray and tan gene modules). Then, the overall SNP effects on phenotypes in $\mathbf{B_{xz}}$ can be decomposed into SNP effects on phenotypes mediated by each gene module $\mathbf{B}_{\mathbf{xz}}^{M_m}$ as follows:

$$\mathbf{B_{xz}} = \sum_{m=1}^{s} \mathbf{B}_{\mathbf{xz}}^{M_m},$$

where

$$\mathbf{B}_{\mathbf{xz}}^{M_m} = \sum_{j\in M_m} [\mathbf{B_{xy}}]_{:,j} [\mathbf{B_{yz}}]_{j,:}.$$

In this decomposition, $[\mathbf{B}_{\mathbf{xz}}^{M_m}]_{k,i}$ quantifies the effect of SNP k on phenotype i through the expression levels of genes in module M_m. If the module size is 1, $[\mathbf{B}_{\mathbf{xz}}^{M_m}]_{k,i}$ corresponds to the SNP effect on the phenotype mediated by the single gene in the module.

- **INF-III for inferring posterior gene network after seeing phenotype data:** INF-III computes $\mathbf{\Lambda}_{\mathbf{y|x,z}} = \mathbf{\Lambda_y} + \mathbf{\Theta_{yz}}\mathbf{\Lambda}_{\mathbf{z}}^{-1}\mathbf{\Theta}_{\mathbf{yz}}^T$ for gene network $\mathbf{\Lambda_y}$ augmented with the component $\mathbf{\Theta_{yz}}\mathbf{\Lambda}_{\mathbf{z}}^{-1}\mathbf{\Theta}_{\mathbf{yz}}^T$ (Fig 1B, blue dashed edge). In this augmented network, additional edges are introduced between two genes if their expression levels influence the same trait or if they both affect traits that are connected in the phenotype network $\mathbf{\Lambda_z}$. The posterior gene network $\mathbf{\Lambda_{y|x,\ z}}$ can be obtained by inferring the posterior distribution of expression levels given phenotypes from the estimated PerturbNet model:

$$p(\mathbf{y}|\mathbf{x},\mathbf{z}) = N(-(\mathbf{z}^T\mathbf{\Theta}_{\mathbf{yz}}^T + \mathbf{x}^T\mathbf{\Theta_{xy}})\mathbf{\Lambda_{y|x,z}}^{-1},\ \mathbf{\Lambda_{y|x,z}}^{-1}).$$

This can also be seen from the alternative representation of the estimated model given as the joint distribution in Eq (6), in which $\mathbf{\Lambda_{y|x,z}}$ defines the network over **y** in the full network $\mathbf{\Lambda_{zy}}$. This process of introducing additional dependencies via $\mathbf{\Theta_{yz}}\mathbf{\Lambda}_{\mathbf{z}}^{-1}\mathbf{\Theta}_{\mathbf{yz}}^T$ in this sCGGM for joint distribution is known as moralization in the probabilistic graphical model literature [26].

SNP perturbation effects on gene modules and trait groups

We summarize the estimated and inferred SNP perturbation effects at the module level and at the trait-group level, using a score function $S([\mathbf{A}]_{I,J}) = \sum_{i\in I,j\in J}|[\mathbf{A}]_{i,j}|$ for perturbation effects **A** of $R \times C$ matrix and for subsets of rows $I \subset \{1, \ldots, R\}$ and columns $J \subset \{1, \ldots, C\}$, as follows:

- **The effects of SNP *i* on gene module *M*:** $S([\mathbf{\Theta_{xy}}]_{i,M})$ for direct effects and $S([\mathbf{B_{xy}}]_{i,M})$ for indirect effects. Similarly, we summarize the overall SNP effects on gene module M as $S([\mathbf{\Theta_{xy}}]_{:,M})$ for direct effects and $S([\mathbf{B_{xy}}]_{:,M})$ for indirect effects.

- **The effects of gene module *M* on trait group *T*:** $S([\mathbf{\Theta}_{\mathbf{yz}}]_{M,T})$ for direct effects and $S([\mathbf{B}_{\mathbf{yz}}]_{M,T})$ for indirect effects.
- **The effects of SNP *i* on trait group *T*:** $S([\mathbf{B}_{\mathbf{xz}}]_{i,T})$. Similarly, we use $S([\mathbf{B}_{\mathbf{xz}}]_{:,T})$ to represent overall effects of SNPs on trait group T.
- **The effects of SNP *i* on trait group *T* mediated by module *M*:** $S([\mathbf{B}^{M}_{\mathbf{xz}}]_{i,T})$. Overall SNP effects on trait group T mediated by module M is obtained as $S([\mathbf{B}^{M}_{\mathbf{xz}}]_{:,T})$.

PerturbNet learning algorithms

We introduce an efficient algorithm for obtaining a sparse estimate of the PerturbNet model parameters with few edges in the graph. Given genotype data $\mathbf{X} \in \mathbb{R}^{n\times p}$ for n samples and p SNPs, expression data $\mathbf{Y} \in \mathbb{R}^{n\times q}$ for q genes, and phenotype data $\mathbf{Z} \in \mathbb{R}^{n\times r}$ for r phenotypes, we estimate a sparse Gaussian chain graph model in Eq (1) by minimizing the negative log-likelihood of data along with sparsity-inducing L_1 penalty:

$$\min_{\mathbf{\Lambda}_{\mathbf{y}}\succ 0,\mathbf{\Theta}_{\mathbf{xy}},\mathbf{\Lambda}_{\mathbf{z}}\succ 0,\mathbf{\Theta}_{\mathbf{yz}}} f(\mathbf{\Lambda}_{\mathbf{y}},\mathbf{\Theta}_{\mathbf{xy}}) + f(\mathbf{\Lambda}_{\mathbf{z}},\mathbf{\Theta}_{\mathbf{yz}}), \tag{7}$$

where

$$\begin{aligned} f(\mathbf{\Lambda}_{\mathbf{y}},\mathbf{\Theta}_{\mathbf{xy}}) &= -\log|\mathbf{\Lambda}_{\mathbf{y}}| + \mathrm{tr}\,(\mathbf{S}_{\mathbf{yy}}\mathbf{\Lambda}_{\mathbf{y}} + 2\mathbf{S}_{\mathbf{xy}}{}^{T}\mathbf{\Theta}_{\mathbf{xy}} + \mathbf{\Lambda}_{\mathbf{y}}^{-1}\mathbf{\Theta}_{\mathbf{xy}}^{T}\mathbf{S}_{\mathbf{xx}}\mathbf{\Theta}_{\mathbf{xy}}) \\ &\quad + \lambda_{\mathbf{\Lambda}_{\mathbf{y}}} \|\, \mathbf{\Lambda}_{\mathbf{y}} \|_1 + \lambda_{\mathbf{\Theta}_{\mathbf{xy}}} \|\, \mathbf{\Theta}_{\mathbf{xy}} \,\|_1 \end{aligned} \tag{8a}$$

$$\begin{aligned} f(\mathbf{\Lambda}_{\mathbf{z}},\mathbf{\Theta}_{\mathbf{yz}}) &= -\log|\mathbf{\Lambda}_{\mathbf{z}}| + \mathrm{tr}\,(\mathbf{S}_{\mathbf{zz}}\mathbf{\Lambda}_{\mathbf{z}} + 2\mathbf{S}_{\mathbf{yz}}{}^{T}\mathbf{\Theta}_{\mathbf{yz}} + \mathbf{\Lambda}_{\mathbf{z}}^{-1}\mathbf{\Theta}_{\mathbf{yz}}^{T}\mathbf{S}_{\mathbf{yy}}\mathbf{\Theta}_{\mathbf{yz}}) \\ &\quad + \lambda_{\mathbf{\Lambda}_{\mathbf{z}}} \|\, \mathbf{\Lambda}_{\mathbf{z}} \|_1 + \lambda_{\mathbf{\Theta}_{\mathbf{yz}}} \|\, \mathbf{\Theta}_{\mathbf{yz}} \,\|_1, \end{aligned} \tag{8b}$$

given data covariance matrices $\mathbf{S}_{\mathbf{xx}} = \frac{1}{n}\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}$, $\mathbf{S}_{\mathbf{xy}} = \frac{1}{n}\tilde{\mathbf{X}}^T\tilde{\mathbf{Y}}$, $\mathbf{S}_{\mathbf{yy}} = \frac{1}{n}\tilde{\mathbf{Y}}^T\tilde{\mathbf{Y}}$, $\mathbf{S}_{\mathbf{yz}} = \frac{1}{n}\tilde{\mathbf{Y}}^T\tilde{\mathbf{Z}}$, and $\mathbf{S}_{\mathbf{zz}} = \frac{1}{n}\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}}$ for mean-centered data matrices $\tilde{\mathbf{X}}$, $\tilde{\mathbf{Y}}$, and $\tilde{\mathbf{Z}}$, and $\|\cdot\|_1$ for non-smooth element-wise L_1 penalty. The regularization parameters $\lambda_{\mathbf{\Lambda}_{\mathbf{y}}}, \lambda_{\mathbf{\Theta}_{\mathbf{xy}}}, \lambda_{\mathbf{\Lambda}_{\mathbf{z}}}, \lambda_{\mathbf{\Theta}_{\mathbf{yz}}} > 0$ are chosen to maximize the Bayesian information criterion (BIC). We do not penalize the diagonal entries of $\mathbf{\Lambda}_{\mathbf{y}}$ and $\mathbf{\Lambda}_{\mathbf{z}}$, following the common practice for sparse inverse covariance estimation. Solving Eq (7) is a convex optimization problem with a guarantee in finding the optimal solution.

To solve the PerturbNet estimation problem in Eq (7), we introduce Fast-sCGGM and Mega-sCGGM for an efficient sCGGM estimation to be used as a key subroutine. The problem in Eq (7) decouples into two subproblems, each containing one of two disjoint sets of parameters $\{\mathbf{\Lambda}_{\mathbf{y}}, \mathbf{\Theta}_{\mathbf{xy}}\}$ and $\{\mathbf{\Lambda}_{\mathbf{z}}, \mathbf{\Theta}_{\mathbf{yz}}\}$, each of which can be solved with an sCGGM optimization algorithm. Below, we introduce Fast-sCGGM for learning an sCGGM that substantially reduces computation time, compared to the previous state-of-the-art method NCD [25]. Then, we describe Mega-sCGGM, a modification of Fast-sCGGM, that performs block-wise computation to learn from a large dataset on a machine with limited memory. Empirically Fast-sCGGM and Meta-sCGGM led to orders-of-magnitude speedup compared to the existing state-of-the-art method.

Fast-sCGGM for improving computation time. We discuss Fast-sCGGM for solving the sCGGM optimization problem in Eq (8a). The same approach can be used to solve Eq (8b). We drop the subscripts and use $\mathbf{\Theta}$ for $\mathbf{\Theta}_{\mathrm{xy}}$ and $\mathbf{\Lambda}$ for $\mathbf{\Lambda}_{\mathbf{y}}$ to simplify the notation. We re-write the problem in Eq (8a) as

$$\min_{\boldsymbol{\Lambda}\succ 0,\boldsymbol{\Theta}} \quad f(\boldsymbol{\Lambda},\boldsymbol{\Theta}) = g(\boldsymbol{\Lambda},\boldsymbol{\Theta}) + h(\boldsymbol{\Lambda},\boldsymbol{\Theta}), \tag{9}$$

where $g(\boldsymbol{\Lambda},\boldsymbol{\Theta}) = -\log|\boldsymbol{\Lambda}| + \text{tr}\,(\mathbf{S}_{\text{yy}}\boldsymbol{\Lambda} + 2\mathbf{S}_{\text{xy}}{}^{T}\boldsymbol{\Theta} + \boldsymbol{\Lambda}^{-1}\boldsymbol{\Theta}^{T}\mathbf{S}_{\text{xx}}\boldsymbol{\Theta})$ is the smooth negative log-likelihood and $h(\boldsymbol{\Lambda},\boldsymbol{\Theta}) = \lambda_{\boldsymbol{\Lambda}}\|\boldsymbol{\Lambda}\|_1 + \lambda_{\boldsymbol{\Theta}}\|\boldsymbol{\Theta}\|_1$ is the non-smooth elementwise L_1 penalty with regularization parameters $\lambda_{\boldsymbol{\Lambda}}$ and $\lambda_{\boldsymbol{\Theta}}$.

Fast-sCGGM uses an alternate Newton coordinate descent method and alternately updates $\boldsymbol{\Lambda}$ and $\boldsymbol{\Theta}$, optimizing Eq (9) over $\boldsymbol{\Lambda}$ given $\boldsymbol{\Theta}$ and vice versa until convergence. Our approach is based on the key observation that with $\boldsymbol{\Lambda}$ fixed, the problem of solving Eq (9) over $\boldsymbol{\Theta}$ becomes simply the well-known Lasso optimization, which can be solved efficiently using a coordinate descent method [32]. On the other hand, optimizing Eq (9) for $\boldsymbol{\Lambda}$ given $\boldsymbol{\Theta}$ requires forming a quadratic approximation to find a generalized Newton direction and performing line search to find a step size. However, this computation is significantly simpler than performing the same type of computation on both $\boldsymbol{\Lambda}$ and $\boldsymbol{\Theta}$ jointly as in the previous approach NCD [25]. Our algorithm iterates between the following two steps until convergence:

- **Coordinate descent optimization for $\boldsymbol{\Theta}$ given $\boldsymbol{\Lambda}$:** With $\boldsymbol{\Lambda}$ fixed, the optimization problem in Eq (9) becomes

$$\underset{\boldsymbol{\Theta}}{\text{argmin}}\; g_{\boldsymbol{\Lambda}}(\boldsymbol{\Theta}) + \lambda_{\boldsymbol{\Theta}}\|\, \boldsymbol{\Theta}\, \|_1, \tag{10}$$

  where $g_{\boldsymbol{\Lambda}}(\boldsymbol{\Theta}) = \text{tr}(2\mathbf{S}_{\mathbf{xy}}{}^{\boldsymbol{T}}\boldsymbol{\Theta} + \boldsymbol{\Lambda}^{\mathbf{-1}}\boldsymbol{\Theta}^{T}\,\mathbf{S}_{\mathbf{xx}}\boldsymbol{\Theta})$. Since $g_{\boldsymbol{\Lambda}}(\boldsymbol{\Theta})$ is a quadratic function, Eq (10) corresponds to the Lasso problem and the coordinate descent method can be used to solve this efficiently.

- **Coordinate descent optimization for $\boldsymbol{\Lambda}$ given $\boldsymbol{\Theta}$:** With $\boldsymbol{\Theta}$ fixed, the problem in Eq (9) becomes

$$\underset{\boldsymbol{\Lambda}\succ 0}{\text{argmin}}\; g_{\boldsymbol{\Theta}}(\boldsymbol{\Lambda}) + \lambda_{\boldsymbol{\Lambda}}\|\, \boldsymbol{\Lambda}\, \|_1, \tag{11}$$

  where $g_{\boldsymbol{\Theta}}(\boldsymbol{\Lambda}) = -\log|\boldsymbol{\Lambda}| + \text{tr}(\mathbf{S}_{\mathbf{yy}}\,\boldsymbol{\Lambda} + \boldsymbol{\Lambda}^{-1}\,\boldsymbol{\Theta}^{T}\,\mathbf{S}_{\mathbf{xx}}\,\boldsymbol{\Theta})$. To solve this, we first find a generalized Newton direction that minimizes the L_1-regularized quadratic approximation $\bar{g}_{\boldsymbol{\Lambda},\boldsymbol{\Theta}}(\Delta_{\boldsymbol{\Lambda}})$ of $g_{\boldsymbol{\Theta}}(\boldsymbol{\Lambda})$:

$$\mathbf{D}_{\boldsymbol{\Lambda}} = \underset{\Delta_{\boldsymbol{\Lambda}}}{\text{argmin}}\; \bar{g}_{\boldsymbol{\Lambda},\boldsymbol{\Theta}}(\Delta_{\boldsymbol{\Lambda}}) + \lambda_{\boldsymbol{\Lambda}}\|\, \boldsymbol{\Lambda} + \Delta_{\boldsymbol{\Lambda}}\, \|_1, \tag{12}$$

  where $\bar{g}_{\boldsymbol{\Lambda},\boldsymbol{\Theta}}(\Delta_{\boldsymbol{\Lambda}})$ is obtained from a second-order Taylor expansion and is given as

$$\bar{g}_{\boldsymbol{\Lambda},\boldsymbol{\Theta}}(\Delta_{\boldsymbol{\Lambda}}) = \text{vec}\,(\nabla_{\boldsymbol{\Lambda}}g(\boldsymbol{\Lambda},\boldsymbol{\Theta}))^{T}\,\text{vec}\,(\Delta_{\boldsymbol{\Lambda}}) + \frac{1}{2}\,\text{vec}\,(\Delta_{\boldsymbol{\Lambda}})^{T}\nabla^{2}_{\boldsymbol{\Lambda}}g(\boldsymbol{\Lambda},\boldsymbol{\Theta})\,\text{vec}\,(\Delta_{\boldsymbol{\Lambda}}).$$

  In the above equation, $\nabla_{\boldsymbol{\Lambda}}\, g(\boldsymbol{\Lambda},\boldsymbol{\Theta}) = \mathbf{S}_{\mathbf{yy}} - \Sigma - \Psi$ and $\nabla^{2}_{\boldsymbol{\Lambda}}g(\boldsymbol{\Lambda},\boldsymbol{\Theta}) = \Sigma \otimes (\Sigma + 2\Psi)$, where $\Sigma = \boldsymbol{\Lambda}^{-1}$ and $\Psi = \Sigma\boldsymbol{\Theta}^{T}\mathbf{S}_{\mathbf{xx}}\,\boldsymbol{\Theta}\Sigma$, are the components of the gradient and Hessian matrices corresponding to $\boldsymbol{\Lambda}$. The problem in Eq (12) is again equivalent to the Lasso problem, which can be solved efficiently via coordinate descent. Given the Newton direction for $\boldsymbol{\Lambda}$, we update $\boldsymbol{\Lambda} \leftarrow \boldsymbol{\Lambda} + \alpha\, \mathbf{D}_{\boldsymbol{\Lambda}}$, where step size $0 < \alpha \leq 1$ is set by line search on the objective in Eq (11) to ensure sufficient decrease in Eq (9) and positive definiteness of $\boldsymbol{\Lambda}$.

In order to further reduce computation time, we adopt the following strategies that have been previously used for sparse GGM and sCGGM optimizations [25, 54]. First, to improve the efficiency of coordinate descent for the Lasso problem in Eqs (10) and (12), we restrict the updates to an active set of variables given as

$$\begin{aligned}\mathcal{S}_{\mathbf{\Lambda}} &= \{[\Delta_{\mathbf{\Lambda}}]_{i,j} : |[\nabla_{\mathbf{\Lambda}} g(\mathbf{\Lambda}, \mathbf{\Theta})]_{i,j}| > \lambda_{\mathbf{\Lambda}} \vee [\mathbf{\Lambda}]_{i,j} \neq 0\} \\ \mathcal{S}_{\mathbf{\Theta}} &= \{[\Delta_{\mathbf{\Theta}}]_{i,j} : |[\nabla_{\mathbf{\Theta}} g(\mathbf{\Lambda}, \mathbf{\Theta})]_{i,j}| > \lambda_{\mathbf{\Theta}} \vee [\mathbf{\Theta}]_{i,j} \neq 0\}.\end{aligned}$$

Because the active set sizes $m_{\mathbf{\Lambda}} = |\mathcal{S}_{\mathbf{\Lambda}}|$, $m_{\mathbf{\Theta}} = |\mathcal{S}_{\mathbf{\Theta}}|$ approach the number of non-zero entries in the sparse solutions for $\mathbf{\Lambda}$ and $\mathbf{\Theta}$ over iterations, this strategy yields a substantial speedup. Second, to further improve the efficiency of coordinate descent, we store intermediate results for the large matrix products that need to be computed repeatedly. We compute and store $\mathbf{U} := \Delta_{\mathbf{\Lambda}}\Sigma$ and $\mathbf{V} := \Delta_{\mathbf{\Theta}}\Sigma$ at the beginning of the optimization. Then, after a coordinate descent update to $[\Delta_{\mathbf{\Lambda}}]_{i,j}$, rows i and j of $\mathbf{U}$ are updated. Similarly, after an update to $[\Delta_{\mathbf{\Theta}}]_{i,j}$, row i of $\mathbf{V}$ is updated. Finally, in each iteration, we warm-start $\mathbf{\Lambda}$ and $\mathbf{\Theta}$ from the results of the previous iteration and make a single pass over the active set. This ensures decrease in the objective in Eq (9), while reducing the overall computation time in practice. The pseudocode for Fast-sCGGM is provided in S2 Text.

Mega-sCGGM for removing memory constraint. Fast-sCGGM as described above is still limited by the space required to store large matrices during coordinate descent computation. Solving Eq (12) for updating $\mathbf{\Lambda}$ requires precomputing and storing $q \times q$ matrices, Σ and $\Psi = \Sigma\mathbf{\Theta}^T\mathbf{S}_{\mathbf{xx}}\mathbf{\Theta}\Sigma$, and solving Eq (10) for updating $\mathbf{\Theta}$ requires Σ and a $p \times p$ matrix $\mathbf{S}_{\mathbf{xx}}$. A naive approach to reduce the memory footprint would be to recompute portions of these matrices on demand for each coordinate update, which would be very expensive.

Here, we describe Mega-sCGGM that combines the alternating Newton coordinate descent algorithm in Fast-sCGGM with block coordinate descent to scale up the optimization to very large problems on a machine with limited memory. During coordinate descent optimization, Mega-sCGGM updates blocks of $\mathbf{\Lambda}$ and $\mathbf{\Theta}$ so that within each block, the computation of the large matrices can be cached and re-used. These blocks are determined automatically by exploiting the sparse stucture. For $\mathbf{\Lambda}$, we extend the block coordinate descent approach in BIG&QUIC [31] developed for sparse GGMs to take into account the conditioning variables in sCGGMs. For $\mathbf{\Theta}$, we describe a new approach for block coordinate descent update. The block-wise update in Mega-sCGGM is described in detail in S2 Text and illustrated in S8 Fig. Our algorithm can, in principle, be applied to problems of any size on a machine with limited memory.

The previous method NCD, Fast-sCGGM, and Mega-sCGGM solve the same optimization problem and produce nearly identical estimate. However, Fast-sCGGM reaches this estimate more efficiently than the NCD and Mega-sCGGM finds the estimate without running out of memory.

Parallelization in Fast-sCGGM and Mega-sCGGM. We parallelize some of the expensive computations in Fast-sCGGM and Mega-sCGGM on multi-core machines. For both methods, we parallelize all matrix-matrix and matrix-vector multiplications. In addition, we parallelize the computation of columns of Σ and Ψ in Fast-sCGGM and the same computation within each block in Mega-sCGGM. In Mega-sCGGM, we parallelize the computation of each row of $\mathbf{S}_{\mathbf{xx}}$ whenever it is recomputed.

Semi-supervised learning. We introduce a modification of our learning algorithm for semi-supervised learning, to handle the situation where expression data are available only for a subset of individuals because of the difficulty of obtaining tissue samples. Our strategy is to use an EM algorithm [55] that imputes the missing expression levels in the E-step and performs our Fast-sCGGM or Mega-sCGGM optimization in the M-step. Given a dataset $\mathcal{D} = \{\mathcal{D}_o, \mathcal{D}_h\}$, where $\mathcal{D}_o = \{\mathbf{X}_o, \mathbf{Y}_o, \mathbf{Z}_o\}$ for the fully-observed data and $\mathcal{D}_h = \{\mathbf{X}_h, \mathbf{Z}_h\}$ for the samples with missing gene-expression levels, our EM algorithm [55] maximizes the expected log-likelihood of data:

$$\mathcal{L}(\mathcal{D}_o;\mathbf{\Theta}) + \mathrm{E}[\mathcal{L}(\mathcal{D}_h, \mathbf{Y}_h;\mathbf{\Theta})],$$

combined with L_1-regularization, where $\mathcal{L}(\mathcal{D}_o;\mathbf{\Theta})$ and $\mathcal{L}(\mathcal{D}_h;\mathbf{\Theta})$ are the log-likelihood of data $\mathcal{D}_o$ and $\mathcal{D}_h$ with respect to the model in Eq (1) and the expectation is taken with respect to:

$$\begin{aligned} &p(\mathbf{y}|\mathbf{z},\mathbf{x}) = N(\mu_{\mathrm{y|x,z}}, \Sigma_{\mathrm{y|x,z}}), \\ \mu_{\mathrm{y|x,z}} = -\Sigma_{\mathrm{y|x,z}}&(\mathbf{\Theta}_{\mathbf{yz}}\mathbf{z} + \mathbf{\Theta}_{\mathbf{xy}}^T\mathbf{x}) \quad \text{and} \quad \Sigma_{\mathrm{y|x,z}} = (\mathbf{\Lambda}_{\mathrm{y}} + \mathbf{\Theta}_{\mathbf{yz}}\mathbf{\Lambda}_{\mathbf{z}}\mathbf{\Theta}_{\mathbf{yz}}^T)^{-1}. \end{aligned} \tag{13}$$

In E-step, a naive inversion of $\mathbf{\Lambda}_{\mathrm{y}} + \mathbf{\Theta}_{\mathbf{yz}}\mathbf{\Lambda}_{\mathbf{z}}^{-1}\mathbf{\Theta}_{\mathbf{yz}}^T$ to obtain $\mathbf{\Sigma}_{\mathbf{y|x,z}}$ is expensive and requires a storage for this dense matrix that may exceed computer memory for large gene expression datasets. To make the EM algorithm efficient in terms of both time and memory, we assume that the number of phenotypes r is relatively small compared to the number of genes (i.e., $r \ll q$), which is typical for most studies. Then, instead of explicitly performing the E-step, we embed the E-step within the M-step by representing the E-step results implicitly to fit in memory and computing them explicitly on-demand as needed in the M-step. Specifically, we implicitly represent $\mathbf{\Lambda}_{\mathrm{y}} + \mathbf{\Theta}_{\mathbf{yz}}\mathbf{\Lambda}_{\mathbf{z}}^{-1}\mathbf{\Theta}_{\mathbf{yz}}^T$ as $\mathbf{\Lambda}_{\mathbf{y}}$+ $\mathbf{KK}^T$, using low-rank component $\mathbf{K} = \mathbf{\Theta}_{\mathbf{yz}}\mathbf{L}_{\mathbf{z}}^T$ and the sparse Cholesky factorization of trait network $\mathbf{L}_{\mathbf{z}}\mathbf{L}_{\mathbf{z}}^T = \mathbf{\Lambda}_{\mathbf{zz}}$. Then, during M-step, we invert $\mathbf{\Lambda}_{\mathbf{y}}$+ $\mathbf{KK}^T$, one column at a time as needed, using the conjugate gradient method. This modified EM algorithm produces the same estimate as the original EM algorithm that iterates between an M-step and an E-step.

Simulation experiments

Simulated data. To assess the accuracy of the computational methods, we generated simulated data from the known ground-truth models and SNPs in the CAMP data. For each type of models, we obtained sensitivities at different FDRs averaged over 10 simulated datasets. To assess accuracy on the recovery of trait-perturbing SNPs and mediator genes, we simulated data from two types of ground-truth models: a two-layer sCGGM as in PerturbNet to mimic the SNP perturbations of clinical traits modulated by gene networks and a two-layer linear regression model to mirror the assumptions of mediator genes acting independently of other genes as in the existing methods. To assess the accuracy of gene network learning, we again simulated data from two types of ground-truth models: a two-layer sCGGM with a gene network perturbed by SNPs and a two-layer GGM [56] with no perturbation of networks.

We set the ground-truth two-layer sCGGM as follows. We assumed gene network $\mathbf{\Lambda}_{\mathbf{y}}$ of size q = 5, 000 and trait network $\mathbf{\Lambda}_{\mathbf{z}}$ of size r = 100. In both networks, we assumed 50 nodes in each module, leading to 100 modules in $\mathbf{\Lambda}_{\mathbf{y}}$ and two modules in $\mathbf{\Lambda}_{\mathbf{z}}$. Then, the algorithm described in [57] was used to generate scale-free networks with average degree 3 and with 90% of the edges within modules and 10% of the edges across modules. Edge weights were randomly generated from normal distribution $\mathcal{N}(0.5, 0.1^2)$. The diagonal elements of $\mathbf{\Lambda}_{\mathbf{z}}$ and $\mathbf{\Lambda}_{\mathbf{y}}$ were set to ensure the minimum eigenvalue of the matrix is 0.3. The perturbation parameters $\mathbf{\Theta}_{\mathbf{yz}}$ and $\mathbf{\Theta}_{\mathbf{xy}}$ were set to represent perturbation of modules. For $\mathbf{\Theta}_{\mathbf{yz}}$, 10 out of 50 gene modules were set to have no influence on the traits. Each half of the other 40 modules were set to influence primarily each of the two trait modules, 90% of the edges connecting to one trait module and 10% to the other trait module. For $\mathbf{\Theta}_{\mathbf{xy}}$, 25 gene modules were perturbed by 1,000 randomly-selected SNPs and the other 25 modules were not perturbed by any SNPs. In both perturbation parameters, each expression and clinical trait was perturbed by three perturbants.

The magnitudes of perturbations were sampled from normal distribution $\mathcal{N}(0.5, 0.1^2)$, with edge signs randomly assigned with probability 0.5 to positive or negative.

For simulation from a two-layer linear regression model, the regression parameters were set using the same approach as $\mathbf{\Theta_{xy}}$ and $\mathbf{\Theta_{yz}}$ above and the noise variances were set to 0.5. For comparison on the task of gene network recovery, we generated data from a model assuming no SNP perturbations, using a two-layer GGM [30] with the same parameters as in the sCGGM network parameters.

Comparison of methods on simulated data. Using simulated data, we compared the accuracy of PerturbNet, PrediXcan [12], eCAVIAR [13], and two-layer Lasso (S1 Text) on the tasks of identifying SNPs perturbing clinical traits and identifying genes mediating these perturbations. PrediXcan tests for an association between a trait and an imputed gene expression value obtained as an elastic-net prediction given SNPs, providing a list of genes mediating overall genetic effects on each trait (p-value < 0.01). From these mediator genes for overall SNP effects, we obtained the mediator genes for an individual SNP by further weighting –log (p- value) for the given gene/trait pair with the elastic-net regression coefficient for each SNP/gene pair. To obtain SNPs with the strongest effects on each trait, we further aggregated these scores by summing over all genes for each SNP/trait pair. eCAVIAR assigns a colocalization posterior probability (CLPP) score, the posterior probability of a GWAS SNP and an eQTL in the same genomic region of M SNPs co-localizing, to each triplet of SNP, gene, and trait, given statistically significant GWAS SNPs and eQTLs (FDR $q < 10^{-3}$). To find SNPs with the strongest effects on traits, we collapsed the CLPP scores for triplets into scores for SNP/trait pairs by taking the maximum CLPP score across all genes for each SNP/trait pair. Then, we identified the corresponding gene as the strongest mediating gene for each SNP/trait pair. To identify genes that are the strongest mediators of overall SNP effects on a trait, we collapsed the CLPP scores for triplets into scores for gene/trait pairs by summing CLPP scores over all SNPs for the given/trait pair and selected the highest-scoring gene. For PerturbNet and two-layer Lasso, we selected the regularization parameters that maximize the BIC.

Analysis of CAMP asthma data

Asthma dataset. We preprocessed the genotype, gene expression, and clinical phenotype data, collected from asthma patients participating in the CAMP study [27–29] for our analysis. We used 174 non-Hispanic Caucasian subjects for whom both genotype and clinical phenotype data were available. For a subset of 140 individuals, gene expression data from primary peripheral blood CD4+ lymphocytes were also available. After removing SNPs with minor allele frequency less than 0.1 and those with missing reference SNP ids, we obtained 495,597 SNPs for autosomal chromosomes. Given expression levels for 22,184 mRNA transcripts profiled with Illumina HumanRef8 v2 BeadChip arrays [29], we removed transcript levels with expression variance less than 0.01, which resulted in the 11,598 transcript levels to be used in our analysis. Then, we converted the expression values to their z-scores. The clinical phenotype data comprised 35 phenotypes (S3 Table), including 25 features related to lung function and 10 features collected via blood testing. The phenotypes were verified to be well-approximated by normal distributions: with the exception of rescue_bd_7day and prevent_bd_7day, which are counts and skewed right, all the other variables represented continuous measurements distributed roughly symmetrically in quantile-quantile (QQ) plots (S7 Fig). The clinical phenotypes were converted to their z-scores within each phenotype so that all phenotypes have equal variance. We then imputed missing values using low-rank matrix completion [58].

Comparison of methods on asthma data. We compared PerturbNet with PrediXcan, eCAVIAR, and two-layer Lasso on the asthma data. We applied PerturbNet with semi-supervised learning to all samples, including the samples with only genotype and clinical trait data. In PerturbNet and two-layer Lasso, we selected the regularization parameters using the BIC. With PrediXcan, we mimiced semi-supervised learning by fitting elastic net to the samples with both expression and genotype data, imputing the missing expression data, and using all samples to perform association tests between the imputed gene expression values and traits. To obtain top trait-associated SNPs, we scored each SNP by first scoring SNP-trait pairs as in the simulation study above and summing these scores across all traits. For eCAVIAR, GWAS SNPs were found using all samples, while eQTLs were found using only the samples with expression data. Then, top asthma-associated SNPs were found by summarizing CLPP scores of all triplets into scores of each SNP by taking maximum over all mediator genes and summing over all traits. We applied the two-layer Lasso only to the samples with all data and obtained top SNPs perturbing asthma traits using the same strategy as in PerturbNet.

Comparison of computation time. To assess the computation time of Mega-sCGGM, PerturbNet, and other existing methods, we used the same hardware setup with comparable software implementations. For Lasso, we used the implementation in GLMNET [59] with a backend written in Fortran. For NCD, we took the implementation written in C++ provided by the authors [25] and sped up this implementation with the Eigen matrix library [60], by employing low-rank matrix representations and using sparse matrix multiplications. For Mega-sCGGM and NCD, we used the same regularization parameters to ensure the resulting estimates are identical, and for Lasso, we chose the regularization parameters so that the L_1-norm of the regression coefficients roughly matched that of inferred indirect SNP effects in PerturbNet. We used the C++ implementation of eCAVIAR provided by the authors [13], and used PrediXcan [12] with elastic-net as implemented in GLMNET [59]. For all methods, the code was compiled and run with OpenMP multi-threading enabled on the same machines with 20Gb memory and 16 cores.

Comparison of SNPs with external annotations. For comparison with the DNase hypersensitivity sites, we used the UES tool [61] to identify the sites enriched in top asthma SNPs from each method and to compute the enrichment p-values for these overlaps with 100 Monte Carlo simulations. For comparison with annotations in RegulomeDB [36], we scored the top 500 asthma SNPs from each method into six categories: category 1 for overlaps with previously reported eQTLs with additional functional annotation on TF binding; categories 2-5 for overlaps with TF binding sites in ChIP-seq, DNase-seq, and motif hits, where lower scores indicate stronger evidence for being functional; and category 6 for little evidence of being functional. Finally, we compared the asthma SNPs found by each method with the previously reported asthma-associated SNPs in the NHGRI GWAS catalog [37]. Out of 557 SNPs in the NHGRI GWAS catalog for asthma, we examined how many of these SNPs are within 10, 20, 50, and 100kb of top k SNPs from each method, k ranging from 1 to 500. To assess the statistical significance of the overlap between two sets of SNPs, we performed a permutation test: for top k SNPs from the given method, we generated 10,000 random sets of k SNPs from the SNPs employed in the analysis to find the distribution of overlaps under the null hypothesis and reported the overlaps with p-value < 0.05.

Supporting information

S1 Text. Two-layer Lasso.
(PDF)

S2 Text. Fast-sCGGM and Mega-sCGGM for efficient sCGGM optimization.
(PDF)

S1 Fig. The PerturbNet model estimated from the CAMP data.
(PDF)

S2 Fig. The PerturbNet posterior gene network for asthma.
(PDF)

S3 Fig. SNP effects on gene modules and gene-module effects on phenotypes in asthma determined by PerturbNet.
(PDF)

S4 Fig. Manhattan plots for overall SNP effects on asthma phenotypes determined by PerturbNet.
(PDF)

S5 Fig. PerturbNet asthma SNPs as eQTLs.
(PDF)

S6 Fig. Overlap with asthma SNPs in NHGRI GWAS catalog.
(PDF)

S7 Fig. The distribution of clinical phenotypes in the CAMP data.
(PDF)

S8 Fig. Schematic of block coordinate descent in Mega-sCGGM.
(PDF)

S1 Table. GO categories enriched in gene modules in the PerturbNet asthma gene network.
(PDF)

S2 Table. GO categories enriched in genes in modules 16-20 connected to modules 13-15 in the PerturbNet asthma posterior gene network.
(PDF)

S3 Table. Description of asthma clinical phenotypes in CAMP data.
(PDF)

Acknowledgments

The authors thank Kathryn Roeder for helpful discussion.

Author Contributions

Conceptualization: Calvin McCarter, Seyoung Kim.

Data curation: Calvin McCarter, Judie Howrylak.

Formal analysis: Calvin McCarter, Seyoung Kim.

Funding acquisition: Seyoung Kim.

Investigation: Calvin McCarter, Seyoung Kim.

Methodology: Calvin McCarter, Seyoung Kim.

Project administration: Calvin McCarter, Seyoung Kim.

Resources: Calvin McCarter.

Software: Calvin McCarter.

Supervision: Seyoung Kim.

Visualization: Calvin McCarter.

Writing – original draft: Calvin McCarter, Seyoung Kim.

Writing – review & editing: Calvin McCarter, Judie Howrylak, Seyoung Kim.

References

1. Cusanovich DA, Pavlovic B, Pritchard JK, Gilad Y. The functional consequences of variation in transcription factor binding. PLoS Genet. 2014; 10(3):e1004226. https://doi.org/10.1371/journal.pgen.1004226
2. Shalem O, Sanjana NE, Zhang F. High-throughput functional genomics using CRISPR–Cas9. Nature Reviews Genetics. 2015; 16(5):299. https://doi.org/10.1038/nrg3899
3. Jansen RC. Studying complex biological systems using multifactorial perturbation. Nature Reviews Genetics. 2003; 4(2):145. https://doi.org/10.1038/nrg996
4. Sohn KA, Kim S. Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization. In: Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS). JMLR W&CP; 2012. p. 1081–1089.
5. Zhang L, Kim S. Learning gene networks under SNP perturbations using eQTL datasets. PLoS Computational Biology. 2014; 10(2):e1003420. https://doi.org/10.1371/journal.pcbi.1003420
6. Frot B, Jostins L, McVean G. Graphical model selection for Gaussian conditional random fields in the presence of latent variables. Journal of the American Statistical Association. 2018; 114(526):723–734.
7. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, et al. Genetics of gene expression and its effect on disease. Nature. 2008; 452:423–428. https://doi.org/10.1038/nature06758 PMID: 18344981
8. Schadt EE. Molecular networks as sensors and drivers of common human diseases. Nature. 2009; 461:218–223. https://doi.org/10.1038/nature08454
9. Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, et al. Mapping the genetic architecture of gene expression in human liver. PLoS biology. 2008; 6(5):e107. https://doi.org/10.1371/journal.pbio.0060107 PMID: 18462017
10. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genetics. 2014; 10(5):e1004383. https://doi.org/10.1371/journal.pgen.1004383 PMID: 24830394
11. He X, Fuller CK, Song Y, Meng Q, Zhang B, Yang X, et al. Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. The American Journal of Human Genetics. 2013; 92(5):667–680. https://doi.org/10.1016/j.ajhg.2013.03.022 PMID: 23643380
12. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, et al. A gene-based association method for mapping traits using reference transcriptome data. Nature Genetics. 2015; 47(9):1091–1098. https://doi.org/10.1038/ng.3367 PMID: 26258848
13. Hormozdiari F, Van De Bunt M, Segre AV, Li X, Joo JWJ, Bilow M, et al. Colocalization of GWAS and eQTL signals detects target genes. The American Journal of Human Genetics. 2016; 99(6):1245–1260. https://doi.org/10.1016/j.ajhg.2016.10.003 PMID: 27866706
14. Greenawalt DM, Dobrin R, Chudin E, Hatoum IJ, Suver C, Beaulaurier J, et al. A survey of the genetics of stomach, liver, and adipose gene expression from a morbidly obese cohort. Genome Research. 2011; 21(7):1008–1016. https://doi.org/10.1101/gr.112821.110 PMID: 21602305
15. Gibson G, Powell JE, Marigorta UM. Expression quantitative trait locus analysis for translational medicine. Genome Medicine. 2015; 7(1):60. https://doi.org/10.1186/s13073-015-0186-7
16. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multi-tissue gene regulation in humans. Science. 2015; 348(6235):648–660. https://doi.org/10.1126/science.1262110
17. Moffatt MF, Gut IG, Demenais F, Strachan DP, Bouzigon E, Heath S, et al. A large-scale, consortium-based genomewide association study of asthma. New England Journal of Medicine. 2010; 363(13):1211–1221. https://doi.org/10.1056/NEJMoa0906312 PMID: 20860503
18. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447(7145):661–678. https://doi.org/10.1038/nature05911

19. Nakka P, Raphael BJ, Ramachandran S. Gene and network analysis of common variants reveals novel associations in multiple complex diseases. Genetics. 2016; 204(2):783–798. https://doi.org/10.1534/genetics.116.188391

20. Mezlini AM, Goldenberg A. Incorporating networks in a probabilistic graphical model to find drivers for complex human diseases. PLoS Computational Biology. 2017; 13(10):e1005580. https://doi.org/10.1371/journal.pcbi.1005580

21. Kim YA, Wuchty S, Przytycka TM. Identifying causal genes and dysregulated pathways in complex diseases. PLoS computational biology. 2011; 7(3):e1001095. https://doi.org/10.1371/journal.pcbi.1001095

22. Botzman M, Nachshon A, Brodt A, Gat-Viks I. POEM: Identifying joint additive effects on regulatory circuits. Frontiers in genetics. 2016; 7:48.

23. Oren Y, Nachshon A, Frishberg A, Wilentzik R, Gat-Viks I. Linking traits based on their shared molecular mechanisms. Elife. 2015; 4:e04346. https://doi.org/10.7554/eLife.04346

24. Peters LA, Perrigoue J, Mortha A, Iuga A, Song Wm, Neiman EM, et al. A functional genomics predictive network model identifies regulators of inflammatory bowel disease. Nature genetics. 2017; 49 (10):1437–1449. https://doi.org/10.1038/ng.3947 PMID: 28892060

25. Wytock M, Kolter JZ. Sparse Gaussian conditional random fields: algorithms, theory, and application to energy forecasting. In: Proceedings of the 30th International Conference on Machine Learning. vol. 28. JMLR W&CP; 2013. p. 1265–1273.

26. Koller D, Friedman N. Probabilistic graphical models: principles and techniques. MIT Press; 2009.

27. Childhood Asthma Management Program Research Group. The childhood asthma management program (CAMP): design, rationale, and methods. Controlled Clinical Trials. 1999; 20(1):91–120. https://doi.org/10.1016/S0197-2456(98)00044-0

28. Childhood Asthma Management Program Research Group. Long-term effects of budesonide or nedocromil in children with asthma. New England Journal of Medicine. 2000; 343(15):1054–1063. https://doi.org/10.1056/NEJM200010123431501

29. Murphy A, Chu JH, Xu M, Carey VJ, Lazarus R, Liu A, et al. Mapping of numerous disease-associated expression polymorphisms in primary peripheral blood CD4+ lymphocytes. Human Molecular Genetics. 2010; 19(23):4745–4757. https://doi.org/10.1093/hmg/ddq392 PMID: 20833654

30. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008; 9(3):432–441. https://doi.org/10.1093/biostatistics/kxm045

31. Hsieh CJ, Sustik MA, Dhillon IS, Ravikumar PK, Poldrack R. BIG & QUIC: Sparse inverse covariance estimation for a million variables. In: Advances in Neural Information Processing Systems 26. Curran Associates; 2013. p. 3165–3173.

32. Friedman J, Hastie T, Höfling H, Tibshirani R. Pathwise coordinate optimization. The Annals of Applied Statistics. 2007; 1(2):302–332. https://doi.org/10.1214/07-AOAS131

33. Karypis G, Kumar V. METIS—unstructured graph partitioning and sparse matrix ordering system, version 2.0. University of Minnesota, Department of Computer Science and Engineering, Army HPC Research Center, Minneapolis, MN; 1995.

34. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research. 2004; 32(suppl_1):D258–D261. https://doi.org/10.1093/nar/gkh036

35. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489(7414):57–74. https://doi.org/10.1038/nature11247

36. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Research. 2012; 22(9):1790–1797. https://doi.org/10.1101/gr.137323.112 PMID: 22955989

37. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Research. 2018; 47(D1):D1005–D1012. https://doi.org/10.1093/nar/gky1120

38. Kichaev G, Bhatia G, Loh PR, Gazal S, Burch K, Freund MK, et al. Leveraging polygenic functional enrichment to improve GWAS power. The American Journal of Human Genetics. 2019; 104(1):65–75. https://doi.org/10.1016/j.ajhg.2018.11.008 PMID: 30595370

39. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491(7422):56–65. https://doi.org/10.1038/nature11632

40. Hai T, Wolford CC, Chang YS. ATF3, a hub of the cellular adaptive-response network, in the pathogenesis of diseases: is modulation of inflammation a unifying component? Gene Expression. 2010; 15 (1):1–11. https://doi.org/10.3727/105221610X12819686555015

41. Wu Z, MacNeil AJ, Junkins R, Li B, Berman JN, Lin TJ. Mast cell FcεRI-induced early growth response 2 regulates CC chemokine ligand 1–dependent CD4+ T cell migration. The Journal of Immunology. 2013; 190(9):4500–4507. https://doi.org/10.4049/jimmunol.1203158

42. Temple R, Allen E, Fordham J, Phipps S, Schneider HC, Lindauer K, et al. Microarray analysis of eosinophils reveals a number of candidate survival and apoptosis genes. American Journal of Respiratory Cell and Molecular Biology. 2001; 25(4):425–433. https://doi.org/10.1165/ajrcmb.25.4.4456 PMID: 11694447

43. Barnig C, Alsaleh G, Jung N, Dembélé D, Paul N, Poirot A, et al. Circulating human eosinophils share a similar transcriptional profile in asthma and other hypereosinophilic disorders. PLoS One. 2015; 10(11): e0141740. https://doi.org/10.1371/journal.pone.0141740 PMID: 26524763

44. Khan MA, Assiri AM, Broering DC. Complement mediators: Key regulators of airway tissue remodeling in asthma. Journal of Translational Medicine. 2015; 13(1):272. https://doi.org/10.1186/s12967-015-0565-2

45. Kay A, Bacon G, Mercer B, Simpson H, Crofton J. Complement components and IgE in bronchial asthma. The Lancet. 1974; 304(7886):916–920. https://doi.org/10.1016/S0140-6736(74)91128-3

46. Jönsson G, Sjöholm A, Truedsson L, Bengtsson A, Braconier JH, Sturfelt G. Rheumatological manifestations, organ damage and autoimmunity in hereditary C2 deficiency. Rheumatology. 2007; 46 (7):1133–1139. https://doi.org/10.1093/rheumatology/kem023

47. Charignon D, Ponard D, de Gennes C, Drouet C, Ghannam A. SERPING1 and F12 combined variants in a hereditary angioedema family. Annals of Allergy, Asthma & Immunology. 2018; 121(4):500–502. https://doi.org/10.1016/j.anai.2018.05.031

48. Luo W, Obeidat M, Di Narzo AF, Chen R, Sin DD, Paré PD, et al. Airway epithelial expression quantitative trait loci reveal genes underlying asthma and other airway diseases. American Journal of Respiratory Cell and Molecular Biology. 2016; 54(2):177–187. https://doi.org/10.1165/rcmb.2014-0381OC PMID: 26102239

49. Buntru A, Kopp K, Voges M, Frank R, Bachmann V, Hauck CR. Phosphatidylinositol 3'-kinase activity is critical for initiating the oxidative burst and bacterial destruction during CEACAM3-mediated phagocytosis. Journal of Biological Chemistry. 2011; 286(11):9555–9566. https://doi.org/10.1074/jbc.M110.216085

50. Lafferty J, McCallum A, Pereira FC. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning. vol. 951; 2001. p. 282–289.

51. Liu J, Yang C, Shi X, Li C, Huang J, Zhao H, et al. Analyzing association mapping in pedigree-based GWAS using a penalized multitrait mixed model. Genetic Epidemiology. 2016; 40(5):382–393. https://doi.org/10.1002/gepi.21975 PMID: 27247027

52. Liu H, Lafferty J, Wasserman L. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. Journal of Machine Learning Research. 2009; 10(Oct):2295–2328.

53. Zhang L, Zeng Z, Ji Q. Probabilistic image modeling with an extended chain graph for human activity recognition and image segmentation. IEEE Transactions on Image Processing. 2011; 20(9):2401–2413. https://doi.org/10.1109/TIP.2011.2128332

54. Hsieh CJ, Dhillon IS, Ravikumar PK, Sustik MA. Sparse inverse covariance matrix estimation using quadratic approximation. In: Advances in Neural Information Processing Systems 24. Curran Associates; 2011. p. 2330–2338.

55. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological). 1977; 39(1):1–38.

56. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008; 9(3):432–441. https://doi.org/10.1093/biostatistics/kxm045

57. Wu JJ, Gao ZY, Sun HJ. Cascade and breakdown in scale-free networks with community structure. Physical Review E. 2006; 74(6):066111. https://doi.org/10.1103/PhysRevE.74.066111

58. Lin Z, Chen M, Ma Y. The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices. Coordinated Science Laboratory, University of Illinois at Urbana-Champaign; 2010. Report no. UILU-ENG-09-2215.

59. Qian J, Hastie T, Friedman J, Tibshirani R, Simon N. Glmnet for Matlab; 2013. Available from: http://www.stanford.edu/~hastie/glmnet_matlab/.

60. Guennebaud G, Jacob B. Eigen v3; 2010. Available from: http://eigen.tuxfamily.org.

61. Hayes JE, Trynka G, Vijai J, Offit K, Raychaudhuri S, Klein RJ. Tissue-specific enrichment of lymphoma risk loci in regulatory elements. PloS One. 2015; 10(9):e0139360. https://doi.org/10.1371/journal.pone.0139360