Supplementary Materials for "Improving SNP Prioritization and Pleiotropic Architecture Estimation by Incorporating Prior Knowledge Using graph-GPA"

Hang J. Kim[1], Zhenning Yu[2], Andrew Lawson[2], Hongyu Zhao[3,4,5], and Dongjun Chung[2*]

1 Department of Mathematical Sciences, University of Cincinnati, Cincinnati, OH, USA.
2 Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC, USA.
3 Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA.
4 Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA.
5 Department of Genetics, Yale School of Medicine, New Haven, CT, USA.

* Correspondence should be addressed to Dongjun Chung (chungd@musc.edu).

Contents

1 Methods

1.1 graph-GPA Model

graph-GPA takes GWAS summary statistics (genotype-phenotype association p-value) for SNP t with phenotype i, denoted as p_{it}, as input, where $i = 1, \ldots, n$ and $t = 1, \ldots, T$. For convenience in modelling and visualization, we transform p_{it} as $y_{it} = \Phi^{-1}(1 - p_{it})$, where Φ is the cumulative distribution of the standard normal variable. We model the density of y_{it} with the latent association indicator e_{it} by a lognormal-normal mixture:

$$p(y_{it}|e_{it}, \mu_i, \sigma_i^2) = e_{it}\ \mathrm{LN}(y_{it}; \mu_i, \sigma_i^2) + (1 - e_{it})\ \mathrm{N}(y_{it}; 0, 1), \tag{1}$$

where $e_{it} = 1$ if SNP t is associated with phenotype i and $e_{it} = 0$ otherwise, LN and N denote the log-normal density and the normal density, respectively.

To model genetic relationship among n phenotypes, we adopt a graphical model based on Markov random field (MRF). Let $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$ denote an MRF graph with nodes $\boldsymbol{V} = (v_1, \ldots, v_n)$ and edges $\boldsymbol{E} = \{E(i, j) : i, j = 1, \ldots, n\}$. We can interpret v_i as phenotype i and $E(i, j) = 1$ as phenotypes i and j are conditionally dependent (i.e., genetically correlated). We model the latent association indicators of SNP t, $\boldsymbol{e}_t = (e_{1t}, \ldots, e_{nt})$, and the graph structure with an auto-logistic spatial scheme:

$$p(\boldsymbol{e}_t|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{G}) = C(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{G})\ \exp\left(\sum_{i=1}^{n} \alpha_i e_{it} + \sum_{i \sim j} \beta_{ij} e_{it} e_{jt}\right) \tag{2}$$

and

$$C(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{G})^{-1} = \sum_{\mathbf{e}^* \in \mathcal{E}^*} \exp\left(\sum_{i=1}^{n} \alpha_i e_i^* + \sum_{i \sim j} \beta_{ij} e_i^* e_j^*\right), \tag{3}$$

where β_{ij} is the MRF coefficient for the pair of phenotypes i and j, the symbol $i \sim j$ denotes that v_i is adjacent to v_j, i.e., $E(i, j) = 1$, and $\mathcal{E}^*$ is the set of all possible values of $\boldsymbol{e}^* = (e_1^*, \ldots, e_n^*)$.

For the log-normal density in (1), we introduce the conjugate prior distribution:

$$\mu_i \sim \mathrm{N}(\theta_\mu, \tau_\mu^2), \qquad \sigma_i^2 \sim \mathrm{IG}(a_\sigma, b_\sigma), \tag{4}$$

where IG denotes the inverse-gamma distribution. For the MRF coefficients in (2), we assume the following prior distributions:

$$\alpha_i \sim \mathrm{N}(\theta_\alpha, \tau_\alpha^2),\ \beta_{ij} \sim E(i, j)\ \Gamma(\beta_{ij}; a_\beta, b_\beta) + \{1 - E(i, j)\}\ \delta_0(\beta_{ij}), \tag{5}$$

where $\Gamma(a, b)$ denotes the gamma distribution with mean a/b and δ_0 denotes Dirac delta function. For the MRF graph $\boldsymbol{G}$, we allow to incorporate prior information from external sources, e.g., those

obtained from a text mining of PubMed literature as described in Section 2. Specifically, we "force in" edges, i.e., set $E(i,j) = 1$, if the external source provides evidence that phenotypes i and j are genetically correlated while other edges are set to have uninformative prior probabilities, i.e., $\Pr\{E(i,j) = 1\} \propto 1$. Weakly informative priors are used for the top level of the Bayesian hierarchical model with the following hyperparameters: $\theta_\mu = 0$, $\tau_\mu^2 = 10000$, $\theta_\alpha = 0$, $\tau_\alpha^2 = 10000$ and $a_\sigma = b_\sigma = 0.5$. We put $a_\beta = 4$ and $b_\beta = 2$ so that most of β_{ij}'s with $E(i,j) = 1$ are *a priori* distinct from zero.

1.2 Posterior Sampling

This section describes full details of the Metropolis-within-Gibbs steps for the Bayesian inferences.

S1. For each phenotype i and SNP t, update $e_{it} \sim \text{Bernoulli}(p_1^*)$ where

$$p_1^* = \left\{ 1 + \frac{\text{N}(y_{it}; 0, 1)}{\exp\left(\alpha_i + \sum_{j \sim i} \beta_{ij}\ e_{jt}\right) \cdot \text{LN}(y_{it}; \mu_i, \sigma_i^2)} \right\}^{-1}.$$

S2. For each i, update

$$\mu_i \sim \text{N}\left(\frac{\sigma_i^2 \theta_\mu + \tau_\mu^2 \sum_{\{t: e_{it}=1\}} \log y_{it}}{\sigma_i^2 + \tau_\mu^2\ n_i}, \frac{\sigma_i^2 \tau_\mu^2}{\sigma_i^2 + \tau_\mu^2\ n_i} \right)$$

where $n_i = \sum_{t=1}^T e_{it}$.

S3. For each i, update

$$\sigma_i^2 \sim \text{IG}\left(a_\sigma + \frac{n_i}{2}, b_\sigma + \frac{\sum_{\{t: e_{it}=1\}} (\log y_{it} - \mu_i)^2}{2} \right)$$

where $n_i = \sum_{t=1}^T e_{it}$.

S4. For each i, update α_i with the Metropolis-Hastings step:

1. Draw α_i^q from $\text{N}(\alpha_i, s_\alpha^2)$. We set $s_\alpha = 0.1$.
2. Update $\alpha_i = \alpha_i^q$ with the acceptance probability

$$\min\left[1, \left\{ \prod_{t=1}^T \frac{\text{C}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{G})}{\text{C}(\boldsymbol{\alpha}^q, \boldsymbol{\beta}, \boldsymbol{G})} \frac{\exp(\alpha_i^q e_{it})}{\exp(\alpha_i e_{it})} \right\} \frac{\text{N}(\alpha_i^q; \theta_\alpha, \tau_\alpha^2)}{\text{N}(\alpha_i; \theta_\alpha, \tau_\alpha^2)} \right]$$

where $\boldsymbol{\alpha}^q = (\alpha_1, \ldots, \alpha_{i-1}, \alpha_i^q, \alpha_{i+1}, \ldots, \alpha_n)$.

S5. For each (i,j) such that $E(i,j) = 1$, update β_{ij} with the Metropolis-Hastings:

1. Draw β_{ij}^q from $\mathrm{N}_+(\beta_{ij}, s_\beta^2)$ where N_+ denotes the truncated normal distribution bounded above zero. We set $s_\beta = 0.1$.
2. Update $\beta_{ij} = \beta_{ij}^q$ with the acceptance probability

$$\min\left[1, \left\{\prod_{t=1}^{T} \frac{\mathrm{C}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{G})}{\mathrm{C}(\boldsymbol{\alpha}, \boldsymbol{\beta}^q, \boldsymbol{G})} \frac{\exp(\beta_{ij}^q e_{it} e_{jt})}{\exp(\beta_{ij} e_{it} e_{jt})}\right\} \frac{\Gamma(\beta_{ij}^q; a_\beta, b_\beta)}{\Gamma(\beta_{ij}; a_\beta, b_\beta)} \frac{\mathrm{N}_+(\beta_{ij}; \beta_{ij}^q, s_\beta^2)}{\mathrm{N}_+(\beta_{ij}^q; \beta_{ij}, s_\beta^2)}\right]$$

where $\boldsymbol{\beta}^q = (\beta_{12}, \beta_{13}, \ldots, \beta_{i,j-1}, \beta_{ij}^q, \beta_{i,j+1}, \ldots, \beta_{n-1,n-2}, \beta_{n-1,n})$.

S6. For a randomly chosen (i, j) among non-forced-in edges, update $(\beta_{ij}, \boldsymbol{G})$ by the reversible jump process: (Note that we do not update the forced-in edges, i.e., we fix $E(i, j) = 1$ for the forced-in edges over the MCMC iterations)

1. Let z denote the number of edges in the current graph $\boldsymbol{G}$, i.e., $z = \sum_{\{(i,j):i\neq j\}} E(i, j)$ and z_{force} denote the number of forced-in edges. Propose the number of edges E^q from the proposal distribution,

$$q(z^q \mid z) = 0.5\ I\left[z^q = z - 1\right] + 0.5\ I\left[z^q = z + 1\right].$$

   If $z = z_{\text{force}}$, set $z^q = z + 1$ with probability 1. If $z = z_{\max}$, set $z^q = z_{\max} - 1$ with probability 1 where $z_{\max}$ denotes the maximum number of possible edges, i.e., $z_{\max} = \binom{n}{2}$.
2. Propose $\boldsymbol{G}^q$ from the proposal distribution $q(\boldsymbol{G}^q|\boldsymbol{G}, z^q)$ and then β_{ij}^q from the proposal distribution $q(\beta_{ij}^q|\boldsymbol{G}^q, z^q)$.
   (a) For the case where $z^q > z$, randomly select a pair of (i, j) such that $E(i, j) = 0$ and let $E(i, j)^q = 1$ with the proposal distribution

$$q(\boldsymbol{G}^q|\boldsymbol{G}, z^q) = \frac{1}{\#\{(i^*, j^*) : G_{i^*j^*} = 0\}} = \frac{1}{z_{\max} - z}$$

   while $G_{i^*j^*}^q = G_{i^*j^*}$ for all other (i^*, j^*). Propose β_{ij}^q from $q(\beta_{ij}^q|E(i, j)^q, z^q) = \Gamma(\beta_{ij}^q; a_{\beta_G}, b_{\beta_G})$. We set $a_{\beta_G} = b_{\beta_G} = 1$.
   (b) For the case where $z^q < z$, randomly select a non-forced-in edge (i, j) such that $E(i, j) = 1$, and let $E(i, j)^q = 0$ with the proposal distribution

$$q(\boldsymbol{G}^q|\boldsymbol{G}, z^q) = \frac{1}{\#\{(i^*, j^*) : G_{i^*j^*} = 1\}} = \frac{1}{z - z_{\text{force}}}$$

   while $G_{i^*j^*}^q = G_{i^*j^*}$ for all other (i^*, j^*). Propose β_{ij}^q from $q(\beta_{ij}^q|E(i, j)^q, z^q) = \delta_0(\beta_{ij}^q)$.

3. Update $(\beta_{ij}, \boldsymbol{G}) = (\beta_{ij}^q, \boldsymbol{G}^q)$ with the acceptance probability

$$\min\left[1, \left\{\prod_{t=1}^{T} \frac{\mathrm{C}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{G})}{\mathrm{C}(\boldsymbol{\alpha}, \boldsymbol{\beta}^q, \boldsymbol{G}^q)} \frac{\exp(\beta_{ij}^q e_{it} e_{jt})}{\exp(\beta_{ij} e_{it} e_{jt})}\right\} \frac{p(\beta_{ij}^q | E(i,j)^q)}{p(\beta_{ij} | E(i,j))} \frac{q(\beta_{ij}|\boldsymbol{G}, z) q(\boldsymbol{G}|\boldsymbol{G}^q, z) q(z|z^q)}{q(\beta_{ij}^q|\boldsymbol{G}^q, z^q) q(\boldsymbol{G}^q|\boldsymbol{G}, z^q) q(z^q|z)}\right]$$

where $\boldsymbol{\beta}^q = (\beta_{12}, \beta_{13}, \ldots, \beta_{i,j-1}, \beta_{ij}^q, \beta_{i,j+1}, \ldots, \beta_{n-1,n-2}, \beta_{n-1,n})$ and
$\boldsymbol{G}^q = (G_{12}, G_{13}, \ldots, G_{i,j-1}, E(i,j)^q, G_{i,j+1}, \ldots, G_{n-1,n-2}, G_{n-1,n})$.

Note that $p(\beta_{ij}|E(i,j)) = q(\beta_{ij}|\boldsymbol{G}, z)$ when $z^q > z$ and $p(\beta_{ij}^q|E(i,j)^q) = q(\beta_{ij}^q|\boldsymbol{G}^q, z^q)$ when $z^q < z$ and, so they are cancelled out from the acceptance probability.

In the Example section of the main text where the GWAS datasets of 228,944 SNPs for 12 diseases were analyzed, the MCMC algorithm takes about 51 minutes per 1,000 iterations using a single 2.2 GHz Intel i7 Core processor. We make the posterior inference for the real data analysis based on the last 40,000 draws from MCMC after tossing out the first 10,000 iterations as burn-in, with the total computation time of about 1.8 days. Note that in this paper, we used the total of 50,000 iterations conservatively to make sure that we could detect weak pleiotropy between diseases. In our MCMC diagnostics, the trace plots suggested that all parameters converge after at most 2,000 iterations.

1.3 Software Implementation

We implemented the graph-GPA model as an R package `'GGPA'`, which is publicly available at our GitHub webpage (`http://dongjunchung.github.io/GGPA/`). In addition, in order to facilitate users' convenience to generate a prior phenotype graph, we developed *DDNet* (`http://www.chunglab.io/ddnet/`), a web interface that allows users to query diseases of interest, investigate their relationships visually, and download the adjacency matrix for the graph-GPA analysis.

Using DDNet, users can generate a prior disease graph as follows. First, if you open the web address `http://www.chunglab.io/ddnet/` in your web browser, you can see the web interface that looks like Figure 1. In the left side, you can a box and you can query diseases of interest. If you want to try an example list of diseases, just click "Try Example" on the top (Figure 2). Alternatively, you can upload a text file of disease names of interest using the "Upload" button. Note that we constructed our disease dictionary using the Disease Ontology database (`http://disease-ontology.org/`). Hence, if you cannot find a disease of your interest, please check the Disease Ontology database. Then, please click the "Submit" button.

Upon clicking the "Submit" button, you will see a network of the diseases you queried in the right side, as depcited in Figure 3. By either using a bar of typing a value below the "Cut-Off Value" section, you can dynamically investigate disease network structures. Here, an edge is connected between a pair of diseases if the corresponding partial correlation coefficient is larger than the

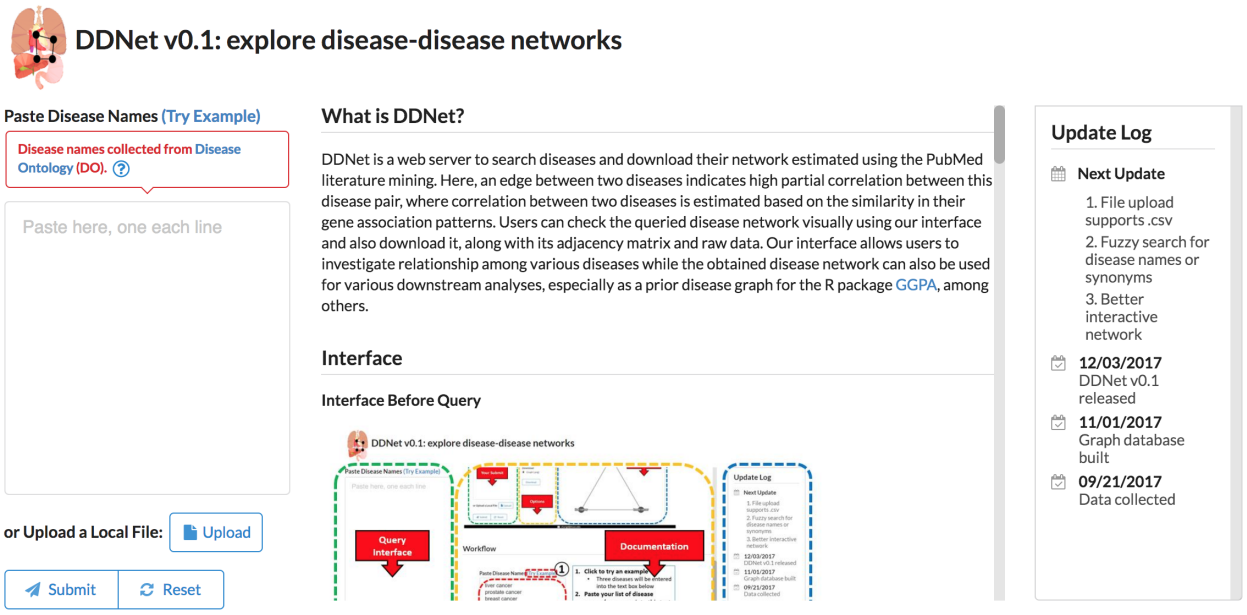

Figure 1: DDNet web interface: Step 1. Enter `http://www.chunglab.io/ddnet/` in your web browser.

specified cut-off. If you click "Download" button, you can also download the disease network plot in PNG file format. If you click the "Table" tab above the disease graph, you can check the adjency matrix corresponding to the disease network for the specified cut-off (Figure 4). You can also check the raw partial correlation coefficient matrix by clicking the "Raw Matrix" tab below the "Table" tab. By clicking "Download" button, you can download the adjacency matrix in the CSV file format and this can be used as a direct input for the `GGPA` package.

The R package `'GGPA'` has three main functions, namely `GGPA()`, `assoc()`, and `plot()`. Specifically, `GGPA()` first fits the graph-GPA model with or without the prior knowledge. Then, `assoc()` and `plot()` implement the association mapping and the phenotype graph estimation, respectively. In order to boost the computational efficiency, the core of the function `GGPA()` is written using the R package `'Rcpp'`, which provides a seamless interface between `R` and `C++`. Suppose that the GWAS association p-value matrix and the downloaded prior disease graph CSV file are loaded to the R environment with object names `pmat` and `pgraph`, respectively. It is assumed that rows and columns of the object `pmat` correspond to SNPs and phenotypes while the objects `pgraph` and `pmat` have the same number of columns and also share the same column names.

First, the following command line fits the graph-GPA model using the downloaded disease network as a prior phenotype graph.

```
R> fit <- GGPA( pmat, pgraph )
```

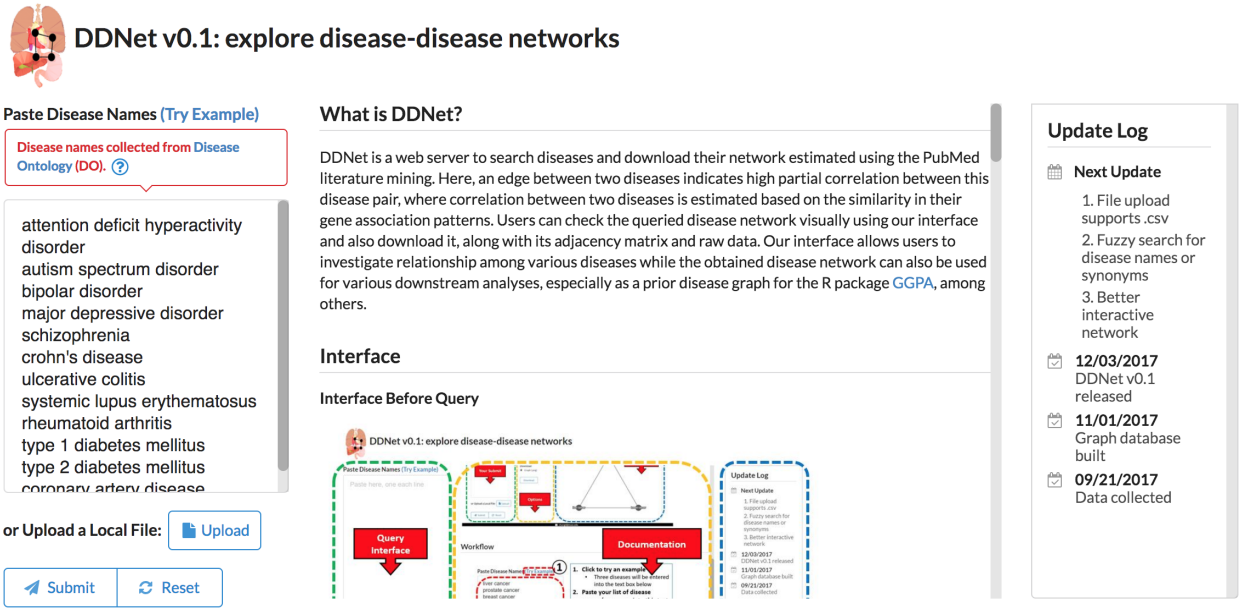

Figure 2: DDNet web interface: Step 2. Enter a list of diseases. Click "Try Example" for an example list of diseases.

Note that if the second argument is missing, then the uninformative prior distribution is used for the phenotype graph. Then, the following command line takes the output of `GGPA()` as an input and generates an estimated phenotype graph plot.

```
R> plot( fit )
```

The association mapping for each of the phenotypes can be implemented using the following command line, where the local false discovery rate (FDR) is controlled at the nominal level of 0.20.

```
R> assoc( fit, FDR=0.20, fdrControl="local" )
```

Similarly, the SNPs that are shared between a pair of phenotypes can also be identified as well using the function `assoc()`. For example, the SNPs that are shared between the fifth and sixth phenotypes (which can be specified using arguments `i` and `j`, respectively) are identified with the following command line.

```
R> assoc( fit, FDR=0.20, fdrControl="local", i=5, j=6 )
```

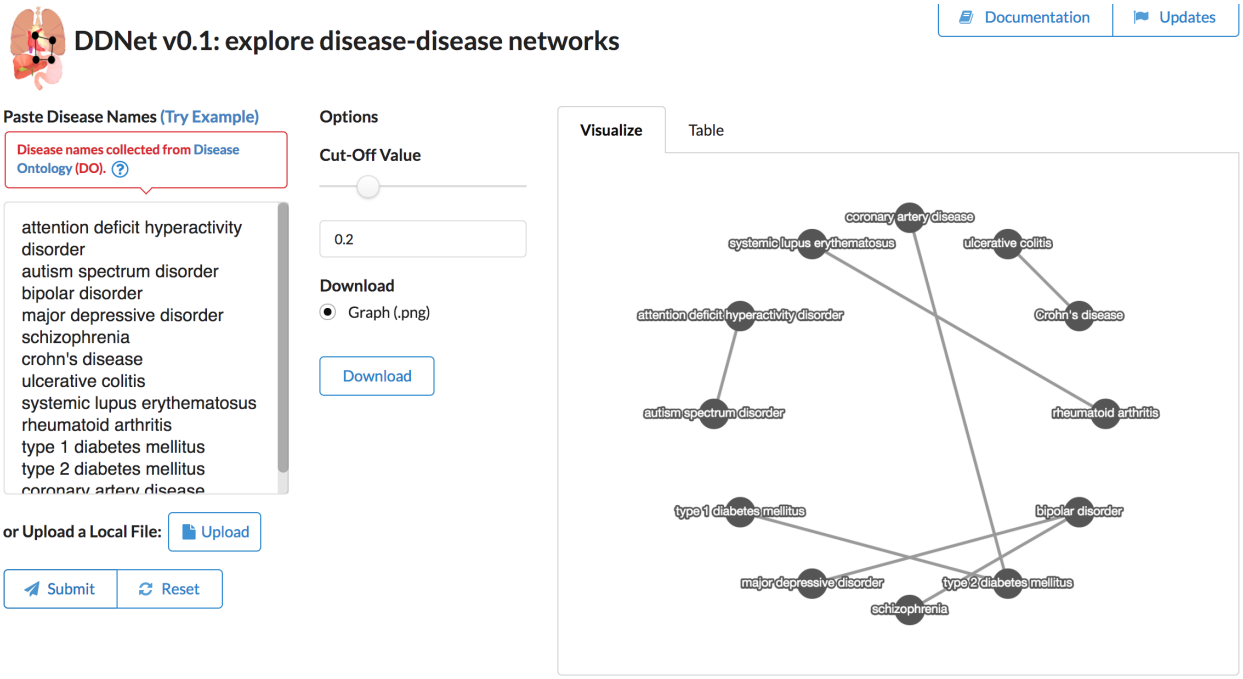

Figure 3: DDNet web interface: Step 3. Investigate a disease-disease network visually.

The R package `GGPA` also provides other useful functions to improve user experience. More information can be found in the R package vignette (`http://dongjunchung.github.io/GGPA/`).

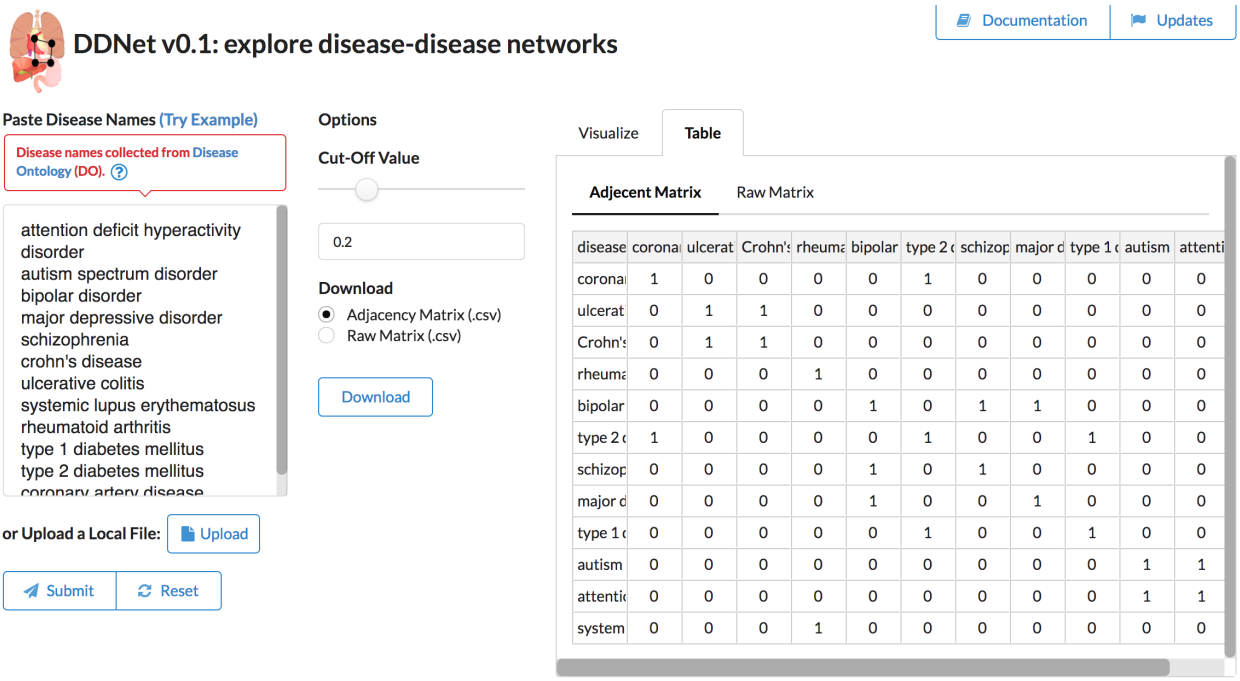

Figure 4: DDNet web interface: Step 4. Download an adjacency matrix for the `graph-GPA` analysis.

2 Strategies to Construct a Prior Disease Graph

We constructed a prior disease graph using a text mining of PubMed literature. Here, we specifically focus on indirect disease-disease relationship mediated by genes. As a result, an edge between two diseases in this disease graph reflects 1) how many genes are shared between two diseases and 2) how many abstracts support this relationship in the literature. We constructed this prior disease graph using the following workflow.

First, we constructed dictionaries of names and aliases for genes and diseases using multiple annotation databases. Specifically, for the gene dictionary, we used 39,819 genes obtained from the *HUGO Gene Nomenclature Committee database* (`http://www.genenames.org/`) on 06/21/2016 and further integrated gene aliases from the *Ensembl* (`http://www.ensembl.org/`) and *UniProt* (`http://www.uniprot.org/`) databases. For the disease dictionary, we used 6,878 symbols and their aliases from the *Disease Ontology* database (`http://disease-ontology.org/`) release 2016-01-07. We add a simple post-processing step to filter out ambiguous keywords in both dictionaries that may refer to other common things (e.g. MICE is a gene name but can also refer to a common model organism).

Next, we used the PubMed Efetch API to check the occurrences of keywords from the dictionaries in the PubMed abstracts or titles. Specifically, using the Python package *Biopython*, we checked the occurrences of each pair of a gene name and a disease name, along with their marginals. In order to infer the association for a given pair of a gene and a disease, we implemented a hypergeometric test to evaluate whether the number of abstracts shared between this pair is significantly larger than what is expected by chance given their marginal counts. Hence, the more abstracts contain the information for this pair, the smaller hypergeometric test p-value we have for this pair. One key strength of this approach is that it takes into account marginal counts, i.e., how much each gene or disease has been studied in the literature. Finally, after taking probit transformation of hypergeometric test p-values, we calculated the correlation coefficient between each pair of diseases and then converted these correlation coefficients to corresponding partial correlation coefficients. The final prior disease graph was constructed by linking edges whose partial correlation coefficients are larger than a specified threshold.

Instead of the workflow described above, alternative approaches can be used to infer relationships among diseases. First, in the context of GWAS, often a pair of diseases are considered to be genetically related if they share a significant number of SNPs identified at the genome-wide significance level, e.g., those reported in the GWAS Catalog (`https://www.ebi.ac.uk/gwas/`). However, this approach has a limitation of missing true positives, especially for complex diseases, because there are a large number of SNPs with weak effect sizes, which do not pass the genome-wide significance level. Second, a disease graph can be constructed using various alternative sources of information, e.g., shared symptoms reported in the biomedical literature (Zhou *et al.* 2014), comor-

bidities reported in medical records (Hidalgo *et al.* 2009), co-expression between genes associated with overlapping diseases, and biological similarities defined using Gene Ontology (GO) annotation (Menche *et al.* 2015), among others. Finally, instead of the partial correlation coefficients that are used in this paper, a disease-disease network can be constructed using other similarity measures such as a cosine similarity based on feature vectors (Zhou *et al.* 2014), relative risk and ϕ-correlation (Hidalgo *et al.* 2009), among others.

3 Real Data Analysis

3.1 GWAS Datasets Used in the Real Data Analysis

In the Example section of the main text, we analyzed GWAS datasets for 12 complex diseases, including attention deficit-hyperactivity disorder (ADHD), autism spectrum disorder (ASD), bipolar disorder (BPD), major depressive disorder (MDD), schizophrenia (SCZ) (Psychiatric Genomics Consortium (PGC); `http://www.med.unc.edu/pgc`; Cross-Disorder Group of the Psychiatric Genomics Consortium *et al.* 2013a,b), Crohn's disease (CD), ulcerative colitis (UC) (International Inflammatory Bowel Disease Genetics Consortium (IIBDGC); `http://ibdgenetics.org`; Franke *et al.* 2010; Anderson *et al.* 2011), systemic lupus erythematosus (SLE) (Hom *et al.* 2008), rheumatoid arthritis (RA) (`http://www.broadinstitute.org/ftp/pub/rheumatoid_arthritis/Stahl_etal_2010NG/`; Stahl *et al.* 2010), type 1 diabetes (T1D) (Barrett *et al.* 2009), type 2 diabetes (T2D) (DIAbetesGenetics Replication and Meta-analysis Consortium (DIAGRAM); `http://diagram-consortium.org`; Morris *et al.* 2012), and coronary artery disease (CAD) (CARDIoGRAM Consortium; `http://www.cardiogramplusc4d.org/downloads/`; Schunkert *et al.* 2011).

3.2 Prior Disease Graphs Obtained from the Literature Mining

Figure 5 shows the prior disease graphs obtained from the literature mining. When we link edges whose partial correlation coefficients are larger than 0.2, neuropsychiatric disorders are linked together, autoimmune diseases are linked together, and type 1 and 2 diabetes are linked together. In addition, we also implemented gene set analyses (for pathway and disease categories) for the top 100 genes that are predicted to be associated with each of the 12 diseases in the literature mining, using the ToppFun tool in the ToppGene Suite (`https://toppgene.cchmc.org/`). Tables 1 - 24 show the top 5 enriched pathways and diseases enriched for the top 100 genes associated with each disease in the gene set enrichment analyses. The results indicate that these genes are enriched for the corresponding disease and also for the pathways related to each disease.

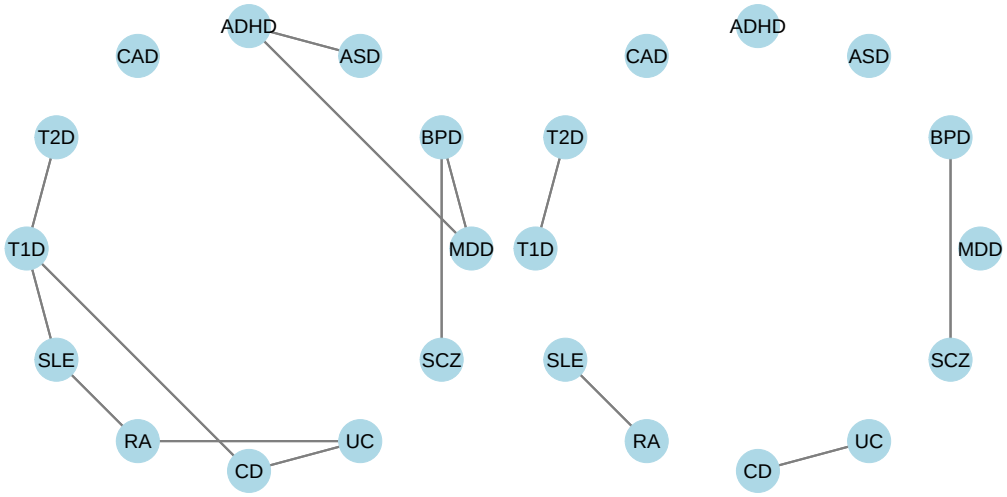

Figure 5: Prior disease graphs obtained the literature mining, where we link edges whose partial correlation coefficients are larger than 0.2 (left) and 0.4 (right).

Table 1: Top 5 pathways enriched for the top 100 genes that are predicted to be associated with ADHD in the literature mining.

Rank	Name	Source	pValue	FDR
1	Dopamine receptor mediated signaling pathway	PantherDB	2.67E-19	1.93E-16
2	Amine ligand-binding receptors	BioSystems: REACTOME	6.87E-17	2.48E-14
3	Adrenaline and noradrenaline biosynthesis	PantherDB	1.45E-13	3.48E-11
4	Dopaminergic synapse	BioSystems: KEGG	4.35E-12	7.55E-10
5	Neuroactive ligand-receptor interaction	BioSystems: KEGG	5.23E-12	7.55E-10

Table 2: Top 5 diseases enriched for the top 100 genes that are predicted to be associated with ADHD in the literature mining.

Rank	Name	Source	pValue	FDR
1	Attention deficit hyperactivity disorder	DisGeNET Curated	7.60E-51	1.92E-47
2	Impulsive character (finding)	DisGeNET BeFree	4.32E-30	5.45E-27
3	Major Depressive Disorder	DisGeNET Curated	5.31E-29	4.47E-26
4	Nicotine Dependence	DisGeNET BeFree	6.29E-28	3.97E-25
5	Unipolar Depression	DisGeNET Curated	9.33E-28	4.71E-25

Table 3: Top 5 pathways enriched for the top 100 genes that are predicted to be associated with ASD in the literature mining.

Rank	Name	Source	pValue	FDR
1	Interactions of neurexins and neuroligins at synapses	BioSystems: REACTOME	8.18E-13	4.36E-10
2	Protein-protein interactions at synapses	BioSystems: REACTOME	7.60E-12	2.03E-09
3	Neuronal System	BioSystems: REACTOME	7.05E-11	1.25E-08
4	Glutamatergic synapse	BioSystems: KEGG	2.27E-07	3.03E-05
5	Vasopressin-like receptors	BioSystems: REACTOME	3.36E-06	3.59E-04

Table 4: Top 5 diseases enriched for the top 100 genes that are predicted to be associated with ASD in the literature mining.

Rank	Name	Source	pValue	FDR
1	Autism Spectrum Disorders	DisGeNET Curated	1.44E-66	4.23E-63
2	Autistic Disorder	DisGeNET Curated	8.34E-55	1.22E-51
3	Pervasive Development Disorder	DisGeNET BeFree	6.02E-44	5.89E-41
4	Neurodevelopmental Disorders	DisGeNET Curated	1.54E-24	1.13E-21
5	Schizophrenia	DisGeNET Curated	8.11E-22	4.76E-19

Table 5: Top 5 pathways enriched for the top 100 genes that are predicted to be associated with BPD in the literature mining.

Rank	Name	Source	pValue	FDR
1	Neuroactive ligand-receptor interaction	BioSystems: KEGG	1.01E-14	8.28E-12
2	Dopaminergic synapse	BioSystems: KEGG	9.06E-14	3.73E-11
3	Cocaine addiction	BioSystems: KEGG	5.13E-13	1.41E-10
4	Serotonergic synapse	BioSystems: KEGG	7.22E-12	1.33E-09
5	Transmission across Chemical Synapses	BioSystems: REACTOME	8.08E-12	1.33E-09

Table 6: Top 5 diseases enriched for the top 100 genes that are predicted to be associated with BPD in the literature mining.

Rank	Name	Source	pValue	FDR
1	Bipolar Disorder	DisGeNET Curated	1.53E-73	3.49E-70
2	Mood Disorders	DisGeNET Curated	6.50E-52	6.53E-49
3	Nonorganic psychosis	DisGeNET Curated	8.58E-52	6.53E-49
4	Psychotic Disorders	DisGeNET Curated	2.47E-50	1.41E-47
5	Mental disorders	DisGeNET Curated	5.05E-48	2.31E-45

Table 7: Top 5 pathways enriched for the top 100 genes that are predicted to be associated with MDD in the literature mining.

Rank	Name	Source	pValue	FDR
1	Serotonergic synapse	BioSystems: KEGG	9.54E-16	6.15E-13
2	Serotonin receptors	BioSystems: REACTOME	1.63E-15	6.15E-13
3	Neuroactive ligand-receptor interaction	BioSystems: KEGG	2.61E-14	6.56E-12
4	Amine ligand-binding receptors	BioSystems: REACTOME	2.04E-13	3.84E-11
5	Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha mediated pathway	PantherDB	1.34E-12	2.02E-10

Table 8: Top 5 diseases enriched for the top 100 genes that are predicted to be associated with MDD in the literature mining.

Rank	Name	Source	pValue	FDR
1	Major Depressive Disorder	DisGeNET Curated	1.09E-54	2.89E-51
2	Unipolar Depression	DisGeNET Curated	7.93E-51	1.06E-47
3	Mental Depression	DisGeNET Curated	7.71E-48	6.86E-45
4	Depressive disorder	DisGeNET Curated	3.75E-47	2.50E-44
5	Mental disorders	DisGeNET Curated	6.05E-40	3.23E-37

Table 9: Top 5 pathways enriched for the top 100 genes that are predicted to be associated with SCZ in the literature mining.

Rank	Name	Source	pValue	FDR
1	Neuroactive ligand-receptor interaction	BioSystems: KEGG	1.30E-18	1.08E-15
2	Cocaine addiction	BioSystems: KEGG	7.83E-15	3.25E-12
3	Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha mediated pathway	PantherDB	2.24E-14	4.65E-12
4	Transmission across Chemical Synapses	BioSystems: REACTOME	2.24E-14	4.65E-12
5	Amine ligand-binding receptors	BioSystems: REACTOME	9.52E-14	1.58E-11

Table 10: Top 5 diseases enriched for the top 100 genes that are predicted to be associated with SCZ in the literature mining.

Rank	Name	Source	pValue	FDR
1	Psychotic Disorders	DisGeNET Curated	3.04E-64	6.87E-61
2	Schizophrenia	DisGeNET Curated	4.99E-57	5.65E-54
3	Nonorganic psychosis	DisGeNET Curated	2.21E-53	1.67E-50
4	Bipolar Disorder	DisGeNET Curated	1.16E-51	6.55E-49
5	Mental disorders	DisGeNET Curated	1.92E-49	8.69E-47

Table 11: Top 5 pathways enriched for the top 100 genes that are predicted to be associated with RA in the literature mining.

Rank	Name	Source	pValue	FDR
1	Cytokine Signaling in Immune system	BioSystems: REACTOME	6.43E-39	4.80E-36
2	Rheumatoid arthritis	BioSystems: KEGG	7.38E-38	2.76E-35
3	Interleukin-10 signaling	BioSystems: REACTOME	3.49E-29	8.68E-27
4	Cytokine-cytokine receptor interaction	BioSystems: KEGG	3.48E-28	6.50E-26
5	Signaling by Interleukins	BioSystems: REACTOME	1.13E-27	1.68E-25

Table 12: Top 5 diseases enriched for the top 100 genes that are predicted to be associated with RA in the literature mining.

Rank	Name	Source	pValue	FDR
1	Arthritis	DisGeNET Curated	4.97E-84	2.18E-80
2	Rheumatoid Arthritis	DisGeNET Curated	1.74E-76	3.80E-73
3	Superficial ulcer	DisGeNET BeFree	6.69E-53	9.78E-50
4	Autoimmune Diseases	DisGeNET Curated	5.77E-51	6.33E-48
5	Diabetes Mellitus, Insulin-Dependent	DisGeNET Curated	1.42E-50	1.24E-47

Table 13: Top 5 pathways enriched for the top 100 genes that are predicted to be associated with CD in the literature mining.

Rank	Name	Source	pValue	FDR
1	Cytokine-cytokine receptor interaction	BioSystems: KEGG	9.01E-10	4.51E-07
2	Inflammatory bowel disease (IBD)	BioSystems: KEGG	5.61E-09	1.40E-06
3	Signaling by Interleukins	BioSystems: REACTOME	1.80E-08	3.00E-06
4	Innate Immune System	BioSystems: REACTOME	3.47E-08	4.34E-06
5	NOD-like receptor signaling pathway	BioSystems: KEGG	8.62E-08	7.64E-06

Table 14: Top 5 diseases enriched for the top 100 genes that are predicted to be associated with CD in the literature mining.

Rank	Name	Source	pValue	FDR
1	Crohn Disease	DisGeNET Curated	2.03E-44	5.43E-41
2	Ulcerative Colitis	DisGeNET Curated	4.68E-39	6.26E-36
3	Inflammatory Bowel Diseases	DisGeNET Curated	4.19E-29	3.73E-26
4	Colitis	DisGeNET Curated	2.13E-22	1.42E-19
5	Crohn's disease	GWAS	1.44E-17	7.70E-15

Table 15: Top 5 pathways enriched for the top 100 genes that are predicted to be associated with UC in the literature mining.

Rank	Name	Source	pValue	FDR
1	Inflammatory bowel disease (IBD)	BioSystems: KEGG	7.12E-17	4.06E-14
2	Defective C1GALT1C1 causes Tn polyagglutination syndrome (TNPS)	BioSystems: REACTOME	7.41E-10	8.57E-08
3	Defective GALNT12 causes colorectal cancer 1 (CRCS1)	BioSystems: REACTOME	7.41E-10	8.57E-08
4	Defective GALNT3 causes familial hyperphosphatemic tumoral calcinosis (HFTC)	BioSystems: REACTOME	7.41E-10	8.57E-08
5	Cytokine-cytokine receptor interaction	BioSystems: KEGG	7.52E-10	8.57E-08

Table 16: Top 5 diseases enriched for the top 100 genes that are predicted to be associated with UC in the literature mining.

Rank	Name	Source	pValue	FDR
1	Ulcerative Colitis	DisGeNET Curated	6.22E-48	1.96E-44
2	Crohn Disease	DisGeNET Curated	7.14E-42	1.13E-38
3	Inflammatory Bowel Diseases	DisGeNET Curated	4.26E-33	4.49E-30
4	Colitis	DisGeNET Curated	5.68E-30	4.48E-27
5	Irritable Bowel Syndrome	DisGeNET Curated	1.65E-18	1.04E-15

Table 17: Top 5 pathways enriched for the top 100 genes that are predicted to be associated with SLE in the literature mining.

Rank	Name	Source	pValue	FDR
1	Systemic lupus erythematosus	BioSystems: KEGG	3.60E-27	2.27E-24
2	Staphylococcus aureus infection	BioSystems: KEGG	3.50E-19	1.11E-16
3	Intestinal immune network for IgA production	BioSystems: KEGG	1.04E-14	2.20E-12
4	Antigen Dependent B Cell Activation	MSigDB C2 BIOCARTA (v6.0)	2.92E-13	4.62E-11
5	Autoimmune thyroid disease	BioSystems: KEGG	1.19E-12	1.51E-10

Table 18: Top 5 diseases enriched for the top 100 genes that are predicted to be associated with SLE in the literature mining.

Rank	Name	Source	pValue	FDR
1	Lupus Erythematosus, Systemic	DisGeNET Curated	3.33E-83	1.01E-79
2	Lupus Vulgaris	DisGeNET BeFree	4.14E-74	6.25E-71
3	Lupus Erythematosus	DisGeNET BeFree	1.80E-73	1.81E-70
4	Lupus Erythematosus, Discoid	DisGeNET Curated	7.51E-73	5.66E-70
5	Autoimmune Diseases	DisGeNET Curated	8.17E-52	4.93E-49

Table 19: Top 5 pathways enriched for the top 100 genes that are predicted to be associated with T1D in the literature mining.

Rank	Name	Source	pValue	FDR
1	Type I diabetes mellitus	BioSystems: KEGG	6.22E-15	4.04E-12
2	Staphylococcus aureus infection	BioSystems: KEGG	1.10E-13	3.58E-11
3	Antigen processing and presentation	BioSystems: KEGG	3.13E-12	6.79E-10
4	Asthma	BioSystems: KEGG	1.14E-10	1.85E-08
5	Phagosome	BioSystems: KEGG	1.78E-10	2.31E-08

Table 20: Top 5 diseases enriched for the top 100 genes that are predicted to be associated with T1D in the literature mining.

Rank	Name	Source	pValue	FDR
1	Diabetes Mellitus, Insulin-Dependent	DisGeNET Curated	2.04E-20	4.43E-17
2	Graves Disease	DisGeNET Curated	2.00E-10	2.17E-07
3	Rheumatoid Arthritis	DisGeNET Curated	4.44E-10	3.22E-07
4	Diabetes Mellitus	DisGeNET Curated	9.16E-10	4.97E-07
5	Diabetes Mellitus, Non-Insulin-Dependent	DisGeNET Curated	3.22E-09	1.39E-06

Table 21: Top 5 pathways enriched for the top 100 genes that are predicted to be associated with T2D in the literature mining.

Rank	Name	Source	pValue	FDR
1	Type II diabetes mellitus	BioSystems: KEGG	4.21E-15	2.07E-12
2	Insulin resistance	BioSystems: KEGG	4.72E-15	2.07E-12
3	Insulin signaling pathway	BioSystems: KEGG	3.60E-12	1.05E-09
4	AMPK signaling pathway	BioSystems: KEGG	2.75E-10	6.02E-08
5	Regulation of lipolysis in adipocytes	BioSystems: KEGG	1.67E-09	2.92E-07

Table 22: Top 5 diseases enriched for the top 100 genes that are predicted to be associated with T2D in the literature mining.

Rank	Name	Source	pValue	FDR
1	Diabetes Mellitus, Non-Insulin-Dependent	DisGeNET Curated	2.29E-33	5.20E-30
2	Impaired glucose tolerance	DisGeNET Curated	1.99E-27	2.25E-24
3	Diabetes	DisGeNET BeFree	1.23E-23	9.26E-21
4	Diabetes Mellitus	DisGeNET Curated	2.58E-23	1.46E-20
5	Hyperglycemia	DisGeNET Curated	5.71E-19	2.59E-16

Table 23: Top 5 pathways enriched for the top 100 genes that are predicted to be associated with CAD in the literature mining.

Rank	Name	Source	pValue	FDR
1	Lipoprotein metabolism	BioSystems: REACTOME	2.51E-17	1.57E-14
2	lipoprotein metabolic	Pathway Ontology	2.82E-15	8.82E-13
3	Lipid digestion, mobilization, and transport	BioSystems: REACTOME	1.59E-14	3.32E-12
4	Chylomicron-mediated lipid transport	BioSystems: REACTOME	8.26E-14	1.29E-11
5	Metabolism of lipids and lipoproteins	BioSystems: REACTOME	1.82E-10	2.27E-08

Table 24: Top 5 diseases enriched for the top 100 genes that are predicted to be associated with CAD in the literature mining.

Rank	Name	Source	pValue	FDR
1	Coronary heart disease	DisGeNET Curated	1.64E-40	5.43E-37
2	Cardiovascular Diseases	DisGeNET Curated	5.50E-40	9.10E-37
3	Cerebrovascular accident	DisGeNET Curated	3.41E-39	3.77E-36
4	Coronary Arteriosclerosis	DisGeNET Curated	2.79E-35	2.31E-32
5	Coronary Artery Disease	DisGeNET Curated	1.90E-33	1.26E-30

3.3 graph-GPA Results Without Incorporating the Prior Disease Graph

Table 25: graph-GPA results without incorporating prior disease graph: Estimates of $p(E(i,j)|\boldsymbol{Y})$. The blanked cell indicates the zero estimated value.

	ADHD	ASD	BPD	CAD	CD	MDD	RA	SCZ	SLE	T1D	T2D	UC
ADHD	–	0.11	0.92	0.03	0.65	0.52	0.05		0.25	0.19	0.11	0.09
ASD	0.11	–		0.42	0.03		1.00	0.26		0.05		
BPD	0.92		–		0.83	0.65		1.00	1.00			
CAD	0.03	0.42		–	1.00		0.04	1.00		0.05	1.00	
CD	0.65	0.03	0.83	1.00	–			0.64		1.00	1.00	1.00
MDD	0.52		0.65			–		1.00	1.00	1.00	0.01	
RA	0.05	1.00		0.04			–			1.00	0.78	1.00
SCZ		0.26	1.00	1.00	0.64	1.00		–	1.00	1.00		
SLE	0.25		1.00			1.00		1.00	–	1.00		
T1D	0.19	0.05		0.05	1.00	1.00	1.00	1.00	1.00	–	0.26	
T2D	0.11			1.00	1.00	0.01	0.78			0.26	–	
UC	0.09				1.00		1.00					–

Table 26: graph-GPA results without incorporating prior disease graph: Posterior mean estimates of β_{ij}. The blanked cell indicates that $p(E(i,j)|\boldsymbol{Y})$ is estimated as zero and the bold number indicates that the 95% credible interval β_{ij} does not contain zero.

	ADHD	ASD	BPD	CAD	CD	MDD	RA	SCZ	SLE	T1D	T2D	UC
ADHD	–	0.09	2.32	0.02	0.76	0.76	0.04		0.27	0.20	0.11	0.15
ASD	0.09	–		0.46	0.01		**1.89**	0.17		0.04		
BPD	2.32		–		0.45	0.84		**1.55**	**0.86**			
CAD	0.02	0.46		–	**0.89**		0.02	**1.00**		0.02	**2.13**	
CD	0.76	0.01	0.45	**0.89**	–			0.27		**1.93**	**0.79**	**2.54**
MDD	0.76		0.84			–		**1.31**	**1.11**	**1.06**	0.01	
RA	0.04	**1.89**		0.02			–			**6.69**	0.82	**1.78**
SCZ		0.17	**1.55**	**1.00**	0.27	**1.31**		–	**1.47**	**1.28**		
SLE	0.27		**0.86**			**1.11**		**1.47**	–	**4.22**		
T1D	0.20	0.04		0.02	**1.93**	**1.06**	**6.69**	**1.28**	**4.22**	–	0.26	
T2D	0.11			**2.13**	**0.79**	0.01	0.82			0.26	–	
UC	0.15				**2.54**		**1.78**					–

Table 27: graph-GPA results without incorporating prior disease graph: Numbers of SNPs identified to be associated with each pair of diseases by controlling the local FDR at nominal level of 20%. Diagonal elements show the number of SNPs to be associated with each disease when the local FDR is controlled at the same level.

	ADHD	ASD	BPD	CAD	CD	MDD	RA	SCZ	SLE	T1D	T2D	UC
ADHD	0	0	0	0	0	0	0	0	0	0	0	0
ASD	0	9	0	0	0	0	4	0	0	4	0	0
BPD	0	0	112	2	41	20	28	75	67	68	2	16
CAD	0	0	2	171	22	2	15	17	9	15	32	15
CD	0	0	41	22	1965	33	294	98	121	350	38	766
MDD	0	0	20	2	33	73	23	61	63	65	2	8
RA	0	4	28	15	294	23	655	131	150	541	29	282
SCZ	0	0	75	17	98	61	131	446	180	233	12	72
SLE	0	0	67	9	121	63	150	180	258	243	11	74
T1D	0	4	68	15	350	65	541	233	243	766	27	284
T2D	0	0	2	32	38	2	29	12	11	27	176	16
UC	0	0	16	15	766	8	282	72	74	284	16	1552

Table 28: graph-GPA results without incorporating prior disease graph: Numbers of SNPs identified to be associated with each pair of diseases by controlling the global FDR at nominal level of 50%. Diagonal elements show the number of SNPs to be associated with each disease when the global FDR is controlled at the same level.

	ADHD	ASD	BPD	CAD	CD	MDD	RA	SCZ	SLE	T1D	T2D	UC
ADHD	27	0	25	0	22	15	6	26	25	25	0	1
ASD	0	192	7	8	61	0	114	31	36	99	8	51
BPD	25	7	1488	38	304	214	183	592	258	277	29	169
CAD	0	8	38	1830	350	26	114	180	58	123	413	209
CD	22	61	304	350	18209	222	993	768	493	1194	428	7443
MDD	15	0	214	26	222	543	217	384	302	348	28	130
RA	6	114	183	114	993	217	1892	532	567	1527	204	955
SCZ	26	31	592	180	768	384	532	5669	646	786	98	510
SLE	25	36	258	58	493	302	567	646	1129	843	64	341
T1D	25	99	277	123	1194	348	1527	786	843	2170	201	947
T2D	0	8	29	413	428	28	204	98	64	201	2149	243
UC	1	51	169	209	7443	130	955	510	341	947	243	15631

3.4 graph-GPA Results Incorporating the Prior Disease Graph

Table 29: graph-GPA results incorporating the prior disease graph obtained from the literature mining, where we linked edges whose partial correlation coefficients are larger than 0.2: Estimates of $p(E(i,j)|\boldsymbol{Y})$. The blanked cell indicates the zero estimated value.

	ADHD	ASD	BPD	CAD	CD	MDD	RA	SCZ	SLE	T1D	T2D	UC
ADHD	–	1.00	1.00	0.10	0.18	1.00	0.04		0.42	0.18	0.01	0.01
ASD	1.00	–	0.09	0.82	0.22		1.00	0.06		0.03	0.07	
BPD	1.00	0.09	–	0.05	1.00	1.00		1.00	0.85			
CAD	0.10	0.82	0.05	–		0.01	0.03	1.00		0.02	1.00	1.00
CD	0.18	0.22	1.00		–	0.01		0.38		1.00	0.44	1.00
MDD	1.00		1.00	0.01	0.01	–		1.00	1.00	1.00	0.01	
RA	0.04	1.00		0.03			–		1.00	1.00	0.05	1.00
SCZ		0.06	1.00	1.00	0.38	1.00		–	1.00	1.00		
SLE	0.42		0.85			1.00	1.00	1.00	–	1.00		
T1D	0.18	0.03		0.02	1.00	1.00	1.00	1.00	1.00	–	1.00	
T2D	0.01	0.07		1.00	0.44	0.01	0.05			1.00	–	
UC	0.01			1.00	1.00		1.00					–

Table 30: graph-GPA results incorporating the prior disease graph obtained from the literature mining, where we linked edges whose partial correlation coefficients are larger than 0.2: Posterior mean estimates of β_{ij}. The blanked cell indicates that $p(E(i,j)|\boldsymbol{Y})$ is estimated as zero and the bold number indicates that the 95% credible interval β_{ij} does not contain zero.

	ADHD	ASD	BPD	CAD	CD	MDD	RA	SCZ	SLE	T1D	T2D	UC
ADHD	–	**0.81**	**2.12**	0.08	0.22	**1.16**	0.04		0.47	0.17	0.01	0.01
ASD	**0.81**	–	0.06	0.85	0.10		**1.82**	0.03		0.02	0.05	
BPD	**2.12**	0.06	–	0.03	**0.61**	**1.23**		**1.54**	0.60			
CAD	0.08	0.85	0.03	–		0.01	0.02	**0.99**		0.01	**2.27**	**1.05**
CD	0.22	0.10	**0.61**		–	0.00		0.15		**2.02**	0.38	**2.52**
MDD	**1.16**		**1.23**	0.01	0.00	–		**1.24**	**1.07**	**1.04**	0.00	
RA	0.04	**1.82**		0.02			–		**0.34**	**6.55**	0.03	**1.71**
SCZ		0.03	**1.54**	**0.99**	0.15	**1.24**		–	**1.51**	**1.33**		
SLE	0.47		0.60			**1.07**	**0.34**	**1.51**	–	**3.93**		
T1D	0.17	0.02		0.01	**2.02**	**1.04**	**6.55**	**1.33**	**3.93**	–	**1.27**	
T2D	0.01	0.05		**2.27**	0.38	0.00	0.03			**1.27**	–	
UC	0.01			**1.05**	**2.52**		**1.71**					–

Table 31: graph-GPA results incorporating the prior disease graph obtained from the literature mining, where we linked edges whose partial correlation coefficients are larger than 0.2: Numbers of SNPs identified to be associated with each pair of diseases by controlling the local FDR at nominal level of 20%. Diagonal elements show the number of SNPs to be associated with each disease when the local FDR is controlled at the same level.

	ADHD	ASD	BPD	CAD	CD	MDD	RA	SCZ	SLE	T1D	T2D	UC
ADHD	0	0	0	0	0	0	0	0	0	0	0	0
ASD	0	9	0	0	0	0	4	0	0	4	0	0
BPD	0	0	120	3	46	36	28	81	71	72	2	17
CAD	0	0	3	183	20	2	16	17	8	20	33	24
CD	0	0	46	20	1941	37	282	92	121	346	26	756
MDD	0	0	36	2	37	82	28	71	70	72	2	14
RA	0	4	28	16	282	28	649	132	155	533	25	275
SCZ	0	0	81	17	92	71	132	450	176	233	13	70
SLE	0	0	71	8	121	70	155	176	262	247	12	78
T1D	0	4	72	20	346	72	533	233	247	770	28	283
T2D	0	0	2	33	26	2	25	13	12	28	180	18
UC	0	0	17	24	756	14	275	70	78	283	18	1553

Table 32: graph-GPA results incorporating the prior disease graph obtained from the literature mining, where we linked edges whose partial correlation coefficients are larger than 0.2: Numbers of SNPs identified to be associated with each pair of diseases by controlling the global FDR at nominal level of 50%. Diagonal elements show the number of SNPs to be associated with each disease when the global FDR is controlled at the same level.

	ADHD	ASD	BPD	CAD	CD	MDD	RA	SCZ	SLE	T1D	T2D	UC
ADHD	25	0	23	0	19	20	6	24	23	24	0	1
ASD	0	195	7	20	64	0	114	30	38	100	12	59
BPD	23	7	1533	39	321	254	189	615	253	279	29	184
CAD	0	20	39	1845	258	28	125	181	62	141	424	329
CD	19	64	321	258	18054	226	987	711	496	1217	294	7392
MDD	20	0	254	28	226	571	222	404	302	348	29	134
RA	6	114	189	125	987	222	1893	537	588	1533	192	945
SCZ	24	30	615	181	711	404	537	5644	645	791	100	510
SLE	23	38	253	62	496	302	588	645	1141	845	69	348
T1D	24	100	279	141	1217	348	1533	791	845	2207	234	958
T2D	0	12	29	424	294	29	192	100	69	234	2111	246
UC	1	59	184	329	7392	134	945	510	348	958	246	15680

3.5 Functional Impact Analysis of Identified SNPs

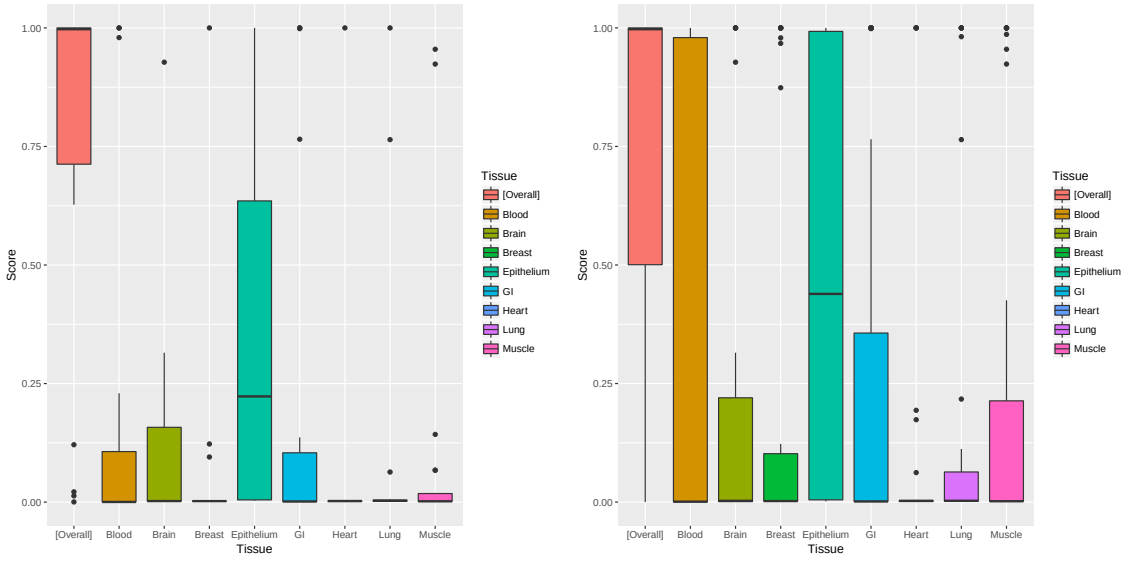

Figure 6: GenoCanyon and GenoSkyline scores for the SNPs shared between BPD and MDD, for the cases without (left) and with (right) incorporating the prior phenotype graph information into graph-GPA. We linked edges whose partial correlation coefficients are larger than 0.2 in the prior disease graph and controlled the local false discovery rate for association mapping at the nominal level of 0.2.

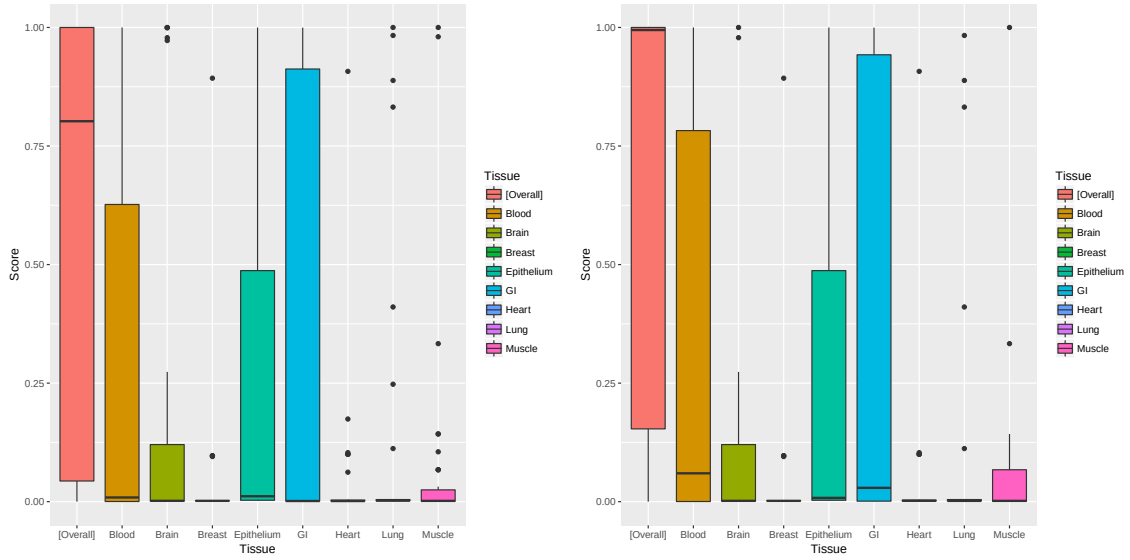

Figure 7: GenoCanyon and GenoSkyline scores for the SNPs shared between CD and T2D, for the cases without (left) and with (right) incorporating the prior phenotype graph information into graph-GPA. We linked edges whose partial correlation coefficients are larger than 0.2 in the prior disease graph and controlled the local false discovery rate for association mapping at the nominal level of 0.2.

4 Reproducibility Analysis

In this section, we check reproducibility of our findings in the real data analysis, by replacing a subset of GWAS datasets with independent validation datasets. Specifically, we replaced the GWAS datasets for RA and T1D with those from the UK Biobank (UKBB) GWAS results (`https://sites.google.com/broadinstitute.org/ukbbgwasresults/home?authuser=0`). For RA and T1D in UKBB, we used GWAS results for the corresponding ICD10 diagnostics phenotypes, i.e., "M05 Seropositive rheumatoid arthritis and Diagnoses" and "E10 Insulin-dependent diabetes mellitus" for RA and T1D, respectively.

Figure 8 shows the pleiotropic architectures estimated using graph-GPA without and with using the prior disease graph obtained from the literature mining and Sections 4.1 and 4.2 provide the graph-GPA analysis results without and with using the prior disease graph., respectively When the prior disease graph is not incorporated, 11 among 17 edges were changed (6 added and 5 lost) by replacing the GWAS datasets for RA and T1D with the UKBB data. In contrast, when the prior disease graph is incorporated, only 6 among 22 edges were changed (3 added and 3 lost) by replacing the GWAS datasets for RA and T1D with the UKBB data. These results indicate that 1) our findings reported in the Example section are reasonably well reproduced in the UKBB data; and 2) higher degree of reproducibility is observed when informative prior disease graph obtained from the literature mining was used.

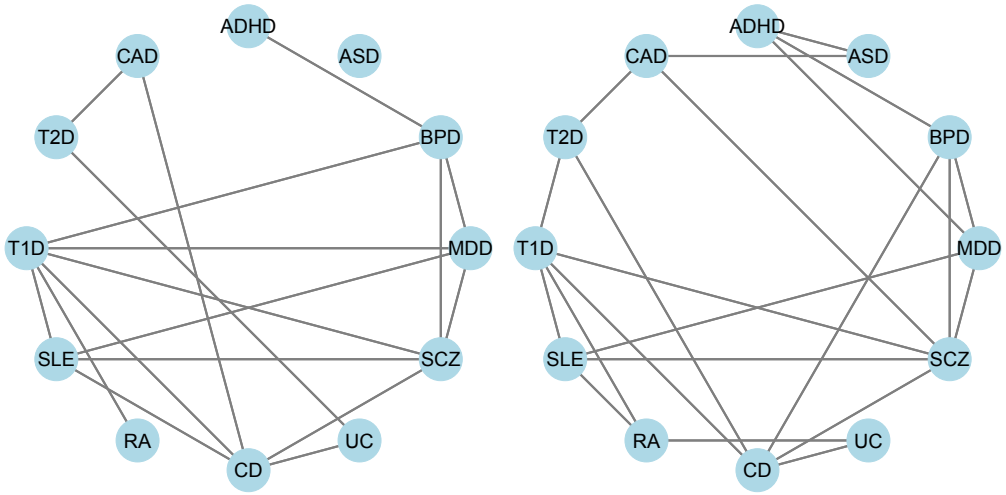

Figure 8: graph-GPA results when GWAS data for RA and T1D are replaced with those from UKBB: pleiotropic architectures estimated using graph-GPA without (left) and with (right) using a prior disease graph obtained from the literature mining.

4.1 graph-GPA Results Without Incorporating the Prior Disease Graph

Table 33: graph-GPA results when GWAS data for RA and T1D are replaced with those from UKBB and the prior disease graph is not used: Estimates of $p(E(i,j)|\boldsymbol{Y})$. The blanked cell indicates the zero estimated value.

	ADHD	ASD	BPD	CAD	CD	MDD	RA	SCZ	SLE	T1D	T2D	UC
ADHD	–	0.02	1.00			0.44		0.03	0.54	0.47	0.03	0.02
ASD	0.02	–	0.84	0.90		0.02	0.93		0.02	0.01	0.18	
BPD	1.00	0.84	–	0.68	0.09	0.99		1.00	0.20	1.00		
CAD		0.90	0.68	–	1.00		0.78	0.64			1.00	
CD	0.00		0.09	1.00	–	0.08	0.36	1.00	1.00	1.00		1.00
MDD	0.44	0.02	0.99		0.08	–		1.00	1.00	1.00	0.07	
RA	0.00	0.93		0.78	0.36		–			1.00	0.58	0.97
SCZ	0.03		1.00	0.64	1.00	1.00		–	1.00	1.00		
SLE	0.54	0.02	0.20		1.00	1.00		1.00	–	1.00	0.01	0.02
T1D	0.47	0.01	1.00		1.00	1.00	1.00	1.00	1.00	–	0.06	0.16
T2D	0.03	0.18		1.00		0.07	0.58		0.01	0.06	–	1.00
UC	0.02				1.00		0.97		0.02	0.16	1.00	–

Table 34: graph-GPA results when GWAS data for RA and T1D are replaced with those from UKBB and the prior disease graph is not used: Posterior mean estimates of β_{ij}. The blanked cell indicates that $p(E(i,j)|\boldsymbol{Y})$ is estimated as zero and the bold number indicates that the 95% credible interval β_{ij} does not contain zero.

	ADHD	ASD	BPD	CAD	CD	MDD	RA	SCZ	SLE	T1D	T2D	UC
ADHD	–	0.01	**2.30**			0.53		0.02	0.79	0.71	0.02	0.01
ASD	0.01	–	0.86	0.94		0.01	1.69		0.01		0.13	
BPD	**2.30**	0.86	–	0.52	0.04	**1.25**		**1.57**	0.15	**1.33**		
CAD	0.00	0.94	0.52	–	**0.95**		1.03	0.58			**2.20**	
CD	0.00		0.04	**0.95**	–	0.02	0.34	**0.61**	**1.04**	**2.40**		**2.60**
MDD	0.53	0.01	**1.25**		0.02	–		**1.21**	**1.38**	**1.13**	0.05	
RA	0.00	1.69		1.03	0.34		–			**4.92**	0.74	1.77
SCZ	0.02		**1.57**	0.58	**0.61**	**1.21**		–	**1.26**	**1.23**		
SLE	0.79	0.01	0.15		**1.04**	**1.38**		**1.26**	–	**4.04**	0.01	0.01
T1D	0.71		**1.33**		**2.40**	**1.13**	**4.92**	**1.23**	**4.04**	–	0.04	0.10
T2D	0.02	0.13		**2.20**		0.05	0.74		0.01	0.04	–	**0.85**
UC	0.01				**2.60**		1.77		0.01	0.10	**0.85**	–

Table 35: graph-GPA results when GWAS data for RA and T1D are replaced with those from UKBB and the prior disease graph is not used: Numbers of SNPs identified to be associated with each pair of diseases by controlling the local FDR at nominal level of 20%. Diagonal elements show the number of SNPs to be associated with each disease when the local FDR is controlled at the same level.

	ADHD	ASD	BPD	CAD	CD	MDD	RA	SCZ	SLE	T1D	T2D	UC
ADHD	0	0	0	0	0	0	0	0	0	0	0	0
ASD	0	10	2	0	5	0	1	0	5	5	0	0
BPD	0	2	128	6	63	59	15	90	79	75	2	17
CAD	0	0	6	169	20	6	9	12	8	8	26	16
CD	0	5	63	20	1897	53	41	94	109	132	19	692
MDD	0	0	59	6	53	83	9	73	71	70	2	16
RA	0	1	15	9	41	9	55	25	30	42	7	36
SCZ	0	0	90	12	94	23	25	351	107	99	5	40
SLE	0	5	79	8	109	71	30	107	156	117	5	40
T1D	0	5	75	8	132	70	42	99	117	164	7	55
T2D	0	0	2	26	19	2	7	5	5	7	163	18
UC	0	0	17	16	692	16	36	40	40	55	18	1458

4.2 graph-GPA Results Incorporating the Prior Disease Graph

Table 36: graph-GPA results when GWAS data for RA and T1D are replaced with those from UKBB and the prior disease graph obtained from the literature mining is used: Estimates of $p(E(i,j)|\boldsymbol{Y})$. The blanked cell indicates the zero estimated value.

	ADHD	ASD	BPD	CAD	CD	MDD	RA	SCZ	SLE	T1D	T2D	UC
ADHD	–	1.00	1.00	0.01	0.16	1.00		0.02	0.40	0.49	0.01	
ASD	1.00	–	0.80	1.00	0.16		0.85		0.03	0.01		
BPD	1.00	0.80	–	0.35	1.00	1.00		1.00	0.66	0.58		
CAD	0.01	1.00	0.35	–		0.02	0.89	0.99		0.03	1.00	0.90
CD	0.16	0.16	1.00		–		0.11	1.00	0.76	1.00	1.00	1.00
MDD	1.00		1.00	0.02		–		1.00	1.00	0.95	0.04	
RA		0.85		0.89	0.11		–		1.00	1.00	0.25	1.00
SCZ	0.02		1.00	0.99	1.00	1.00		–	1.00	1.00		0.03
SLE	0.40	0.03	0.66		0.76	1.00	1.00	1.00	–	1.00		
T1D	0.49	0.01	0.58	0.03	1.00	0.95	1.00	1.00	1.00	–	1.00	0.11
T2D	0.01			1.00	1.00	0.04	0.25			1.00	–	0.18
UC				0.90	1.00		1.00	0.03		0.11	0.18	–

Table 37: graph-GPA results when GWAS data for RA and T1D are replaced with those from UKBB and the prior disease graph obtained from the literature mining is used: Posterior mean estimates of β_{ij}. The blanked cell indicates that $p(E(i,j)|\boldsymbol{Y})$ is estimated as zero and the bold number indicates that the 95% credible interval β_{ij} does not contain zero.

	ADHD	ASD	BPD	CAD	CD	MDD	RA	SCZ	SLE	T1D	T2D	UC
ADHD	–	**0.83**	**2.25**		0.20	**1.12**		0.02	0.47	0.66	0.01	
ASD	**0.83**	–	0.80	**1.01**	0.08		1.40		0.01	0.01		
BPD	**2.25**	0.80	–	0.20	**0.48**	**1.19**		**1.51**	0.60	0.61		
CAD	0.00	**1.01**	0.20	–		0.01	1.30	**0.93**		0.02	**2.17**	0.78
CD	0.20	0.08	**0.48**		–		0.09	**0.54**	0.77	**2.51**	**0.90**	**2.60**
MDD	**1.12**		**1.19**	0.01		–		**1.23**	**1.31**	1.19	0.02	
RA	0.00	1.40		1.30	0.09		–		**1.01**	**4.12**	0.26	**1.92**
SCZ	0.02		**1.51**	**0.93**	**0.54**	**1.23**		–	**1.22**	**1.32**		0.01
SLE	0.47	0.01	0.60		0.77	**1.31**	**1.01**	**1.22**	–	**4.07**		
T1D	0.66	0.01	0.61	0.02	**2.51**	1.19	**4.12**	**1.32**	**4.07**	–	**0.50**	0.06
T2D	0.01			**2.17**	**0.90**	0.02	0.26			**0.50**	–	0.07
UC	0.00			0.78	**2.60**		**1.92**	0.01		0.06	0.07	–

Table 38: graph-GPA results when GWAS data for RA and T1D are replaced with those from UKBB and the prior disease graph obtained from the literature mining is used: Numbers of SNPs identified to be associated with each pair of diseases by controlling the local FDR at nominal level of 20%. Diagonal elements show the number of SNPs to be associated with each disease when the local FDR is controlled at the same level.

	ADHD	ASD	BPD	CAD	CD	MDD	RA	SCZ	SLE	T1D	T2D	UC
ADHD	0	0	0	0	0	0	0	0	0	0	0	0
ASD	0	9	0	0	4	0	0	0	4	4	0	0
BPD	0	0	128	6	63	55	12	92	80	73	6	17
CAD	0	0	6	171	20	4	9	15	8	9	30	20
CD	0	4	63	20	1888	55	38	100	106	131	29	696
MDD	0	0	55	4	44	82	10	73	70	69	2	14
RA	0	0	12	9	38	10	54	25	30	41	7	38
SCZ	0	0	92	15	100	73	25	355	106	100	7	44
SLE	0	4	80	8	106	70	30	106	153	114	7	41
T1D	0	4	73	9	131	69	41	100	114	165	11	57
T2D	0	0	6	30	29	2	7	7	7	11	169	15
UC	0	0	17	20	696	14	38	44	41	57	15	1444

5 Multicollinearity Study

5.1 Simulation Studies

We generated our simulation data as follows. First, we assumed the true phenotype graph ($\mathbf{G}_0$) as depicted in Figure 9. The graph consists of a group of tightly linked phenotypes (P1, P2, and P3), a group of weakly linked phenotypes (P3, P4, and P5), and an isolated phenotypes (P6) as negative control. Given this graph, we set $\alpha_1 = -4.7$, $\alpha_2 = -3.0$, $\alpha_3 = -5.5$, $\alpha_4 = -4.8$, $\alpha_5 = -3.6$, and $\alpha_6 = -2.5$ with $\beta_{12} = 3.2$, $\beta_{13} = 1.8$, $\beta_{23} = 2.3$, $\beta_{34} = 2.5$, and $\beta_{45} = 5.0$ (all the remaining β_{ij} were set to zeros). Then, given the MRF coefficients, we generated association status of 20,000 common SNPs, $\boldsymbol{e}_t = (e_{1t}, e_{2t}, e_{3t}, e_{4t}, e_{5t}, e_{6t})$, from the model of Equation (2) in Section 1.1, by running the Gibbs sampler for 1,000 iterations (Mitra *et al.* 2013). Finally, given the association status of SNPs, we generated y_{it} from $\mathrm{N}(\mu_i, \sigma_i^2)$ if $e_{it} = 1$, and from $\mathrm{N}(0, 1)$ if $e_{it} = 0$, where $\boldsymbol{\mu} = (0.55, 0.5, 0.6, 0.6, 0.65, 0.55)$ and $\boldsymbol{\sigma} = (0.4, 0.3, 0.35, 0.3, 0.45, 0.4, 0.3)$.

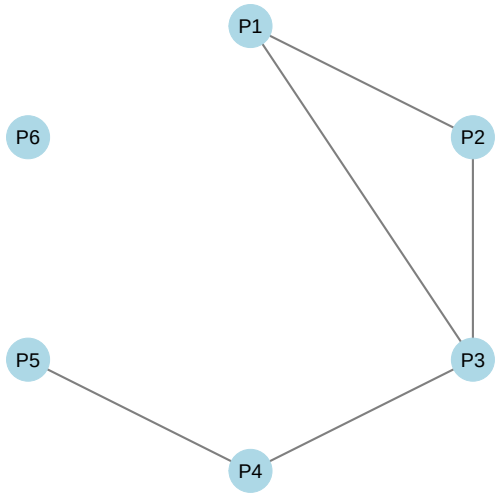

Figure 9: True phenotype graph $\boldsymbol{G}_0$ when no multicollinearity exists, which was accurately recovered by both graph-GPA models with uninformative prior and with informative prior graph using force-in edges.

We applied the graph-GPA model introduced in Section 1.1 to the simulated data with two different approaches:

1. without using an informative prior graph $\boldsymbol{G}$, i.e., $\Pr\{E(i,j) = 1\} \propto 1$ for all $i \neq j$,

2. incorporating the informative prior graph $\boldsymbol{G}$ by forcing in edges which are correlated in the true phenotype graph $\mathbf{G}_0$, i.e., $f\{E(i,j)\} = \delta_0\{E(i,j)-1\}$ for $(i,j) = (1,2), (1,3), (2,3), (3,4), (4,5)$, where $\delta_0(\cdot)$ denotes the Dirac delta function, and $\Pr\{E(i,j) = 1\} \propto 1$ for all other $i \neq j$.

In this setting, graph-GPA models with both approaches correctly identified the true phenotype graph $\boldsymbol{G}_0$ in Figure 9. This result imply that if the data provides sufficient information about the correlation structure, the graph-GPA models can identify the true genetic correlation among phenotypes regardless how a prior phenotype graph is imposed.

Now, we introduce a perfect multicollinearity in the association status by adding P7 whose association status is identical to P1, i.e., $\boldsymbol{e}_t = (e_{1t}, e_{2t}, e_{3t}, e_{4t}, e_{5t}, e_{6t}, e_{7t})$ where $e_{7t} = e_{1t}$. We again applied the graph-GPA model without and with incorporating the informative prior graph with five forced-in edges. Under this multicollinearity setting, the graph-GPA model without the informative prior graph resulted in the estimated graph in Figure 10(a) where some true edges are lost. In contrast, the graph-GPA model with force-in edges still correctly identified the true phenotype graph $\boldsymbol{G}_0$ as shown in Figure 10(b).

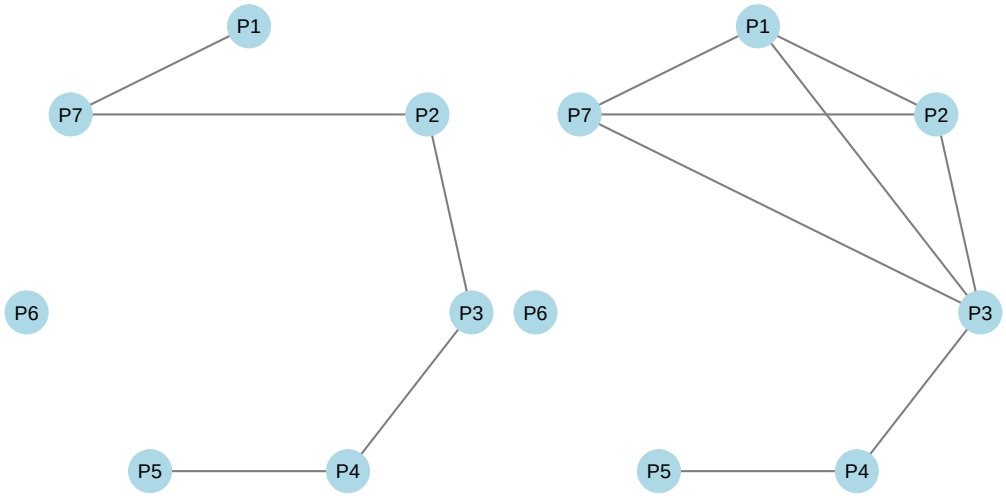

Figure 10: Simulation study when multicollinearity exists: (a) Phenotype graph identified using the graph-GPA model with uninformative prior and (b) phenotype graph identified using the graph-GPA model with informative prior graph $\boldsymbol{G}_0$.

The two simulation studies suggest that the informative prior graph helps the graph-GPA model to find the true graph structure when a strong multicollinearity exists. In practice, we might not be sure whether an informative prior graph $\boldsymbol{G}$ is similar to the unknown true graph $\boldsymbol{G}_0$ and one may be concerned about false positives or negatives due to a *mis*-informative prior graph. We marginally relieve the concern by making inference about the pleiotropic architecture not only based on the posterior probability of $E(i, j)$ but also based on that of β_{ij}. Hence, even when a wrong edge between i and j is forced in the graph-GPA model by the misinformative prior graph, data can still inform that β_{ij} is close to zero, so prevent the false positive.

5.2 Real Data Analysis

In this section, we implement empirical studies where some phenotypes are strongly correlated genetically and the multicollinearity issue is suspected to happen in the pleiotropic structure inference. For this purpose, in addition to the GWAS datasets analyzed in the main text, we further considered GWAS datasets for high-density lipoprotein (HDL), total cholesterol (TC), triglycerides (TG) (Global Lipids Consortium; `http://csg.sph.umich.edu/abecasis/public/lipids2010/`; Teslovich *et al.* 2010), fasting glucose (FG), log of fasting insulin (LFI) (Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC); `https://www.magicinvestigators.org/downloads/`; Scott *et al.* 2012), and systolic blood pressure (SBP) (International Consortium for Blood Pressure; `http://www.georgehretlab.org/icbp_088023401234-9812599.html`; International Consortium for Blood Pressure Genome-Wide Association Studies *et al.* 2011).

To study the impact of the high correlations between some phenotypes on the pleiotropic structure inference, we fitted the graph-GPA model to GWAS datasets for three different combinations of phenotypes:

- Setting 1: T2D, CAD, HDL, TC, TG, FG, LFI, SBP
- Setting 2: ADHD, ASD, BPD, MDD, SCZ, T2D, CAD, HDL, TC, TG, FG, LFI, SBP
- Setting 3: RA, CD, UC, T2D, CAD, HDL, TC, TG, FG, LFI, SBP.

In these three settings, we apply the graph-GPA models without and with incorporating the informative prior graph as used in the Example section of the main text, where we obtained the informative prior phenotype graph from the literature mining and linked edges whose partial correlation coefficients from literature mining data are larger than 0.2. Figures 11(a), 12(a), and 13(a) show the prior phenotype graphs for these three settings.

Figures 11, 12, and 13 show the graph-GPA results for Settings 1, 2, and 3, respectively. When we used the informative prior graph, the graph-GPA models found three edges linked to HDL – (HDL, TC), (HDL, TG), (HDL, CAD) – concordantly for all three settings. In contrast, when we do not use the informative prior graph, the graph-GPA model found two edges linked to HDL – (HDL,TC) and (HDL, T2D) – for Setting 1 while found only one edge linked to HDL, (HDL, TG), for Setting 2 and two edges linked to HDL – (HDL,TG) and (HDL, UC) – for Setting 3. Similarly, when we used the informative prior graph, the graph-GPA models found four edges linked to CAD – (CAD, HDL), (CAD, TC), (CAD, T2D), and (CAD, SBP) – concordantly for all three settings. When we do not use the uninformative prior graph, the graph-GPA model found two edges linked to CAD – (CAD, T2D) and (CAD, SBP) – for Setting 1, a different set of of two edges linked to CAD – (CAD, TC) and (CAD, SBP) – for Setting 2, and three edges linked to CAD – (CAD, TC), (CAD, T2D), and (CAD, SBP) – for Setting 3. These empirical studies confirm the findings of the simulation studies in Section 5.1. Specifically, multicollinearity among phenotypes may result

in incongruous pleiotropic structure inferences when an informative prior graph is not used. In contrast, the graph-GPA models with the informative prior graph lead to concordant inferences for different combinations of phenotypes included in the model.

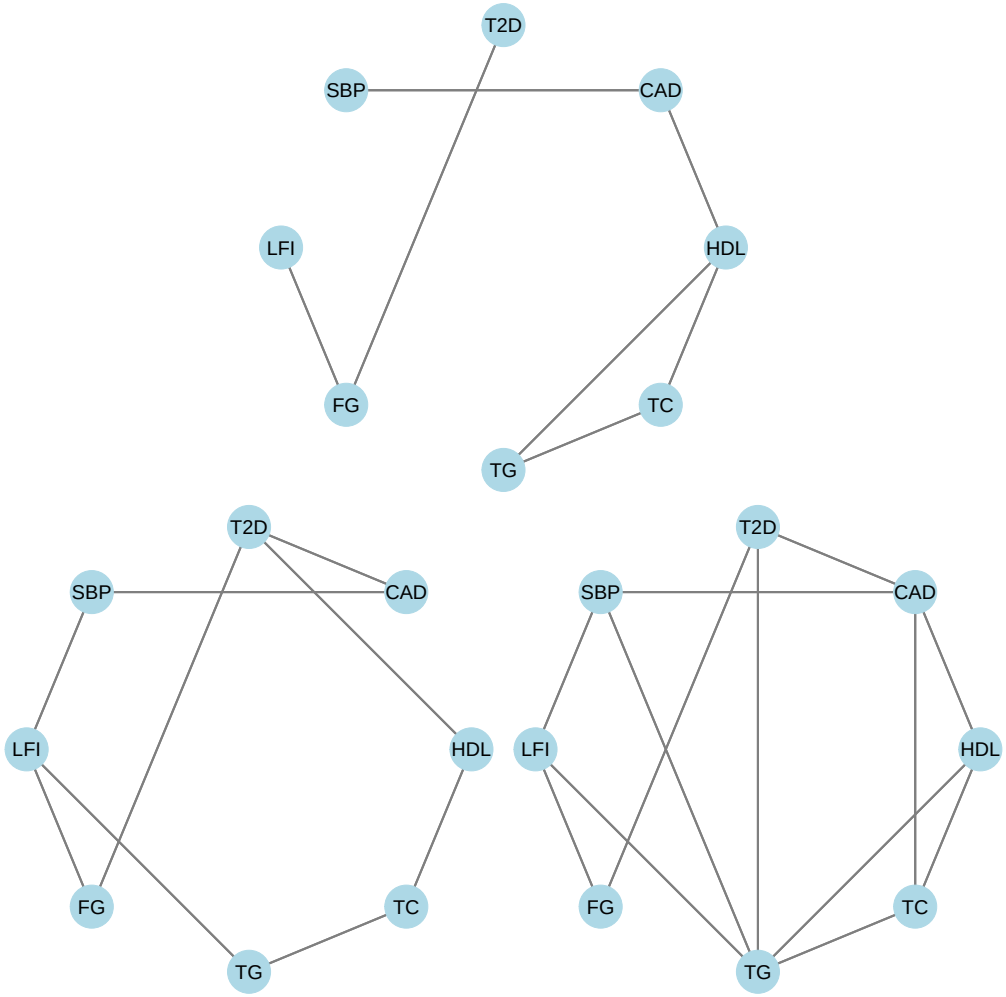

Figure 11: Setting #1: (a) A prior phenotype graph obtained the literature mining, where we link edges whose partial correlation coefficients are larger than 0.2. (b) graph-GPA results obtained with uninformative prior. (c) graph-GPA results obtained with informative prior.

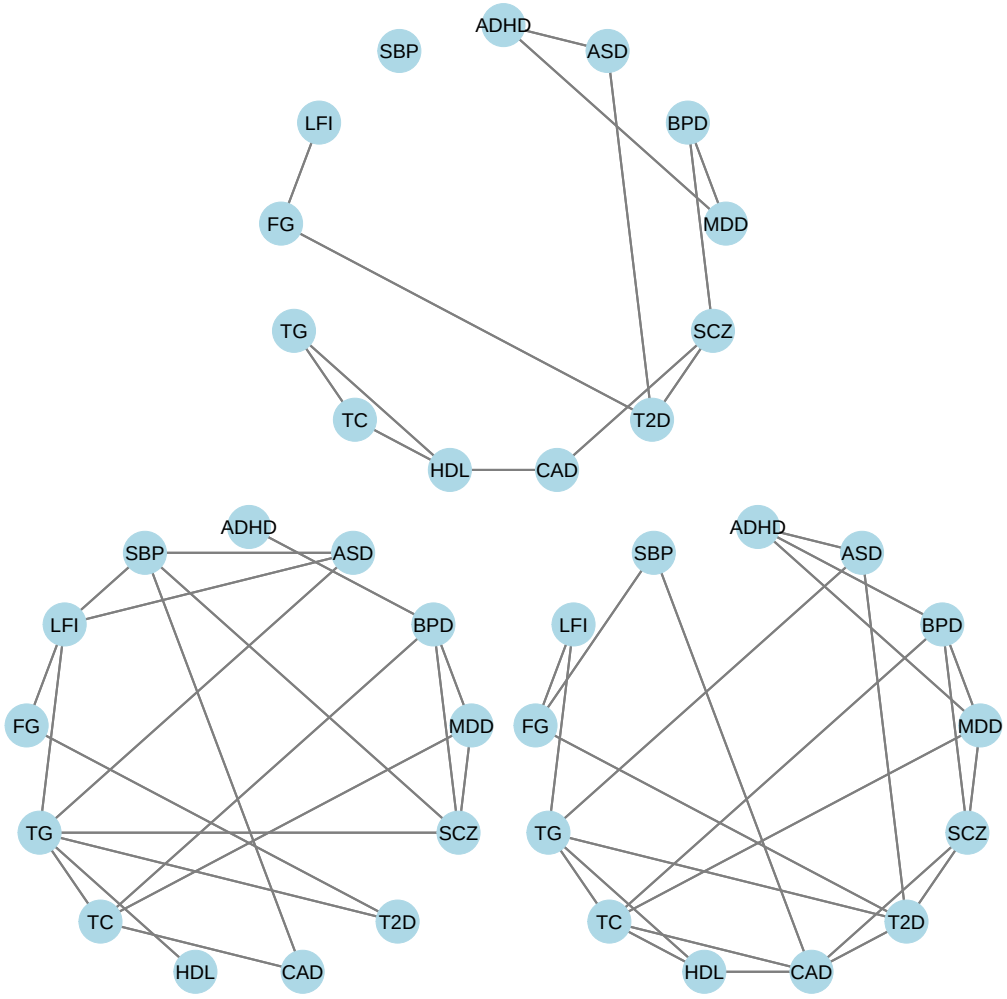

Figure 12: Setting #2: (a) A prior phenotype graph obtained the literature mining, where we link edges whose partial correlation coefficients are larger than 0.2. (b) graph-GPA results obtained with uninformative prior. (c) graph-GPA results obtained with informative prior.

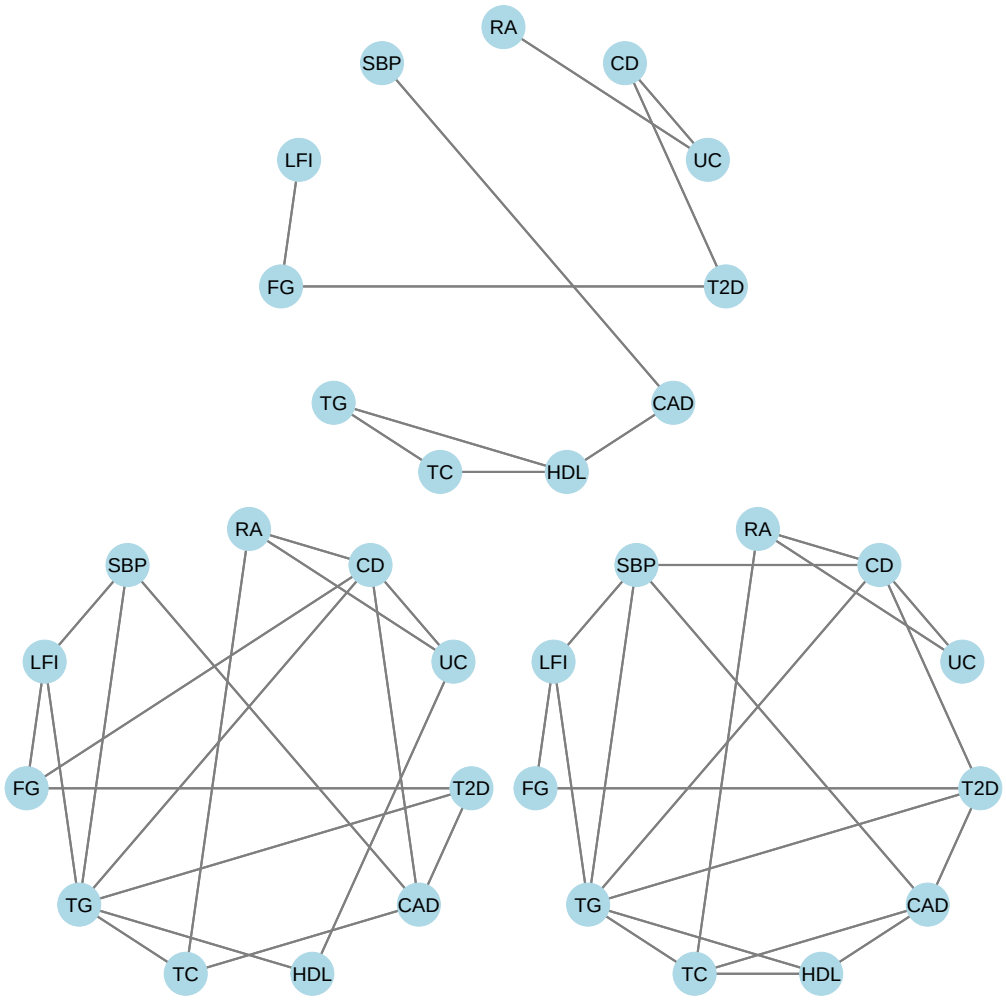

Figure 13: Setting #3: (a) A prior phenotype graph obtained the literature mining, where we link edges whose partial correlation coefficients are larger than 0.2. (b) graph-GPA results obtained with uninformative prior. (c) graph-GPA results obtained with informative prior.

References

Anderson, C. A., Boucher, G., Lees, C. W., Franke, A., D'Amato, M., Taylor, K. D., Lee, J. C., Goyette, P., Imielinski, M., Latiano, A., *et al.* (2011). Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nature Genetics*, **43**(3), 246–252.

Barrett, J. C., Clayton, D. G., Concannon, P., Akolkar, B., Cooper, J. D., Erlich, H. A., Julier, C., Morahan, G., Nerup, J., Nierras, C., *et al.* (2009). Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature Genetics*, **41**(6), 703–707.

Cross-Disorder Group of the Psychiatric Genomics Consortium *et al.* (2013a). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature Genetics*, **45**(9), 984–994.

Cross-Disorder Group of the Psychiatric Genomics Consortium *et al.* (2013b). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *The Lancet*, **381**(9875), 1371–1379.

Franke, A., McGovern, D. P., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., Lees, C. W., Balschun, T., Lee, J., Roberts, R., *et al.* (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genetics*, **42**(12), 1118–1125.

Hidalgo, C. A., Blumm, N., Barabási, A.-L., and Christakis, N. A. (2009). A dynamic network approach for the study of human phenotypes. *PLOS Computational Biology*, **5**(4), e1000353.

Hom, G., Graham, R. R., Modrek, B., Taylor, K. E., Ortmann, W., Garnier, S., Lee, A. T., Chung, S. A., Ferreira, R. C., Pant, P. K., *et al.* (2008). Association of systemic lupus erythematosus with C8orf13–BLK and ITGAM–ITGAX. *New England Journal of Medicine*, **358**(9), 900–909.

International Consortium for Blood Pressure Genome-Wide Association Studies *et al.* (2011). Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*, **478**(7367), 103–109.

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., and Barabási, A.-L. (2015). Uncovering disease-disease relationships through the incomplete interactome. *Science*, **347**(6224), 1257601.

Mitra, R., Müller, P., Liang, S., Yue, L., and Ji, Y. (2013). A Bayesian Graphical Model for ChIP-Seq Data on Histone Modifications. *Journal of the American Statistical Association*, **108**(501), 69–80.

Morris, A. P., Voight, B. F., Teslovich, T. M., Ferreira, T., Segre, A. V., Steinthorsdottir, V., Strawbridge, R. J., Khan, H., Grallert, H., Mahajan, A., *et al.* (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics*, **44**(9), 981.

Schunkert, H., König, I. R., Kathiresan, S., Reilly, M. P., Assimes, T. L., Holm, H., Preuss, M., Stewart, A. F., Barbalic, M., Gieger, C., *et al.* (2011). Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature Genetics*, **43**(4), 333–338.

Scott, R. A., Lagou, V., Welch, R. P., Wheeler, E., Montasser, M. E., Luan, J., Mägi, R., Strawbridge, R. J., Rehnberg, E., Gustafsson, S., *et al.* (2012). Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nature Genetics*, **44**(9), 991–1005.

Stahl, E. A., Raychaudhuri, S., Remmers, E. F., Xie, G., Eyre, S., Thomson, B. P., Li, Y., Kurreeman, F. A., Zhernakova, A., Hinks, A., *et al.* (2010). Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nature Genetics*, **42**(6), 508–514.

Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., Pirruccello, J. P., Ripatti, S., Chasman, D. I., Willer, C. J., *et al.* (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, **466**(7307), 707–713.

Zhou, X., Menche, J., Barabási, A.-L., and Sharma, A. (2014). Human symptoms–disease network. *Nature Communications*, **5**, 4212.