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

Hang J. Kim\textsuperscript{1}, Zhenning Yu\textsuperscript{2}, Andrew Lawson\textsuperscript{2}, Hongyu Zhao\textsuperscript{3,4,5}, and Dongjun Chung\textsuperscript{2}\* 

\textsuperscript{1} Department of Mathematical Sciences, University of Cincinnati, Cincinnati, OH, USA. 
\textsuperscript{2} Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC, USA. 
\textsuperscript{3} Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA. 
\textsuperscript{4} Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA. 
\textsuperscript{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 3
   1.1 graph-GPA Model  ................................................. 3
   1.2 Posterior Sampling  ............................................. 4
   1.3 Software Implementation  ..................................... 6

2 Strategies to Construct a Prior Disease Graph 11

3 Real Data Analysis 13
   3.1 GWAS Datasets Used in the Real Data Analysis  ............. 13
   3.2 Prior Disease Graphs Obtained from the Literature Mining  13
   3.3 graph-GPA Results Without Incorporating the Prior Disease Graph  27
   3.4 graph-GPA Results Incorporating the Prior Disease Graph  .... 29
   3.5 Functional Impact Analysis of Identified SNPs  ............... 31

4 Reproducibility Analysis 33
   4.1 graph-GPA Results Without Incorporating the Prior Disease Graph  34
   4.2 graph-GPA Results Incorporating the Prior Disease Graph  .... 36

5 Multicollinearity Study 38
   5.1 Simulation Studies  ............................................. 38
   5.2 Real Data Analysis  ........................................... 40
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 $\Phi$ 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:

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

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 $G = (V, E)$ denote an MRF graph with nodes $V = (v_1, \ldots, v_n)$ and edges $E = \{E(i, j) : i, j = 1, \ldots, n\}$. 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 $t$, $e_t = (e_{1t}, \ldots, e_{nt})$, and the graph structure with an auto-logistic spatial scheme:

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

and

$$
C(\alpha, \beta, G)^{-1} = \sum_{e^* \in E^*} \exp \left( \sum_{i=1}^{n} \alpha_i e^*_i + \sum_{i \sim j} \beta_{ij} e^*_i e^*_j \right),
$$

where $\beta_{ij}$ 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 $E^*$ is the set of all possible values of $e^* = (e^*_1, \ldots, e^*_n)$.

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

$$
\mu_i \sim \text{N}(\theta_{\mu}, \tau_{\mu}^2), \quad \sigma_i^2 \sim \text{IG}(a_\sigma, b_\sigma),
$$

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

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

where $\Gamma(a, b)$ denotes the gamma distribution with mean $a/b$ and $\delta_0$ denotes Dirac delta function. For the MRF graph $G$, 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 $\beta_{ij}$’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 $\alpha_i$ with the Metropolis-Hastings step:

1. Draw $\alpha_i^q$ 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

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

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

S5. For each $(i, j)$ such that $E(i,j) = 1$, update $\beta_{ij}$ with the Metropolis-Hastings:
1. Draw $\beta_{ij}^q$ from $N_+(\beta_{ij}, s^2_\beta)$ 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

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

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

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

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

$$
q(z^q | z) = 0.5 I [z^q = z - 1] + 0.5 I [z^q = z + 1].
$$

If $z = z_{\text{force}}$, set $z^q = z + 1$ with probability 1. If $z = z_{\text{max}}$, set $z^q = z_{\text{max}} - 1$ with probability 1 where $z_{\text{max}}$ denotes the maximum number of possible edges, i.e., $z_{\text{max}} = \binom{n}{2}$.

2. Propose $G^q$ from the proposal distribution $q(G^q|G, z^q)$ and then $\beta_{ij}^q$ from the proposal distribution $q(\beta_{ij}^q|G^q, z^q)$.

(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

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

while $G^q_{i^*, j^*} = G_{i^*, j^*}$ for all other $(i^*, j^*)$. Propose $\beta_{ij}^q$ from $q(\beta_{ij}^q|E(i,j)^q, z^q) = \Gamma(\beta_{ij}; a_\beta, b_\beta)$. We set $a_\beta = b_\beta = 1$.

(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

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

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

5
3. Update \((\beta_{ij}, G) = (\beta_{ij}^q, G^q)\) with the acceptance probability

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

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

Note that \(p(\beta_{ij} | E(i, j)) = q(\beta_{ij} | G, z)\) when \(z^q > z\) and \(p(\beta_{ij}^q | E(i, j)^q) = q(\beta_{ij}^q | G^q, z^q)\) 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 depicted 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 \url{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 adjacency 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 \texttt{GGPA()}, \texttt{assoc()}, and \texttt{plot()}. Specifically, \texttt{GGPA()} first fits the graph-GPA model with or without the prior knowledge. Then, \texttt{assoc()} and \texttt{plot()} implement the association mapping and the phenotype graph estimation, respectively. In order to boost the computational efficiency, the core of the function \texttt{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 \texttt{pmat} and \texttt{pgraph}, respectively. It is assumed that rows and columns of the object \texttt{pmat} correspond to SNPs and phenotypes while the objects \texttt{pgraph} and \texttt{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.

\begin{verbatim}
R> fit <- GGPA( pmat, pgraph )
\end{verbatim}
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.

\[
\text{R} \ (> \ \text{plot}( \ \text{fit} \ ))
\]

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.

\[
\text{R} \ (> \ \text{assoc}( \ \text{fit}, \ \text{FDR}=0.20, \ \text{fdrControl}="\text{local}" \ ))
\]

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.

\[
\text{R} \ (> \ \text{assoc}( \ \text{fit}, \ \text{FDR}=0.20, \ \text{fdrControl}="\text{local}" , \ \text{i}=5, \ \text{j}=6 \ ))
\]
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 Efetech 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 $\phi$-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.

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

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

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Source</th>
<th>pValue</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Attention deficit hyperactivity disorder</td>
<td>DisGeNET Curated</td>
<td>7.60E-51</td>
<td>1.92E-47</td>
</tr>
<tr>
<td>2</td>
<td>Impulsive character (finding)</td>
<td>DisGeNET BeFree</td>
<td>4.32E-30</td>
<td>5.45E-27</td>
</tr>
<tr>
<td>3</td>
<td>Major Depressive Disorder</td>
<td>DisGeNET Curated</td>
<td>5.31E-29</td>
<td>4.47E-26</td>
</tr>
<tr>
<td>4</td>
<td>Nicotine Dependence</td>
<td>DisGeNET BeFree</td>
<td>6.29E-28</td>
<td>3.97E-25</td>
</tr>
<tr>
<td>5</td>
<td>Unipolar Depression</td>
<td>DisGeNET Curated</td>
<td>9.33E-28</td>
<td>4.71E-25</td>
</tr>
</tbody>
</table>
Table 3: Top 5 pathways enriched for the top 100 genes that are predicted to be associated with ASD in the literature mining.

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

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

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Source</th>
<th>pValue</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Autism Spectrum Disorders</td>
<td>DisGeNET Curated</td>
<td>1.44E-66</td>
<td>4.23E-63</td>
</tr>
<tr>
<td>2</td>
<td>Autistic Disorder</td>
<td>DisGeNET Curated</td>
<td>8.34E-55</td>
<td>1.22E-51</td>
</tr>
<tr>
<td>3</td>
<td>Pervasive Development Disorder</td>
<td>DisGeNET BeFree</td>
<td>6.02E-44</td>
<td>5.89E-41</td>
</tr>
<tr>
<td>4</td>
<td>Neurodevelopmental Disorders</td>
<td>DisGeNET Curated</td>
<td>1.54E-24</td>
<td>1.13E-21</td>
</tr>
<tr>
<td>5</td>
<td>Schizophrenia</td>
<td>DisGeNET Curated</td>
<td>8.11E-22</td>
<td>4.76E-19</td>
</tr>
</tbody>
</table>
Table 5: Top 5 pathways enriched for the top 100 genes that are predicted to be associated with BPD in the literature mining.

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

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

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Source</th>
<th>pValue</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bipolar Disorder</td>
<td>DisGeNET Curated</td>
<td>1.53E-73</td>
<td>3.49E-70</td>
</tr>
<tr>
<td>2</td>
<td>Mood Disorders</td>
<td>DisGeNET Curated</td>
<td>6.50E-52</td>
<td>6.53E-49</td>
</tr>
<tr>
<td>3</td>
<td>Nonorganic psychosis</td>
<td>DisGeNET Curated</td>
<td>8.58E-52</td>
<td>6.53E-49</td>
</tr>
<tr>
<td>4</td>
<td>Psychotic Disorders</td>
<td>DisGeNET Curated</td>
<td>2.47E-50</td>
<td>1.41E-47</td>
</tr>
<tr>
<td>5</td>
<td>Mental disorders</td>
<td>DisGeNET Curated</td>
<td>5.05E-48</td>
<td>2.31E-45</td>
</tr>
</tbody>
</table>
Table 7: Top 5 pathways enriched for the top 100 genes that are predicted to be associated with MDD in the literature mining.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Source</th>
<th>pValue</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Serotonergic synapse</td>
<td>BioSystems: KEGG</td>
<td>9.54E-16</td>
<td>6.15E-13</td>
</tr>
<tr>
<td>3</td>
<td>Neuroactive ligand-receptor interaction</td>
<td>BioSystems: KEGG</td>
<td>2.61E-14</td>
<td>6.56E-12</td>
</tr>
<tr>
<td>4</td>
<td>Amine ligand-binding receptors</td>
<td>BioSystems: REACTOME</td>
<td>2.04E-13</td>
<td>3.84E-11</td>
</tr>
<tr>
<td>5</td>
<td>Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha mediated pathway</td>
<td>PantherDB</td>
<td>1.34E-12</td>
<td>2.02E-10</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Source</th>
<th>pValue</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Major Depressive Disorder</td>
<td>DisGeNET Curated</td>
<td>1.09E-54</td>
<td>2.89E-51</td>
</tr>
<tr>
<td>2</td>
<td>Unipolar Depression</td>
<td>DisGeNET Curated</td>
<td>7.93E-51</td>
<td>1.06E-47</td>
</tr>
<tr>
<td>3</td>
<td>Mental Depression</td>
<td>DisGeNET Curated</td>
<td>7.71E-48</td>
<td>6.86E-45</td>
</tr>
<tr>
<td>4</td>
<td>Depressive disorder</td>
<td>DisGeNET Curated</td>
<td>3.75E-47</td>
<td>2.50E-44</td>
</tr>
<tr>
<td>5</td>
<td>Mental disorders</td>
<td>DisGeNET Curated</td>
<td>6.05E-40</td>
<td>3.23E-37</td>
</tr>
</tbody>
</table>
Table 9: Top 5 pathways enriched for the top 100 genes that are predicted to be associated with SCZ in the literature mining.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Source</th>
<th>pValue</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Neuroactive ligand-receptor interaction</td>
<td>BioSystems: KEGG</td>
<td>1.30E-18</td>
<td>1.08E-15</td>
</tr>
<tr>
<td>3</td>
<td>Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha mediated pathway</td>
<td>PantherDB</td>
<td>2.24E-14</td>
<td>4.65E-12</td>
</tr>
<tr>
<td>4</td>
<td>Transmission across Chemical Synapses</td>
<td>BioSystems: REACTOME</td>
<td>2.24E-14</td>
<td>4.65E-12</td>
</tr>
<tr>
<td>5</td>
<td>Amine ligand-binding receptors</td>
<td>BioSystems: REACTOME</td>
<td>9.52E-14</td>
<td>1.58E-11</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Source</th>
<th>pValue</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Psychotic Disorders</td>
<td>DisGeNET Curated</td>
<td>3.04E-64</td>
<td>6.87E-61</td>
</tr>
<tr>
<td>2</td>
<td>Schizophrenia</td>
<td>DisGeNET Curated</td>
<td>4.99E-57</td>
<td>5.65E-54</td>
</tr>
<tr>
<td>3</td>
<td>Nonorganic psychosis</td>
<td>DisGeNET Curated</td>
<td>2.21E-53</td>
<td>1.67E-50</td>
</tr>
<tr>
<td>4</td>
<td>Bipolar Disorder</td>
<td>DisGeNET Curated</td>
<td>1.16E-51</td>
<td>6.55E-49</td>
</tr>
<tr>
<td>5</td>
<td>Mental disorders</td>
<td>DisGeNET Curated</td>
<td>1.92E-49</td>
<td>8.69E-47</td>
</tr>
</tbody>
</table>
Table 11: Top 5 pathways enriched for the top 100 genes that are predicted to be associated with RA in the literature mining.

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

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

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Source</th>
<th>pValue</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Arthritis</td>
<td>DisGeNET Curated</td>
<td>4.97E-84</td>
<td>2.18E-80</td>
</tr>
<tr>
<td>2</td>
<td>Rheumatoid Arthritis</td>
<td>DisGeNET Curated</td>
<td>1.74E-76</td>
<td>3.80E-73</td>
</tr>
<tr>
<td>3</td>
<td>Superficial ulcer</td>
<td>DisGeNET BeFree</td>
<td>6.69E-53</td>
<td>9.78E-50</td>
</tr>
<tr>
<td>4</td>
<td>Autoimmune Diseases</td>
<td>DisGeNET Curated</td>
<td>5.77E-51</td>
<td>6.33E-48</td>
</tr>
<tr>
<td>5</td>
<td>Diabetes Mellitus, Insulin-Dependent</td>
<td>DisGeNET Curated</td>
<td>1.42E-50</td>
<td>1.24E-47</td>
</tr>
</tbody>
</table>
Table 13: Top 5 pathways enriched for the top 100 genes that are predicted to be associated with CD in the literature mining.

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

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

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Source</th>
<th>pValue</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Crohn Disease</td>
<td>DisGeNET Curated</td>
<td>2.03E-44</td>
<td>5.43E-41</td>
</tr>
<tr>
<td>2</td>
<td>Ulcerative Colitis</td>
<td>DisGeNET Curated</td>
<td>4.68E-39</td>
<td>6.26E-36</td>
</tr>
<tr>
<td>3</td>
<td>Inflammatory Bowel Diseases</td>
<td>DisGeNET Curated</td>
<td>4.19E-29</td>
<td>3.73E-26</td>
</tr>
<tr>
<td>4</td>
<td>Colitis</td>
<td>DisGeNET Curated</td>
<td>2.13E-22</td>
<td>1.42E-19</td>
</tr>
<tr>
<td>5</td>
<td>Crohn’s disease</td>
<td>GWAS</td>
<td>1.44E-17</td>
<td>7.70E-15</td>
</tr>
</tbody>
</table>
Table 15: Top 5 pathways enriched for the top 100 genes that are predicted to be associated with UC in the literature mining.

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

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

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Source</th>
<th>pValue</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ulcerative Colitis</td>
<td>DisGeNET Curated</td>
<td>6.22E-48</td>
<td>1.96E-44</td>
</tr>
<tr>
<td>2</td>
<td>Crohn Disease</td>
<td>DisGeNET Curated</td>
<td>7.14E-42</td>
<td>1.13E-38</td>
</tr>
<tr>
<td>3</td>
<td>Inflammatory Bowel Diseases</td>
<td>DisGeNET Curated</td>
<td>4.26E-33</td>
<td>4.49E-30</td>
</tr>
<tr>
<td>4</td>
<td>Colitis</td>
<td>DisGeNET Curated</td>
<td>5.68E-30</td>
<td>4.48E-27</td>
</tr>
<tr>
<td>5</td>
<td>Irritable Bowel Syndrome</td>
<td>DisGeNET Curated</td>
<td>1.65E-18</td>
<td>1.04E-15</td>
</tr>
</tbody>
</table>
Table 17: Top 5 pathways enriched for the top 100 genes that are predicted to be associated with SLE in the literature mining.

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

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

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Source</th>
<th>pValue</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Lupus Erythematosus, Systemic</td>
<td>DisGeNET Curated</td>
<td>3.33E-83</td>
<td>1.01E-79</td>
</tr>
<tr>
<td>2</td>
<td>Lupus Vulgaris</td>
<td>DisGeNET BeFree</td>
<td>4.14E-74</td>
<td>6.25E-71</td>
</tr>
<tr>
<td>3</td>
<td>Lupus Erythematosus</td>
<td>DisGeNET BeFree</td>
<td>1.80E-73</td>
<td>1.81E-70</td>
</tr>
<tr>
<td>4</td>
<td>Lupus Erythematosus, Discoid</td>
<td>DisGeNET Curated</td>
<td>7.51E-73</td>
<td>5.66E-70</td>
</tr>
<tr>
<td>5</td>
<td>Autoimmune Diseases</td>
<td>DisGeNET Curated</td>
<td>8.17E-52</td>
<td>4.93E-49</td>
</tr>
</tbody>
</table>
Table 19: Top 5 pathways enriched for the top 100 genes that are predicted to be associated with T1D in the literature mining.

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

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

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Source</th>
<th>pValue</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Diabetes Mellitus, Insulin-Dependent</td>
<td>DisGeNET Curated</td>
<td>2.04E-20</td>
<td>4.43E-17</td>
</tr>
<tr>
<td>2</td>
<td>Graves Disease</td>
<td>DisGeNET Curated</td>
<td>2.00E-10</td>
<td>2.17E-07</td>
</tr>
<tr>
<td>3</td>
<td>Rheumatoid Arthritis</td>
<td>DisGeNET Curated</td>
<td>4.44E-10</td>
<td>3.22E-07</td>
</tr>
<tr>
<td>4</td>
<td>Diabetes Mellitus</td>
<td>DisGeNET Curated</td>
<td>9.16E-10</td>
<td>4.97E-07</td>
</tr>
<tr>
<td>5</td>
<td>Diabetes Mellitus, Non-Insulin-Dependent</td>
<td>DisGeNET Curated</td>
<td>3.22E-09</td>
<td>1.39E-06</td>
</tr>
</tbody>
</table>
Table 21: Top 5 pathways enriched for the top 100 genes that are predicted to be associated with T2D in the literature mining.

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

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

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Source</th>
<th>pValue</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Diabetes Mellitus, Non-Insulin-Dependent</td>
<td>DisGeNET Curated</td>
<td>2.29E-33</td>
<td>5.20E-30</td>
</tr>
<tr>
<td>2</td>
<td>Impaired glucose tolerance</td>
<td>DisGeNET Curated</td>
<td>1.99E-27</td>
<td>2.25E-24</td>
</tr>
<tr>
<td>3</td>
<td>Diabetes</td>
<td>DisGeNET BeFree</td>
<td>1.23E-23</td>
<td>9.26E-21</td>
</tr>
<tr>
<td>4</td>
<td>Diabetes Mellitus</td>
<td>DisGeNET Curated</td>
<td>2.58E-23</td>
<td>1.46E-20</td>
</tr>
<tr>
<td>5</td>
<td>Hyperglycemia</td>
<td>DisGeNET Curated</td>
<td>5.71E-19</td>
<td>2.59E-16</td>
</tr>
</tbody>
</table>
Table 23: Top 5 pathways enriched for the top 100 genes that are predicted to be associated with CAD in the literature mining.

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

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

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Source</th>
<th>pValue</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Coronary heart disease</td>
<td>DisGeNET Curated</td>
<td>1.64E-40</td>
<td>5.43E-37</td>
</tr>
<tr>
<td>2</td>
<td>Cardiovascular Diseases</td>
<td>DisGeNET Curated</td>
<td>5.50E-40</td>
<td>9.10E-37</td>
</tr>
<tr>
<td>3</td>
<td>Cerebrovascular accident</td>
<td>DisGeNET Curated</td>
<td>3.41E-39</td>
<td>3.77E-36</td>
</tr>
<tr>
<td>4</td>
<td>Coronary Arteriosclerosis</td>
<td>DisGeNET Curated</td>
<td>2.79E-35</td>
<td>2.31E-32</td>
</tr>
<tr>
<td>5</td>
<td>Coronary Artery Disease</td>
<td>DisGeNET Curated</td>
<td>1.90E-33</td>
<td>1.26E-30</td>
</tr>
</tbody>
</table>
### 3.3 graph-GPA Results Without Incorporating the Prior Disease Graph

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

<table>
<thead>
<tr>
<th></th>
<th>ADHD</th>
<th>ASD</th>
<th>BPD</th>
<th>CAD</th>
<th>CD</th>
<th>MDD</th>
<th>RA</th>
<th>SCZ</th>
<th>SLE</th>
<th>T1D</th>
<th>T2D</th>
<th>UC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADHD</td>
<td>–</td>
<td>0.11</td>
<td>0.92</td>
<td>0.03</td>
<td>0.65</td>
<td>0.52</td>
<td>0.05</td>
<td>0.25</td>
<td>0.19</td>
<td>0.11</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>ASD</td>
<td>0.11</td>
<td>–</td>
<td>0.42</td>
<td>0.03</td>
<td>1.00</td>
<td>0.26</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BPD</td>
<td>0.92</td>
<td>–</td>
<td></td>
<td>0.83</td>
<td>0.65</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAD</td>
<td>0.03</td>
<td>0.42</td>
<td>–</td>
<td>1.00</td>
<td>0.04</td>
<td>1.00</td>
<td>0.05</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD</td>
<td>0.65</td>
<td>0.03</td>
<td>0.83</td>
<td>1.00</td>
<td>–</td>
<td>0.64</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MDD</td>
<td>0.52</td>
<td>0.05</td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RA</td>
<td>0.05</td>
<td>1.00</td>
<td>0.04</td>
<td>–</td>
<td>–</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCZ</td>
<td>0.26</td>
<td>1.00</td>
<td>1.00</td>
<td>0.64</td>
<td>1.00</td>
<td>–</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLE</td>
<td>0.25</td>
<td>1.00</td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1D</td>
<td>0.19</td>
<td>0.05</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.10</td>
<td>0.01</td>
<td>0.78</td>
<td>0.26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2D</td>
<td>0.11</td>
<td></td>
<td>1.00</td>
<td>1.00</td>
<td>0.01</td>
<td>0.78</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UC</td>
<td>0.09</td>
<td></td>
<td></td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

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

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

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.

<table>
<thead>
<tr>
<th></th>
<th>ADHD</th>
<th>ASD</th>
<th>BPD</th>
<th>CAD</th>
<th>CD</th>
<th>MDD</th>
<th>RA</th>
<th>SCZ</th>
<th>SLE</th>
<th>T1D</th>
<th>T2D</th>
<th>UC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADHD</td>
<td>27</td>
<td>0</td>
<td>25</td>
<td>0</td>
<td>22</td>
<td>15</td>
<td>6</td>
<td>26</td>
<td>25</td>
<td>25</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ASD</td>
<td>0</td>
<td>192</td>
<td>7</td>
<td>8</td>
<td>61</td>
<td>0</td>
<td>114</td>
<td>31</td>
<td>36</td>
<td>99</td>
<td>8</td>
<td>51</td>
</tr>
<tr>
<td>BPD</td>
<td>25</td>
<td>7</td>
<td>1488</td>
<td>38</td>
<td>304</td>
<td>214</td>
<td>183</td>
<td>592</td>
<td>258</td>
<td>277</td>
<td>29</td>
<td>169</td>
</tr>
<tr>
<td>CAD</td>
<td>0</td>
<td>8</td>
<td>1830</td>
<td>350</td>
<td>26</td>
<td>114</td>
<td>180</td>
<td>58</td>
<td>123</td>
<td>413</td>
<td>209</td>
<td></td>
</tr>
<tr>
<td>CD</td>
<td>22</td>
<td>61</td>
<td>304</td>
<td>350</td>
<td>18209</td>
<td>222</td>
<td>993</td>
<td>768</td>
<td>493</td>
<td>1194</td>
<td>428</td>
<td>7443</td>
</tr>
<tr>
<td>MDD</td>
<td>15</td>
<td>0</td>
<td>214</td>
<td>26</td>
<td>222</td>
<td>543</td>
<td>217</td>
<td>384</td>
<td>302</td>
<td>348</td>
<td>28</td>
<td>130</td>
</tr>
<tr>
<td>RA</td>
<td>6</td>
<td>114</td>
<td>183</td>
<td>114</td>
<td>993</td>
<td>217</td>
<td>1892</td>
<td>532</td>
<td>567</td>
<td>1527</td>
<td>204</td>
<td>955</td>
</tr>
<tr>
<td>SCZ</td>
<td>26</td>
<td>31</td>
<td>592</td>
<td>180</td>
<td>768</td>
<td>384</td>
<td>532</td>
<td>5669</td>
<td>646</td>
<td>786</td>
<td>98</td>
<td>510</td>
</tr>
<tr>
<td>SLE</td>
<td>25</td>
<td>36</td>
<td>258</td>
<td>58</td>
<td>493</td>
<td>302</td>
<td>567</td>
<td>646</td>
<td>1129</td>
<td>843</td>
<td>64</td>
<td>341</td>
</tr>
<tr>
<td>T1D</td>
<td>25</td>
<td>99</td>
<td>277</td>
<td>123</td>
<td>1194</td>
<td>348</td>
<td>1527</td>
<td>786</td>
<td>843</td>
<td>2170</td>
<td>201</td>
<td>947</td>
</tr>
<tr>
<td>T2D</td>
<td>0</td>
<td>8</td>
<td>29</td>
<td>413</td>
<td>428</td>
<td>28</td>
<td>204</td>
<td>98</td>
<td>64</td>
<td>201</td>
<td>2149</td>
<td>243</td>
</tr>
<tr>
<td>UC</td>
<td>1</td>
<td>51</td>
<td>169</td>
<td>209</td>
<td>7443</td>
<td>130</td>
<td>955</td>
<td>510</td>
<td>341</td>
<td>947</td>
<td>243</td>
<td>15631</td>
</tr>
</tbody>
</table>
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 $p(E(i,j)|Y)$. The blanked cell indicates the zero estimated value.

<table>
<thead>
<tr>
<th></th>
<th>ADHD</th>
<th>ASD</th>
<th>BPD</th>
<th>CAD</th>
<th>CD</th>
<th>MDD</th>
<th>RA</th>
<th>SCZ</th>
<th>SLE</th>
<th>T1D</th>
<th>T2D</th>
<th>UC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADHD</td>
<td>1.00</td>
<td>0.09</td>
<td>0.82</td>
<td>0.22</td>
<td>1.00</td>
<td>0.06</td>
<td>0.03</td>
<td>0.07</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASD</td>
<td>0.09</td>
<td>0.82</td>
<td>0.22</td>
<td>1.00</td>
<td>0.06</td>
<td>0.03</td>
<td>0.07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BPD</td>
<td>0.10</td>
<td>0.05</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.85</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAD</td>
<td>0.18</td>
<td>0.22</td>
<td>1.00</td>
<td>0.01</td>
<td>0.38</td>
<td>1.00</td>
<td>0.44</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD</td>
<td>0.18</td>
<td>0.22</td>
<td>1.00</td>
<td>0.01</td>
<td>0.38</td>
<td>1.00</td>
<td>0.44</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MDD</td>
<td>1.00</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RA</td>
<td>0.04</td>
<td>0.03</td>
<td>0.01</td>
<td>0.01</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCZ</td>
<td>0.42</td>
<td>0.85</td>
<td>0.01</td>
<td>0.01</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLE</td>
<td>0.18</td>
<td>0.03</td>
<td>0.02</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1D</td>
<td>0.01</td>
<td>0.07</td>
<td>0.44</td>
<td>0.01</td>
<td>0.44</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2D</td>
<td>0.01</td>
<td>0.07</td>
<td>0.44</td>
<td>0.01</td>
<td>0.44</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UC</td>
<td>0.01</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 $\beta_{ij}$. The blanked cell indicates that $p(E(i,j)|Y)$ is estimated as zero and the bold number indicates that the 95% credible interval $\beta_{ij}$ does not contain zero.

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

<table>
<thead>
<tr>
<th></th>
<th>ADHD</th>
<th>ASD</th>
<th>BPD</th>
<th>CAD</th>
<th>CD</th>
<th>MDD</th>
<th>RA</th>
<th>SCZ</th>
<th>SLE</th>
<th>T1D</th>
<th>T2D</th>
<th>UC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADHD</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ASD</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>BPD</td>
<td>0</td>
<td>0</td>
<td>120</td>
<td>3</td>
<td>46</td>
<td>36</td>
<td>28</td>
<td>81</td>
<td>71</td>
<td>72</td>
<td>2</td>
<td>17</td>
</tr>
<tr>
<td>CAD</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>183</td>
<td>20</td>
<td>2</td>
<td>16</td>
<td>17</td>
<td>8</td>
<td>20</td>
<td>33</td>
<td>24</td>
</tr>
<tr>
<td>CD</td>
<td>0</td>
<td>0</td>
<td>46</td>
<td>20</td>
<td>1941</td>
<td>37</td>
<td>282</td>
<td>92</td>
<td>121</td>
<td>346</td>
<td>26</td>
<td>756</td>
</tr>
<tr>
<td>MDD</td>
<td>0</td>
<td>0</td>
<td>36</td>
<td>2</td>
<td>37</td>
<td>82</td>
<td>28</td>
<td>71</td>
<td>70</td>
<td>72</td>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>RA</td>
<td>0</td>
<td>4</td>
<td>28</td>
<td>16</td>
<td>282</td>
<td>28</td>
<td>649</td>
<td>132</td>
<td>155</td>
<td>533</td>
<td>25</td>
<td>275</td>
</tr>
<tr>
<td>SCZ</td>
<td>0</td>
<td>0</td>
<td>81</td>
<td>17</td>
<td>92</td>
<td>71</td>
<td>132</td>
<td>450</td>
<td>176</td>
<td>233</td>
<td>13</td>
<td>70</td>
</tr>
<tr>
<td>SLE</td>
<td>0</td>
<td>7</td>
<td>8</td>
<td>121</td>
<td>70</td>
<td>155</td>
<td>176</td>
<td>262</td>
<td>247</td>
<td>12</td>
<td>78</td>
<td></td>
</tr>
<tr>
<td>T1D</td>
<td>0</td>
<td>4</td>
<td>72</td>
<td>20</td>
<td>346</td>
<td>72</td>
<td>533</td>
<td>233</td>
<td>247</td>
<td>770</td>
<td>28</td>
<td>283</td>
</tr>
<tr>
<td>T2D</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>33</td>
<td>26</td>
<td>2</td>
<td>25</td>
<td>13</td>
<td>12</td>
<td>28</td>
<td>180</td>
<td>18</td>
</tr>
<tr>
<td>UC</td>
<td>0</td>
<td>0</td>
<td>17</td>
<td>24</td>
<td>756</td>
<td>14</td>
<td>275</td>
<td>70</td>
<td>78</td>
<td>283</td>
<td>18</td>
<td>1553</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>ADHD</th>
<th>ASD</th>
<th>BPD</th>
<th>CAD</th>
<th>CD</th>
<th>MDD</th>
<th>RA</th>
<th>SCZ</th>
<th>SLE</th>
<th>T1D</th>
<th>T2D</th>
<th>UC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADHD</td>
<td>25</td>
<td>0</td>
<td>23</td>
<td>0</td>
<td>19</td>
<td>20</td>
<td>6</td>
<td>24</td>
<td>23</td>
<td>24</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ASD</td>
<td>0</td>
<td>195</td>
<td>7</td>
<td>20</td>
<td>64</td>
<td>0</td>
<td>114</td>
<td>30</td>
<td>38</td>
<td>100</td>
<td>12</td>
<td>59</td>
</tr>
<tr>
<td>BPD</td>
<td>23</td>
<td>7</td>
<td>1533</td>
<td>39</td>
<td>321</td>
<td>254</td>
<td>189</td>
<td>615</td>
<td>253</td>
<td>279</td>
<td>29</td>
<td>184</td>
</tr>
<tr>
<td>CAD</td>
<td>0</td>
<td>20</td>
<td>39</td>
<td>1845</td>
<td>258</td>
<td>28</td>
<td>125</td>
<td>181</td>
<td>62</td>
<td>141</td>
<td>424</td>
<td>329</td>
</tr>
<tr>
<td>CD</td>
<td>19</td>
<td>64</td>
<td>321</td>
<td>258</td>
<td>18054</td>
<td>226</td>
<td>987</td>
<td>711</td>
<td>496</td>
<td>1217</td>
<td>294</td>
<td>7392</td>
</tr>
<tr>
<td>MDD</td>
<td>20</td>
<td>0</td>
<td>254</td>
<td>28</td>
<td>226</td>
<td>571</td>
<td>222</td>
<td>404</td>
<td>302</td>
<td>348</td>
<td>29</td>
<td>134</td>
</tr>
<tr>
<td>RA</td>
<td>6</td>
<td>114</td>
<td>189</td>
<td>125</td>
<td>987</td>
<td>222</td>
<td>1893</td>
<td>537</td>
<td>588</td>
<td>1533</td>
<td>192</td>
<td>945</td>
</tr>
<tr>
<td>SCZ</td>
<td>24</td>
<td>30</td>
<td>615</td>
<td>181</td>
<td>711</td>
<td>404</td>
<td>537</td>
<td>5644</td>
<td>645</td>
<td>791</td>
<td>100</td>
<td>510</td>
</tr>
<tr>
<td>SLE</td>
<td>23</td>
<td>38</td>
<td>253</td>
<td>62</td>
<td>496</td>
<td>302</td>
<td>588</td>
<td>645</td>
<td>1141</td>
<td>845</td>
<td>69</td>
<td>348</td>
</tr>
<tr>
<td>T1D</td>
<td>24</td>
<td>100</td>
<td>279</td>
<td>141</td>
<td>1217</td>
<td>348</td>
<td>1533</td>
<td>791</td>
<td>845</td>
<td>2207</td>
<td>234</td>
<td>958</td>
</tr>
<tr>
<td>T2D</td>
<td>0</td>
<td>12</td>
<td>29</td>
<td>424</td>
<td>294</td>
<td>29</td>
<td>192</td>
<td>100</td>
<td>69</td>
<td>234</td>
<td>2111</td>
<td>246</td>
</tr>
<tr>
<td>UC</td>
<td>1</td>
<td>59</td>
<td>184</td>
<td>329</td>
<td>7392</td>
<td>134</td>
<td>945</td>
<td>510</td>
<td>348</td>
<td>958</td>
<td>246</td>
<td>15680</td>
</tr>
</tbody>
</table>
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 $p(E(i,j)|Y)$. The blanked cell indicates the zero estimated value.

<table>
<thead>
<tr>
<th></th>
<th>ADHD</th>
<th>ASD</th>
<th>BPD</th>
<th>CAD</th>
<th>CD</th>
<th>MDD</th>
<th>RA</th>
<th>SCZ</th>
<th>SLE</th>
<th>T1D</th>
<th>T2D</th>
<th>UC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADHD</td>
<td>–</td>
<td>0.02</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASD</td>
<td>0.02</td>
<td>–</td>
<td>0.84</td>
<td>0.90</td>
<td>0.44</td>
<td>0.03</td>
<td>0.54</td>
<td>0.47</td>
<td>0.03</td>
<td>0.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BPD</td>
<td>1.00</td>
<td>0.84</td>
<td>–</td>
<td>0.68</td>
<td>0.09</td>
<td>0.99</td>
<td>1.00</td>
<td>0.78</td>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAD</td>
<td>0.90</td>
<td>0.68</td>
<td>–</td>
<td>1.00</td>
<td></td>
<td>0.01</td>
<td>0.86</td>
<td>0.36</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MDD</td>
<td>0.44</td>
<td>0.02</td>
<td>0.93</td>
<td>0.08</td>
<td></td>
<td>0.01</td>
<td>0.02</td>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCZ</td>
<td>0.03</td>
<td></td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLE</td>
<td>0.54</td>
<td>0.02</td>
<td>0.20</td>
<td>0.08</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1D</td>
<td>0.47</td>
<td>0.01</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2D</td>
<td>0.03</td>
<td>0.18</td>
<td>1.00</td>
<td>0.07</td>
<td>0.58</td>
<td>0.01</td>
<td>0.16</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UC</td>
<td>0.02</td>
<td></td>
<td>0.97</td>
<td>0.02</td>
<td>0.02</td>
<td>0.16</td>
<td>1.00</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 $\beta_{ij}$. The blanked cell indicates that $p(E(i,j)|Y)$ is estimated as zero and the bold number indicates that the 95% credible interval $\beta_{ij}$ does not contain zero.

<table>
<thead>
<tr>
<th></th>
<th>ADHD</th>
<th>ASD</th>
<th>BPD</th>
<th>CAD</th>
<th>CD</th>
<th>MDD</th>
<th>RA</th>
<th>SCZ</th>
<th>SLE</th>
<th>T1D</th>
<th>T2D</th>
<th>UC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADHD</td>
<td>–</td>
<td>0.01</td>
<td></td>
<td>2.30</td>
<td></td>
<td>0.53</td>
<td></td>
<td>0.02</td>
<td>0.79</td>
<td>0.71</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>ASD</td>
<td>0.01</td>
<td>–</td>
<td>0.86</td>
<td>0.94</td>
<td>0.01</td>
<td>1.69</td>
<td>0.01</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
<td>0.13</td>
</tr>
<tr>
<td>BPD</td>
<td></td>
<td></td>
<td></td>
<td>2.30</td>
<td>0.04</td>
<td>0.52</td>
<td>0.04</td>
<td>1.25</td>
<td>1.57</td>
<td>0.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAD</td>
<td>0.00</td>
<td>0.94</td>
<td>0.52</td>
<td></td>
<td>0.04</td>
<td>1.00</td>
<td>1.57</td>
<td>0.58</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD</td>
<td>0.00</td>
<td></td>
<td>0.04</td>
<td>0.95</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.34</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MDD</td>
<td>0.53</td>
<td>0.01</td>
<td></td>
<td>2.30</td>
<td>0.02</td>
<td></td>
<td>0.02</td>
<td>0.15</td>
<td>1.26</td>
<td>1.26</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>RA</td>
<td>0.00</td>
<td>1.69</td>
<td></td>
<td></td>
<td>0.03</td>
<td>0.03</td>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCZ</td>
<td>0.02</td>
<td></td>
<td>1.57</td>
<td>0.58</td>
<td>0.61</td>
<td>1.21</td>
<td></td>
<td>0.16</td>
<td>4.92</td>
<td>0.74</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>SLE</td>
<td>0.79</td>
<td>0.01</td>
<td>1.15</td>
<td></td>
<td>1.04</td>
<td>1.38</td>
<td></td>
<td>1.26</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1D</td>
<td>0.71</td>
<td></td>
<td>1.33</td>
<td>2.40</td>
<td>1.13</td>
<td>1.38</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2D</td>
<td>0.02</td>
<td>0.13</td>
<td></td>
<td>2.30</td>
<td>0.05</td>
<td>0.74</td>
<td>0.01</td>
<td>0.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UC</td>
<td>0.01</td>
<td></td>
<td></td>
<td>2.30</td>
<td>0.05</td>
<td>0.74</td>
<td>0.01</td>
<td>0.04</td>
<td></td>
<td></td>
<td></td>
<td>0.85</td>
</tr>
</tbody>
</table>

34
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.

<table>
<thead>
<tr>
<th></th>
<th>ADHD</th>
<th>ASD</th>
<th>BPD</th>
<th>CAD</th>
<th>CD</th>
<th>MDD</th>
<th>RA</th>
<th>SCZ</th>
<th>SLE</th>
<th>T1D</th>
<th>T2D</th>
<th>UC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADHD</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ASD</td>
<td>0</td>
<td>10</td>
<td>2</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>BPD</td>
<td>0</td>
<td>2</td>
<td>128</td>
<td>6</td>
<td>63</td>
<td>59</td>
<td>15</td>
<td>90</td>
<td>79</td>
<td>75</td>
<td>2</td>
<td>17</td>
</tr>
<tr>
<td>CAD</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>169</td>
<td>20</td>
<td>6</td>
<td>9</td>
<td>12</td>
<td>8</td>
<td>8</td>
<td>26</td>
<td>16</td>
</tr>
<tr>
<td>CD</td>
<td>0</td>
<td>5</td>
<td>63</td>
<td>20</td>
<td>1897</td>
<td>53</td>
<td>41</td>
<td>94</td>
<td>109</td>
<td>132</td>
<td>19</td>
<td>692</td>
</tr>
<tr>
<td>MDD</td>
<td>0</td>
<td>0</td>
<td>59</td>
<td>6</td>
<td>53</td>
<td>83</td>
<td>9</td>
<td>73</td>
<td>71</td>
<td>70</td>
<td>2</td>
<td>16</td>
</tr>
<tr>
<td>SCZ</td>
<td>0</td>
<td>1</td>
<td>15</td>
<td>9</td>
<td>94</td>
<td>23</td>
<td>25</td>
<td>351</td>
<td>99</td>
<td>5</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>SLE</td>
<td>0</td>
<td>0</td>
<td>90</td>
<td>12</td>
<td>94</td>
<td>23</td>
<td>25</td>
<td>351</td>
<td>99</td>
<td>5</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>RA</td>
<td>0</td>
<td>1</td>
<td>15</td>
<td>9</td>
<td>94</td>
<td>23</td>
<td>25</td>
<td>351</td>
<td>99</td>
<td>5</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>T1D</td>
<td>0</td>
<td>5</td>
<td>79</td>
<td>8</td>
<td>109</td>
<td>71</td>
<td>30</td>
<td>107</td>
<td>156</td>
<td>117</td>
<td>5</td>
<td>40</td>
</tr>
<tr>
<td>T2D</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>19</td>
<td>2</td>
<td>7</td>
<td>5</td>
<td>163</td>
<td>183</td>
<td>5</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>UC</td>
<td>0</td>
<td>0</td>
<td>17</td>
<td>16</td>
<td>692</td>
<td>16</td>
<td>36</td>
<td>40</td>
<td>40</td>
<td>55</td>
<td>18</td>
<td>1458</td>
</tr>
</tbody>
</table>

35
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 $p(E(i,j)|Y)$. The blanked cell indicates the zero estimated value.

<table>
<thead>
<tr>
<th></th>
<th>ADHD</th>
<th>ASD</th>
<th>BPD</th>
<th>CAD</th>
<th>CD</th>
<th>MDD</th>
<th>RA</th>
<th>SCZ</th>
<th>SLE</th>
<th>T1D</th>
<th>T2D</th>
<th>UC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADHD</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.01</td>
<td>1.00</td>
<td>0.02</td>
<td>0.40</td>
<td>0.49</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASD</td>
<td>1.00</td>
<td>0.80</td>
<td>1.00</td>
<td>0.16</td>
<td>1.00</td>
<td>0.04</td>
<td>0.03</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BPD</td>
<td>1.00</td>
<td>0.80</td>
<td>0.35</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.66</td>
<td>0.58</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAD</td>
<td>1.00</td>
<td>0.02</td>
<td>0.89</td>
<td>0.99</td>
<td>0.03</td>
<td>1.00</td>
<td>0.90</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD</td>
<td>0.16</td>
<td>0.26</td>
<td>0.01</td>
<td>1.00</td>
<td>1.00</td>
<td>0.76</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MDD</td>
<td>1.00</td>
<td>0.02</td>
<td>1.00</td>
<td>0.11</td>
<td>1.00</td>
<td>0.95</td>
<td>0.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCZ</td>
<td>0.02</td>
<td>0.85</td>
<td>0.11</td>
<td>0.01</td>
<td>1.00</td>
<td>0.25</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLE</td>
<td>0.40</td>
<td>1.00</td>
<td>0.76</td>
<td>1.00</td>
<td>0.99</td>
<td>0.03</td>
<td>0.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1D</td>
<td>0.49</td>
<td>0.03</td>
<td>0.66</td>
<td>1.00</td>
<td>0.01</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2D</td>
<td>0.01</td>
<td>1.00</td>
<td>0.04</td>
<td>0.25</td>
<td>1.00</td>
<td>0.18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UC</td>
<td>0.90</td>
<td>1.00</td>
<td>0.11</td>
<td>0.18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 $\beta_{ij}$. The blanked cell indicates that $p(E(i,j)|Y)$ is estimated as zero and the bold number indicates that the 95% credible interval $\beta_{ij}$ does not contain zero.

<table>
<thead>
<tr>
<th></th>
<th>ADHD</th>
<th>ASD</th>
<th>BPD</th>
<th>CAD</th>
<th>CD</th>
<th>MDD</th>
<th>RA</th>
<th>SCZ</th>
<th>SLE</th>
<th>T1D</th>
<th>T2D</th>
<th>UC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADHD</td>
<td>0.83</td>
<td>0.25</td>
<td>0.12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASD</td>
<td>0.83</td>
<td>0.80</td>
<td>1.01</td>
<td>0.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BPD</td>
<td>1.19</td>
<td>1.19</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAD</td>
<td>0.00</td>
<td>1.01</td>
<td>0.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD</td>
<td>0.00</td>
<td>0.85</td>
<td>0.12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MDD</td>
<td>1.12</td>
<td>0.01</td>
<td>1.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCZ</td>
<td>0.02</td>
<td>1.51</td>
<td>0.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLE</td>
<td>0.47</td>
<td>0.01</td>
<td>0.60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T1D</td>
<td>0.66</td>
<td>0.01</td>
<td>0.61</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2D</td>
<td>0.01</td>
<td>0.78</td>
<td>0.60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UC</td>
<td>0.00</td>
<td>0.07</td>
<td>2.60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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.

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

5.1 Simulation Studies

We generated our simulation data as follows. First, we assumed the true phenotype graph \((G_0)\) 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 \(\beta_{ij}\) were set to zeros). Then, given the MRF coefficients, we generated association status of 20,000 common SNPs, \(e_t = (e_{1t}, e_{2t}, e_{3t}, e_{4t}, e_{5t}, e_{6t})\), from the model of Equation (2) in Section 1.1, by running the Gibbs sampler for 1,000 iterations (Mitra \textit{et al.} 2013). Finally, given the association status of SNPs, we generated \(y_{it}\) from \(N(\mu, \sigma^2)\) if \(e_{it} = 1\), and from \(N(0,1)\) if \(e_{it} = 0\), where \(\mu = (0.55, 0.5, 0.6, 0.6, 0.65, 0.55)\) and \(\sigma = (0.4, 0.3, 0.35, 0.3, 0.45, 0.4, 0.3)\).

![Figure 9: True phenotype graph \(G_0\) 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.](image)

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 \(G\), i.e., \(\Pr\{E(i,j) = 1\} \propto 1\) for all \(i \neq j\),

2. incorporating the informative prior graph \(G\) by forcing in edges which are correlated in the true phenotype graph \(G_0\), i.e., \(\delta_0\{E(i,j) - 1\}\) 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 \( G_0 \) 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 \( P_7 \) whose association status is identical to \( P_1 \), i.e., \( e_t = (e_{1t}, e_{2t}, e_{3t}, e_{4t}, e_{5t}, e_{6t}, e_{7t}) \) 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 \( G_0 \) 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 \( G_0 \).](image)

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 \( G \) is similar to the unknown true graph \( G_0 \) 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 \( \beta_{ij} \). 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 \( \beta_{ij} \) 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.georgehrelab.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


