Constructing genotype and phenotype network helps reveal disease heritability and phenome-wide association studies

Xuewei Cao[1,#], Lirong Zhu[1,#], Xiaoyu Liang[2], Shuanglin Zhang[1], Qiuying Sha[1,*]

[1]Department of Mathematical Sciences, Michigan Technological University, Houghton, Michigan, USA

[2]Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, MI, USA

[#] Both authors contributed equally

[*]Corresponding author: QIUYING SHA, Department of Mathematical Sciences, Michigan Technological University, Houghton, Michigan 49931, USA. E-mail: qsha@mtu.edu

Abstract

Analyses of a bipartite Genotype and Phenotype Network (GPN), linking the genetic variants and phenotypes based on statistical associations, provide an integrative approach to elucidate the complexities of genetic relationships across diseases and identify pleiotropic loci. In this study, we first assess contributions to constructing a well-defined GPN with a clear representation of genetic associations by comparing the network properties with a random network, including connectivity, centrality, and community structure. Next, we construct network topology annotations of genetic variants that quantify the possibility of pleiotropy and apply stratified linkage disequilibrium (LD) score regression to 12 highly genetically correlated phenotypes to identify enriched annotations. The constructed network topology annotations are informative for disease heritability after conditioning on a broad set of functional annotations from the baseline-LD model. Finally, we extend our discussion to include an application of bipartite GPN in phenome-wide association studies (PheWAS). The community detection method can be used to obtain a priori grouping of phenotypes detected from GPN based on the shared genetic architecture, then jointly test the association between multiple phenotypes in each network module and one genetic variant to discover the cross-phenotype associations and pleiotropy. Significance thresholds for PheWAS are adjusted for multiple testing by applying the false discovery rate (FDR) control approach. Extensive simulation studies and analyses of 633 electronic health record (EHR)-derived phenotypes in the UK Biobank GWAS summary dataset reveal that most multiple phenotype association tests based on GPN can well-control FDR and

identify more significant genetic variants compared with the tests based on UK Biobank categories.

Keywords: genotype and phenotype network, network topology annotation, disease heritability, phenome-wide association studies, GWAS summary statistics

Introduction

The studies utilizing biological networks have proven to be successful in providing a comprehensive understanding of the complex relationships within the biological systems, such as gene regulatory networks[1; 2], protein-protein interaction networks[3], human disease networks[4], et al. One of the commonly used biological networks is the bipartite network, which is defined as a network that consists of two distinct sets of nodes, with nodes in one set only connected to nodes in the other set and not within the same set. The human disease network usually describes the biological system as a bipartite network, where diseases and genes are represented as two distinct sets of nodes and disease nodes are only connected to their associated gene nodes. Rather than simply identifying the association between a genetic variant and a specific disease, the construction of a bipartite network can reveal the integrated molecular underpinnings of diseases[5]. Therefore, a bipartite network can be used to explore whether human diseases or complex traits and the corresponding genetic variants are related to each other at a higher level of cellular and organization[6; 7]. In addition, due to many complex diseases being affected by a shared set of pleiotropic variants, the construction of a bipartite network can also be used to determine the pathobiological relationship of one disease to other diseases[5] and elucidate the complexities of genetic correlations across diseases[6].

Over the past decade, genome-wide association studies (GWAS) have generated an impressive list of genetic variant and phenotype association pairs[8; 9], which offer a great opportunity to establish a bipartite network connecting genetic

variants and phenotypes, referred to as a genotype and phenotype network (GPN)[7]. GPN provides integrative analyses that allow for the characterization of complex relationships between genetic variants and phenotypes, which are reproducible and accurately represent biological relationships. Therefore, it has become increasingly important in recent years[10; 11]. Notably, a well-defined GPN is crucial as it provides a clear representation of the genetic association between genetic variants and phenotypes, including factors such as connectivity, centrality, and community structure. Meanwhile, the real-world biological network, including GPN, often exhibits a scale-free degree distribution[12; 13], which means that a small number of nodes (genetic variants and phenotypes) have a much larger number of connections than the majority of nodes. In a random network, the nodes are connected randomly without any preferential attachment, resulting in a network with a relatively uniform degree distribution[14]. Therefore, comparing the degree distribution of a bipartite GPN to that of a random network can reveal important insights into the underlying mechanisms driving the construction of the network. Additionally, random networks can serve as a useful null model for testing the significance of network properties observed in the bipartite GPN.

The centralities of a bipartite GPN are one of the most important statistics to measure the importance of genetic variants (phenotypes) across phenotypes (genetic variants) based on the connectivity in the network[15]. The nodes with high centralities often act as hubs for information flow within the network[16]. For example, a genetic variant with high centrality accounting for all phenotypes is more likely to be a pleiotropic variant, as it is highly connected to multiple phenotypes in a

bipartite GPN. Therefore, these centralities can be used to define the network topology annotations of genetic variants that quantify the possibility of a genetic variant being a pleiotropic variant. To study whether these network topology annotations are enriched for disease heritability, we apply stratified linkage disequilibrium (LD) score regression (S-LDSC)[17] along with the leave-one-phenotype-out strategy to quantify the contribution of these annotations to disease heritability. We condition our analyses of the network topology annotations on the baseline-LD model, which includes a broad set of coding, conserved, regulatory, and LD-related functional annotations[18]. Additionally, in a bipartite GPN, a phenotype with a higher centrality accounting for all genetic variants is more likely to have a higher heritability, as it is connected to multiple genetic variants or with higher association evidence.

With the widespread availability of electronic health records (EHR) data, phenome-wide association studies (PheWAS) have been used to systematically examine the impact of one genetic variant across a broad range of phenotypes. Phenotypes in the whole phenome can be grouped by digitized codes (e.g., ICD-10 code) to represent the common clinical factors underlying the diseases. However, the taxonomy of digitized codes is based on their etiology rather than their genetic architecture, but applying the community detection method for GPN allows us to identify network modules that provide an integrative approach to understanding the complex genetic relationships across phenotypes[7]. A network module is loosely defined as a subnetwork with high local link density so that the phenotypes within a network module share more genetic architecture across all

genetic variants than phenotypes outside the network module[19; 20]. Therefore, the network modules can serve as a priori grouping of phenotypes in PheWAS, allowing for jointly testing multiple phenotypes in each network module and a genetic variant to identify the cross-phenotype associations and pleiotropy. For multiple testing corrections, we apply a refined false discovery rate (FDR) control approach to obtain the significance thresholds for PheWAS.

Material and Methods

In this section, we first describe our approach to constructing Genotype and Phenotype Networks (GPN) and defining the network topology annotations for genetic variants and phenotypes. The construction of GPN does not require access to individual-level genotype and phenotype data and only requires the marginal association evidence between each genetic variant and each phenotype (e.g., z-scores or estimated effect sizes from GWAS summary statistics). We first identify differences in denser representation and sparse representations of GPN with various sparsity approaches, then provide details of the implementation of constructed GPN, such as heritability enrichment of network topology annotations, estimation of the genetic correlation of multiple phenotypes, community detection of phenotypes, and phenome-wide association studies. **Figure 1** shows the workflow of this study.

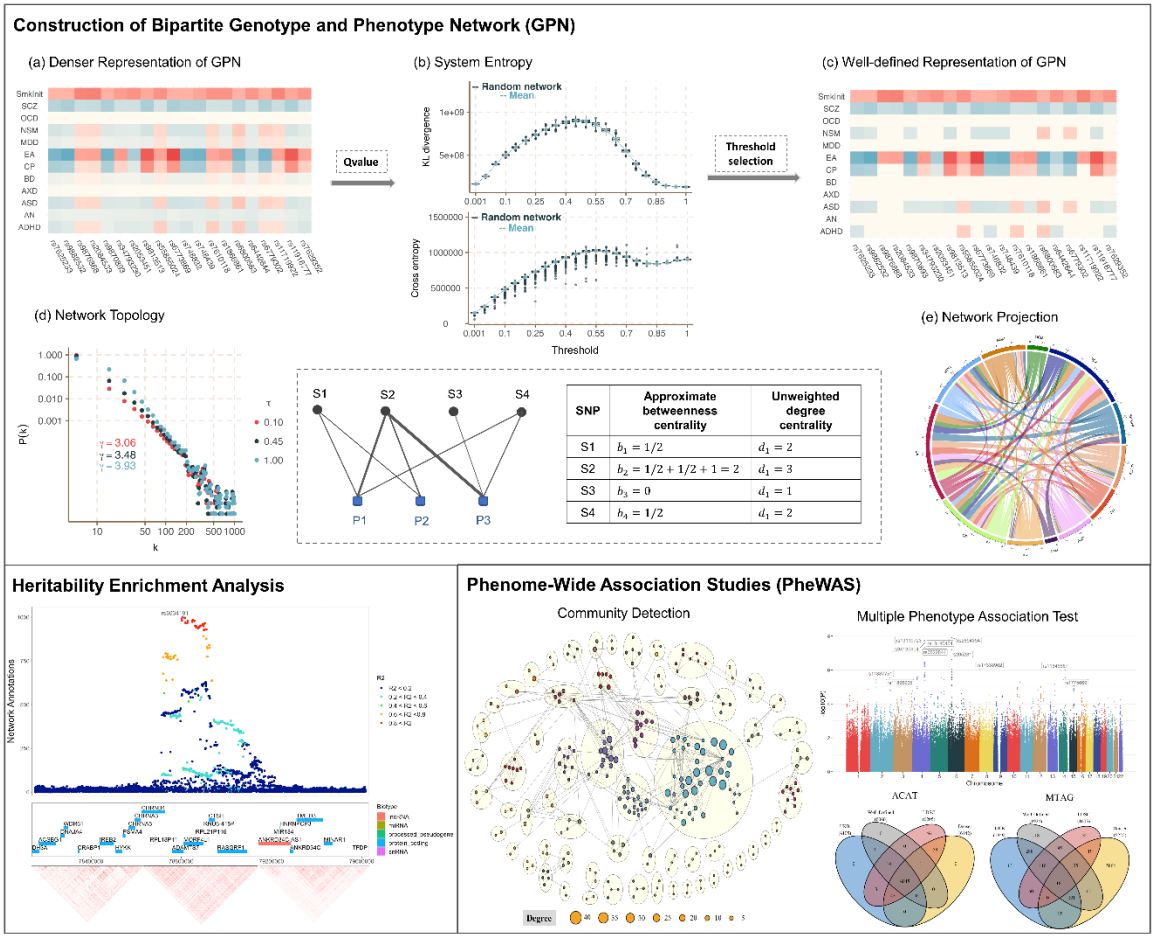


Figure 1. Graphical abstract. Construction of bipartite genotype and phenotype network (GPN) includes: (a) – (c) Construction of the denser and well-defined representations of GPN by comparing the network properties with the random networks, including connectivity, centrality, and system entropy; (d) The weighted degree distributions with different thresholds and the examples of two network topology annotations, approximate betweenness centrality and degree centrality, used in the heritability enrichment analysis; (e) The one-mode projection of GPN onto phenotypes that are linked through shared genetic architecture. Based on the constructed well-defined GPN, heritability enrichment analysis and phenome-wide

association studies are introduced as two important applications of the constructed GPN.

Bipartite genotype and phenotype networks construction

We consider GWAS summary statistical results from the same or different study cohorts with K phenotypic traits. Assume that the GWAS summary results for the k^{th} ($k=1,\cdots,K$) phenotype are calculated by testing the marginal association between a genetic variant and the k^{th} phenotype based on a sample with N_k unrelated individuals. Note that $N_k = N_l$ ($k \neq l$) if the GWAS summary statistics of the k^{th} phenotype and l^{th} phenotype are calculated from the same study cohort, otherwise, $N_k \neq N_l$. For simplicity, we assume the generalized linear regression[7], $g\left(E\left(y_{ik} \mid g_{im}\right)\right) = \alpha_{0mk} + \boldsymbol{\alpha}_{mk}^T \boldsymbol{X}_{ik} + \beta_{mk} g_{im}$, where y_{ik} is the k^{th} phenotype value and $\boldsymbol{X}_{ik}$ is the vector of covariates, for example, used to account for population stratification in the study, for the i^{th} ($1 \le i \le N_k$) individual and the k^{th} phenotype. Assuming that there are M_k genetic variants in the GWAS summary statistics for the k^{th} phenotype and g_{im} is the genotype of the m^{th} ($1 \le m \le M_k$) genetic variant taking values from 0, 1, and 2 that counts the number of copies of the minor allele. Here, $g(\bullet)$ is either the identity link function for quantitative phenotypes or the logit link for binary phenotypes.

The GWAS summary results are calculated for testing the genetic association between the k^{th} phenotype and the m^{th} genetic variant under the null

hypothesis $H_{0,mk}:\beta_{mk}=0$. The commonly used Wald-type statistic is defined as $Z_{mk}=\hat{\beta}_{mk}\Big/\widehat{se}\left(\hat{\beta}_{mk}\right)$ under the generalized linear regression model, where $\hat{\beta}_{mk}$ is the maximum likelihood estimation (MLE) of β_{mk} and $\widehat{se}\left(\hat{\beta}_{mk}\right)$ is its estimated standard error[21]. The p-value p_{mk} may also be calculated by assuming $Z_{mk}\sim N(0,1)$ in the GWAS summary results. In this study, we assume that only GWAS summary results ($\hat{\beta}_{mk}$ and p_{mk}) are available.

Let M be the total number of unique SNPs included in the GWAS summary statistics for K phenotypes with the property of $\max_{k=1,\cdots,K}\{M_k\}\le M\le\sum_{k=1}^{K}M_k$. In particular, $M=\max_{k=1,\cdots,K}\{M_k\}$ if and only if there is at least one GWAS summary data containing all unique genetic variants and $M=\sum_{k=1}^{K}M_k$ if and only if there are no variants included in different GWAS summary data. We can exclude the case $M=\sum_{k=1}^{K}M_k$ from our analyses since it rarely occurs in most GWAS summary datasets.

Denser representation of GPN

Same as the network construction introduced by Gaynor et al.[11], the denser representation of GPN allows us to capture the fact that we have no prior knowledge of precisely which genetic variants and phenotypes might have an association. Here, we construct the denser representation of GPN, an adjacency matrix that includes all the associations between genetics variants and phenotypes.

We first define a signed bipartite GPN, $\mathscr{G}_{GPN} = (Y, G, E)$, where $Y = \{Y_1, \cdots, Y_K\}$ and $G = \{G_1, \cdots, G_M\}$ denote two disjoint and independent sets of phenotypes and genetic variants, and E denotes the set of edges in GPN. Similar to our previous work[7], we denote $\mathbf{T} = (T_{mk})$ as an $M \times K$ adjacency matrix of the denser representation of GPN, where $T_{mk} = \text{sign}\left(\hat{\beta}_{mk}\right) F_{Chi}^{-1}(1 - p_{mk})$ is the weight of the edge between the m^{th} genetic variant and the k^{th} phenotype. $F_{Chi}(\bullet)$ denotes the cumulative distribution function (CDF) of χ_1^2; $\text{sign}\left(\hat{\beta}_{mk}\right) = 1$ if $\hat{\beta}_{mk} > 0$; $\text{sign}\left(\hat{\beta}_{mk}\right) = 0$ if $\hat{\beta}_{mk} = 0$; otherwise, $\text{sign}\left(\hat{\beta}_{mk}\right) = -1$. Note that $|T_{mk}|$ represents the strength of the association between the m^{th} genetic variant and k^{th} phenotype and $\text{sign}\left(\hat{\beta}_{mk}\right)$ represents the direction of the association. The denser representation includes all associations and does not involve thresholding.

Sparse representations of GPN

Given that even disease-associated genetic variants typically have a small effect size and are unlikely to exert their influence across the genome[11], utilizing a sparsity-based approach makes biologically sense. Therefore, we introduce the false discovery rate (FDR) based sparse representations of GPN, in which the networks only include edges with associations meet a certain level of significance (i.e., p-value below a threshold) from the denser representation of GPN. Let $\mathscr{G}_{GPN}^{\tau} = (Y, G, E^{\tau})$ be a sparse representation of a bipartite GPN for a specific threshold τ, where E^{τ} denotes the set of edges in the sparse representation of

GPN. $\mathbf{T}^{\tau}=\left(T_{mk}^{\tau}\right)$ is an $M\times K$ adjacency matrix of GPN, where $T_{mk}^{\tau}=T_{mk}\cdot \mathrm{I}\left(p_{mk}^{*}<\tau\right)$ with T_{mk} from the denser representation of GPN. $\mathrm{I}(\bullet)$ is an indicator function that takes value 1 when $p_{mk}^{*}<\tau$, otherwise, it takes value 0. p_{mk}^{*} is a measure of the significance of genetic association between the m^{th} genetic variant and the k^{th} phenotype by correcting for multiple comparisons in each GWAS summary data. We use q-value[22; 23] to define p_{mk}^{*} in our main analyses, but other adjustment methods for multiple comparisons can also be used, such as local FDR (LFDR)[24; 25] and an adaptation of Benjamini-Hochberg (BH) FDR[26]. We use different thresholds $\tau\in[0,1]$, where $\tau=1$ represents the denser representation of GPN since all edges are included; $\tau=0$ represents the empty network with no edges between genetic variants and phenotypes.

Well-defined sparse representation of GPN

Selecting the appropriate threshold, τ, is very important in constructing GPN. The threshold is a sort of information filter, as decreasing τ, the resulting network will change from a denser network to a very sparse one. An overly dense network can be challenging in understanding and interpreting the most biologically informative interactions between genetic variants and phenotypes due to the abundance of information. Conversely, an excessively sparse network may lead to the loss of important information. The construction of a well-defined sparse representation of GPN can be presented to determine the optimal threshold ($\hat{\tau}$) of $\mathcal{G}_{GPN}^{\tau}$, which can retain the key information about the interactions between genetic variants and phenotypes[27]. Therefore, we propose an approach to determine the optimal

threshold by comparing the network properties with a corresponding random network, including connectivity, centrality, and community structure.

More specifically, we first calculate the network "connectance" for each $\mathcal{G}_{GPN}^{\tau}$, which is defined as the ratio of the number of edges in GPN to the total number of possible edges[28; 29]. Mathematically, it can be expressed as: $connectance^{\tau} = \#\left\{E^{\tau}\right\} / \left(M \times K\right)$, where $\#\{\bullet\}$ is the counting measure, that is, $\#\left\{E^{\tau}\right\}$ represents the number of edges included in $\mathcal{G}_{GPN}^{\tau}$. The degree of "connectance" in GPN can provide insight into the structure and functionality of the interactions between genetic variants and phenotypes. As decreasing τ, the resulting network will change from a dense network ($connectance^{\tau=1} \approx 1$) to a sparse one ($connectance^{\tau=0} = 0$). For a specific τ, we then construct a corresponding random network by shuffling the edges of the original network $\mathcal{G}_{GPN}^{\tau}$. Let $\mathcal{G}_{GPN}^{random} = \left(Y, G, E^{random}\right)$ be the corresponding random network, where $conectance^{\tau}$ equals to $conectance^{random}$. We also build an adjacency matrix $\mathbf{T}^{random}$ by keeping the same weights of the edges in E^{τ}. Then, we compute the following network properties of $\mathcal{G}_{GPN}^{\tau}$ and $\mathcal{G}_{GPN}^{random}$, respectively.

Weighted and unweighted degree. The unweighted degree of a genetic variant (phenotype) in a bipartite GPN is defined as the number of edges across all phenotypes (genetic variants)[6]. The unweighted degree of the m^{th} genetic variant and the k^{th} phenotype are defined as $d_m^{G,unweight} = \sum_{k=1}^{K} \mathrm{I}\left(T_{mk}^{\tau} \neq 0\right)$ and $d_k^{P,unweight} = \sum_{m=1}^{M} \mathrm{I}\left(T_{mk}^{\tau} \neq 0\right)$, respectively. The weighted degree is reflecting the

strength of the associations of edges, which are defined as $d_m^{G,weight} = \sum_{k=1}^{K} |T_{mk}|$ and $d_k^{P,weight} = \sum_{m=1}^{M} |T_{mk}|$.

Kullback–Leibler (KL) divergence. We define KL divergences[30; 31] of degree of genetic variant and phenotypes between $\mathcal{G}_{GPN}^{\tau}$ and $\mathcal{G}_{GPN}^{random}$ to determine the diversities between a bipartite GPN and a random bipartite network, which are given by

$$KL\left(D_\tau^G \,||\, D_\tau^{G,random}\right) = \sum_{m=1}^{M} \bar{d}_m^G \log\left(\bar{d}_m^G / \bar{d}_m^{G,random}\right),$$
$$KL\left(D_\tau^P \,||\, D_\tau^{P,random}\right) = \sum_{k=1}^{K} \bar{d}_k^P \log\left(\bar{d}_k^P / \bar{d}_k^{P,random}\right),$$

where $\bar{d}_m^G$ and $\bar{d}_k^P$ are the min-max standardized degree (either weighted or unweighted) which is defined as $\bar{d}_m^G = \left(d_m^G - \min_m\left\{d_m^G\right\}\right) \Big/ \left(\max_m\left\{d_m^G\right\} - \min_m\left\{d_m^G\right\}\right)$ for the m^{th} genetic variant and $\bar{d}_k^P = \left(d_k^P - \min_k\left\{d_k^P\right\}\right) \Big/ \left(\max_k\left\{d_k^P\right\} - \min_m\left\{d_k^P\right\}\right)$ for the k^{th} phenotype. $KL\left(D_\tau^G \,||\, D_\tau^{G,random}\right)$ and $KL\left(D_\tau^P \,||\, D_\tau^{P,random}\right)$ are used to measure the difference between degree distributions of genetic variants and phenotypes in $\mathcal{G}_{GPN}^{\tau}$ and $\mathcal{G}_{GPN}^{random}$. $KL\left(D_\tau^G \,||\, D_\tau^{G,random}\right)$ will equal 0 if the degree of genetic variants are the same in $\mathcal{G}_{GPN}^{\tau}$ and $\mathcal{G}_{GPN}^{random}$; it will be negative if most degrees in $\mathcal{G}_{GPN}^{random}$ are greater than those in $\mathcal{G}_{GPN}^{\tau}$; and it will be positive if most degrees in $\mathcal{G}_{GPN}^{\tau}$ are greater than those in $\mathcal{G}_{GPN}^{random}$. $KL\left(D_\tau^P \,||\, D_\tau^{P,random}\right)$ has same properties. We also

define a global KL divergence of a bipartite network as $KL\left(D_\tau \| D_\tau^{random}\right) = KL\left(D_\tau^G \| D_\tau^{G,random}\right) + KL\left(D_\tau^P \| D_\tau^{P,random}\right)$.

Without loss of the generality, the optimal threshold τ should be selected by maximizing $KL\left(D_\tau^G \| D_\tau^{G,random}\right)$ and $KL\left(D_\tau^P \| D_\tau^{P,random}\right)$. Meanwhile, considering the equivalent numbers and weights of edges in the original network and the corresponding random network, the greater the difference in network topologies between $\mathcal{G}_{GPN}^{\tau}$ and $\mathcal{G}_{GPN}^{random}$, the more information $\mathcal{G}_{GPN}^{\tau}$ includes. To investigate the stability of the diversities, $KL\left(D_\tau^G \| D_\tau^{G,random}\right)$ and $KL\left(D_\tau^P \| D_\tau^{P,random}\right)$, we construct 1,000 random networks for each $\mathcal{G}_{GPN}^{\tau}$. We thus can estimate the standard error of KL divergence and then obtain the stability by computing their 95% confidence intervals (CIs). We also evaluate two other network properties, degree entropy and cross entropy of degree (details in **Text S1**).

Network topology annotations

For both denser and sparse representations of GPN, we constructed two probabilistic annotations based on the following network centralities. The centralities of a bipartite network are measuring the importance of genetic variants (phenotypes) across phenotypes (genetic variants) in the network. To simplify the notation, we use $\mathbf{T}$ to denote the adjacency matrix of GPN, which can be constructed by either a denser or sparse representation of GPN.

Degree centrality

In the context of bipartite GPN, a genetic variant with a high degree across phenotypes is more likely to be pleiotropic, owing to its strong connections with multiple phenotypes. Similarly, a phenotype with a high degree across genetic variants is more likely to have higher heritability and be associated with polygenic inheritance, as it is connected to multiple genetic variants or has stronger association evidence. The weighted degree of the m^{th} genetic variant or the k^{th} phenotype is defined as

$$d_m^G = \sum_{k=1}^{K} \left| T_{mk} \right| \text{ or } d_k^P = \sum_{m=1}^{M} \left| T_{mk} \right|.$$

Approximate betweenness centrality

In a bipartite GPN, we define an approximate betweenness centrality of a genetic variant which can be used to measure its importance in connecting different phenotypes. A genetic variant with high approximate betweenness can be considered an important connector between phenotypes. The approximate betweenness centrality of the m^{th} genetic variant is defined as

$$b_m = \sum_{(k,l) \in Y} \sigma_{k,l}(m) \Big/ \max\left\{\sigma_{k,l}, 1\right\},$$

where $\sigma_{k,l}$ is the number of shortest paths between the k^{th} phenotype and the l^{th} phenotype and $\sigma_{k,l}(m)$ is the number of the shortest path between the k^{th} phenotype and the l^{th} phenotype that pass through the m^{th} genetic variant. Note that there are no direct edges between phenotypes in the bipartite GPN. Therefore, the shortest path $\sigma_{k,l}$ is the number of genetic variants that are associated with

both the k^{th} phenotype and the l^{th} phenotype; the shortest path $\sigma_{k,l}(m)$ only takes the value 0 or 1, where $\sigma_{k,l}(m)=1$ if the m^{th} genetic variant is associated with both the k^{th} phenotype and the l^{th} phenotype, otherwise, $\sigma_{k,l}(m)=0$.

Heritability enrichment of network annotations

Note that the network topology annotations of genetic variants quantify the possibility of a genetic variant being a pleiotropic variant. To study whether these annotations are enriched for disease heritability of the highly correlated phenotype, we first perform a leave-one-phenotype-out (LOPO) approach to construct the network topology annotations. Then, we use stratified LD score regression (S-LDSC) to estimate the enrichment and the standardized effect size of the annotation[32; 33].

Leave-one-phenotype-out (LOPO)

We consider K highly genetically correlated phenotypes. To simplify the notation, we use $\tilde{\mathbf{T}}_k$ to denote the adjacency matrix of GPN by removing the k^{th} phenotype. $\tilde{\mathbf{T}}_k$ can be constructed by either denser or one of the sparse representations. Then, we use one of the network topology annotations based on the degree centrality and approximate centrality to assign the numeric value to each genetic variant for the evaluation of the k^{th} phenotype. Assigning a network topology annotation to each genetic variant is a way to quantify its potential for pleiotropy. The LOPO approach can assist in determining whether genetic variants have

highly evidenced impacts on other $K-1$ phenotypes through pleiotropy and can also contribute to estimate the heritability of the k^{th} phenotype.

Stratified LD score regression (S-LDSC)

S-LDSC is a method to assess the contribution of the annotation to disease heritability[32; 33] conditional on other functional annotations. We use 86 functional annotations in the baseline-LD model (v2.1)[34], including regulatory annotations (e.g., promoter, enhancer, histone marks, TF binding sites), LD-related annotations, et al. In this section, we omit the index k to simplify the notations. Let a_{mc} be the annotation value of the m^{th} genetic variant for the c^{th} annotation, where $m=1,\cdots,M_k$ and $c=0,\cdots,C$. In particular, a_{m0} represent the network topology annotation of the m^{th} genetic variant constructed by the LOPO approach.

S-LDSC assumes that the per-SNP heritability or variance of the effect size of each genetic variant is given by $Var(\beta_m)=\sum_{c=0}^{C} a_{mc}\phi_c$, where ϕ_c is the per-SNP contribution of the c^{th} annotation to disease heritability. We can estimate ϕ_c using S-LDSC,

$$E(\chi_m^2)=N\sum_{c=0}^{C} l(m,c)\phi_c+1,$$

where χ_m^2 is the chi-square test statistic for testing the association between the m^{th} genetic variant and a phenotype in GWAS summary data, $l(m,c)=\sum_{\tilde{m}} a_{mc}r_{m,\tilde{m}}^2$ is the LD score of the m^{th} genetic variant to the c^{th} annotation, and $r_{m,\tilde{m}}$ is the genotypic correlation between the m^{th} and the $\tilde{m}^{th}$ genetic variants.

We only focus on the network topology annotation ϕ_0. As demonstrated by Finucane et al.[35], ϕ_0 will be positive if the network annotation increases per-SNP heritability, accounting for all other factors. Let $sd(\boldsymbol{a}_0)$ be the standard deviation of the network topology annotation. The standardized effect size ϕ_0^* is defined by

$$\phi_0^* = \frac{\phi_0 sd(\boldsymbol{a}_0)}{\sum_m Var(\beta_m)/M_k}.$$

Note that ϕ_0^* is defined as the proportionate change in per-SNP heritability associated with a one-standard-deviation increase in the network topology annotation conditioning on all other annotations[33]. The standard error on the estimate of ϕ_0^*, $sd(\phi_0^*)$, is computed using a block jackknife[32]. Then, we can compute the p-value to test if $\phi_0^* > 0$ by assuming $\phi_0^*/sd(\phi_0^*) \sim N(0,1)$.

We also calculate the enrichment of the network topology annotation, which is defined as the proportion of the heritability explained by genetic variants in the annotation divided by the proportion of genetic variants in the annotation.

$$Enrichment = \frac{h_g^2(\phi_0)/h_g^2}{\sum_m a_{m0}/M_k},$$

where $h_g^2 = \sum_m Var(\beta_m)$ is the estimated heritability and $h_g^2(\phi_0)$ is the heritability captured by the network annotation. $Enrichment > 1$ represents the network annotation enriched for the disease heritability. Same as ϕ_0^*, the significance for $Enrichment$ is computed using a block jackknife[32]. The inclusion of the 86

functional annotations in the baseline-LD model can minimize the risk of bias in enrichment estimates and can also estimate the effect size ϕ_0 conditional on the known functional annotations[32].

Community detection methods

Community detection methods are essential in comprehending the global and local structures of associations between genetic variants and phenotypes, and in shedding light on association connections that may not be easily visible in the network topology[15]. Calculating the projection of GPN onto phenotypes that are linked through shared genetic variants is a very important step in community detection. Let $\mathcal{G}_{PPN} = \left(Y, E^P\right)$ be the one-mode projection of GPN, called Phenotype and Phenotype Network (PPN), where E^P denotes the set of edges between phenotypes in PPN. Denote $\mathbf{W} = \left(W_{kl}\right)$ as an $K \times K$ adjacency matrix of PPN, where W_{kl} is the weight of the edge between the k^{th} phenotype and the l^{th} phenotype. In this study, we perform community detection methods to partition K phenotypes into L disjoint network modules based on the adjacency matrix of PPN.

Community detection method for the denser representation of GPN

For the denser representation of GPN, one straightforward way to define the adjacency matrix $\mathbf{W}$ is to use the correlation of $\mathbf{T}$, $\mathbf{W} = cor\left(\mathbf{T}\right)$[7]. The elements of $\mathbf{W}$ can be both positive and negative, implying that the PPN represented by the adjacency matrix of $\mathbf{W}$ is a signed network. Inspired by our previously proposed

modularity-based community detection method[36], we introduce a community detection method for the signed network in this study. Let $\mathbf{W}^{+}=\left(W_{kl}^{+}\right)$ and $\mathbf{W}^{-}=\left(W_{kl}^{-}\right)$ be adjacency matrices of the positive and negative weights, respectively, where $W_{kl}^{+}=\max\left\{W_{kl},0\right\}$ and $W_{kl}^{-}=-\min\left\{W_{kl},0\right\}$ such that $\mathbf{W}=\mathbf{W}^{+}-\mathbf{W}^{-}$. First, we assume K phenotypes can be divided into k_0 network modules using a hierarchical clustering method with similarity matrix $\mathbf{W}$ for $k_0=1,\cdots,K$. Let $\mathbf{C}^{(k_0)}=\left(C_{k,l}^{(k_0)}\right)$ be a $K\times K$ connectivity matrix, where $C_{k,l}^{(k_0)}=1$ if the k^{th} phenotype and the l^{th} phenotype are in the same network module, otherwise, $C_{k,l}^{(k_0)}=0$. Then, we calculate the modularity of network with only positive weights, $\mathbf{W}^{+}$, as $Q_{k_0}^{+}=\frac{1}{2D^{+}}\sum_{k,l=1}^{K}\left(W_{kl}^{+}-\frac{d_k^{+}d_l^{+}}{2D^{+}}\right)C_{k,l}^{(k_0)}$ for each k_0, where $d_k^{+}=\sum_{l=1}^{K}W_{kl}^{+}$ and $D^{+}=\sum_{k=1}^{K}d_k^{+}/2$ represent the degree of the k^{th} phenotype and overall degree of $\mathbf{W}^{+}$. Similarly, we calculate the modularity of $\mathbf{W}^{-}$ as $Q_{k_0}^{-}$. Therefore, we define the modularity for the signed network as $Q_{k_0}=\frac{2D^{+}}{2D^{+}+2D^{-}}Q_{k_0}^{+}-\frac{2D^{-}}{2D^{+}+2D^{-}}Q_{k_0}^{-}$. Note that a network's modularity value indicates the density of connections within network modules and sparsity of connections between phenotypes in different models[15]. Then, we determine the optimal number of network modules as $L=\arg\max\left\{Q_1,Q_2,\cdots,Q_K\right\}$.

Community detection method for the sparse representation of GPN

To eliminate the biases in projections caused by a large number of genetic variants that are unlikely to exert their influence across the whole genome[11], we also provide a weighted projection approach by only focusing on the shared genetic variants between two phenotypes in the (well-defined) sparse representations of GPN, $\mathbf{T}^{sparse}$. Let S_{kl}^{*} be the set of genetic variants that are connected with the k^{th} phenotype and the l^{th} phenotype. We define $W_{kl}=\sum_{m\in S_{kl}^{*}}\left|T_{mk}^{sparse}\right|\Big/d_{k}^{sparse}$ and $W_{lk}=\sum_{m\in S_{kl}^{*}}\left|T_{ml}^{sparse}\right|\Big/d_{l}^{sparse}$, where d_{k}^{sparse} and d_{l}^{sparse} are the weighted degree of the k^{th} and the l^{th} phenotypes, respectively. More specifically, W_{kl} is a proportion of degree of the k^{th} phenotype explained by the shared associations between the k^{th} and the l^{th} phenotypes; similarly, W_{lk} is a proportion of degree of the l^{th} phenotype explained by the shared associations between the k^{th} and the l^{th} phenotypes. Therefore, $W_{kl}\neq W_{lk}$ indicates that the projected PPN is a directed network. If $W_{kl}>W_{lk}$, the shared associations between the k^{th} and the l^{th} phenotypes are more important to the k^{th} phenotype than the l^{th} phenotype. In particular, $W_{kl}=1$ if and only if the k^{th} phenotype only links with the genetic variants in S_{kl}^{*}. The modularity is easily generalized to both weighted and directed network, where the modularity based on LinkRank is given by[37; 38]: $Q_{k_0}=\sum_{k,l=1}^{K}\left(\pi_k G_{k,l}-\pi_k\pi_l\right)C_{k,l}^{(k_0)}$. Let $W_{k}^{out}=\sum_{l=1}^{K}W_{kl}$ be the out-degree of the k^{th} phenotype for a directed PPN. Then, $\pi_1,\cdots,\pi_K$ is the PageRank vector indicating the probability of a phenotype being visited by a random surfer.

$G_{k,l} = \alpha \cdot W_{kl} / W_k^{out} + 1/K \cdot (\alpha g_k + 1 - \alpha)$ is the Google Matrix, where α is the damping parameter for PageRank[37] (with probability $1-\alpha$ random surfer jumps to a random phenotype) and $g_k = \mathrm{I}(W_k^{out} = 0)$ is an indicator of dangling phenotype. Same as the community detection method for the denser representation of GPN, we also determine the optimal number of network modules as $L = \arg\max\{Q_1, Q_2, \cdots, Q_K\}$.

Phenome-wide association studies (PheWAS)

The community detection method for PPN based on $\mathbf{W}$ has potential applications in PheWAS and multiple phenotype association studies. We extend our discussion to include the application of GPN in PheWAS. By using the community detection method of PPN, we can obtain a priori grouping of phenotypes and then jointly test the association between genetic variant and multiple phenotypes in each network module to discover the cross-phenotype associations and pleiotropy.

Assume that K is the total number of phenotypes in the whole phenome, which can be partitioned into L disjoint network modules by community detection. Let $K = K_1 + \cdots + K_L$, where K_l is the number of phenotypes in the l^{th} network module. We apply four commonly used GWAS summary-based multiple phenotype association tests to identify the association between genetic variant and phenotypes in the l^{th} network module, including minP[39], ACAT[40], MTAG[41], SHom[42] (details in **Text S2**). Then, we refine our previous approach to evaluate FDR by thresholding the p-values obtained from the multiple phenotype association tests[43]. Let $\left\{p_m^{(1)}, \cdots, p_m^{(L)}\right\}$ be a sequence of p-values for testing the

association between phenotypes in each of the network modules and the m^{th} genetic variant. For a given nominal FDR level $\alpha \in (0,1)$, the optimal threshold for the m^{th} genetic variant is given by

$$\hat{p}_m = \sup\left\{ p \in [0,1] : p \le \frac{\alpha \max\left\{1, \sum_{l=1}^{L} \mathrm{I}\left(p_m^{(l)} \le p\right)\right\}}{m_0} \right\},$$

where m_0 is the number of network modules under the null hypothesis that phenotypes in the network module and the m^{th} genetic variant have no association. We use $m_0 = L - m_1$, where $m_1 = \sum_{l=1}^{L} \mathrm{I}\left(p_m^{(l)} \le 0.05/L\right)$ is the number of identified network modules that are associated with the m^{th} genetic variant based on the Bonferroni Correction.

Empirical GWAS summary datasets

In our analyses, we consider two publicly available GWAS summary datasets to evaluate the performance of constructed bipartite GPN, heritability enrichment of network annotations, community detection methods, and the applications of PheWAS.

GWAS summary statistics for correlated phenotypes

To perform the heritability enrichment analysis of network annotations, we obtain publicly available GWAS summary data for 12 highly genetically correlated phenotypes in individuals of European ancestry, including attention deficit/hyperactivity disorder (ADHD), smoking initiation (SmkInit), autism spectrum disorder (ASD), neuroticism (NSM), anxiety disorder (AXD), major

462 depressive disorder (MDD), obsessive-compulsive disorder (OCD), anorexia
463 nervosa (AN), bipolar disorder (BD), schizophrenia (SCZ), educational attainment
464 (EA), and cognitive performance (CP). The details of GWAS summary data for the
465 12 phenotypes are summarized in **Table S1**. As demonstrated by Zhang et al.[44],
466 the global genetic correlations among the 12 phenotypes estimated by their
467 proposed SUPERGNOVA are ranging from -0.41 to 0.69. 51 out of 66 pairs of
468 phenotypes have significant non-zero global genetic correlations (right upper
469 triangle of **Table S2**). Meanwhile, they also reported the proportions of correlated
470 regions between two phenotypes that are ranging from 0.11% to 93%. 46 pairs of
471 phenotypes contain at least one significantly correlated region after Bonferroni
472 correction (left lower triangle of **Table S2**). We only include the genetic variants in
473 22 autosomes.

474 *GWAS summary statistics in the UK Biobank*

475 The UK Biobank is a population-based cohort study with a wide variety of genetic
476 and phenotypic information[45]. It recently released GWAS data on ~ 500K
477 individuals throughout the United Kingdom[46; 47]. For our study, we obtain the
478 publicly available GWAS summary data for 633 EHR-derived phenotypes with
479 main ICD10 diagnoses from Neale lab (Data availability). These GWAS summary
480 data are calculated based on score tests on ~337,000 unrelated individuals of
481 British ancestry. We utilize the LD score regression (LDSC)[48] on each of these 633
482 phenotypes, excluding 45 phenotypes from our analyses since the heritability
483 estimators for them are out of bounds. There are 588 phenotypes across
484 1,096,648 genetic variants in autosomes in our analyses.

Results

Construction of GPNs for 12 genetically correlated phenotypes

We construct three bipartite GPNs for 12 genetically correlated phenotypes listed in **Table S1**, including a denser representation, an arbitrary sparse representation, and a well-defined representation. There are a total of 17,585,432 unique genetic variants from 12 GWAS summary datasets. The global genetic correlations and proportions of correlated regions among the 12 phenotypes estimated by SUPERGNOVA[44] are shown in **Table S2**. We also perform LDSC[48] to estimate phenotypic correlation (right upper triangle of **Table S3**) and genetic correlation (left lower triangle of **Table S3**) among the 12 phenotypes. Among a total of 66 pairs of phenotypes, 45 pairs of phenotypes have significant non-zero genetic correlations ($\text{p-values} < 0.05/66 = 7.58\times10^{-4}$). In particular, MDD has significant genetic correlations with all of the other 11 phenotypes, NSM has significant genetic correlations with 10 phenotypes except for BD, and SCZ and EA have significant genetic correlations with 10 other phenotypes but do not have significant genetic correlations with each other.

The denser representation of GPN is constructed without using any thresholds. Since the 12 GWAS summary datasets contain different numbers of the 17,585,432 unique genetic variants, the connectance of the denser representation of GPN is 0.5123 (**Figure S1(a)**). The well-defined sparse representation of GPN is constructed by comparing the network properties with the

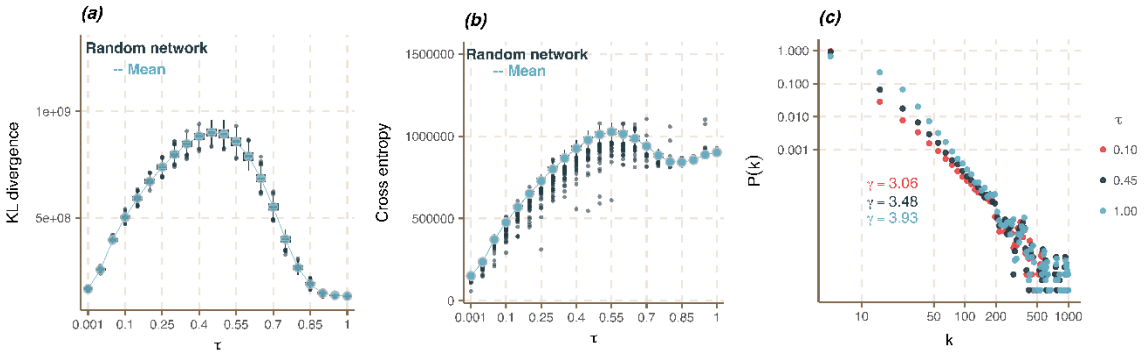
corresponding random networks. Since we have only 12 phenotypes in this analysis, we only consider the network properties for genetic variants of the constructed GPN and the corresponding random networks. For each $\tau \in (0,1)$, we generate 1,000 corresponding random networks. **Figure 2 (a)** shows the comparisons of the KL divergence for genetic variants across 1,000 random networks. The KL divergence increases from 0 to a specific value of the threshold and then decreases from that value to 1, indicating that the difference between the original and random network reaches the maximum at the specific value. We also calculate the cross entropy of the weighted degree of genetic variants compared to the corresponding random network (**Figure 2 (b)**).

Note that the weighted degree of genetic variants in a corresponding random network becomes more different than the original one if the original network retains the key information about the interactions between genetic variants[27]. The network properties, KL divergence and cross entropy, will reach the maximum value at the most informative network. In our analysis, we prioritize choosing the optimal threshold with respect to KL divergence and then check the cross entropy and weighted degree entropy at that optimal threshold. The maximum mean of KL divergence equals 9.02×10^8 at $\tau = 0.45$, where the mean of cross entropy equals a larger value (9.83×10^5) even though it does not reach the maximum value. Therefore, we constructed the well-defined sparse representation of GPN with $\tau = 0.45$. This optimal threshold is much larger than the significant level for the association testing (e.g., $\tau = 0.05$ for controlling FDR at the nominal level of 0.05). The optimal threshold in the construction of GPN does not represent

the significant associations between genetic variants and phenotypes. It is only used to ensure that the constructed GPN is more informative than a random network.

As a comparison, we also construct an arbitrary sparse representation of GPN by using the threshold $\tau = 0.1$. **Figure 2(c)** shows the weighted degree distribution of genetic variants for three GPNs, denser representation ($\tau = 1$), well-defined sparse representation ($\tau = 0.45$), and an arbitrary threshold sparse representation ($\tau = 0.1$). We observe that the degree distributions of all three networks follow the power law with different scale parameters γ, indicating that a small number of genetic variants have a much larger number of connections than the majority of genetic variants. In particular, the degree of genetic variants in the denser representation of GPN is greater than those in a sparser GPN, resulting in the scale parameter increases with increasing the threshold τ.

We also calculate the network properties of the unweighted GPNs by comparing them with the corresponding random networks (**Figure S2**). Furthermore, the adjacency matrix of the projected PPN, $\mathbf{W}$ can be considered as the phenotypic correlation among 12 phenotypes based on the shared genetic architecture. **Figure S3** shows the comparisons of the adjacency matrix of PPN constructed by the denser and well-defined sparse representations of GPN with the genetic correlation matrix estimated by SUPERGNOVA[44] (**Table S2**) and LDSC[48] (**Table S3**).



551

552 **Figure 2**. Network properties of the weighted bipartite GPNs for 12 genetically
553 correlated phenotypes. (a) KL divergence for genetic variants. The blue line is the
554 mean of KL divergencies across 1,000 random network comparisons. The boxplots
555 show the scaled distributions of the KL divergence for each threshold. (b) Cross
556 entropy for genetic variants. Blue lines are the means of cross entropy across
557 1,000 random network comparisons. The boxplot shows the scaled distribution of
558 the cross entropy for each threshold. (c) Plot of the weighted degree distribution of
559 genetic variants for three GPNs on the log-log scale, denser representation ($\tau = 1$
560), well-defined sparse representation ($\tau = 0.45$), and an arbitrary threshold sparse
561 representation ($\tau = 0.1$).

562

563 ***Heritability enrichment analysis of network annotations***

564 For each of the three bipartite GPNs for the 12 phenotypes, we perform S-LDSC
565 along with LOPO to evaluate whether the network topology annotations are
566 enriched for disease heritability. We consider both degree centrality and
567 betweenness centrality of genetic variants, conditioning on 86 functional
568 annotations in the baseline-LD model (v2.1)[34]. These 86 existing functional

annotations have been demonstrated to be highly informative by capturing functionality and LD-related features, thus, we evaluate the added value of our network topology annotations in capturing disease heritability, contributed by the pleiotropic variants with other genetically correlated phenotypes.

Table 1 shows the heritability enrichment analysis results for degree centrality calculated from denser, arbitrary sparse, and well-defined sparse representations of GPN, respectively. From the LDSC results (**Table S3**), MDD has significant non-zero genetic correlations with all other 11 phenotypes. **Table 1** shows that the degree centrality annotation is significantly enriched for the heritability of phenotype MDD based on all of the three constructed GPNs ($\text{p-values} < 0.05/12 \approx 0.0042$). Specifically, the network topology annotation of each genetic variant quantifies its possibility for pleiotropy among other correlated phenotypes. After we use the LOPO approach to construct the network annotations of MDD, the significance enrichment indicates that the network annotation can contribute more information to disease heritability if it is computed based on other highly genetically correlated phenotypes. In particular, even though the arbitrary sparse representation of GPN ($\tau = 0.1$) contains less information than the denser and well-defined GPN, the degree centrality annotation is still significantly enriched in heritability of MDD ($\text{p-value} = 2.79 \times 10^{-5}$) conditioned on the 86 functional annotations. Meanwhile, the degree annotation is also significantly enriched in heritability of CP ($\text{p-value} = 2.76 \times 10^{-6}$) and SCZ ($\text{p-value} = 0.0021$) for the arbitrary sparse representation of GPN. SCZ has significant non-zero genetic correlations with 9 phenotypes except for EA and CP

(**Table S3**); CP has significant proportions of correlated regions with 9 phenotypes in which there are over 15% of correlated regions with 8 phenotypes (**Table S2**).

The network annotation based on degree centrality obtained by the denser representation of a bipartite GPN includes the complete information for explaining the associations between phenotypes and genetic variants. It is significantly enriched to disease heritability of 11 out of 12 phenotypes as expected, except for AXD, with enrichment estimates ranging from 1.4457 (OCD with $\text{p-value} = 0.0016$) to 2.2894 (ASD with $\text{p-value} = 8.69 \times 10^{24}$). We identify the most significant enrichment of network annotations based on degree centrality for CP ($\text{Enrichment} = 2.2026$ with $\text{p-value} = 6.33 \times 10^{-54}$) and EA ($\text{Enrichment} = 2.0406$ with $\text{p-value} = 1.14 \times 10^{-52}$). These two phenotypes have a significant proportion of correlated regions, 93%, estimated by SUPERGNOVA[44]. **Figures S4(a) and S4(b)** show the QQ-plot of EA versus CP based on the weight of the denser and the well-defined sparse representations of GPN. Most of the genetic variants have similar weights for both EA and CP, lying in the diagonal line, but there exist some genetic variants that have the largest weights for only one phenotype. The same relationship between EA and CP is shown in the marginal associations from GWAS summary datasets (**Figures S4(c) and Figure S4(d)**).

The network topology annotations obtained by the well-defined sparse representation of GPN ($\tau = 0.45$) perform similarly on the heritability enrichment compared to the denser representation of GPN. Even though some information is excluded from the well-defined GPN, the annotations obtained by the well-defined

GPN contribute similar effects to disease heritability. **Table 1** and **Table S4** show that the annotations from both denser and well-defined sparse representations of GPN can significantly enrich disease heritability of the same phenotypes. However, the network topology annotations obtained by the arbitrary sparse representation of GPN ($\tau = 0.1$) are not enriched to most disease heritability. We can conclude that a more informative network can be used to understand heritability rather than an arbitrary one with a smaller threshold. For example, if we use the significance level of the associations (e.g., $\tau = 0.1$ or $\tau = 0.05$) to construct a GPN, it may lose more information and key connections even though its edges represent the significant associations between genetic variants and phenotypes.

Table 1. Heritability enrichment analyses of network topology annotation (degree centrality) based on denser and sparse representations of bipartite GPN for each of the 12 phenotypes.

Trait	Denser		Sparse ($\tau = 0.45$)		Sparse ($\tau = 0.1$)	
	Enrichment (Standard error) *p-value*	Effect τ^* ($se(\tau^*)$) *z-score*	Enrichment (Standard error) *p-value*	Effect τ^* ($se(\tau^*)$) *z-score*	Enrichment (Standard error) *p-value*	Effect τ^* ($se(\tau^*)$) *z-score*
ADHD	2.2175 (0.1697) ***8.26e-24***	3.5434 (0.3247) *10.8870*	3.3012 (0.3209) ***8.49e-22***	3.5192 (0.3423) *10.2797*	3.4734 (0.9173) *0.0072*	2.6504 (0.9882) *2.6820*
AN	1.7796 (0.1097) ***4.31e-21***	1.5274 (0.1694) *9.0145*	2.5216 (0.2174) ***3.73e-19***	1.5866 (0.1823) *8.7030*	2.5594 (0.9810) *0.1119*	1.1405 (0.7423) *1.5364*
ASD	2.2894 (0.2640) ***8.69e-24***	2.2771 (0.2373) *9.5973*	3.4316 (0.4836) ***6.52e-21***	2.3124 (0.2580) *8.9614*	6.1025 (1.9961) *0.0118*	3.5573 (1.4359) *2.4773*
AXD	1.5678 (0.5801) *0.0754*	0.2486 (0.1613) *1.5382*	2.1892 (1.1815) *0.0653*	0.2913 (0.1703) *1.7102*	5.6798 (5.0946) *0.2467*	0.7908 (0.6693) *1.1816*
BD	2.0745 (0.1184) ***7.61e-31***	3.8595 (0.3194) *12.0837*	3.2647 (0.2417) ***1.25e-30***	4.3352 (0.3547) *12.2213*	2.9583 (0.7146) *0.0043*	2.5911 (0.9309) *2.7835*
CP	2.2026 (0.0562) ***6.33e-54***	3.4031 (0.1680) *20.2517*	3.9373 (0.1260) ***2.63e-55***	4.1757 (0.1972) *21.0983*	4.6075 (0.7325) ***2.76e-06***	3.3237 (0.6999) *4.7485*
EA	2.0406 (0.0459) ***1.14e-52***	1.9705 (0.1001) *19.5241*	3.7963 (0.1204) ***1.24e-50***	2.4471 (0.1267) *19.3187*	3.5526 (0.8799) *0.0045*	1.2735 (0.4486) *2.8389*

MDD	1.9550 (0.0715) ***4.40e-32***	0.7342 (0.0580) *12.6561*	3.0106 (0.1537) ***1.19e-29***	0.7761 (0.0615) *12.1223*	3.6246 (0.6172) ***2.79e-05***	0.6783 (0.1609) *4.2153*
NSM	1.8706 (0.1088) ***1.06e-19***	1.0423 (0.1147) *9.0888*	2.8629 (0.2225) ***9.01e-20***	1.1485 (0.1243) *9.2426*	4.1886 (1.0518) *0.0097*	1.3055 (0.5086) *2.5669*
OCD	1.4457 (0.2218) ***0.0016***	1.3711 (0.5976) *2.2942*	1.8569 (0.4276) ***0.0022***	1.4454 (0.6231) *2.3197*	0.6951 (2.1090) *0.8867*	-0.5192 (3.1212) *-0.1663*
SCZ	1.9353 (0.0668) ***2.65e-36***	5.4211 (0.3765) *14.3994*	3.0742 (0.1512) ***1.38e-33***	5.6948 (0.4217) *13.5116*	3.2212 (0.7209) ***0.0021***	4.0283 (1.3343) *3.0190*
SmkInit	1.6750 (0.0918) ***9.76e-21***	0.5857 (0.0675) *8.6809*	2.3947 (0.1866) ***8.62e-20***	0.6398 (0.0731) *8.5610*	2.1556 (0.8704) *0.1839*	0.3691 (0.2839) *1.2888*

Notes: The estimated effect size and its estimated standard error, τ^* and $se(\tau^*)$, are scaled by dividing 10^{-9}. Z-score of the effect size is reported to test the null hypothesis that either $\tau \leq 0$ (one-sided) or $\tau = 0$ (two-sided). P-value of enrichment is reported to test the null hypothesis that $Enrichment > 1$. The bold-faced p-values indicate the annotation significantly enriched in the disease heritability after accounting for multiple testing (p-value$< 0.05/12 \approx 0.0041$).

However, the network annotation based on approximate betweenness centrality performs differently on the heritability enrichment analysis than the annotation based on degree centrality. **Table S4** shows the heritability enrichment analysis results for betweenness centrality calculated from denser, arbitrary sparse, and well-defined sparse representations of GPN, respectively. We observe that the betweenness centrality calculated by the denser representation of GPN significantly enriches the disease heritability of only seven phenotypes, whereas the annotation calculated by the well-defined GPN can significantly enrich the heritability of 10 phenotypes. The strength of the associations between genetic variants and phenotypes is not considered in the betweenness centrality and the denser representation of GPN includes all edges. Therefore, the betweenness centrality of GPN is not an important feature that can be considered in the heritability enrichment analysis. Alternatively, it is an important network property for the sparse representation of GPN since only the edges with strong evidence of

associations are included in the GPN. A genetic variant with high approximate betweenness can be considered an important connector between phenotypes. Therefore, the network annotations based on the approximate betweenness centrality calculated from the well-defined ($\tau = 0.45$) and the arbitrary ($\tau = 0.1$) sparse representation of GPN are significantly enriched to 10 phenotypes' heritability. Meanwhile, the network annotation calculated by a well-defined GPN has stronger evidence than that calculated by the arbitrary one.

According to heritability enrichment results, we observe that network annotations are not enriched to the disease heritability of AXD and OCD. **Figure S5** shows the heatmap of edge weights in the well-defined sparse representation of GPN for the top 100 and the top 1000 genetic variants with the highest degree of centrality, respectively. We observe that these top genetic variants have smaller weights on AXD and OCD, which means that the genetic variants with the highest degree of centrality are not associated with AXD and OCD. Therefore, the network annotation is not enriched to their heritability. In particular, there are no edges between OCD and genetic variants if the threshold is smaller than 0.4.

Construction of GPNs for 588 EHR-derived phenotypes in the UK Biobank

For a total of 1,096,648 genetic variants and 588 EHR-derived phenotypes with main ICD10 diagnoses after preprocessing, we construct two bipartite GPNs including a denser representation and the well-defined sparse representation. Different from the previous 12 GWAS summary datasets obtained from different studies, GWAS summary datasets of these 588 phenotypes are calculated based on score tests on the same ~337,000 unrelated individuals of British ancestry.

Therefore, the connectance of the denser representation of GPN equals 1, that is, all genetic variants link with all phenotypes with strength of the associations (**Figure S1(b)**).

We consider the network properties for both genetic variants and phenotypes of constructed GPN and the corresponding random networks. For each $\tau \in (0,1)$, we generate 1,000 corresponding random networks. **Figures 3(a) and 3(b)** show the KL divergence for genetic variants and phenotypes across 1,000 random network comparisons, respectively. The KL divergence increases from 0 to a specific value of the threshold and then decreases from that value to 1, indicating that the difference between the original and random network reaches the maximum at the specific value. We also calculate the cross entropy and degree entropy of the weighted degree of genetic variants compared to the corresponding random network (**Figure S6**). The maximum mean of KL divergence equals 1.14×10^{8} at $\tau = 0.6$, where the mean of cross entropy equals 3.90×10^{4} with the largest standard error (17.08) compared with other thresholds. Therefore, we constructed the well-defined sparse representation of GPN with $\tau = 0.6$. We also compare degree distributions of the well-defined network with a denser representation ($\tau = 0.8$) and two arbitrary threshold sparse representations ($\tau = 0.2$ and $\tau = 0.4$) of GPN. Similar to the constructed GPN of 12 genetically correlated phenotypes, the degree distributions of all four networks follow the power law with different scale parameters γ, indicating that a small number of genetic variants have a much larger number of connections than the majority of genetic variants. In particular, the degree of genetic variants in the denser

representation of GPN is greater than those in the sparser GPNs, resulting in the scale parameter increases with increasing the threshold τ. Meanwhile, we calculate the network properties of the unweighted GPNs by comparing them with the corresponding random networks (**Figure S7**).

We calculate three network topology annotations of genetic variants in the constructed GPNs with $\tau = 0.2, 0.4, 0.6, 0.8$, including weighted degree centrality, unweighted degree centrality, and approximate betweenness centrality (**Figure S8 and S9**). **Figure S8** illustrates the relationship between the approximate betweenness centrality of genetic variants and the weighted degree centrality of genetic variants. The top five genetic variants with the highest degree and centrality are marked, respectively. These variants have mostly been associated with multiple phenotypes in the GWAS Catalog, and they overlap considerably under different parameter τ. Using the optimal parameter ($\tau = 0.6$), we have summarized the number of significantly associated phenotypes in **Table S5**. Additionally, the top five genetic variants with the highest weighted degree centrality are almost entirely located in the same LD blocks. However, the top five genetic variants with the highest approximate betweenness centrality are associated with multiple phenotypes and display a pleiotropic effect among them. Similarly, we also compare the relationship between the approximate betweenness centrality of genetic variants and the unweighted degree centrality of genetic variants (**Figure S9**). **Table S6** shows the top five genetic variants with highest unweighted degree and approximate centralities.

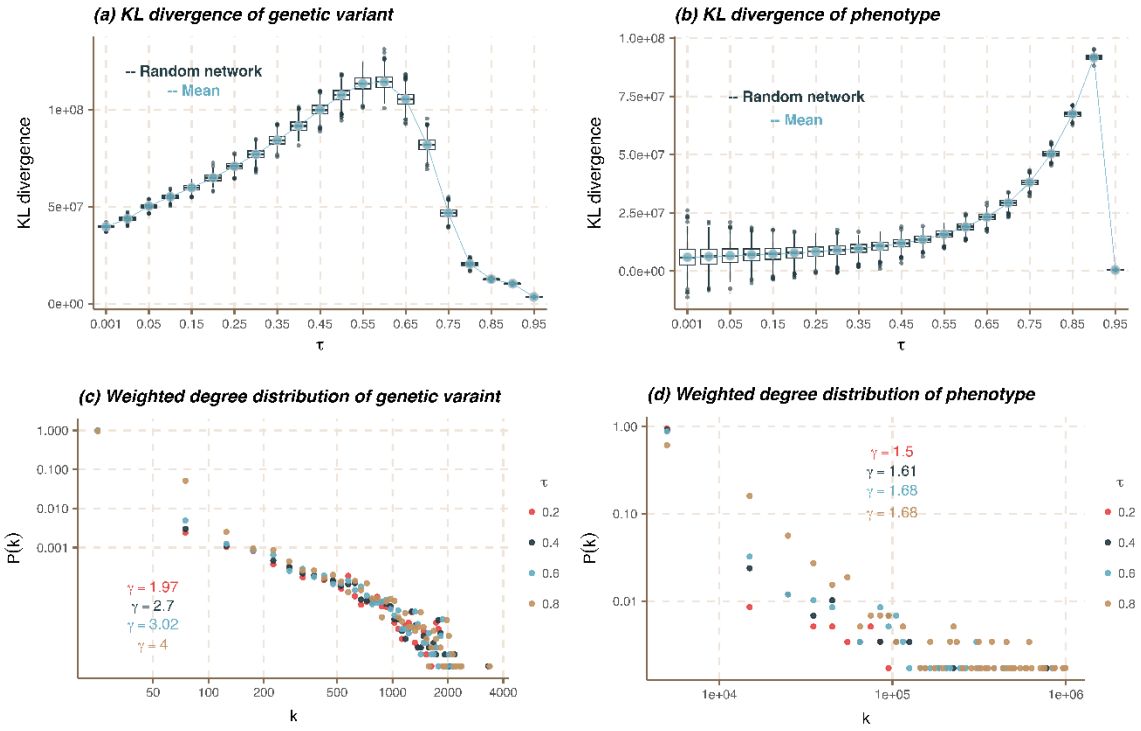


Figure 3. Network properties of the bipartite GPNs for 588 EHR-derived phenotypes in the UK Biobanks. (a) and (b) KL divergence for genetic variants and phenotypes. The blue line is the mean of KL divergencies across 1,000 random network comparisons. The boxplots show the scaled distribution of KL divergence for each threshold. (c) and (d) Weighted degree distribution of genetic variants and phenotypes for four GPNs on log-log scale, denser representation ($\tau = 0.8$), well-defined sparse representation ($\tau = 0.6$), and two arbitrary threshold sparse representations ($\tau = 0.2$ and $\tau = 0.4$).

Community detection for phenotypes

For the denser representation of GPN, we construct the one-mode projected PPN by taking the correlation of the adjacency matrix of GPN. After applying the modularity-based community detection method to the signed PPN, we partition 588

728 EHR-derived phenotypes into 132 disjoint network modules. The number of
729 phenotypes in each network module ranges from 1 to 87. For the well-defined
730 sparse representation of GPN, we also construct a directed PPN by only focusing
731 on the shared genetic variants between two phenotypes. In the sparse
732 representation of GPN, phenotypes link with multiple genetic variants, but different
733 phenotypes may not share a link with the same genetic variants. That is, we define
734 the adjacency matrix for the k^{th} phenotype as $W_{kl} = 0$ for all $l = 1, \cdots, K$ if the k^{th}
735 phenotype does not share the same genetic variants with other phenotypes.
736 Therefore, we first isolate 125 phenotypes without sharing any genetic variants
737 with other phenotypes as 125 network modules for a single phenotype. Then, we
738 partition the remaining 463 phenotypes into 71 network modules using the
739 community detection method introduced in method. The number of phenotypes in
740 the 71 network modules ranges from 2 to 37, and there are a total of 196 network
741 modules. For comparison, we also apply our proposed community detection
742 method based on the denser representation of GPN to LDSC phenotypic
743 correlation. 588 phenotypes are divided into 114 categories with the number of
744 phenotypes ranging from 2 to 82.

745

746 ***PheWAS for 588 EHR-derived phenotypes in the UK Biobank***

747 In PheWAS, a priori grouping (network module) of phenotypes in whole phenome
748 can be obtained by the community detection of PPN. For each network module,
749 we jointly test the phenotypes within this module and a genetic variant to discover
750 the cross-phenotype associations and potential pleiotropy. In this study, we

perform four most commonly used GWAS summary-based multiple phenotype association tests to identify the association between phenotype in each network module and each of genetic variants, including minP[39], ACAT[40], MTAG[41], and SHom[42] (details in **Text S2**). Then, we use the refined FDR controlling approach to evaluate FDR by thresholding the p-values obtained from the multiple phenotype association tests.

Simulation studies

We first conduct extensive simulation studies to evaluate whether these four multiple phenotype association tests used in our study can well-control FDR. We consider two simulation settings: 500 phenotypes with 50 phenotypic categories and 1,000 phenotypes with 100 phenotypic categories (details in **Text S3**). We assume that only the phenotypes within the same phenotypic category are correlated with each other. Similar to Lee et al.[49], we consider two scenarios of correlations among phenotypes within the same category: 1) same correlation between each pair of phenotypes (SAME); 2) different correlation between each pair of phenotypes that is generated by using an autoregressive (AR(1)) model. **Table S7** and **Table S8** show the average FDR in the simulation studies for 500 phenotypes and 1,000 phenotypes, respectively. FDR is evaluated using 10 Monte-Carlo (MC) runs, equivalent to 1,000 replications at a nominal FDR level of 5% (**Text S3**). The 95% confidence interval (CI) is (0.0365, 0.0635). Note that we directly generate z-scores instead of effect sizes of genetic variants on phenotypes without considering LD, therefore, we do not consider MTAG in our simulation studies. The correlations among phenotypes are estimated by the method

774 introduced in Kim et al.[39]. We observe that minP cannot control FDR in all
775 scenarios but ACAT, and SHom can well control FDR as expected.

776 *PheWAS based on 165 UK Biobank level 1 categories*

777 As benchmarked categories, 588 EHR-derived phenotypes are grouped into 165
778 UK Biobank level 1 categories defined in data-field 41202
779 (https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=41202). The number of
780 phenotypes in each category ranges from 1 to 20: there are 43 categories
781 containing only one phenotype; 35 and 31 categories contain 2 and 3 phenotypes,
782 respectively; only 7 categories contain more than 10 phenotypes. In our real data
783 analyses, we only apply three multiple phenotype association tests (ACAT, SHom,
784 and MTAG) to test the association between phenotypes in each category and each
785 genetic variant. minP is not considered here since it cannot control FDR evaluated
786 in our simulation studies. We use the commonly used genome-wide nominal FDR
787 level 5×10^{-8}. After applying our refined FDR controlling approach for the tests of
788 each genetic variant, ACAT can identify 6,105 genetic variants associated with at
789 least one category. We observe that most genetic variants are associated with only
790 one category. SHom can identify 2,701 genetic variants and MTAG can identify
791 2,980 genetic variants (**Figure 4**).

792 *PheWAS based on 114 phenotypic categories from LDSC*

793 As a comparison, there are 114 phenotypic categories of the 588 EHR-derived
794 phenotypes detected from the phenotypic correlation estimated by LDSC. We also
795 apply three multiple phenotype association tests to 114 categories. ACAT identifies
796 6,205 genetic variants, SHom identifies 2,237 genetic variants, and MTAG

identifies 1,603 genetic variants. Compared with the association tests based on the phenotypic categories in the UK Biobank, ACAT based on the LDSC can identify all of the 6,105 genetic variants identified by ACAT based on the UK Biobank (**Figure 4**). Meanwhile, 100 genetic variants are uniquely identified by ACAT based on the LDSC. **Figure S10** shows the heatmap of -log10(p-value) from GWAS summary datasets of these 100 genetic variants. We observe that all of these 100 genetic variants are significantly associated with at least one phenotype at the GWAS significance level 5×10^{-8}. According to results from SHom and MTAG, tests based on the UK Biobank identify more genetic variants than the tests based on the LDSC.

PheWAS based on 132 network modules from the denser representation of GPN

Based on the denser representations of GPN, 588 EHR-derived phenotypes are partitioned into 132 disjoint network modules According to these 132 network modules, ACAT can identify 6,142 genetic variants associated with at least one network module and SHom can identify 6,139 genetic variants. In the application of MTAG, it is time-consuming and out of memory for one network module with 87 phenotypes. Therefore, we perform MTAG on the other 131 network modules and MTAG identifies 6,220 genetic variants. **Figure 4** shows the Venn plot for genetic variants identified by three multiple phenotype association tests based on different phenotypic categories and network modules. Based on the network modules detected from the denser representation of GPN, all three methods (ACAT, SHom, and MTAG) can identify ~6,000 genetic variants associated with at least one network module.

PheWAS based on 196 network modules from the well-defined representation of GPN

Based on the well-defined representation of GPN, 588 EHR-derived phenotypes are partitioned into 196 network modules. According to the 196 network modules, ACAT can identify 6,060 genetic variants associated with at least one network module; SHom can identify 2,385 genetic variants; and MTAG can identify 1,934 genetic variants. From ACAT results, 6,060 genetic variants are identified by ACAT based on at least two other grouping of phenotypes, even if it identifies a smaller number of genetic variants. According to results from SHom and MTAG, tests based on the network modules detected from well-defined GPN identify more genetic variants than the tests based on the LDSC and the UK Biobank, but they identify fewer genetic variants than the tests based on the network modules detected from denser GPN.

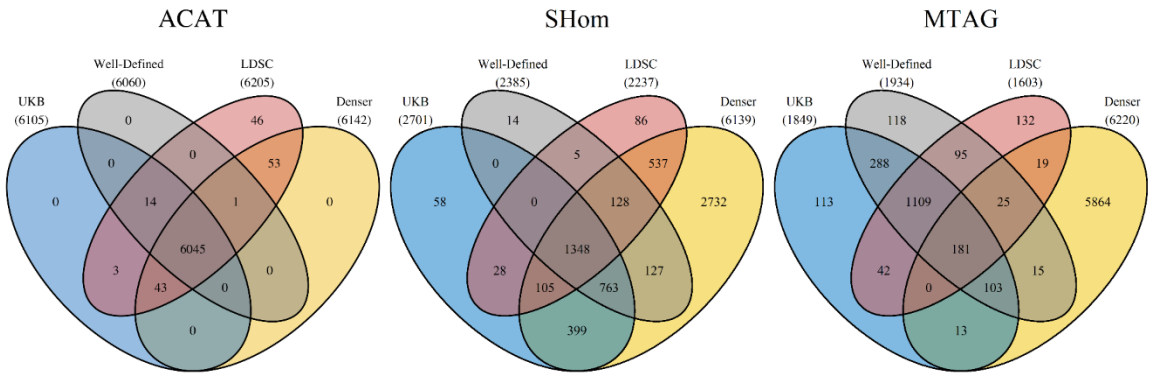


Figure 4. Venn plots for genetic variants identified by three multiple phenotype association tests based on different phenotypic categories and network modules.

Discussion

In this paper, we conduct a comprehensive analysis to build the GPNs, which can be a routine procedure in post-GWAS investigations. Owing to increasingly accessible to GWAS summary statistics, the construction of GPN only requires the marginal association evidence between each genetic variant and each phenotype in GWAS summary data instead of individual-level genotypes and phenotypes data. The denser representation of the bipartite GPN can be directly constructed by linking all genetic variants and phenotypes in GWAS summary datasets. Although a denser representation of bipartite GPN contains information on all pairwise associations between genetic variants and phenotypes, pruning the network is both biologically meaningful and computationally efficient[11]. However, the thresholding approach used to prune is significantly influenced by the network size (connectance). To address this issue, we propose to construct a well-defined GPN with a clear representation of genetic associations by comparing the network properties with a random network, including connectivity, centrality, and community structure. Our findings indicate that a well-defined network with an optimal threshold can preserve crucial information on the associations between genetic variants and phenotypes.

Based on the construction of the denser and well-defined representation of bipartite GPNs, we further propose two network topology annotations based on the degree centrality and the approximate betweenness centrality. Both of the annotations can be used to quantify the possibility of pleiotropy for genetic variants. We highlight one of our significant discoveries that link pleiotropy and disease heritability through the utilization of heritability enrichment analysis using the

stratified LD score regression. We analyze 12 genetically correlated phenotypes to show that the genetic variants with high degree centrality and approximate betweenness centrality are enriched for disease heritability conditioning on known functional annotations from the baseline LD model. First, in the analyses of the degree centrality based on the denser and the well-defined GPNs, we identify 10 phenotypes with significant heritability enrichment after using the LOPO approach. The significant enrichment indicates that the degree annotation can contribute more information to disease heritability if it is computed based on other highly genetically correlated phenotypes. We also observe that the denser GPN provides more information in the degree centrality as the degree centrality contains the strength of marginal association evidence. Second, we determine that network annotation based on the approximate betweenness centrality calculated from the well-defined GPN is strongly enriched for disease heritability. However, the disease heritability of some phenotypes is fully explained by annotations from the baseline-LD model in the analysis of the approximate betweenness centrality calculated from the denser GPN.

Construction of the bipartite GPNs also has important implications for the PheWAS. In particular, detecting the network modules of phenotypes from the constructed GPN is essential in understanding the global and local structures of associations between genetic variants and phenotypes, and in shedding light on association connections that may not be easily visible in the network topology. The detected network modules can be used as a priori grouping of phenotypes in PheWAS, then jointly testing of multiple phenotypes in each network module and

884 one genetic variant can be performed to discover the cross-phenotype
885 associations and pleiotropy. Significance thresholds for PheWAS are adjusted for
886 multiple testing by applying the false discovery rate (FDR) control approach. First,
887 we discover that the three multiple phenotype association tests (ACAT, SHom, and
888 MTAG) applied in this study can well-control FDR as demonstrated by extensive
889 simulation studies. Second, we analyze 633 EHR-derived phenotypes in the UK
890 Biobank GWAS summary datasets. Based on the network modules detected from
891 the denser representation of GPN, all three tests can identify more genetic variants
892 associated with at least one network module (~6,000 genetic variants) compared
893 with the tests based on the UK Biobank, LDSC, and well-defined GPN.

894 There are some limitations to the work presented here. First, genetic effects
895 can be heterogenous across phenotypes and studies based on different GWAS
896 summary statistics[50; 51] due to different sample sizes, genetic architectures, and
897 quality of the genotyping and phenotyping data, et al. In our current analyses, we
898 ignore the influence of different sample sizes for different GWAS summary
899 statistics in the construction of GPN. However, larger sample sizes are typically
900 associated with smaller standard errors and more precise effect size estimates,
901 which can help to reduce bias and increase the stability of effect size estimates.
902 To construct a GPN with stable evidence of the associations in the edges, we
903 suggest that the sample sizes used to calculate the GWAS summary results in
904 each study are sufficiently large (e.g., $N_k > 10,000$). Second, we use the marginal
905 association between each genetic variant and each phenotype to define the edge
906 of GPN. The challenge in validating our proposed construction of GPNs is that

907 there is no source of "ground truth" of GWAS. There may exist spurious
908 associations between multiple genetic variants and a phenotype due to LD[9]. For
909 example, a genetic variant in high LD with a true causal variant may be detected
910 instead of the causal variant itself. However, several powerful fine-mapping and
911 colocalization approaches have been developed to leverage information on LD to
912 identify the putative causal variants in a specific genomic region[52-54], which
913 provides a great opportunity to construct a more informative GPN for future
914 studies. Third, we do not consider the functional relationships between genetic
915 variants and phenotypes. Filtering candidate (functional) regions based on the
916 strength of gene-based associations may reduce multiple testing burdens and
917 consequently improve statistical power in the construction of GPN. For example,
918 transcriptome-wide association studies can combine genetic and transcriptomic
919 data in a specific tissue to identify functional variants and genomic regions, which
920 provide insights into biological pathways[55].

921

922 Declaration of interests

923 The authors declare no competing interests.

924

925 Acknowledgments

926 Part of this research has been conducted using the UK Biobank resource under
927 application number 102999 and the NHGRI-EBI GWAS Catalog. The work was in
928 part funded by the Portage Health Foundation Graduate Assistantship, and

Michigan Technological University Graduate Dean Awards . High-Performance Computing Shared Facility (Superior) at Michigan Technological University was used in obtaining results presented in this publication.

Author contributions

Formal analysis and Methodology: XC, LZ, XL, SZ, and QS; Data curation and Visualization: XC and LZ; Writing original draft: XC, LZ, and QS; Writing review and editing: XC, LZ, XL, SZ, and QS.

Web resources

Data

GWAS summary statistics for 12 highly correlated phenotypes can be downloaded from the corresponding consortium websites reported in Zhang et al.[44].

GWAS summary statistics for 633 EHR-derived phenotypes with main ICD10 diagnoses can be found from Neale lab: http://www.nealelab.is/blog/2017/7/19/rapid-gwas-of-thousands-of-phenotypes-for-337000-samples-in-the-uk-biobank.

Software

PLINK version 1.9 can be downloaded from https://www.cog-genomics.org/plink/1.9/ [56].

LDSC: the command line tool for estimating heritability and genetic correlation from GWAS summary statistics can be downloaded from https://github.com/bulik/ldsc [57].

952 Cytoscape: an open-source software platform for visualizing complex networks
953 which can be accessed via https://cytoscape.org/ [58].
954
955 ## Data and code availability
956 This study does not generate new data. The codes generated during this study
957 are available at a public repository https://github.com/xueweic/GPN.
958
959

References

1. Karlebach, G., and Shamir, R. (2008). Modelling and analysis of gene regulatory networks. Nature reviews Molecular cell biology 9, 770-780.

2. Sharma, A., Gulbahce, N., Pevzner, S.J., Menche, J., Ladenvall, C., Folkersen, L., Eriksson, P., Orho-Melander, M., and Barabasi, A.-L. (2013). Network-based analysis of genome wide association data provides novel candidate genes for lipid and lipoprotein traits. Molecular & Cellular Proteomics 12, 3398-3408.

3. Vinayagam, A., Gibson, T.E., Lee, H.-J., Yilmazel, B., Roesel, C., Hu, Y., Kwon, Y., Sharma, A., Liu, Y.-Y., Perrimon, N., et al. (2016). Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets. Proceedings of the National Academy of Sciences 113, 4976-4981.

4. Goh, K.-I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., and Barabási, A.-L. (2007). The human disease network. Proceedings of the National Academy of Sciences 104, 8685-8690.

5. Loscalzo, J. (2017). Network medicine.(Harvard University Press).

6. Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. Nature reviews genetics 12, 56-68.

7. Cao, X., Zhang, S., and Sha, Q. (2023). A novel method for multiple phenotype association studies based on genotype and phenotype network. bioRxiv, 2023.2002. 2023.529687.

8. Abdellaoui, A., Yengo, L., Verweij, K.J., and Visscher, P.M. (2023). 15 years of GWAS discovery: Realizing the promise. The American Journal of Human Genetics.

9. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 years of GWAS discovery: biology, function, and translation. The American Journal of Human Genetics 101, 5-22.

10. Korte, A., and Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. Plant methods 9, 1-9.

11. Gaynor, S.M., Fagny, M., Lin, X., Platig, J., and Quackenbush, J. (2022). Connectivity in eQTL networks dictates reproducibility and genomic properties. Cell Reports Methods 2, 100218.

12. Barabasi, A.-L., and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. Nature reviews genetics 5, 101-113.

13. Barabási, A.-L., and Albert, R. (1999). Emergence of scaling in random networks. science 286, 509-512.

14. Erdős, P., and Rényi, A. (1960). On the evolution of random graphs. Publ Math Inst Hung Acad Sci 5, 17-60.

15. Newman, M. (2018). Networks.(Oxford university press).

16. Borgatti, S.P. (2005). Centrality and network flow. Social networks 27, 55-71.

17. (!!! INVALID CITATION !!! 17; 18).

18. Kim, S.S., Dai, C., Hormozdiari, F., van de Geijn, B., Gazal, S., Park, Y., O'Connor, L., Amariuta, T., Loh, P.-R., and Finucane, H. (2019). Genes

1005 with high network connectivity are enriched for disease heritability. The
1006 American Journal of Human Genetics 104, 896-913.
1007 19. Girvan, M., and Newman, M.E. (2002). Community structure in social and
1008 biological networks. Proceedings of the national academy of sciences 99,
1009 7821-7826.
1010 20. Fortunato, S. (2010). Community detection in graphs. Physics reports 486,
1011 75-174.
1012 21. Liu, Z., and Lin, X. (2018). Multiple phenotype association tests using
1013 summary statistics in genome-wide association studies. Biometrics 74,
1014 165-175.
1015 22. Storey, J.D. (2002). A direct approach to false discovery rates. Journal of the
1016 Royal Statistical Society: Series B (Statistical Methodology) 64, 479-498.
1017 23. Storey, J.D., and Tibshirani, R. (2003). Statistical significance for
1018 genomewide studies. Proceedings of the National Academy of Sciences
1019 100, 9440-9445.
1020 24. Efron, B., Tibshirani, R., Storey, J.D., and Tusher, V. (2001). Empirical Bayes
1021 analysis of a microarray experiment. Journal of the American statistical
1022 association 96, 1151-1160.
1023 25. Storey, J.D. (2003). The positive false discovery rate: a Bayesian
1024 interpretation and the q-value. The annals of statistics 31, 2013-2035.
1025 26. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a
1026 practical and powerful approach to multiple testing. Journal of the Royal
1027 statistical society: series B (Methodological) 57, 289-300.

27. Cao, X., Shi, Y., Wang, P., Chen, L., and Wang, Y. (2018). The evolution of network topology structure of Chinese stock market. In 2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA). (IEEE), pp 329-333.

28. Easley, D., and Kleinberg, J. (2010). Networks, crowds, and markets: Reasoning about a highly connected world.(Cambridge university press).

29. Newman, M.E. (2003). The structure and function of complex networks. SIAM review 45, 167-256.

30. Kullback, S., and Leibler, R.A. (1951). On information and sufficiency. The annals of mathematical statistics 22, 79-86.

31. Murphy, K.P. (2012). Machine learning: a probabilistic perspective.(MIT press).

32. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., and Farh, K. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nature genetics 47, 1228-1235.

33. Gazal, S., Finucane, H.K., Furlotte, N.A., Loh, P.-R., Palamara, P.F., Liu, X., Schoech, A., Bulik-Sullivan, B., Neale, B.M., and Gusev, A. (2017). Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. Nature genetics 49, 1421-1427.

34. Dey, K.K., Gazal, S., van de Geijn, B., Kim, S.S., Nasser, J., Engreitz, J.M., and Price, A.L. (2022). SNP-to-gene linking strategies reveal contributions

of enhancer-related and candidate master-regulator genes to autoimmune disease. Cell genomics 2, 100145.

35. Finucane, H.K., Reshef, Y.A., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., Gazal, S., Loh, P.-R., Lareau, C., and Shoresh, N. (2018). Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nature genetics 50, 621-629.

36. Xie, H., Cao, X., Zhang, S., and Sha, Q. (2023+). Joint analysis of multiple phenotypes for extremely unbalanced case-control association studies using multi-layer network. submitted.

37. Kim, Y., Son, S.-W., and Jeong, H. (2010). Finding communities in directed networks. Physical Review E 81, 016103.

38. Mikhail, D., Anton, K., and Denis, T. (2016). Parallel modularity computation for directed weighted graphs with overlapping communities. Труды Института системного программирования РАН 28, 153-170.

39. Kim, J., Bai, Y., and Pan, W. (2015). An adaptive association test for multiple phenotypes with GWAS summary statistics. Genetic epidemiology 39, 651-663.

40. Liu, Y., Chen, S., Li, Z., Morrison, A.C., Boerwinkle, E., and Lin, X. (2019). ACAT: a fast and powerful p value combination method for rare-variant analysis in sequencing studies. The American Journal of Human Genetics 104, 410-421.

41. Turley, P., Walters, R.K., Maghzian, O., Okbay, A., Lee, J.J., Fontana, M.A., Nguyen-Viet, T.A., Wedow, R., Zacher, M., and Furlotte, N.A. (2018). Multi-

trait analysis of genome-wide association summary statistics using MTAG. Nature genetics 50, 229-237.

42. Zhu, X., Feng, T., Tayo, B.O., Liang, J., Young, J.H., Franceschini, N., Smith, J.A., Yanek, L.R., Sun, Y.V., and Edwards, T.L. (2015). Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. The American Journal of Human Genetics 96, 21-36.

43. Liang, X., Cao, X., Sha, Q., and Zhang, S. (2022). HCLC-FC: A novel statistical method for phenome-wide association studies. Plos one 17, e0276646.

44. Zhang, Y., Lu, Q., Ye, Y., Huang, K., Liu, W., Wu, Y., Zhong, X., Li, B., Yu, Z., and Travers, B.G. (2021). SUPERGNOVA: local genetic correlation analysis reveals heterogeneous etiologic sharing of complex traits. Genome biology 22, 1-30.

45. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., and O'Connell, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature 562, 203-209.

46. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., and Landray, M. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. Plos med 12, e1001779.

47. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., and O'Connell, J. (2017).

Genome-wide genetic data on~ 500,000 UK Biobank participants. BioRxiv, 166298.

48. Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.-R., Consortium, R., Consortium, P.G., 3, G.C.f.A.N.o.t.W.T.C.C.C., and Duncan, L. (2015). An atlas of genetic correlations across human diseases and traits. Nature genetics 47, 1236-1241.

49. Lee, C.H., Shi, H., Pasaniuc, B., Eskin, E., and Han, B. (2021). PLEIO: a method to map and interpret pleiotropic loci with GWAS summary statistics. The American Journal of Human Genetics 108, 36-48.

50. Han, B., and Eskin, E. (2011). Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. The American Journal of Human Genetics 88, 586-598.

51. Lee, S., Abecasis, G.R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. The American Journal of Human Genetics 95, 5-23.

52. Claussnitzer, M., Cho, J.H., Collins, R., Cox, N.J., Dermitzakis, E.T., Hurles, M.E., Kathiresan, S., Kenny, E.E., Lindgren, C.M., and MacArthur, D.G. (2020). A brief history of human disease genetics. Nature 577, 179-189.

53. Wallace, C. (2021). A more accurate method for colocalisation analysis allowing for multiple causal variants. PLoS genetics 17, e1009440.

54. Zou, Y., Carbonetto, P., Wang, G., and Stephens, M. (2022). Fine-mapping from summary data with the “Sum of Single Effects” model. PLoS Genetics 18, e1010299.

55. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W., Jansen, R., De Geus, E.J., Boomsma, D.I., and Wright, F.A. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nature genetics 48, 245-252.

56. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4, s13742-13015-10047-13748.

57. Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.-R., Duncan, L., Perry, J.R., Patterson, N., and Robinson, E.B. (2015). An atlas of genetic correlations across human diseases and traits. Nature genetics 47, 1236.

58. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research 13, 2498-2504.