Received: 3 July 2023 | Revised: 2 March 2024 | Accepted: 24 April 2024

DOI: 10.1002/sim.10101

RESEARCH ARTICLE

Statistics in Medicine WILEY

A multivariate to multivariate approach for voxel-wise genome-wide association analysis

Qiong Wu[1] | **Yuan Zhang**[2] | **Xiaoqi Huang**[3] | **Tianzhou Ma**[4,5] | **L. Elliot Hong**[6] | **Peter Kochunov**[6] | **Shuo Chen**[5,6,7,8]

[1]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[2]Department of Statistics, Ohio State University, Columbus, Ohio, USA

[3]Department of Mathematics, Louisiana State University, Baton Rouge, Louisiana, USA

[4]Department of Epidemiology and Biostatistics, School of Public Health, University of Maryland, College Park, Maryland, USA

[5]Maryland Psychiatric Research Center, Department of Psychiatry, University of Maryland School of Medicine, Baltimore, Maryland, USA

[6]Faillace Department of Psychiatry and Behavioral Sciences at McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, Texas, USA

[7]Division of Biostatistics and Bioinformatics, Department of Epidemiology and Public Health, University of Maryland, Baltimore, Maryland, USA

[8]The University of Maryland Institute for Health Computing, University of Maryland, North Bethesda, USA

Correspondence
Shuo Chen, 55 Wade Ave, Catonsville, MD, 21228, USA.
Email: shuochen@som.umaryland.com

Funding information
National Science Foundation, Grant/Award Number: DMS-2311109; National Institutes of Health, Grant/Award Numbers: 1DP1DA048968, R01EB015611, R01MH094520

The joint analysis of imaging-genetics data facilitates the systematic investigation of genetic effects on brain structures and functions with spatial specificity. We focus on voxel-wise genome-wide association analysis, which may involve trillions of single nucleotide polymorphism (SNP)-voxel pairs. We attempt to identify underlying organized association patterns of SNP-voxel pairs and understand the polygenic and pleiotropic networks on brain imaging traits. We propose a *bi-clique* graph structure (ie, a set of SNPs highly correlated with a cluster of voxels) for the systematic association pattern. Next, we develop computational strategies to detect latent SNP-voxel *bi-cliques* and an inference model for statistical testing. We further provide theoretical results to guarantee the accuracy of our computational algorithms and statistical inference. We validate our method by extensive simulation studies, and then apply it to the whole genome genetic and voxel-level white matter integrity data collected from 1052 participants of the human connectome project. The results demonstrate multiple genetic loci influencing white matter integrity measures on splenium and genu of the corpus callosum.

KEYWORDS
bi-clique, imaging-genetics, ultra-high dimensionality, voxel-wise GWAS, white matter integrity

1 | INTRODUCTION

Imaging-genetics has garnered increased interest in the field of neuropsychiatric research as it provides a viable pathway to understand brain diseases by integrating genetic, brain imaging, and environmental factors. Compared to clinical descriptions of symptoms in psychiatry, brain imaging measurements assess brain structures and functions quantitatively with reproducibility, which are reported to be associated with psychiatric disorders including schizophrenia,[1] Alzheimer's disease,[2] major depressive disorder.[3] More importantly, neuroimaging signals can serve as intermediate phenotypes resulting in increased power in the detection of genetic loci. Recent studies have been focused on the joint analysis of imaging-genetics data that reveals the genetic effects on spatially specific brain functions and structures.[4–10] Identifying genetic effects on objectively measured high-resolution imaging traits can not only enhance understanding the complex genetic and neurological mechanisms of neuropsychiatric disorders, but further impact early diagnosis and treatment of psychiatric disorders.

In imaging-genetics studies, both brain imaging data and genome sequence are measured for each participant. The genetic measurements can characterize genetic variations using single nucleotide polymorphism (SNP) and copy number variants (CNVs). The non-invasive brain imaging techniques assess the brain structures by magnetic resonance imaging (MRI), diffusion tensor imaging (DTI), and brain functions by functional magnetic resonance imaging (fMRI). The recent development of neuroimaging technology provides high-resolution imaging data with improved spatial specificity and thus can better assess the genetic effects on brain structures and functions.

The statistical analysis of imaging-genetics data is computationally intensive and methodologically challenging. These challenges mainly rise from the combination of two sets of high-dimensional features: multivariate imaging traits with multivariate genetic variants (Figure 1). Moreover, both imaging traits and genetic variants exhibit complex and organized dependence structure reflecting the underlying neurophysiological mechanisms and linkage disequilibrium patterns.[6] For example, a typical imaging-genetics study collects up to 10^7 SNPs and 10^5 voxels, jointly contributing trillions (10^{12}) of SNP-voxel pairs.[11,12] The direct application of classic voxel-wise genome-wide association analysis (vGWAS) could require an enormous sample size (eg, multiple millions of participants) to control the false positive error rate while maintaining adequate statistical power.[13–16]

Furthermore, advanced methods have been developed to leverage group sparsity by techniques including regularization, low rank techniques and projection of high-dimensional features.[10,17–25] However, while these methods could gain statistical power by jointly modeling genetic variants and imaging traits through a multivariate regression model, the high dimensionality of imaging-genetics data remains challenging due to computational burdens and/or over-fittings. For instance, the analysis can only be applied on imaging data at an regional-level or genetic data with filtered to thousands of SNP loci. Besides, the results from summarized measures as a few latent variables or a coarser scale are less interpretable or lacking the spatial specificity.[5]

In this study, we propose a new multivariate to multivariate method to systematically investigate the SNP-(imaging)voxel association patterns with four aims: (i) identify voxel clusters as genetically correlated imaging traits, (ii) detect functionally related SNP sets, (iii) understand the SNP-voxel association patterns as *polygenic* and *pleiotropic* relationships, and (iv) test the association patterns while controlling multiplicity. In our study, a polygenic trait refers to a voxel influenced by multiple SNPs while pleiotropy indicates that one gene can affect multiple voxel traits. Specifically, we consider genetic variants and imaging voxels as two disjoint sets of nodes, correspondingly, and associations between all SNP-voxel pairs as edges in a bipartite graph. We model the polygenic and pleiotropic SNP-voxel association structure as an imaging-genetics *dense* bi-clique (IGDB). IGDB is a node-induced subgraph consisting of a subset of SNPs and a subset of voxels, where the possibility of a SNP associated with a voxel is much elevated than the rest of graph. Within an IGDB, each voxel can be considered as a polygenic imaging trait, and a SNP as a pleiotropic genetic variant. Therefore, our method contributes as a new GWAS tool for voxel level neuroimaging traits which alleviates the burden of ultra stringent threshold (eg, $p < 5 \times 10^{-12}$ in vGWAS) and uncover the systematic SNP-trait association patterns.

With the specified IGDB structure of polygenic and pleiotropic association pattern, the current study makes several contributions. First, we develop computationally efficient algorithms to identify the IGDB structure with the scalability for analyzing the whole genome-whole brain data. Second, the proposed greedy algorithm is presented with the approximation bounds for the true optimal as well as its asymptotically full recovery of IGDB-based network structure. Last, we formulate the existence of a polygenic and pleiotropic SNP-voxel association structure against a random bipartite graph, which can be evaluated through likelihood-based statistics.

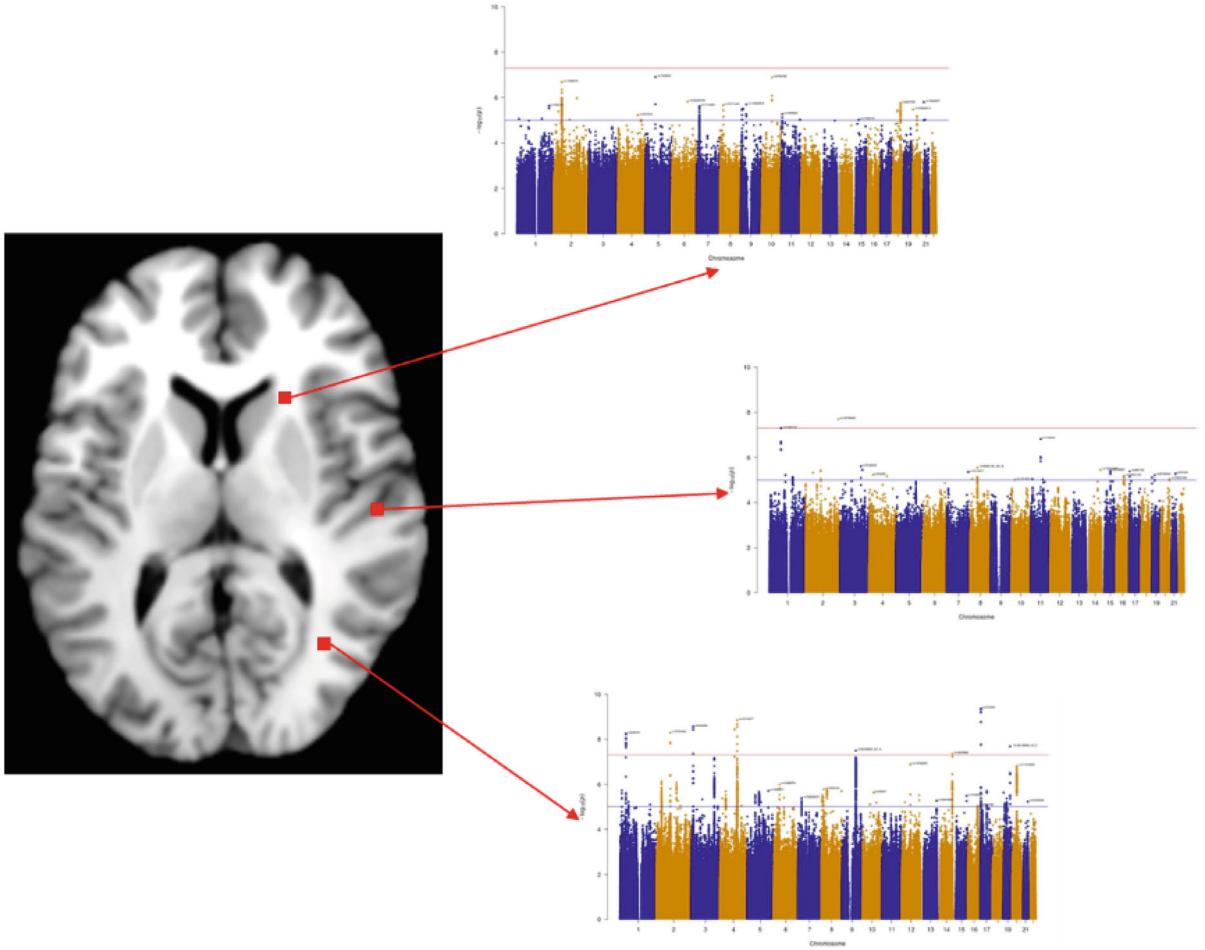


FIGURE 1 Data structure for vGWAS. For imaging-genetics data, we can perform GWAS analysis on each voxel of 3D brain imaging data for the study cohort. The vGWAS analyses generate billions of association results, which raises challenges of result interpretation and comprehension.

2 | MOTIVATING DATA EXAMPLE

The human connectome project (HCP) sponsored by National Institutes of Health (NIH) aims to construct the underlying neural pathways of healthy human brain functions. It is an important public resource for structural and functional brain connectivity data, accompanied by demographic, behavioral, genetic and other data. In this study, we focus on the brain imaging and genetics data in the HCP surveyed from 1052 participants (F/M 483/569; age 28.1 ± 3.7), for whom the scans and data were released in June 2014 (https://humanconnectome.org) that passed the HCP and ENIGMA quality control and assurance standards.[26] The participants in the HCP study were recruited from a large population-based study named "the Missouri Family and Twin Registry."[27]

The fractional anisotropy (FA) measure, derived from diffusion tensor imaging (DTI), is a widely-used metric characterizing the localized white matter microstructural integrity.[28] Previous studies have investigated the heritability through variance components method of pedigrees.[29] They find that 70% to 80% of the total phenotypic variance of trait-wise FA measures can be explained by additive genetic factors.[30] The significantly and reliably heritable FA measurements are qualified as a set of endophenotypes which suggests further exploration on associated genetic variants. Hence, the genetic analysis is desirable to detect the genetic effect from specific loci on imaging traits with statistical inference. Moreover, it is reported that FA measurements at multiple brain locations can be affected by a common set of genetic variates.[9] FA is a complex trait determined by multiple alleles. It stimulates the identification of functionally-related genetic variants. This investigation naturally invokes the search for polygenicity and pleiotropy of networks as the focus of this study. Voxel-level association analysis between imaging traits and genetic variants can provide the maximal spatial resolution.

Nevertheless, the implementation is challenging because it requires a multivariate to multivariate association analysis to extract SNP-voxel subnetworks with polygenic and pleiotropic structures and further to provide sound statistical inference. To close this gap, we develop an IGDB-based framework to perform voxel-vise GWAS and systematically identify polygenic and pleiotropic structures.

3 | METHODS

3.1 | Background and notations

We consider an imaging-genetics data set collected from L independent subjects. We let V be the set of brain imaging voxels with $|V| = n$ and U be the set of genetic variants (ie, SNPs) with $|U| = m$. For each participant $l \in \{1, \dots, L\}$, define $\boldsymbol{x}_l = (x_{1,l}, \dots, x_{m,l})^T$ to be the genetic variants for the participant l and $\boldsymbol{y}_l = (y_{1,l}, \dots, y_{n,l})^T$ to be the vector of multivariate imaging traits. Let $\boldsymbol{z}_l$ denote a p-dimensional vector of individual-level profiling covariates. We model the associations between multivariate imaging traits and multivariate genetic variants using a generalized linear regression model:

$$\mathbb{E}(\boldsymbol{y}_l|\boldsymbol{x}_l) = g^{-1}(\boldsymbol{B}^T\boldsymbol{x}_l + \boldsymbol{\alpha}^T\boldsymbol{z}_l),$$

where $g(\cdot)$ is a known link function with inverse $g^{-1}(\cdot)$. The coefficient $\boldsymbol{B} = \{\beta_{uv}\}_{u\in U, v\in V} \in \mathbb{R}^{m\times n}$ is called the *SNP-voxel association matrix*. Without loss of generality, we consider the association matrix based on GWAS analysis (eg, using open-source whole genome association analysis toolset).[31] The goal of our statistical inference is to accurately identify the subset of significant associations $\{(u, v) : \beta_{uv} \neq 0\}$ from billions of entries of $\boldsymbol{B}$ by multivariate to multivariate hypothesis testing[32, 33]:

$$H_0^{(u,v)} : \beta_{uv} = 0, \text{ vs } H_1^{(u,v)} : \beta_{uv} \neq 0, \quad \text{for all } u \in U, v \in V.$$

Conventional statistical inference methods (eg, multiple testing correction or regression shrinkage) work by regularizing vectorized $\boldsymbol{B}$. However, this strategy may only capture individual association pairs β_{uv} without recognizing systematic patterns (eg, the pleiotropic and polygenic structure). A prominent example is that a cluster of SNPs may jointly influence the observations through a cluster of neighboring voxels. To address this challenge, we propose a new multivariate to multivariate inference framework that extracts the joint structure in $\boldsymbol{B}$, which we call *imaging-genetics dense bi-clique (IGDB)*. Next, we introduce the IGDB structure, based on which, we then formally propose a novel estimation and inference procedure on this structure.

3.2 | IGDB in a multivariate to multivariate graph structure

We characterize the vGWAS association as a bipartite graph $G = (U, V, E)$, where U and V are distinct node sets representing SNPs and voxels, respectively. The set of binary edges E describes the locations of significant SNP-voxel associations: $e_{uv} \in E$ if and only if $\beta_{uv} \neq 0$ in the association matrix $\boldsymbol{B} = \{\beta_{uv}\}_{u\in U, v\in V}$. In contrast to conventional approaches that treat edges e_{uv} individually, our proposal provides a succinct description of pleiotropic (one SNP to multiple image voxels) and polygenic (multiple SNPs to one voxel) relationships. To this end, we now formally propose IGDB as a subgraph structure of G. Denote an arbitrary subgraph of G by $G[S, T] = (S, T, E[S, T])$, where $S \subset U$, $T \subset V$ and $E[S, T] = \{e_{uv} \in E | i \in S, j \in T\}$. Our proposed IGDB will be defined based on some particular subgraph $G[S_0, T_0]$ such that most β_{uv}'s are nonzero for $e_{uv} \in G[S_0, T_0]$, while most $\beta_{u'v'}$'s elsewhere are zero. We illustrate the IGDB structure of a bipartite graph in Figure 2.

Our core intuition can be quantified into the following formulation:

$$\frac{\sum_{u,v} I(\beta_{uv} \neq 0 | \delta_{uv} = 1)}{\sum_{u,v} I(\delta_{uv} = 1)} > \frac{\sum_{u,v} I(\beta_{uv} \neq 0 | \delta_{uv} = 0)}{\sum_{u,v} I(\delta_{uv} = 0)}, \tag{1}$$

where δ_{uv} is a binary variable indicating the IGDB-based network structure, that is,

$$\delta_{uv} \equiv \delta_{uv}(S_0, T_0) = I(e_{uv} \in G[S_0, T_0]).$$

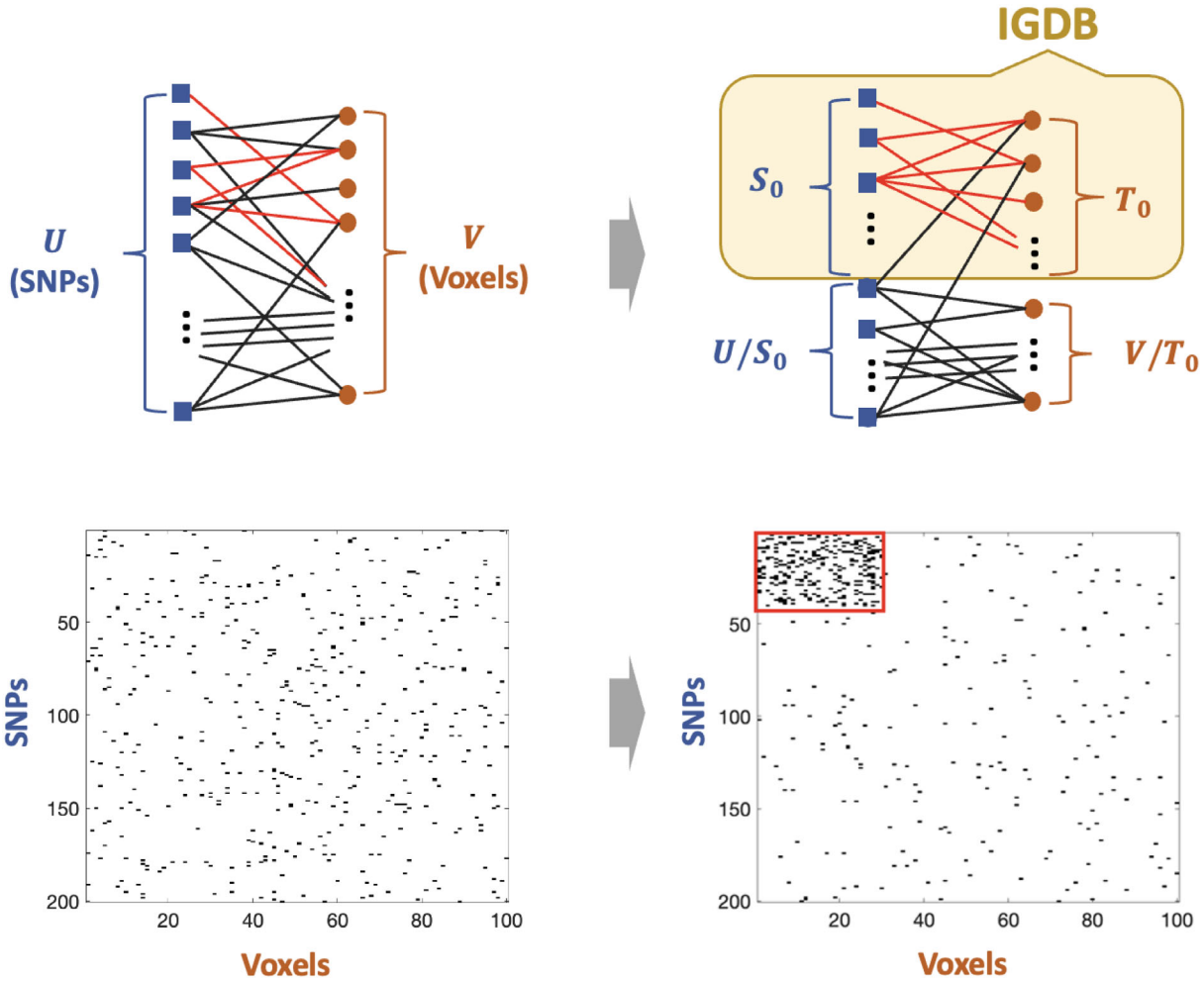


FIGURE 2 Illustration of a bipartite graph with IGDB structure $G[S_0, T_0]$ that reveals underlying patterns of massive SNP-voxel association. In the top-left bipartite graph, each node (square) on the left side represents an SNP, while each node (circle) represents a location-specific voxel. The edges connecting the SNPs and voxels illustrate the associations, with red edges indicating pairs of associated SNPs and voxels in an IGDB structure. The bottom-left 2D figure provides an alternative representation of SNP-voxel associations, where associated pairs are depicted as black dots. The SNP-voxel association patterns in the left figures appear to be random. The bottom-right figure showcases the patterns that can be unveiled through the proposed IGDB method, suggesting systematic associations between imaging features and genetic variants. Note that traditional statistical methods, such as bi-clustering, face limitations in accurately identifying these patterns (see Figure A1 in the Appendix).

This reflects that imaging features (T_0) are polygenic traits and the genetic variants (S_0) are pleiotropic alleles. The genetically correlated imaging features and functionally related SNPs jointly compose a functional biclique $G[S_0, T_0]$. In neuroimaging studies, findings are often reported for spatially contiguous brain areas (ie, connected voxels) because of the biological interpretability and inference advantages.[34] This is reflected in our proposed IGDB structure by further formulating S_0 and T_0 as disjoint vertex neighborhoods, as follows:

$$S_0 = \mathcal{N}_1^{S_0} \cup \ldots \cup \mathcal{N}_{K_1}^{S_0}, \quad \text{and} \quad T_0 = \mathcal{N}_1^{T_0} \cup \ldots \cup \mathcal{N}_{K_2}^{T_0},$$

where each $\mathcal{N}_k^{T_0}$ ($k \in \{1, \ldots, K_2\}$) is a spatially contiguous voxel cluster, and accordingly $\mathcal{N}_k^{S_0}$ ($k \in \{1, \ldots, K_1\}$) is a set of functionally related SNPs associated with one or multiple spatially-contiguous voxel clusters (eg, $\mathcal{N}_k^{T_0}$). In the next subsection, we articulate that the IGDB enjoys several statistical advantages supported by graph and combinatorics theory.

3.3 | Graph properties of IGDB

Without loss of generality, we consider the following two cases regarding the underlying network structure of G:

$$\begin{aligned}
&\textbf{Case 0}: \quad G \text{ is observed from a random bipartite graph } G(m, n, \mu_0),\\
&\textbf{Case 1}: \quad \text{There exists at least one non-trivial IGDB } G[S_0, T_0] \text{ such that} G \text{is observed from}\\
&\qquad e_{uv} = I(\beta_{uv} \neq 0) \sim \begin{cases} \text{Bernoulli}(\mu_1), & \text{if } u \in S_0 \ \& \ v \in T_0 \\ \text{Bernoulli}(\mu_0), & \text{otherwise} \end{cases}\\
&\qquad \text{with } \mu_1 > \mu_0.
\end{aligned} \tag{2}$$

In Case 0 (ie, no polygenic and pleiotropic patterns), we can directly implement the conventional multiple testing corrections and regression shrinkage methods to determine individual associations between genetic variants and imaging traits. If Case 1 presents, our primary goal becomes to extract and test the underlying IGDB subgraphs as polygenic and pleiotropic subnetworks.

In practice, the estimated IGDB from a sample can be used to distinguish Case 0 vs Case 1 because the observed network behave differently under two cases on the size of the maximal "dense" subgraph. For convenience, we call a subgraph $G[S, T]$ a *γ-quasi biclique*, if it contains at least $\gamma \cdot |S| \cdot |T|$ edges. Then, asymptotically, if $|S_0|, |T_0| \to \infty$ as $m, n \to \infty$, with high probability, the true IGDB subgraph $G[S_0, T_0]$ would be a γ-quasi biclique for any fixed $\gamma \in (\mu_0, \mu_1)$. In contrast, under Case 0, there would rarely exist a γ-quasi biclique of decent size with high density as the following lemma.

Lemma 1. *Suppose G is observed from a random bipartite graph $G(m, n, \mu_0)$ as Case 0. $G[S, T]$ is any subgraph with edge density $\frac{|E[S,T]|}{|S||T|} \geq \gamma \in (\mu_0, 1)$ (ie, γ-quasi biclique). Let $m_0, n_0 = \Omega(\max\{m^\epsilon, n^\epsilon\})$ for some $0 < \epsilon < 1$. Then for sufficiently large m, n with $c(\gamma, \mu_0)m_0 \geq 8 \log n$ and $c(\gamma, \mu_0)n_0 \geq 8 \log m$, we have*

$$\mathbb{P}(|S| \geq m_0, |T| \geq n_0) \leq 2mn \cdot \exp\left(-\frac{1}{4}c(\gamma, \mu_0)m_0 n_0\right),$$

where $c(a, b) = \left\{ \frac{1}{(a-b)^2} + \frac{1}{3(a-b)} \right\}^{-1}$.

4 | ESTIMATION AND INFERENCE

Let $\boldsymbol{W}_{m\times n}$ denote the inference result matrix (eg, test statistics $w_{uv} = t_{uv}$ or $-\log(p_{uv})$) for the regression coefficients $\widehat{\boldsymbol{B}}_{m\times n}$. Then, our goal becomes to extract and test the IGDB structure from a weighted bipartite graph $G = (U, V, \boldsymbol{W})$. Similar to Reference 33, as a natural consequence of our model set up in Section 3.2, edge weights in $\boldsymbol{W}$ follow a mixture marginal distribution:

$$w_{uv} \sim \begin{cases} f_1(\cdot; \boldsymbol{\theta}_1), & \text{if } \beta_{uv} \neq 0 \\ f_0(\cdot; \boldsymbol{\theta}_0), & \text{if } \beta_{uv} = 0. \end{cases} \tag{3}$$

where $w_{uv}|\delta_{uv} = 1 \sim \mu_1 f_1 + (1 - \mu_1)f_0$, while $w_{uv}|\delta_{uv} = 0 \sim \mu_0 f_1 + (1 - \mu_0)f_0$. Empirically, we have the central tendency of $f_1(\cdot; \boldsymbol{\theta}_1)$ being greater than $f_0(\cdot; \boldsymbol{\theta}_0)$, in the sense that $\mathbb{E}_{\theta_1}[w_{uv}|\beta_{uv} \neq 0] > \mathbb{E}_{\theta_0}[w_{uv}|\beta_{uv} = 0]$.

4.1 | IGDB estimation

Motivated by the nature of IGDB as a subgraph of elevated mean edge weights, we estimate it by looking for the maximal subgraph of G with a density constraint. Inspired by Lemma 1, we estimate the IGDB $G[S_0, T_0]$ based on the edge weight matrix $\boldsymbol{W}$ by optimizing:

Algorithm 1. Direct optimization of objective function (5)

Input: $G = (U, V, \boldsymbol{W})$, λ, pre-specified ratio $h \in \{h_1, h_2, \ldots, h_{\max}\}$;**Output:** $G[\tilde{S}_\lambda, \tilde{T}_\lambda]$
procedure Algorithm
  for $h \in \{h_1, h_2, \ldots, h_{\max}\}$ **do**
    $S_1 \leftarrow U$, $T_1 \leftarrow V$
    for k = 1 to $n + m - 1$ **do**
      Let $i \in S_k$ be the node with smallest degree: $i = \arg\min_{i' \in S_k} \deg_X(i'; S_k, T_k)$;
      Let $j \in T_k$ be the node with smallest degree: $j = \arg\min_{j' \in T_k} \deg_Y(j'; S_k, T_k)$;
      if $\sqrt{h}\deg_X(i; S_k, T_k) \le \frac{1}{\sqrt{h}}\deg_Y(j; S_k, T_k)$ **then**
        $S_{k+1} \leftarrow S_k/\{i\}$ and $T_{k+1} \leftarrow T_k$;
      else
        $S_{k+1} \leftarrow S_k$ and $T_{k+1} \leftarrow T_k/\{j\}$;
      end if
    end for
    Output $G[S^h, T^h]$ with largest objective function in $G[S_1, T_1], \ldots, G[S_{n+m-1}, T_{n+m-1}]$;
  end for
  Output $G[\tilde{S}_\lambda, \tilde{T}_\lambda]$ with largest objective function in $G[S^{h_1}, T^{h_1}], \ldots, G[S^{h_{\max}}, T^{h_{\max}}]$;
end procedure

$$\max_{S \subseteq U, T \subseteq V} |S||T| \qquad \text{subject to } \frac{\|\boldsymbol{W}[S,T]\|_{1,1}}{|S||T|} \ge \gamma' \tag{4}$$

or the Lagrangian form after taking logarithm on both terms:

$$\max_{S \subseteq U, T \subseteq V} \log(|S||T|) + \lambda \log\left(\frac{\|\boldsymbol{W}[S,T]\|_{1,1}}{|S||T|}\right), \tag{5}$$

where $\|\cdot\|_{1,1}$ refers to the entry-wise ℓ_1 norm such that $\|\boldsymbol{W}[S,T]\|_{1,1} = \sum_{u \in S, v \in T} |w_{uv}|$, γ' is the density constraint and the tuning parameter $\lambda \in (1, \infty)$.

The direct optimization of the objective function (5) is challenging because it is a nondeterministic polynomial (NP) problem.[35,36] We propose a computationally efficient greedy algorithm to approximately carry out the optimization of (5). We describe the greedy algorithm as Algorithm 1 in the following. In designing it, we extended the greedy algorithms for dense subgraph discovery[36] in an adjacency matrix to a large bipartite matrix to extract dense bi-cliques. Algorithm 1 removes nodes with the smallest degrees iteratively, which is a deterministic algorithm that does not depend on initial values. The computational complexity of Algorithm 1 is $O(C_1 mn)$, where C_1 is determined by the grid search of h, that is, $h = |S|/|T|$, representing the aspect ratio of a dense subgraph, in the following Algorithm 1.

Now we establish approximation accuracy results of Algorithm 1 and its estimation of IGDB. Let S^*_λ and T^*_λ be the true optimal solution to (5):

$$(S^*_\lambda, T^*_\lambda) = \arg\max_{S \subset U, T \subset V} d_\lambda(S, T),$$

and $(\tilde{S}_\lambda, \tilde{T}_\lambda)$ is from Algorithm 1 with

$$(\tilde{S}_\lambda, \tilde{T}_\lambda) = \arg\max_h \arg\max_{(S_1,T_1), \ldots, (S_{m+n-1}, T_{m+n-1})} d_\lambda(S, T),$$

where $d_\lambda(S,T) := \log(|S||T|) + \lambda \log\left(\frac{\|\boldsymbol{W}[S,T]\|_{1,1}}{|S||T|}\right)$.

The greedy algorithm with average-degree based density (or equivalently $\lambda = 2$) is said to have a 2-approximation guarantee for the true optimal,[35] namely, $2d_2(\tilde{S}_2, \tilde{T}_2) > d_2(S^*_2, T^*_2)$. In this article, we present the approximation bounds for the proposed objective function (5) in terms of a parameter λ as the following Theorem 1.

Algorithm 2. Determine tuning parameter λ by likelihood function

Input: $G = (U, V, \boldsymbol{W})$, a grid of tuning parameters: $\lambda_1, \lambda_2, \ldots, \lambda_J$, a sequence of cutoffs $r_1, r_2, \ldots, r_R$ and its mass $q(r_1), \ldots, q(r_R)$; **Output:** $G[\tilde{S}_{\hat{\lambda}}, \tilde{T}_{\hat{\lambda}}]$ and $\hat{\lambda}$
procedure Algorithm
  while $\lambda \in \{\lambda_1, \ldots, \lambda_J\}$ **do**
    Return the IGDB $(\tilde{S}_\lambda, \tilde{T}_\lambda)$ of $\boldsymbol{W}$ from Algorithm 1
    for $r = r_1$ to r_R **do**
      calculate the likelihood defined in 4.2: $\mathcal{L}_\lambda(\hat{\boldsymbol{\pi}}; \tilde{S}_\lambda, \tilde{T}_\lambda, \boldsymbol{W}(r))$ (We refer to Section 4.2 for detailed definition of the likelihood function.)
    end for
    integrate w.r.t. r:
    $\mathcal{L}_\lambda(\boldsymbol{W}) = \sum_{i=1}^{R} \mathcal{L}_\lambda(\hat{\boldsymbol{\pi}}; \tilde{S}_\lambda, \tilde{T}_\lambda, \boldsymbol{W}(r_i))q(r_i)$
  end while
  Output $\hat{\lambda}$ and $(\tilde{S}_{\hat{\lambda}}, \tilde{T}_{\hat{\lambda}})$ with maximized $\mathcal{L}_\lambda(\boldsymbol{W})$
end procedure

Theorem 1. *For a given bipartite graph $G = (U, V, E)$, with $(S^*_\lambda, T^*_\lambda)$ and $(\tilde{S}_\lambda, \tilde{T}_\lambda)$ defined in Section 3.1.1, the greedy algorithm 1 has a $\rho(\lambda, m, n)$-approximation, that is, $d_\lambda(S^*_\lambda, T^*_\lambda) \le \rho(\lambda, m, n) d_\lambda(\tilde{S}_\lambda, \tilde{T}_\lambda)$ with*

$$\rho(\lambda, m, n) = \begin{cases} 2(mn)^{\frac{1}{\lambda}\left(1-\frac{2}{\lambda}\right)} & \text{if } \lambda \ge 2 \\ 2(mn)^{\left(\frac{1}{\lambda}-\frac{1}{2}\right)} & \text{if } \frac{4}{3} < \lambda < 2 \\ (mn)^{\left(1-\frac{1}{\lambda}\right)} & \text{if } 1 < \lambda \le \frac{4}{3}. \end{cases}$$

In Theorem 2, we state that the optimization of the proposed objective function (5) asymptotically leads to almost full recovery of the IGDB-based network structure.

Theorem 2. *We assume the graph $G = (U, V, E)$ with an IGDB $G[S_0, T_0] = (S_0, T_0, E[S_0, T_0])$ is generated from a mixture of Bernoulli distributions*: $e_{uv} \sim \delta_{uv}\ \text{Bernoulli}(\pi_1) + (1 - \delta_{uv})\ \text{Bernoulli}(\pi_0)$, $\delta_{uv} = I(e_{uv} \in G[S_0, T_0])$ *and $\pi_1 > \pi_0$. For simplicity, we let $m = \Theta(n)$. Assume $|S_0| = O(|m|^{1/2+\epsilon})$ and $|T_0| = O(|n|^{1/2+\epsilon})$ as $n \to \infty$ for some $\epsilon > 0$. Denote*

$$e_S = \left(1 - \frac{\tilde{S}_\lambda \cap S_0}{S_0}\right) + \left(1 - \frac{\tilde{S}^c_\lambda \cap S^c_0}{S^c_0}\right)$$

and

$$e_T = \left(1 - \frac{\tilde{T}_\lambda \cap T_0}{T_0}\right) + \left(1 - \frac{\tilde{T}^c_\lambda \cap T^c_0}{T^c_0}\right)$$

to be the error rates of node memberships based on $(\tilde{S}_\lambda, \tilde{T}_\lambda)$ from Algorithm 1. Then, there exists some λ such that we will get almost full recovery in Algorithm 1, that is, for any fixed $a \in (0, 1)$, as $n \to \infty$, we have

$$\mathbb{P}(e_S + e_T \ge a) \to 1.$$

In practice, we select the tuning parameter λ by a grid search based on the likelihood criterion,[37] and describe the details in Algorithm 2. Based on each dense subgraph $G[S, T]$, we further identify spatially-contiguous voxel clusters (ie, $\tilde{\mathcal{N}}^T_k$, $k = 1, \ldots, \tilde{K}_2$), and a corresponding set of SNPs (ie, $\tilde{\mathcal{N}}^S_k$, $k = 1, \ldots, \tilde{K}_1$) that are functionally associated with voxel clusters (see Supplement A). Last, multiple IGDBs can be extracted by performing algorithms repeatedly with the detected IGDBs masked.[38]

4.2 | Statistical inference of the IGDB

Recall that the purpose of this study is to perform statistical inference on the pleiotropic and polygenic association pattern or the IGDB. We investigate the significance of the presence of an IGDB against a random bipartite graph (Case 1 vs Case 0) as illustrated in Section 3.3.

Let r be a sound cutoff that dichotomize the weighted graph G into a binary graph $G^r = (U, V, \boldsymbol{A})$ using $a_{uv} = I(|w_{uv}| > r)$. Then, under IGDB structure indexed by node sets (S_0, T_0), the edges in G^r follow a mixture of two Bernoulli distributions:

$$a_{uv}|(S_0, T_0) \sim \text{Bernoulli}(\pi_{uv}), \tag{6}$$

where $\pi_{uv} = \delta_{uv}\pi_1 + (1-\delta_{uv})\pi_0$, $\pi_1 = \mu_1 \int_r^\infty f_1(w, \boldsymbol{\theta}_1)dw + (1-\mu_1)\int_r^\infty f_0(w, \boldsymbol{\theta}_0)dw$, $\pi_0 = \mu_0 \int_r^\infty f_1(w, \boldsymbol{\theta}_1)dw + (1-\mu_0)\int_r^\infty f_0(w, \boldsymbol{\theta}_0)dw$, and $\pi_1 > \pi_0$.[39] Then, a hypothesis testing to distinguish Cases 0 and 1 can be proposed:

$$H_0 : \pi_1 = \pi_0 = \pi \quad \text{vs} \quad H_1 : \pi_1 > \pi_0,$$

based on our mixture distribution model (6).

We propose a likelihood-based statistic for the IGDB test. For a binarized graph G^r, let

$$t_G = \log \frac{\sup\limits_{H_0 \cup H_1} \mathcal{L}(\boldsymbol{\pi}; S, T, \boldsymbol{A})}{\sup\limits_{H_0} \mathcal{L}(\pi; \boldsymbol{A})},$$

with likelihood given by Bernoulli distributions in (6). Specifically,

$$\mathcal{L}(\boldsymbol{\pi}; S, T, \boldsymbol{A}) = \prod_{u \in S \text{ and } v \in T} \pi_1^{a_{uv}}(1-\pi_1)^{1-a_{uv}} \times \prod_{u \in U/S \text{ or } v \in V/T} \pi_0^{a_{uv}}(1-\pi_0)^{1-a_{uv}}$$
$$\text{and } \mathcal{L}(\pi; \boldsymbol{A}) = \prod_{u \in U \text{ and } v \in V} \pi^{a_{uv}}(1-\pi)^{1-a_{uv}}.$$

Then, the asymptotic power is ensured using the likelihood-based statistic through the following Theorem 3.

Theorem 3 (Under IGDB alternative hypothesis H_1). *Assume $m = \Theta(n)$ and the underlying IGDB $G[S_0, T_0]$ with generating probabilities $\pi_1 > \pi_0$ satisfies $|S_0| = m_0$, $|T_0| = n_0$ and $m_0, n_0 = \Omega(n^\epsilon)$ for some $\epsilon > 0$. Then for any $\eta > 1$, as $n \to \infty$, we have*

$$\Pr(t_G > \eta) \to 1.$$

In determining the significance of IGDBs, the simultaneous testing needs to be accounted for all potential IGDBs. Besides, a rejection region (η) should be determined based on the distribution of t_G under null model. Hence, we employ the commonly used permutation test procedure in the field of neuroimaging[40,41] to empirically approximate the distribution of the likelihood-based statistic t_G under the IGDB null and control the family-wise error rates (FWER).

Let $\phi(\cdot)$ be the vectorization of a matrix, such that $\phi(\boldsymbol{A})$ is an mn vector of the adjacency matrix $\boldsymbol{A}$. Denote τ as a permutation of mn elements, and P_τ is the corresponding permutation matrix. Let $G_\tau = (U, V, E_\tau)$ an edge-permuted graph from G. Then, under random bipartite graph (Case 0), the edge-permuted graph G_τ would be a realization from the same null model. We let $\tau(1), \ldots, \tau(M)$ be M random permutations and the corresponding edge-permuted adjacency matrices are given by $A_{\tau(1)}, \ldots, A_{\tau(M)}$. The test statistics associated with edge-permuted adjacency matrices $A_{\tau(1)}, \ldots, A_{\tau(M)}$ forms a random sample of t_G under null hypothesis, which can be utilized to obtain the empirical distribution of t_G under null hypothesis. We illustrate whole procedure of the permutation test in Algorithm 3, while the P-values of multiple IGDBs can be observed by considering each IGDB individually.

To dichotomize the weighted graph G, rather than setting r as a fixed value, which could lead to an arbitrary selection, we consider r as a random variable with a distribution $q(r)$. This allows us to integrate the likelihood function over r,

Algorithm 3. Implementation of likelihood ratio statistic via permutation tests

Input: $G = (U, V, \boldsymbol{A})$, $\hat{S}$, $\hat{T}$; **Output:** p-value
procedure Algorithm
  calculate the test statistic on G with subgraph $G[\hat{S}, \hat{T}]$ and denote as: t_0
  for $b = 1$ to M **do**
    generate permutation matrix P_b on mn elements
    observe adjacency matrix of edge-permuted graph G_b: $\boldsymbol{A}_b = \phi^{-1}(P_b \phi(A))$
    calculate the test statistic on G_b as: t_b
  end for
end procedure

utilizing the prior distribution $q(r)$, thereby making our optimization process robust to the specific choice of r. We implement a discrete distribution for $q(r)$, defined by a set of possible values $r_1, \ldots, r_R$ and their corresponding probabilities $q(r_1), \ldots, q(r_R)$. In practice, our algorithm demonstrates robustness to the choice of the prior distribution, given that a reasonable range for the support of r is selected.

5 | RESULTS

We applied the IGDB approach to the motivating data set. The FA measures of DTI at 117 139 voxels were used in this study to characterize the white matter integrity.[30,42] The image acquisition parameters are described in the Supplement A. Regarding genetic variants, 10 595 779 SNPs passed the quality control filters in HCP data set (MAF < 0.01; HQE < 1e−6; r-squared > 0.03; call rate > 0.95) after imputation on the Michigan Imputation Server Minimac3 (https://imputationserver.sph.umich.edu) using the 1000 Genomes Project (phase 1 v3) reference set.[43]

We preprocessed the diffusion weighted images following the ENIGMA-DTI workflow (http://enigma.ini.usc.edu/protocols/dti-protocols/). We further applied the Sequential Oligogenic Linkage Analysis Routines (SOLAR)-Eclipse software (https://www.nitrc.org/projects/se_linux) for the heritability analysis, of which imaging voxels were kept with significant heritability, based on the Fast and Powerful Heritability Inference (FPHI) function of SOLAR-Eclipse ($P <$ 0.05) in both the HCP and Amish Connectome Project (ACP). For these voxels, we performed vGWAS using *PLINK* while adjusting covariates including sex, age, BWI, and population characteristics using the first 10 principal components in our application.[31] We then performed sure independence screening on SNPs with multiple imaging responses through a direct extension of univariate screening procedure.[44] 13 498 SNPs across 22 chromosomes survive into further analysis. The details are described in the Supplement A.

We tested the imaging-genetic associations between SNPs across 22 chromosomes and voxel-level imaging traits using our proposed method. Based on the procedures described in Sections 4.1 and 4.2, we extracted IGDBs and performed permutation tests to determine its statistical significance while controlling family-wise error rate ($q < 0.05$). We observe different brain areas being influenced by distinct genetic loci. A Manhattan plot for all SNPs across 22 chromosomes with selected imaging-genetic associations highlighted and tables for SNP and voxels across all 22 chromosomes are included in the Supplement A.

In this section, we focus on SNPs on chromosome 1 to demonstrate their systematic association patterns with voxel-traits, and then annotate the genes in the detected IGDB. Based on the matrix of association strength $\boldsymbol{W}_{1178\times29627}$ (ie, Figure 3A), we detected an IGDB with 384 SNPs and 3803 voxels as Figure 3B by maximizing the objective function (5), which is achieved by implementing Algorithm 2 utilizing a grid search for h across the range $\{1/20, 1/19, \ldots, 1, 2, \ldots, 19, 20\}$, and for λ within the interval 0.5 to 1.2, with an incremental step of 0.02. We further calculated the p value for the IGDB statistical inference via the permutation test, which results in a significant existence of an IGDB with P value < 0.001. Although the IGDB is an irreducible subgraph, it can be further refined based on data-driven algorithms and spatial information of imaging data. We applied the existing community detection algorithms[45] on similarity matrices observed from the detected IGDB. The refined pattern in Figure 3C displays 6 distinct SNP-voxel association clusters. Note that the refined structure cannot be identified without revealing the IGDB by the proposed algorithm.

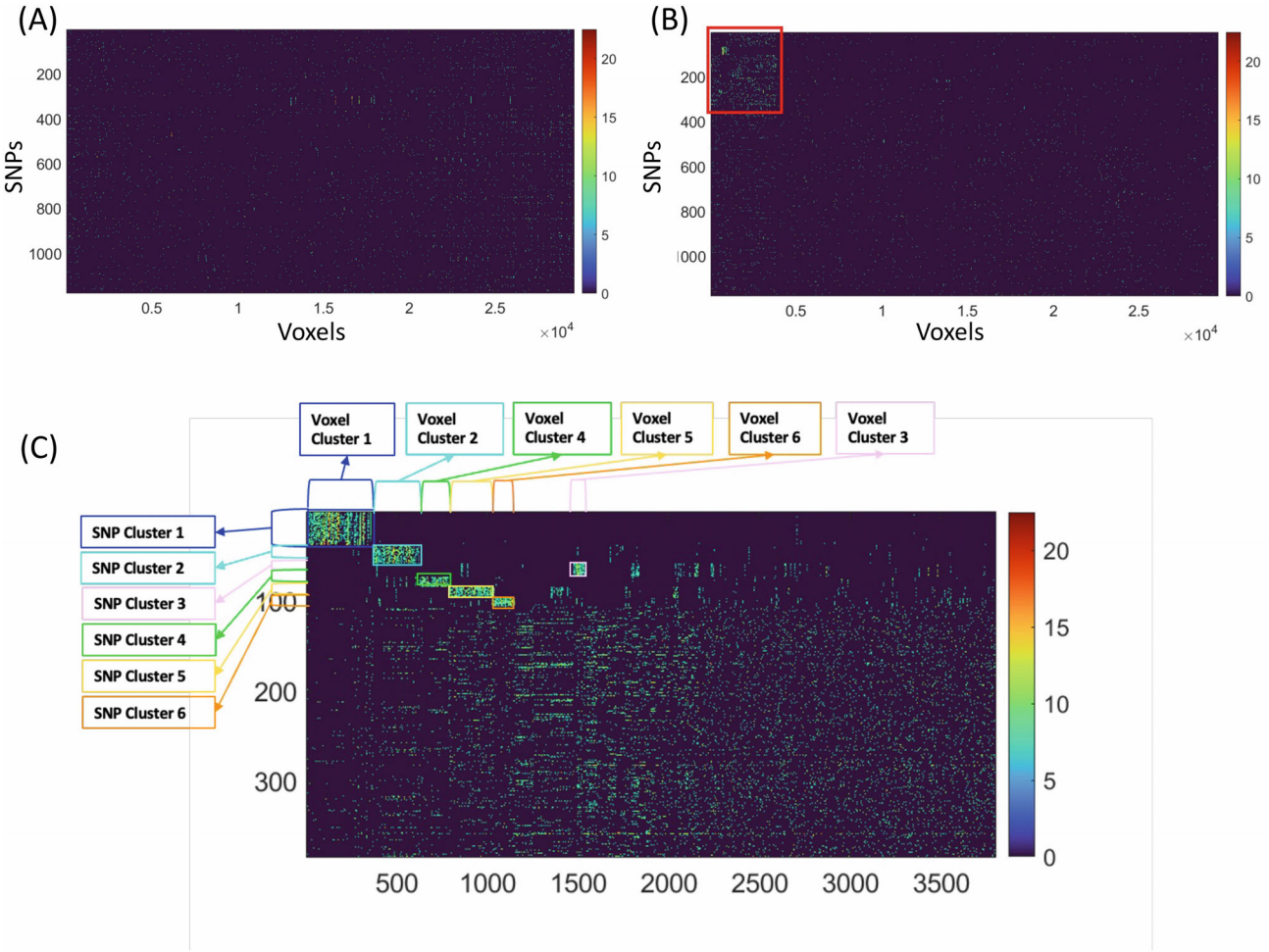


FIGURE 3 IGDB procedure on chromosome 1: (A) is the input matrix $\boldsymbol{W}$, derived from vGWAS using *PLINK* while adjusting covariates including sex, age, BWI, and population characteristics using the first 10 principal components in our application.[31] Each entry in the matrix is $-\log(p_{ij})$ of the association between an SNP and imaging voxel pair (ie, a hotter entry indicates a higher level of SNP-voxel association). Although $\boldsymbol{W}$ is obtained after screening (eg, by voxel-level heritability analysis), it remains challenging to directly recognize the patterns of imaging-genetics associations; (B) demonstrates the detected IGDB which reveals *dense* blocks of imaging-genetics associations; (C) displays the refined pattern of the IGDB. In panels (B) and (C), we have reordered the SNPs and voxels to better illustrate their patterns of association.

As a greedy algorithm, the computational complexity of Algorithm 1 is linear in the size of the original graph. By determining the tuning parameters through the likelihood function, as outlined in Algorithm 2, the computation remains efficient, which took 20 minutes on a PC with an i7 CPU 3.60 GHz and 64 GB memory to detect the IGDB of the SNP-voxel association graph in chromosome 1. The computation of the p-value is dependent on the number of permutations, which can be easily parallelized for efficient computation.

We illustrate the voxel clusters and corresponding SNP sets in Figure 4. For example, the voxel cluster 2 (colored cyan) includes voxels mainly from the splenium of corpus callosum (SCC), part of one of the largest white matter tracts that connects many parts of the brain, and which lesions to often result in many varied neurological issues.[46] To annotate the SNPs in the identified clusters, we queried the SNPs in the QTLbase (http://mulinlab.org/qtlbase/index.html,[47]) for potential expression quantitative trait locus (eQTL) and examined the genes being regulated by these variants in a tissue-specific pattern. The summary of associated genes related with brain tissues is displayed in the Supplement A as supporting information. In cluster 1, multiple SNPs are linked with the LEPR gene, a protein coding gene for leptin receptor generation that has been shown to be associated with obesity. It has been known the white matter integrity is highly associated with obese disorder and body mass index.[48] Therefore, this cluster reveals the marginal association of (obesity-related) LEPR gene and white matter integrity. In clusters 2 to 5, the associated genes, for example, S100A1, TAF1A, CFH, CFHR3,

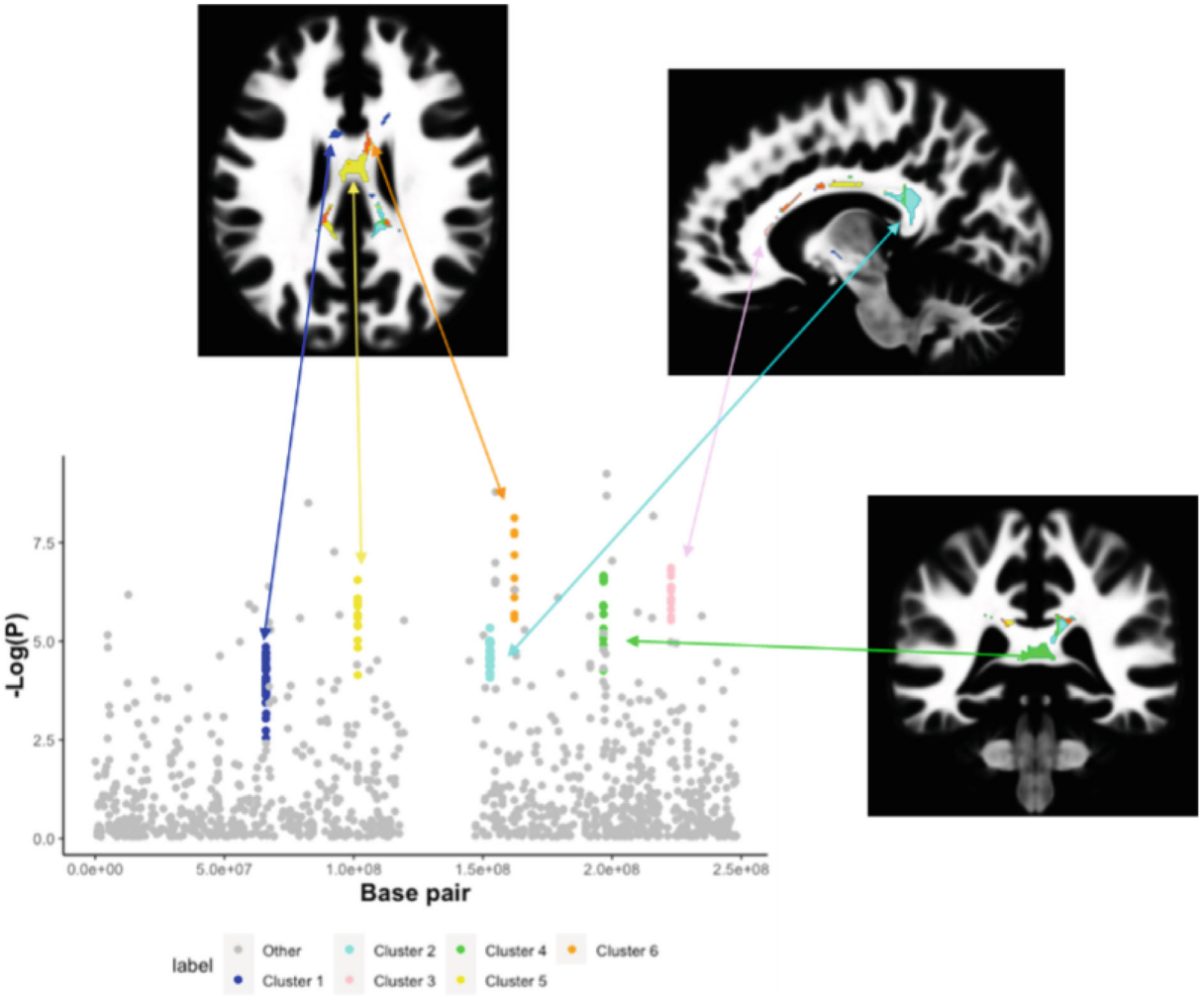


FIGURE 4 An illustration of the association patterns between SNP and voxel clusters on Chromosome 1. We demonstrate the systematic imaging-genetics associations in an integrated Manhattan plot based on the results of our analysis by IGDB. The highlighted subsets of SNPs are systematically associated with corresponding areas of the white matter tracks. The dual localized association patterns provide a straightforward interpretation of the genetic effects on location-specific brain areas.

and DPH5 are associated with immune system functions (http://immunet.princeton.edu/, https://www.innatedb.com/moleculeSearch.do). White matter integrity can be influenced by the immune system functions and systematic inflammation. In cluster 6, the NOS1AP gene has been found to be associated with white matter microstructure in previous studies.[8] In addition, the NOS1AP gene is identified to be a risk factor for schizophrenia,[49] while the alterations of white matter integrity for patients with schizophrenia were studied in Kubicki et al.[50] In summary, our findings provided insights into the complex neurogenetic mechanisms of how genetic variants influence imaging traits in a systematic fashion potentially via regulating gene expression and generated hypotheses to be further confirmed in future multi-omics studies.

6 | SIMULATION

6.1 | Synthetic data

We evaluate the finite-sample performance of our proposed method based on simulation studies. We generate the input matrix $\boldsymbol{W}_{m\times n}$ based on the two sets of multivariate variables representing genetic variants $\boldsymbol{X}_{m\times L}$ and imaging voxels $\boldsymbol{Y}_{n\times L}$. We let the pattern of $\boldsymbol{W}_{m\times n}$ be determined by a graph $G = (U, V, E)$. Specifically, we assume there exists an IGDB

$G[S_0, T_0] = (S_0, T_0, E[S_0, T_0])$ with higher proportion of edges as significant imaging-genetics associations (ie, μ_1) than the rest of graph (ie, μ_0). Then, we let the entries of $\boldsymbol{W}_{m\times n}$ follow mixture distributions according to G as $w_{uv}|\delta_{uv} = 1 \sim \mu_1 t_{df}(\nu) + (1-\mu_1)t_{df}(0)$, $w_{uv}|\delta_{uv} = 0 \sim \mu_0 t_{df}(\nu) + (1-\mu_0)t_{df}(0)$, where δ_{uv} is an indicator variable with $\delta_{uv} = 1$ for edges in the IGDB and 0 otherwise. $t_{df}(\nu)$ and $t_{df}(0)$ are the non-null and null distributions of imaging-genetics associations respectively. $t_{df}(\nu)$ is a t distribution with the degree of freedom $L - p$ (p covariates) and non-central parameter $\nu = \frac{\theta}{\sqrt{4/L}}$, where θ is standardized effect size (eg, Cohen's d). μ_1 and μ_0 are the proportions of the non-null distribution within the IGDB and otherwise. We use $m = 200, n = 100$, and $L = 60$. We simulate data sets with multiple settings by varying the size of IGDB (ie, $(|S_0|, |T_0|) = (50, 40)$ and $(30, 20)$), standard effect size (ie, $\theta = 0.8$, 1, and 1.2), and proportions of noisy edges (ie, $(\mu_1, \mu_0) = (0.8, 0.2)$ and $(0.9, 0.1)$). Additional simulation settings with larger graph and sample sizes are included in the Appendix.

6.2 | Performance metrics and results

We evaluate the performance of proposed method at several levels. At the subgraph-level, we assess the accuracy of IGDB inference by examining if we can reject the null (ie, no systematic imaging-genetics association). At the edge-level, we evaluate the accuracy of detected IGDB by comparing it with ground truth in terms of edge differences. We also evaluated the node-assignment accuracy of the proposed method using synthetic data (see Section 1.5 of Supplement A in the Supporting information for details). The performance was only compared to Charikar's algorithm[35] for dense component extraction instead of bi-clustering algorithms. As bi-clustering algorithms tend to assign all SNPs and voxels into clusters, they are not well suited to the IGDB structure extraction (see demonstration in Appendix).

For IGDB inference, we consider a detected IGDB $G[\hat{S}, \hat{T}]$ is a recovery of the underlying IGDB $G[S_0, T_0]$ if it is rejected in the proposed likelihood-ratio test and has high similarity with $G[S_0, T_0]$. Specifically, we consider $G[\hat{S}, \hat{T}]$ is a true positive detection of $G[S_0, T_0]$ if $J_{\boldsymbol{X}} \wedge J_{\boldsymbol{Y}}$ is no less than the cutoff with

$$J_{\boldsymbol{X}} = \frac{S_0 \cap \hat{S}}{S_0 \cup \hat{S}} \text{ and } J_{\boldsymbol{Y}} = \frac{T_0 \cap \hat{T}}{T_0 \cup \hat{T}},$$

and we succeed to reject the IGDB null hypothesis in the permutation test. We display the results with cutoff of 0.8 and 0.9 on the $J_{\boldsymbol{X}} \wedge J_{\boldsymbol{Y}}$. Therefore, the detected IGDB leads to a false negative finding if the P-value in the permutation test is not lower than the a significant level (ie, 0.05). Besides, we observe a false positive error if $G[\hat{S}, \hat{T}]$ has low similarity to $G[S_0, T_0]$ even we rejected the IGDB null hypothesis. We report the accuracy of inference by false positive rate (FPR) and false negative rate (FNR) among replications.

Furthermore, we compare IGDB to commonly-used multivariate testing methods at the edge-level: positive false discovery rate (pFDR) by Storey[51] and Bonferroni correction. These correction methods are commonly used in GWAS and vGWAS analysis in practice. We evaluate the true $\boldsymbol{\Delta} = \{\delta_{uv}\}_{u\in U, v\in V}$ with estimated $\hat{\boldsymbol{\Delta}} = \{\hat{\delta}_{uv}\}_{u\in U, v\in V}$ from varied methods. For the proposed method, we obtain the $\hat{\boldsymbol{\Delta}}$ based on the extracted IGDB $G[\hat{S}, \hat{T}]$ and the hypothesis testing. Particularly, if we reject the IGDB null hypothesis with a detected IGDB $G[\hat{S}, \hat{T}]$, we let $\hat{\boldsymbol{\Delta}} = \{\hat{\delta}_{uv}\} = \{I(e_{uv} \in G[\hat{S}, \hat{T}])\}$. In the case that we fails to reject, we consider $\hat{S}, \hat{T}$ as empty sets such that $\hat{\boldsymbol{\Delta}} = \boldsymbol{0}_{m\times n}$. The FDR threshold of 0.2 and corrected α level of 0.05 are used in the pFDR and Bonferroni correction respectively.

Subsequently, based on the $\hat{\delta}_{uv}$ observed from different methods, and true parameters δ_{uv}, we calculate true positive rate (TPR) and true negative rate (TNR) as:

$$\text{TPR} = \frac{\sum_{u,v} I(\delta_{uv} = \hat{\delta}_{uv} = 1)}{\sum_{u,v} I(\delta_{uv} = 1)}, \quad \text{TNR} = \frac{\sum_{u,v} I(\delta_{uv} = \hat{\delta}_{uv} = 0)}{\sum_{u,v} I(\delta_{uv} = 0)}.$$

The associated means and standard deviations are reported based on 100 replications for each simulation scenario.

The results from the IGDB inference are summarized in Table 1. The power of the IGDB inference relies on the size and SNR (by different standard effect sizes) of the underlying IGDB $G[S_0, T_0]$, which concurs with our theoretical results. We fails to reject the IGDB null hypothesis for one simulated data set with a smaller size (30, 20) and effect size 0.8, and higher noise (0.8, 0.2).

TABLE 1 IGDB inference results under varied SNRs and noises.

$(\lvert S_0\rvert, \lvert T_0\rvert)$	(q_1, q_2)	Metrics	0.8	1.0	1.2
(50, 40)	(0.9, 0.1)	FPR (0.8)	0 (0)	0 (0)	0 (0)
		FPR (0.9)	0 (0)	0 (0)	0 (0)
		FNR	0 (0)	0 (0)	0 (0)
	(0.8, 0.2)	FPR (0.8)	0 (0)	0 (0)	0 (0)
		FPR (0.9)	0 (0)	0 (0)	0 (0)
		FNR	0 (0)	0 (0)	0 (0)
(30, 20)	(0.9, 0.1)	FPR (0.8)	0 (0)	0 (0)	0 (0)
		FPR (0.9)	0 (0)	0 (0)	0 (0)
		FNR	0 (0)	0 (0)	0 (0)
	(0.8, 0.2)	FPR (0.8)	0 (0)	0 (0)	0 (0)
		FPR (0.9)	0.2100 (0.4073)	0.0400 (0.1960)	0 (0)
		FNR	0.0600 (0.2375)	0 (0)	0 (0)

Note: We summarize the FPR (with cutoff of 0.8 and 0.9 on the $J_X \wedge J_Y$) and FNR to evaluate the estimated IGDB. The results suggest robust and accurate performance of our method at a *bi-clique* level (ie, revealing patterns).

The comparative edge-level results from the proposed method and competing methods are displayed in Table 2 for different sizes of IGDB. All three methods have improved performance with higher SNRs and lower noise levels. The proposed method outperforms pFDR and Bonferroni correction methods for both TPR and TNR under different scenarios. Both pFDR and Bonferroni methods have high TNR but low TPR indicating a stringent cutoff, while the proposed method achieves a higher TPR maintaining a similar or even higher TNR than the others. The Bonferroni method is even more stringent where the TPR is even smaller than 10% when we have low SNRs (eg, 0.8) for all cases.

7 | DISCUSSION

Imaging-genetics studies aim to model the predictive mechanism of genetic variants on quantitative imaging measures. However, high dimensionality and complex association patterns between genetic variants and imaging traits raise a considerable challenge for statistical estimation and inference. For example, purely region-level inference erases local voxel heterogeneity, thus may be ineffective in learning spatial specificity of imaging voxels. In this article, we have developed an IGDB multivariate to multivariate analysis tool to identify systematic associations between multivariate voxel-level imaging features and multivariate genetic variants. Our method focuses on the systematic polygenic and pleiotropic patterns rather than individual pairwise associations, and thus mitigates the challenges of ultra-high dimensionality due to multivariate to multivariate association analysis. Besides, our high-resolution voxel-level genome wide association analysis is not constrained by pre-specified regions of interest, hence fully accounts for the variability between voxels, and yields data-driven brain regions associated with functionally related genetic loci. Therefore, our findings are more biologically interpretable and meaningful.

We develop a new optimization solution to extract IGDB by leveraging its graph properties that we discovered in theoretical study. Our IGDB extraction algorithm is computationally efficient and scalable. The input data for our method could be either individual-level or GWAS summary statistics. The IGDB inference method controls the family-wise error rate for IGDB-level findings. We provide theoretical results to guarantee the numerical performance of IGDB extraction and accuracy of the inference model. Although initially proposed in analyzing systematic association patterns between SNPs and voxels, this approach is also well-suited for analyzing region-level imaging data, where spatial constraints are not necessary.

In real data applications, we applied our method to the HCP data set to study the genetic effects on white matter microstructure integrity. The results revealed a variety of functionally related genetic loci that are associated

TABLE 2 Edge-wise accuracy under varied IGDB sizes, SNRs and noises.

$(\lvert S_0\rvert, \lvert T_0\rvert)$	(q_1, q_2)	Metrics		0.8	1.0	1.2
(50,40)	(0.9, 0.1)	IGDB	TPR	0.9879 (0.0184)	0.9942 (0.0124)	0.9968 (0.0097)
			TNR	1 (0)	1 (0)	1 (0)
		pFDR	TPR	0.7453 (0.0090)	0.8686 (0.0045)	0.8995 (0.0023)
			TNR	0.8858 (0.0020)	0.8667 (0.0018)	0.8619 (0.0018)
		Bonferroni	TPR	0.0520 (0.0048)	0.1739 (0.0092)	0.3941 (0.0096)
			TNR	0.9942 (0.0005)	0.9806 (0.0008)	0.9562 (0.0012)
	(0.8, 0.2)	IGDB	TPR	0.9938 (0.0126)	0.9982 (0.0064)	0.9984 (0.0061)
			TNR	0.9998 (0.0006)	1.0000 (0.0003)	1.0000 (0.0004)
		pFDR	TPR	0.7032 (0.0067)	0.7903 (0.0039)	0.8095 (0.0027)
			TNR	0.7842 (0.0021)	0.7577 (0.0019)	0.7517 (0.0018)
		Bonferroni	TPR	0.0458 (0.0043)	0.1557 (0.0084)	0.3506 (0.0097)
			TNR	0.9884 (0.0007)	0.9612 (0.0014)	0.9125 (0.0020)
(30,20)	(0.9, 0.1)	IGDB	TPR	0.9987 (0.0081)	0.9992 (0.0060)	1 (0)
			TNR	1.0000 (0.0001)	1 (0)	1(0)
		pFDR	TPR	0.7043 (0.0176)	0.8537 (0.0085)	0.8954 (0.0042)
			TNR	0.9017 (0.0019)	0.8799 (0.0015)	0.8741 (0.0014)
		Bonferroni	TPR	0.0517 (0.0082)	0.1741 (0.0163)	0.3946 (0.0175)
			TNR	0.9942 (0.0005)	0.9807 (0.0009)	0.9561 (0.0012)
	(0.8, 0.2)	IGDB	TPR	0.8527 (0.2248)	0.9645 (0.0398)	0.9778 (0.0287)
			TNR	0.9996 (0.0009)	0.9995 (0.0009)	0.9997 (0.0005)
		pFDR	TPR	0.6891 (0.0114)	0.7857 (0.0075)	0.8069 (0.0045)
			TNR	0.7952 (0.0022)	0.7661 (0.0017)	0.7596 (0.0019)
		Bonferroni	TPR	0.0473 (0.0095)	0.1563 (0.0144)	0.3525 (0.0173)
			TNR	0.9884 (0.0008)	0.9610 (0.0013)	0.9123 (0.0017)

Note: We compare the performance of IGDB with multiple testing correction methods in terms of the accuracy of individual SNP-voxel pairs. The extracted IGDB patterns dramatically improve the SNP-voxel pair level inference accuracy by allowing pairs to borrow strengths from each other.

with sub-regions of white matter area tracts on posterior corpus callosum. These novel findings are consistent with previous findings.[30] Our annotation analysis further provide evidence that selected SNPs are associated with white matter microstructures through gene expression. The overall computational load for imaging-genetics analysis remains heavy regardless improved algorithms and computational facilities. Since our initial vGWAS is performed using GWAS analysis tools (eg, plink), the analysis is limited on individual SNPs. Regardless, the input of our method is vGWAS analysis results and thus suits for any vGWAS analysis methods. Our IGDB algorithm can also be extended to further constrain the IGDB structure by leveraging the functional annotation of genetic variants.[52]

In summary, we have developed a new neuroimaging-GWAS tool to identify systematic associations between multivariate imaging features and multivariate genetic variants. Our IGDB method is computationally efficient and improves the accuracy and power through revealing systematic polygenic and pleiotropic patterns.

ACKNOWLEDGEMENTS

This work was partially supported by the National Institute on Drug Abuse of the National Institutes of Health under Award Number 1DP1DA048968-01, R01EB015611, R01MH094520. The second author Dr. Yuan Zhang was supported by NSF Grant DMS-2311109.

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions. Software in the form of Matlab code, together with a sample input data set and complete documentation is available on github through link https://github.com/qwu1221/multi2multi.

ORCID

Qiong Wu https://orcid.org/0000-0001-5747-1899
Xiaoqi Huang https://orcid.org/0000-0002-4410-9185

REFERENCES

1. Meisenzahl E, Koutsouleris N, Bottlender R, et al. Structural brain alterations at different stages of schizophrenia: a voxel-based morphometric study. *Schizophr Res*. 2008;104(1-3):44-60.
2. Lee S, Viqar F, Zimmerman ME, et al. White matter hyperintensities are a core feature of Alzheimer's disease: evidence from the dominantly inherited Alzheimer network. *Ann Neurol*. 2016;79(6):929-939.
3. Savitz JB, Drevets WC. Imaging phenotypes of major depressive disorder: genetic correlates. *Neuroscience*. 2009;164(1):300-330.
4. Ge T, Schumann G, Feng J. Imaging genetics-towards discovery neuroscience. *Quant Biol*. 2013;1(4):227-245.
5. Liu J, Calhoun VD. A review of multivariate analyses in imaging genetics. *Front Neuroinform*. 2014;8:29.
6. Nathoo FS, Kong L, Zhu H, Initiative ADN. A review of statistical methods in imaging genetics. *Can J Stat*. 2019;47(1):108-131.
7. Smith SM, Douaud G, Chen W, et al. An expanded set of genome-wide association studies of brain imaging phenotypes in UK biobank. *Nat Neurosci*. 2021;24(5):737-745.
8. Zhao B, Zhang J, Ibrahim JG, et al. Large-scale GWAS reveals genetic architecture of brain white matter microstructure and genetic overlap with cognitive and mental health traits (n= 17,706). *Mol Psychiatry*. 2019;26:3943-3955.
9. Zhao B, Li T, Yang Y, et al. Common genetic variation influencing human white matter microstructure. *Science*. 2021;372(6548).
10. Zhu H, Khondker Z, Lu Z, Ibrahim JG. Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers. *J Am Stat Assoc*. 2014;109(507):977-990.
11. Huang M, Nichols T, Huang C, et al. FVGWAS: fast voxelwise genome wide association analysis of large-scale imaging genetic data. *Neuroimage*. 2015;118:613-627.
12. Huang C, Thompson P, Wang Y, et al. FGWAS: functional genome wide association analysis. *Neuroimage*. 2017;159:107-121.
13. Ge T, Feng J, Hibar DP, Thompson PM, Nichols TE. Increasing power for voxel-wise genome-wide association studies: the random field theory, least square kernel machines and fast permutation procedures. *Neuroimage*. 2012;63(2):858-873.
14. Ge T, Nichols TE, Ghosh D, et al. A kernel machine method for detecting effects of interaction between multidimensional variable sets: an imaging genetics application. *Neuroimage*. 2015;109:505-514.
15. Hibar DP, Stein JL, Kohannim O, et al. Voxelwise gene-wide association study (vGeneWAS): multivariate gene-based association testing in 731 elderly subjects. *Neuroimage*. 2011;56(4):1875-1891.
16. Stein JL, Hua X, Lee S, et al. Voxelwise genome-wide association study (vGWAS). *Neuroimage*. 2010;53(3):1160-1174.
17. Chi EC, Allen GI, Zhou H, Kohannim O, Lange K, Thompson PM. Imaging genetics via sparse canonical correlation analysis. *2013 IEEE 10th International Symposium on Biomedical Imaging*. New York: IEEE; 2013:740-743.
18. Greenlaw K, Szefer E, Graham J, Lesperance M, Nathoo FS, Initiative ADN. A Bayesian group sparse multi-task regression model for imaging genetics. *Bioinformatics*. 2017;33(16):2513-2522.
19. Hardoon DR, Ettinger U, Mourão-Miranda J, et al. Correlation-based multivariate analysis of genetic influence on brain volume. *Neurosci Lett*. 2009;450(3):281-286.
20. Kong D, An B, Zhang J, Zhu H. L2RM: low-rank linear regression models for high-dimensional matrix responses. *J Am Stat Assoc*. 2020;115(529):403-424.
21. Le Floch É, Guillemot V, Frouin V, et al. Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse partial least squares. *Neuroimage*. 2012;63(1):11-24.
22. Liu J, Pearlson G, Windemuth A, Ruano G, Perrone-Bizzozero NI, Calhoun V. Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ICA. *Hum Brain Mapp*. 2009;30(1):241-255.
23. Wang H, Nie F, Huang H, et al. Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort. *Bioinformatics*. 2012;28(2):229-237.
24. Vounou M, Nichols TE, Montana G. Discovering genetic associations with high-dimensional neuroimaging phenotypes: a sparse reduced-rank regression approach. *Neuroimage*. 2010;53(3):1147-1159.
25. Vounou M, Janousova E, Wolz R, et al. Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease. *Neuroimage*. 2012;60(1):700-716.
26. Marcus DS, Harms MP, Snyder AZ, et al. Human connectome project informatics: quality control, database services, and data visualization. *Neuroimage*. 2013;80:202-219.

27. Van Essen DC, Smith SM, Barch DM, et al. The WU-Minn human connectome project: an overview. *Neuroimage*. 2013;80:62-79.
28. Jahanshad N, Kochunov PV, Sprooten E, et al. Multi-site genetic analysis of diffusion images and voxelwise heritability analysis: a pilot project of the ENIGMA–DTI working group. *Neuroimage*. 2013;81:455-469.
29. Kochunov P, Jahanshad N, Sprooten E, et al. Multi-site study of additive genetic effects on fractional anisotropy of cerebral white matter: comparing meta and megaanalytical approaches for data pooling. *Neuroimage*. 2014;95:136-150.
30. Kochunov P, Kochunov P. Heritability of fractional anisotropy in human white matter: a comparison of human connectome project and ENIGMA-DTI data. *Neuroimage*. 2015;111:300-311.
31. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. 2015;4(1):7.
32. Benjamini Y, Hochberg Y. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J Educ Behav Stat*. 2000;25(1):60-83.
33. Efron B. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Vol 1. Cambridge, UK: Cambridge University Press; 2012.
34. Woo CW, Krishnan A, Wager TD. Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. *Neuroimage*. 2014;91:412-419.
35. Charikar M. Greedy approximation algorithms for finding dense components in a graph. *International Workshop on Approximation Algorithms for Combinatorial Optimization*. Berlin: Springer; 2000:84-95.
36. Khuller S, Saha B. On finding dense subgraphs. *International Colloquium on Automata, Languages, and Programming*. Berlin: Springer; 2009:597-608.
37. Amini AA, Chen A, Bickel PJ, Levina E. Pseudo-likelihood methods for community detection in large sparse networks. *Ann Stat*. 2013;41(4):2097-2122.
38. Cheng Y, Church GM. Biclustering of expression data. *Proc Int Conf Intell Syst Mol Biol*. 2000;8:93-103.
39. Xu M, Jog V, Loh PL. Optimal rates for community estimation in the weighted stochastic block model. *Ann Stat*. 2020;48(1):183-204.
40. Zalesky A, Fornito A, Bullmore ET. Network-based statistic: identifying differences in brain networks. *Neuroimage*. 2010;53(4):1197-1207.
41. Nichols TE. Multiple testing corrections, nonparametric methods, and random field theory. *Neuroimage*. 2012;62(2):811-815.
42. Kochunov P, Rowland LM, Fieremans E, et al. Diffusion-weighted imaging uncovers likely sources of processing-speed deficits in schizophrenia. *Proc Natl Acad Sci*. 2016;113(47):13504-13509.
43. Das S, Forer L, Schönherr S, et al. Next-generation genotype imputation service and methods. *Nat Genet*. 2016;48(10):1284-1287.
44. Zou H, He D, Zhou Y. On sure screening with multiple responses. *Stat Sin*. 2021;31:1749-1777. doi:10.5705/ss.202018.0462
45. Chen S, Kang J, Xing Y, Zhao Y, Milton DK. Estimating large covariance matrix with network topology for high-dimensional biomedical data. *Comput Stat Data Anal*. 2018;127:82-95.
46. Park MK, Hwang SH, Jung S, Hong SS, Kwon SB. Lesions in the splenium of the corpus callosum: clinical and radiological implications. *Neurol Asia*. 2014;19(1):79-88.
47. Zheng Z, Huang D, Wang J, et al. QTLbase: an integrative resource for quantitative trait loci across multiple human molecular phenotypes. *Nucleic Acids Res*. 2020;48(D1):D983-D991.
48. Verstynen TD, Weinstein AM, Schneider WW, Jakicic JM, Rofey DL, Erickson KI. Increased body mass index is associated with a global and distributed decrease in white matter microstructural integrity. *Psychosom Med*. 2012;74(7):682.
49. Brzustowicz LM, Simone J, Mohseni P, et al. Linkage disequilibrium mapping of schizophrenia susceptibility to the CAPON region of chromosome 1q22. *Am J Hum Genet*. 2004;74(5):1057-1063.
50. Kubicki M, Park H, Westin CF, et al. DTI and MTR abnormalities in schizophrenia: analysis of white matter integrity. *Neuroimage*. 2005;26(4):1109-1118.
51. Storey JD. A direct approach to false discovery rates. *J R Stat Soc Series B Stat Methodology*. 2002;64(3):479-498.
52. Li X, Li Z, Zhou H, et al. Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat Genet*. 2020;52(9):969-983.
53. Dhillon IS. Co-clustering documents and words using bipartite spectral graph partitioning. *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM; 2001:269-274.

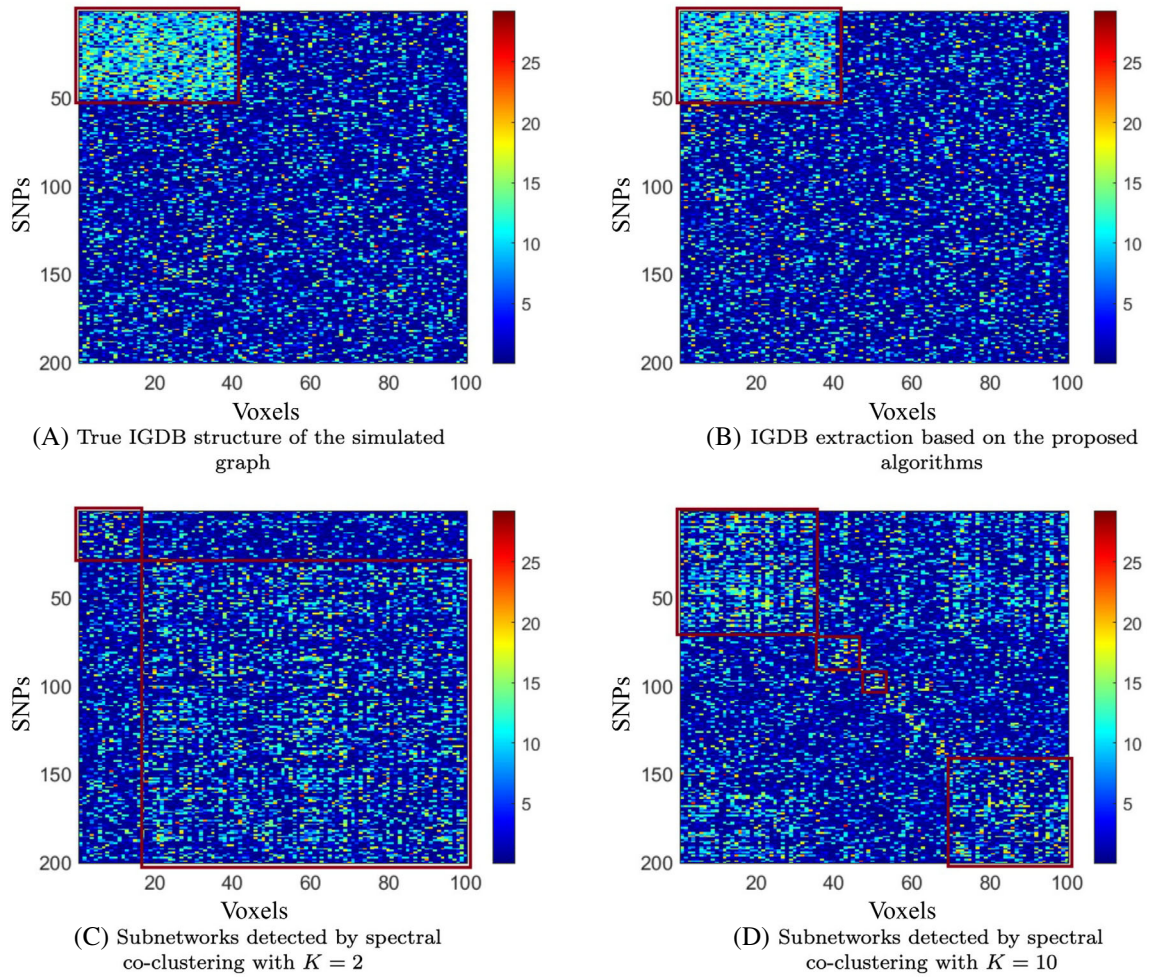
SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Wu Q, Zhang Y, Huang X, et al. A multivariate to multivariate approach for voxel-wise genome-wide association analysis. *Statistics in Medicine*. 2024;43(20):3862-3880. doi: 10.1002/sim.10101

APPENDIX A. ADDITIONAL NUMERICAL RESULTS

Comparisons with bi-clustering algorithms. In our simulation analysis, we only compared our method to Charikar's algorithm instead of bi-clustering algorithms because these methods are not well suited to dense *bi-clique* extraction. To demonstrate this, we applied the classic spectral bi-clustering algorithm[53] to a simulated data set. Specifically, we generated a bipartite graph with $m = 200, n = 100, L = 60$, the IGDB size $(|S_0|, |T_0|) = (50, 40)$, and standard effect size $\theta = 0.8$, and proportions of noisy edges $(\mu_1, \mu_0) = (0.8, 0.2)$. The true structure of simulated bipartite graph, detected sub-networks from competing methods are displayed in Figure A1. The convectional bi-clustering algorithms can miss the dense bi-cliques.



(A) True IGDB structure of the simulated graph

(B) IGDB extraction based on the proposed algorithms

(C) Subnetworks detected by spectral co-clustering with $K = 2$

(D) Subnetworks detected by spectral co-clustering with $K = 10$

FIGURE A1 Comparison with other biclustering algorithm in a simulated data set. True and detected subnetworks are highlighted in red. (A) displays the true bipartite graph with an IGDB; (B) shows the IGDB structure extracted by Algorithms 1 and 2; (C) shows subnetworks detected by spectral co-clustering algorithm with $K = 2$; (D) highlights several subnetworks detected spectral co-clustering algorithm with $K = 10$.

Simulation results from large graphs. We extended our simulation studies by considering larger graphs by setting $m = 800, n = 500$, and $L = 200$. The synthetic data was generated with an IGDB (ie, $(|S_0|, |T_0|) = (100, 80)$). The results are displayed in Table A1 with the same setting of standard effect size (ie, $\theta = 0.8$, 1, and 1.2), and proportions of noisy edges (ie, $(\mu_1, \mu_0) = (0.8, 0.2)$ and $(0.9, 0.1)$) as in the main analysis.

TABLE A1 Edge-wise accuracy under varied SNRs and noises with $(|S_0|, |T_0|) = (100, 80)$.

(q_1, q_2)	Methods		0.8	1.0	1.2
(0.9, 0.1)	IGDB	TPR	0.9600 (0.0000)	0.9600 (0.0000)	0.9600 (0.0000)
		TNR	0.9998 (0.0000)	0.9998 (0.0000)	0.9998 (0.0000)
	pFDR	TPR	0.9025 (0.0005)	0.9029 (0.0006)	0.9029 (0.0006)
		TNR	0.8747 (0.0003)	0.8746 (0.0003)	0.8747 (0.0003)
	Bonferroni	TPR	0.6060 (0.0048)	0.8692 (0.0021)	0.8994 (0.0003)
		TNR	0.9326 (0.0002)	0.9035 (0.0001)	0.9001 (0.0000)
(0.8, 0.2)	IGDB	TPR	0.9545 (0.0101)	0.9598 (0.0024)	0.9598 (0.0024)
		TNR	0.9998 (0.0001)	0.9998 (0.0000)	0.9998 (0.0000)
	pFDR	TPR	0.8100 (0.0011)	0.8100 (0.0011)	0.8101 (0.0011)
		TNR	0.7598 (0.0004)	0.7597 (0.0004)	0.7597 (0.0004)
	Bonferroni	TPR	0.5385 (0.0044)	0.7724 (0.0018)	0.7994 (0.0001)
		TNR	0.8652 (0.0003)	0.8069 (0.0001)	0.8001 (0.0001)