bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467021; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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

[1]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania

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

[3]Department of Mathematics, University of Maryland, College Park, Maryland

[4]School of Public Health, University of Maryland, College Park, Maryland

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

[6]Division of Biostatistics and Bioinformatics, Department of Epidemiology and Public Health, School of Medicine, University of Maryland, Baltimore, Maryland

Abstract

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 (i.e., 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 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 (HCP). The results demonstrate multiple genetic loci influencing white matter integrity measures on splenium and genu of the corpus callosum.

Keywords: bi-clique, imaging-genetics, graph, ultra-high dimensionality, voxel-wise GWAS, white matter integrity

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467021; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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. The joint analysis of imging-genetics data reveals the genetic effects on spatially specific brain functions and structures (Ge et al., 2013; Liu and Calhoun, 2014; Nathoo et al., 2019; Smith et al., 2021; Zhao et al., 2019, 2021; Zhu et al., 2014). Identifying genetic effects on objectively measured high-resolution imaging traits can enhance understanding the complex genetic and neurological mechanisms of neuropsychiatric 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. Moreover, both imaging traits and genetic variants exhibit complex and organized dependence structure reflecting the underlying neurophysiological mechanisms and linkage disequilibrium patterns (Nathoo et al., 2019). 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 (Huang et al., 2015, 2017). The direct application of classic voxel-wise genome-wide association analysis (vGWAS) may require an enormous sample size (e.g., multiple millions of participants) to control the false positive error rate while maintaining adequate statistical

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467021; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

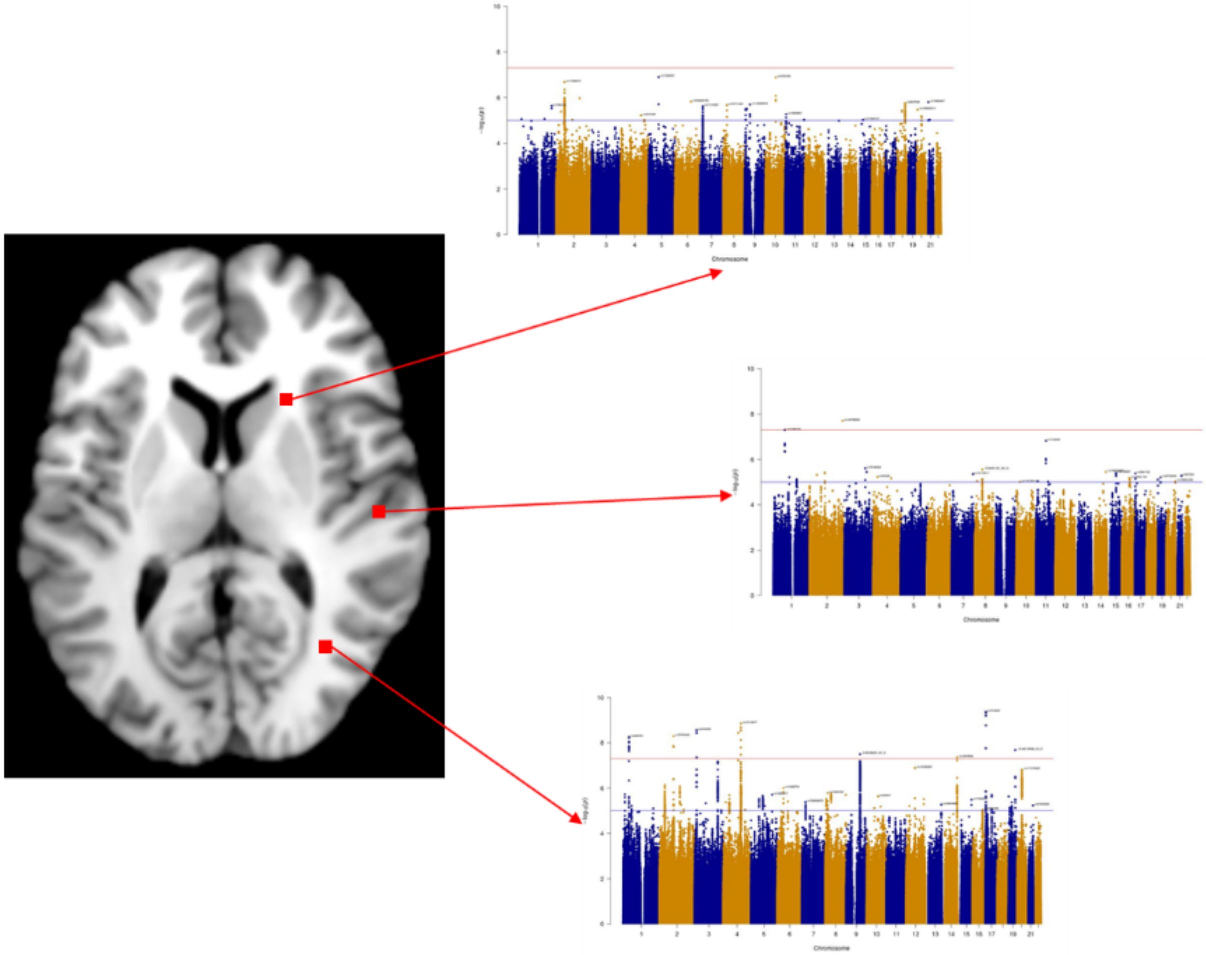
power (Ge et al., 2012, 2015; Hibar et al., 2011; Stein et al., 2010).

Furthermore, advanced methods have been developed to leverage group sparsity by techniques including regularization, low rank techniques and projection of high-dimensional features (Chi et al., 2013; Greenlaw et al., 2017; Hardoon et al., 2009; Kong et al., 2020; Le Floch et al., 2012; Liu et al., 2009; Wang et al., 2012; Vounou et al., 2010, 2012; Zhu et al., 2014). However, while these methods could gain statistical power by jointly modelling 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 overfittings. The results from summarized measures as a few latent variables or a coarser scale are less interpretable or lacking the spatial specificity (Liu and Calhoun, 2014).

In this study, we propose a new multivariate to multivariate method to systematically investigate the SNP-voxel association patterns with four aims: identify voxel clusters as genetically correlated imaging traits, detect functionally related SNP sets, understand the SNP-voxel association patterns as polygenic and pleiotropic relationships, and test the association patterns while controlling multiplicity. 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. The existence of the polygenic and pleiotropic SNP-voxel association structure can be evaluated against a random bipartite graph. We then develop computationally efficient algorithms to extract the IGDB structure from the bipartite graph mixture model and thus provide sound estimates of parameters in the mixture model. Our inference on IGDB is constructed via likelihood based statistic on the bipartite graph mixture model, and thus can improve statistical power with controlled family-wise error rate.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467021; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 1: Data structure for vGWAS



bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467021; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2 Motivating Data Example

The Human Connectome Project (HCP) sponsored by National Institutes of Health (NIH) aims to construct the underlying neuro pathways with 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 (`humanconnectome.org`) that passed the HCP and ENIGMA quality control and assurance standards (Marcus et al., 2013). The participants in the HCP study were recruited from a large population-based study named "the Missouri Family and Twin Registry" (Van Essen et al., 2013).

The fractional anisotropy (FA) measure, derived from diffusion tensor imaging (DTI), is a widely-used brain structural connectivity metric for studying the white matter microstructure. Previous studies have investigated the heritability quantitatively through variance components method of pedigrees (Jahanshad et al., 2013; Kochunov et al., 2014). They find that 70% to 80% of the total phenotypic variance of tract-wise FA measures can be explained by additive genetic factors (Kochunov et al., 2015). The significantly and reliably hertiable FA measurements are qualified as a set of endophenotypes which suggests to further specify genetic variants associated with these traits. 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 (Zhao et al., 2021). 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 polygenity and pleiotropy 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 statisti-

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467021; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

cal 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 (i.e., SNPs) with $|U| = m$. For each participant $l \in \{1, ..., L\}$, define $\boldsymbol{x}_l = (x_{1,l}, ..., x_{m,l})^T$ to be the genetic variants for the participant l and $\boldsymbol{y}_l = (y_{1,l}, ..., y_{n,l})^T$ 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 $g^{-1}(\cdot)$, and 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.* The goal of our statistical inference is to accurately identify the subset of significant associations $\{(u, v) : \beta_{uv} \neq 0\}$ based on multivariate to multivariate hypothesis testing (Benjamini and Hochberg, 2000; Efron, 2012):

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

Conventional statistical inference methods (e.g., 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 (e.g., the pleiotropic and polygenic structure). A prominent example is that a cluster of SNPs may jointly influence the observations on a cluster of neighboring voxels. To address this challenge, we propose a

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467021; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 succicint 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. 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, i.e.,

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

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 (i.e., connected voxels) because of the biological

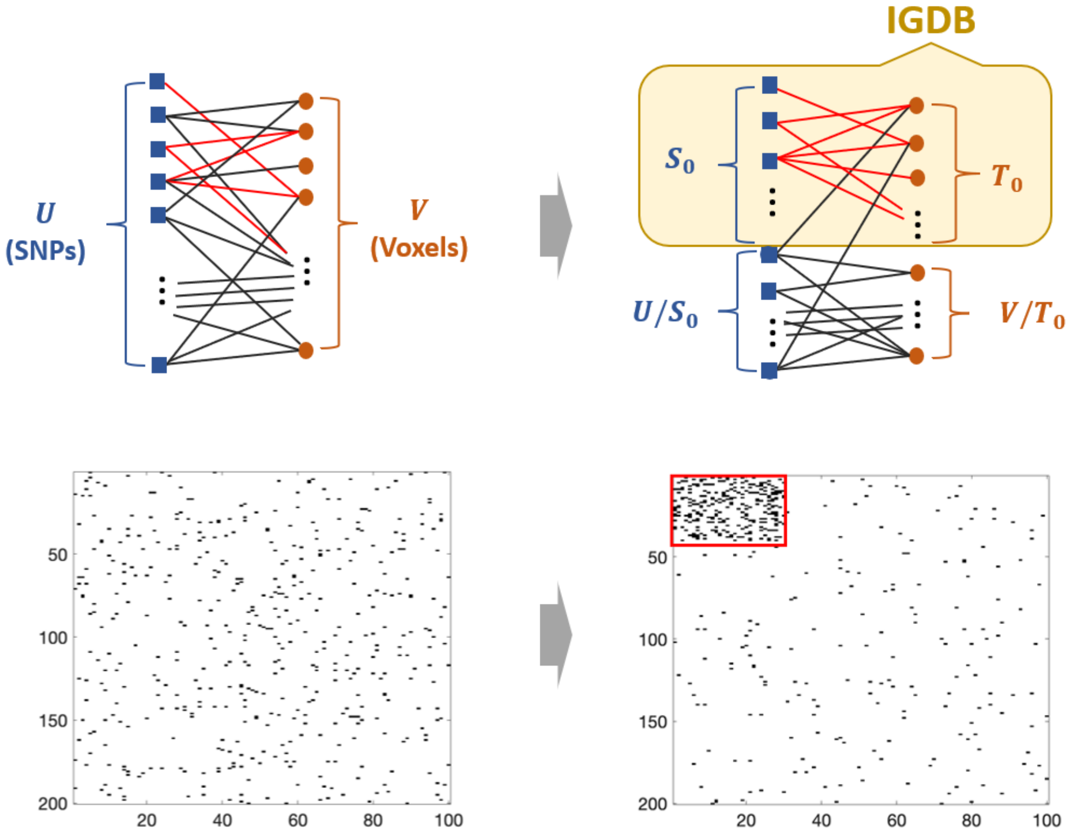
bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467021; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

interpretability and inference advantages (Woo et al., 2014). This is reflected in our proposed IGDB structure by further formulating S_0 and T_0 as disjoint vertex neighborhoods, as follows:

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

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

Figure 2: Illustration of a bipartite graph with IGDB structure $G[S_0, T_0]$. The right subfigure highlights $G[S_0, T_0]$ in G with nodes reordered.



bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467021; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3.3 Graph properties of IGDB

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

Case 0 : G is observed from a random bipartite graph $G(m, n, \mu_0)$,

Case 1 : There exists at least one non-trivial IGDB $G[S_0, T_0]$ such that G is observed from

$$e_{uv} = I(\beta_{uv} \neq 0) \sim \begin{cases} \text{Bernoulli}(\mu_1), & \text{if } u \in S_0 \;\&\; v \in T_0 \\ \text{Bernoulli}(\mu_0), & \text{otherwise} \end{cases} \quad \text{with } \mu_1 > \mu_0.$$

In Case 0 (i.e., 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 versus 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 $\frac{|E[S,T]|}{|S||T|} \geq \gamma \in (\mu_0, 1)$ (i.e., 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*

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467021; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$c(\gamma,\mu_0)m_0 \geq 8\log n$ and $c(\gamma,\mu_0)n_0 \geq 8\log m$, we have

$$\mathbb{P}\left(|S| \geq m_0, |T| \geq n_0\right) \leq 2mn \cdot \exp\left(-\frac{1}{4}c(\gamma,\mu_0)m_0n_0\right),$$

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 (e.g., 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 Efron (2012), as a natural consequence of our model set up in Section 3.2, edge weights in $\boldsymbol{W}$ follow a mixture marginal distribution:

$$w_{uv} \sim \begin{cases} f_1(\cdot;\boldsymbol{\theta}_1), & \text{if } \beta_{uv} \neq 0 \\ f_0(\cdot;\boldsymbol{\theta}_0), & \text{if } \beta_{uv} = 0. \end{cases} \tag{2}$$

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 $\mathbb{E}_{\boldsymbol{\theta}_1}[w_{uv}|\beta_{uv} \neq 0] > \mathbb{E}_{\boldsymbol{\theta}_0}[w_{uv}|\beta_{uv} = 0]$.

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:

$$\max_{S\subseteq U, T\subseteq V} |S||T| \qquad \text{subject to } \frac{\|\boldsymbol{W}[S,T]\|_{1,1}}{|S||T|} \geq \gamma' \tag{3}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467021; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

or the Lagrangian form after taking logarithm on both terms:

$$\max_{S \subseteq U, T \subseteq V} \log(|S||T|) + \lambda \log \left(\frac{\|\boldsymbol{W}[S,T]\|_{1,1}}{|S||T|} \right), \tag{4}$$

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 (4) is challenging because it is a nondeterministic polynomial (NP) problem (Charikar, 2000; Khuller and Saha, 2009). We propose a computationally efficient greedy algorithm to approximately carry out the optimization of (4). We describe the greedy algorithm as Algorithm 1 in the following. In designing it, we extended the greedy algorithms for dense subgraph discovery (Khuller and Saha, 2009) in an adjacency matrix to a large bipartite matrix to extract dense bi-cliques. The computational complexity of Algorithm 1 is $O(C_1 mn)$, where C_1 is determined by the grid search of h (i.e., $|S|/|T|$) in the following Algorithm 1.

Algorithm 1 Direct optimization of objective function (4)

```
    Input: $G = (U, V, E, \boldsymbol{W})$, $\lambda$
    Output: $G[\tilde{S}_\lambda, \tilde{T}_\lambda]$
procedure ALGORITHM
    for $h \in \{h_1, h_2, ..., h_L\}$ 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 $\sqrt{d} \deg_X(i; S_k, T_k) \leq \frac{1}{\sqrt{d}} \deg_Y(j; S_k, T_k)$ 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 $G[S_1, T_1], ..., G[S_{n+m-1}, T_{n+m_1}]$;
    end for
    Output $G[\tilde{S}_\lambda, \tilde{T}_\lambda]$ with largest objective function in $G[S^{h_1}, T^{h_1}], ..., G[S^{h_L}, T^{h_L}]$;
end procedure
```

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467021; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (4):

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

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 (Charikar, 2000), 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 (4) in terms of a parameter λ as the following Theorem 1.

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, i.e., $d_\lambda(S^*_\lambda, T^*_\lambda) \leq \rho(\lambda, m, n) d_\lambda(\tilde{S}_\lambda, \tilde{T}_\lambda)$ with*

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

In Theorem 2, we state that the optimization of the proposed objective function (4) asymptotically leads to almost full recovery of the IGDB-based network structure.

Theorem 2. *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 mixture of Bernoulli distributions: $e_{uv} \sim \delta_{uv} Bernoulli(\pi_1) + (1-\delta_{uv}) Bernoulli(\pi_0)$,*

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467021; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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

and

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

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, i.e. for any fixed $a \in (0, 1)$, as $n \to \infty$, we have

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

In practice, the tuning parameter λ can be objectively selected by a likelihood method (see the web Appendix A for details). Based on each dense subgraph $G[S, T]$, we further identify spatially-contiguous voxel clusters (i.e., $\tilde{\mathcal{N}}_k^T$, $k = 1, , , , \tilde{K}_2$), and a corresponding set of SNPs (i.e., $\tilde{\mathcal{N}}_k^S$, $k = 1, , , , \tilde{K}_1$) that are functionally associated with voxel clusters (see Web Appendix A). Last, multiple IGDBs can be extracted by performing algorithms repeatedly with the detected IGDBs masked (Cheng and Church, 2000).

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467021; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(S_0, T_0), the edges in G^r follow a mixture of two Bernoulli distributions:

$$a_{uv}|(S_0, T_0) \sim \text{Bernoulli}(\pi_{uv}) \tag{5}$$

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

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

based on our mixture distribution model (5).

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

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

with likelihood given by Bernoulli distributions in (5). 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 (Zalesky et al., 2010; Nichols, 2012) to empirically approximate the distribution of the likelihood-based statistic t_G under the IGDB null and

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467021; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

control the family-wise error rates (FWER). We describe the detailed testing procedure in the Web Appendix A. The p-values of multiple IGDBs can be observed by considering each IGDB individually.

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 (Kochunov et al., 2015, 2016). The image acquisition parameters are described in the Web Appendix B. 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 (Das et al., 2016).

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 while adjusting covariates including sex, age, BWI, and population characteristics using the first 10 principal components in our application. We then performed sure independence screening on SNPs with multiple imaging responses through a direct extension of univariate screening procedure (Zou et al., 2021). 13,498 SNPs survive into further analysis. The details are described in the Web Appendix B.

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467021; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

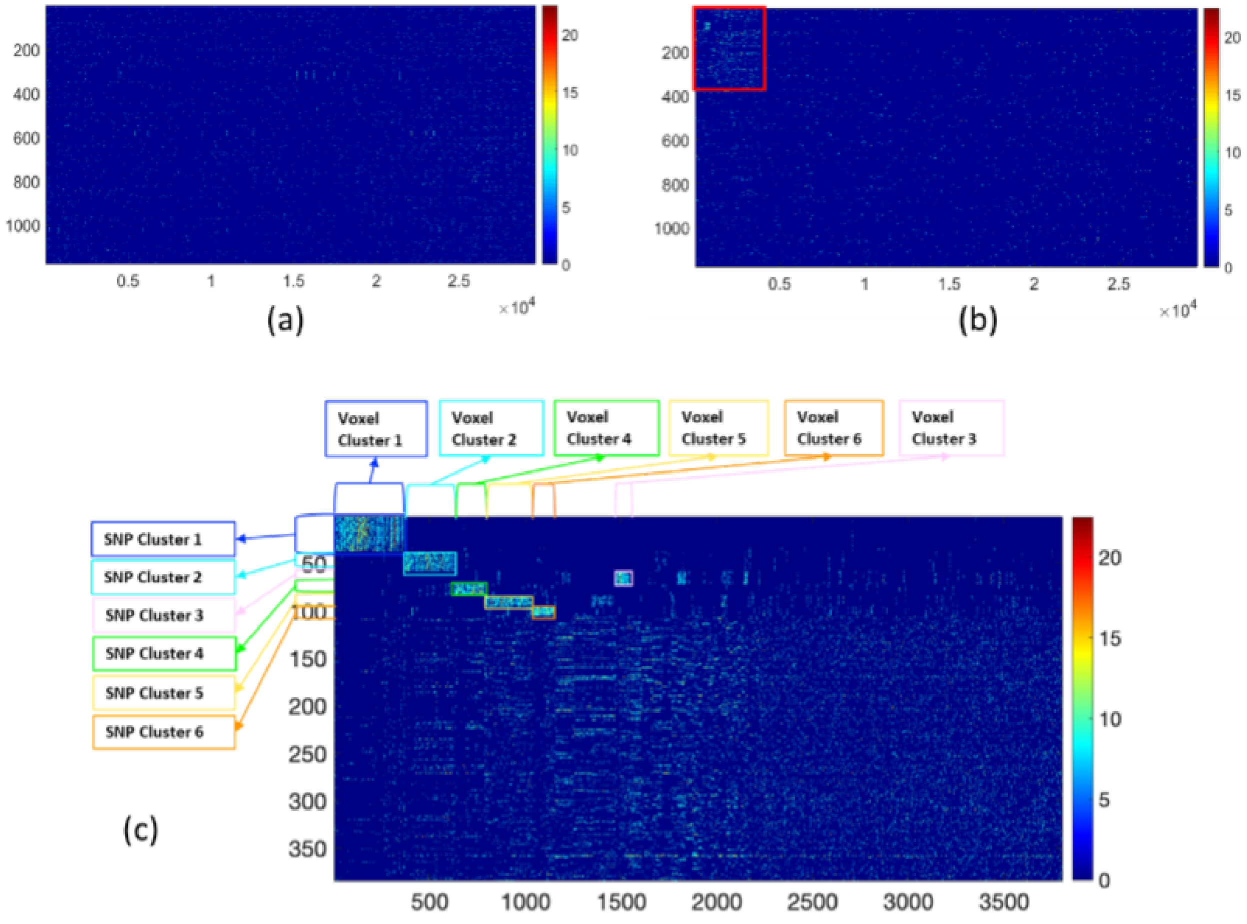
in section 4.1 and 4.2, we extracted IGDBs and performed permutation tests to determine its statistical significance while controlling family-wise error rate ($q < 0.05$). We observe different brain areas being influenced by distinct genetic loci. A Manhattan plot for all SNPs across 22 chromosomes with selected imaging-genetic associations highlighted and tables for snp and voxels across all 22 chromosomes are included in the Web Appendix B.

In this section, we focus on SNPs on chromsome 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 $\boldsymbol{W}_{1178\times 29627}$ (i.e., Figure 3 (a)), we detected an IGDB with 384 SNPs and 3803 voxels as Figure 3 (b) by maximizing the objective function (4). The computation is efficient, which took 20 minutes on a PC with an i7 CPU 3.60 GHz and 64GB memory. 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 (Chen et al., 2018) on similarity matrices observed from the detected IGDB. The refined pattern in Figure 3 (c) displays 6 distinct SNP-voxel association clusters. Note that the refined structure can not be identified without revealing the IGDB by the proposed algorithm.

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 (Park et al., 2014). To annotate the SNPs in the identified clusters, we queried the SNPs in the QTLbase (`http://mulinlab.org/qtlbase/index.html`, (Zheng et al., 2020)) 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 Web Table 4 as supporting information. In cluster 1, multiple SNPs are linked with

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467021; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 3: IGDB procedure on chromosome 1: (a) is the input matrix $\boldsymbol{W}$; (b) demonstrates the detected IGDB; (c)displays the refined pattern of the IGDB



bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467021; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (Verstynen et al., 2012). Therefore, this cluster reveals the marginal association of (obesity-related) LEPR gene and white matter integrity. In clusters 2, 3, 4, and 5, the associated genes, for example, S100A1, TAF1A, CFH, CFHR3, 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 (Zhao et al., 2019). In addition, the NOS1AP gene is identified to be a risk factor for schizophrenia (Brzustowicz et al., 2004), while the alterations of white matter integrity for patients with schizophrenia were studied in Kubicki et al. (2005). 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.

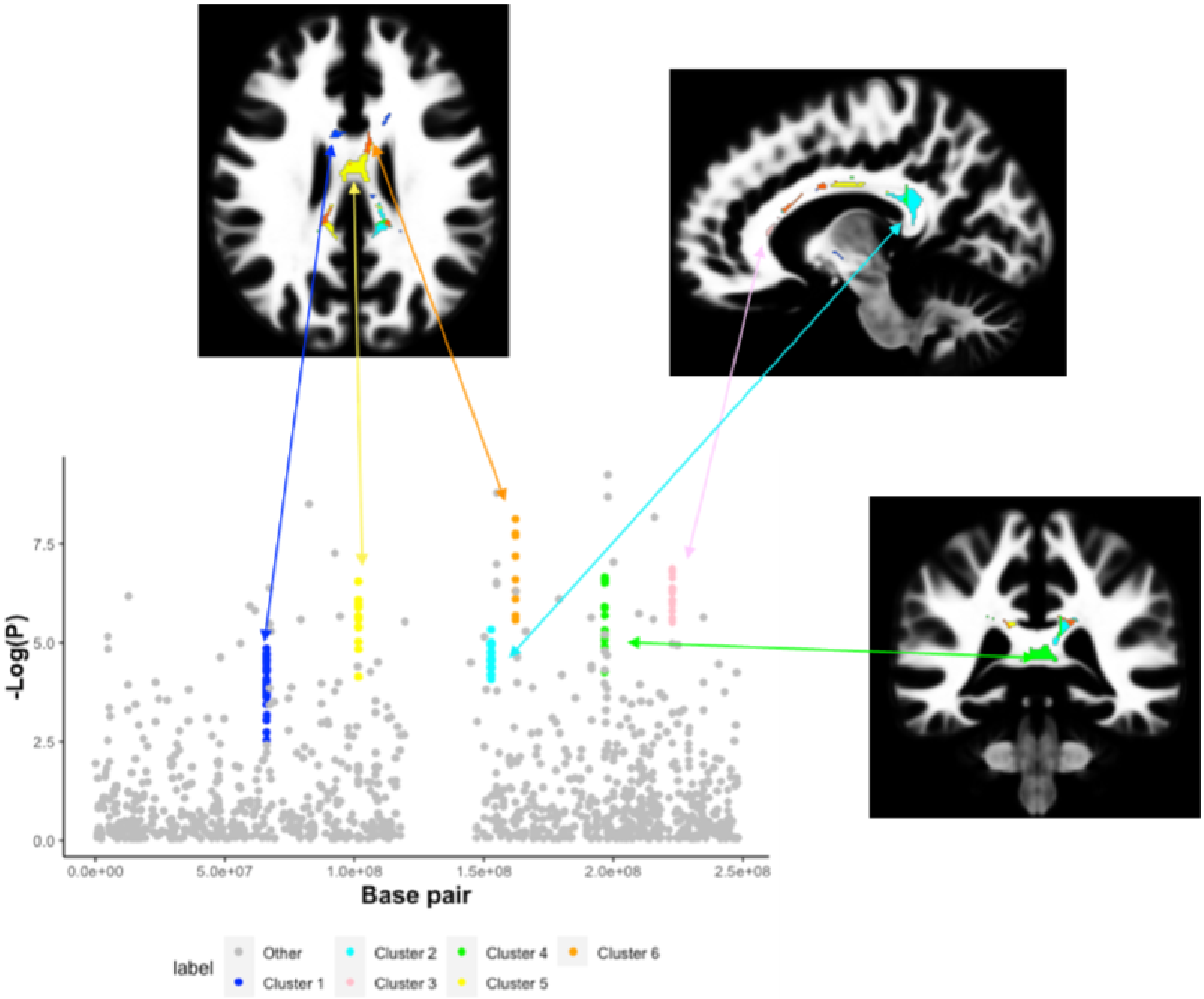
6 Simulation Studies

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 (i.e., μ_1) than the rest of graph (i.e., 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

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467021; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 4: An illustration of the association patterns between SNP and voxel clusters on chromosome 1.



bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467021; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (e.g., 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 (i.e., $(|S_0|, |T_0|) = (50, 40)$ and $(30, 20)$), standard effect size (i.e., $\theta = 0.8$, 1, and 1.2), and proportions of noisy edges (i.e., $(\mu_1, \mu_0) = (0.8, 0.2)$ and $(0.9, 0.1)$). Additional simulation settings with larger graph and sample sizes are included in the Web Appendix B.

6.2 Performance metrics and results

We evaluate the performance of proposed method at two levels. At the subgraph-level, we assess the accuracy of IGDB inference by examining if we can reject the null (i.e., 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.

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 0.8 and 0.9. 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 (i.e., 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

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467021; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (2002) and Bonferroni correction. These correction methods are commonly used in GWAS and vGWAS analysis in practice. We evaluate the true $\mathbf{\Delta} = \{\delta_{uv}\}_{u\in U, v\in V}$ with estimated $\hat{\mathbf{\Delta}} = \{\hat{\delta}_{uv}\}_{u\in U, v\in V}$ from varied methods. For the proposed method, we obtain the $\hat{\mathbf{\Delta}}$ 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 $\hat{\mathbf{\Delta}} = \{\hat{\delta}_{uv}\} = \{I(e_{uv} \in G[\hat{S}, \hat{T}])\}$. In the case that we fails to reject, we consider $\hat{S}, \hat{T}$ as empty sets such that $\hat{\mathbf{\Delta}} = \mathbf{0}_{m\times n}$. 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 $(0.8, 0.2)$.

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,

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467021; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1: IGDB inference results under varied SNRs and noises

			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)

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 (e.g., 0.8) for all cases.

7 Discussion

We have developed an IGDB mulivariate 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.

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 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. In real data applications,

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467021; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

$(\|S_0\|, \|T_0\|)$	(q_1, q_2)	Methods		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)

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467021; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

we identify significant IGDBs where voxels are spatially contiguous and SNPs are functionally correlated confirmed by eQTL. Our IGDB algorithm can also be extended to further constrain the IGDB structure by leveraging the functional annotation of genetic variants (Li et al., 2020).

References

Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of educational and Behavioral Statistics* **25,** 60–83.

Brzustowicz, L. M., Simone, J., Mohseni, P., Hayter, J. E., Hodgkinson, K. A., Chow, E. W., and Bassett, A. S. (2004). Linkage disequilibrium mapping of schizophrenia susceptibility to the capon region of chromosome 1q22. *The American Journal of Human Genetics* **74,** 1057–1063.

Charikar, M. (2000). Greedy approximation algorithms for finding dense components in a graph. In *International Workshop on Approximation Algorithms for Combinatorial Optimization*, pages 84–95. Springer.

Chen, S., Kang, J., Xing, Y., Zhao, Y., and Milton, D. K. (2018). Estimating large covariance matrix with network topology for high-dimensional biomedical data. *Computational Statistics & Data Analysis* **127,** 82–95.

Cheng, Y. and Church, G. M. (2000). Biclustering of expression data. in intelligent systems for molecular biology.

Chi, E. C., Allen, G. I., Zhou, H., Kohannim, O., Lange, K., and Thompson, P. M. (2013). Imaging genetics via sparse canonical correlation analysis. In *2013 IEEE 10th International Symposium on Biomedical Imaging*, pages 740–743. IEEE.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467021; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. *Nature genetics* **48,** 1284–1287.

Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press.

Ge, T., Feng, J., Hibar, D. P., Thompson, P. M., and Nichols, T. E. (2012). Increasing power for voxel-wise genome-wide association studies: the random field theory, least square kernel machines and fast permutation procedures. *Neuroimage* **63,** 858–873.

Ge, T., Nichols, T. E., Ghosh, D., Mormino, E. C., Smoller, J. W., Sabuncu, M. R., Initiative, A. D. N., et al. (2015). A kernel machine method for detecting effects of interaction between multidimensional variable sets: An imaging genetics application. *NeuroImage* **109,** 505–514.

Ge, T., Schumann, G., and Feng, J. (2013). Imaging genetics—towards discovery neuroscience. *Quantitative Biology* **1,** 227–245.

Greenlaw, K., Szefer, E., Graham, J., Lesperance, M., Nathoo, F. S., and Initiative, A. D. N. (2017). A bayesian group sparse multi-task regression model for imaging genetics. *Bioinformatics* **33,** 2513–2522.

Hardoon, D. R., Ettinger, U., Mourão-Miranda, J., Antonova, E., Collier, D., Kumari, V., Williams, S. C., and Brammer, M. (2009). Correlation-based multivariate analysis of genetic influence on brain volume. *Neuroscience letters* **450,** 281–286.

Hibar, D. P., Stein, J. L., Kohannim, O., Jahanshad, N., Saykin, A. J., Shen, L., Kim, S., Pankratz, N., Foroud, T., Huentelman, M. J., et al. (2011). Voxelwise gene-wide association study (vgenewas): multivariate gene-based association testing in 731 elderly subjects. *Neuroimage* **56,** 1875–1891.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467021; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Huang, C., Thompson, P., Wang, Y., Yu, Y., Zhang, J., Kong, D., Colen, R. R., Knickmeyer, R. C., Zhu, H., Initiative, A. D. N., et al. (2017). Fgwas: Functional genome wide association analysis. *NeuroImage* **159,** 107–121.

Huang, M., Nichols, T., Huang, C., Yu, Y., Lu, Z., Knickmeyer, R. C., Feng, Q., Zhu, H., Initiative, A. D. N., et al. (2015). Fvgwas: Fast voxelwise genome wide association analysis of large-scale imaging genetic data. *Neuroimage* **118,** 613–627.

Jahanshad, N., Kochunov, P. V., Sprooten, E., Mandl, R. C., Nichols, T. E., Almasy, L., Blangero, J., Brouwer, R. M., Curran, J. E., de Zubicaray, G. I., et al. (2013). Multi-site genetic analysis of diffusion images and voxelwise heritability analysis: A pilot project of the enigma–dti working group. *Neuroimage* **81,** 455–469.

Khuller, S. and Saha, B. (2009). On finding dense subgraphs. In *International Colloquium on Automata, Languages, and Programming*, pages 597–608. Springer.

Kochunov, P., Jahanshad, N., Marcus, D., Winkler, A., Sprooten, E., Nichols, T. E., Wright, S. N., Hong, L. E., Patel, B., Behrens, T., et al. (2015). Heritability of fractional anisotropy in human white matter: a comparison of human connectome project and enigma-dti data. *Neuroimage* **111,** 300–311.

Kochunov, P., Jahanshad, N., Sprooten, E., Nichols, T. E., Mandl, R. C., Almasy, L., Booth, T., Brouwer, R. M., Curran, J. E., de Zubicaray, G. I., et al. (2014). Multi-site study of additive genetic effects on fractional anisotropy of cerebral white matter: comparing meta and megaanalytical approaches for data pooling. *Neuroimage* **95,** 136–150.

Kochunov, P., Rowland, L. M., Fieremans, E., Veraart, J., Jahanshad, N., Eskandar, G., Du, X., Muellerklein, F., Savransky, A., Shukla, D., et al. (2016). Diffusion-weighted imaging uncovers likely sources of processing-speed deficits in schizophrenia. *Proceedings of the National Academy of Sciences* **113,** 13504–13509.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467021; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Kong, D., An, B., Zhang, J., and Zhu, H. (2020). L2rm: Low-rank linear regression models for high-dimensional matrix responses. *Journal of the American Statistical Association* **115,** 403–424.

Kubicki, M., Park, H., Westin, C.-F., Nestor, P. G., Mulkern, R. V., Maier, S. E., Niznikiewicz, M., Connor, E. E., Levitt, J. J., Frumin, M., et al. (2005). Dti and mtr abnormalities in schizophrenia: analysis of white matter integrity. *Neuroimage* **26,** 1109–1118.

Le Floch, É., Guillemot, V., Frouin, V., Pinel, P., Lalanne, C., Trinchera, L., Tenenhaus, A., Moreno, A., Zilbovicius, M., Bourgeron, T., et al. (2012). Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse partial least squares. *Neuroimage* **63,** 11–24.

Li, X., Li, Z., Zhou, H., Gaynor, S. M., Liu, Y., Chen, H., Sun, R., Dey, R., Arnett, D. K., Aslibekyan, S., et al. (2020). Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nature genetics* **52,** 969–983.

Liu, J. and Calhoun, V. D. (2014). A review of multivariate analyses in imaging genetics. *Frontiers in neuroinformatics* **8,** 29.

Liu, J., Pearlson, G., Windemuth, A., Ruano, G., Perrone-Bizzozero, N. I., and Calhoun, V. (2009). Combining fmri and snp data to investigate connections between brain function and genetics using parallel ica. *Human brain mapping* **30,** 241–255.

Marcus, D. S., Harms, M. P., Snyder, A. Z., Jenkinson, M., Wilson, J. A., Glasser, M. F., Barch, D. M., Archie, K. A., Burgess, G. C., Ramaratnam, M., et al. (2013). Human connectome project informatics: quality control, database services, and data visualization. *Neuroimage* **80,** 202–219.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467021; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Nathoo, F. S., Kong, L., Zhu, H., and Initiative, A. D. N. (2019). A review of statistical methods in imaging genetics. *Canadian Journal of Statistics* **47,** 108–131.

Nichols, T. E. (2012). Multiple testing corrections, nonparametric methods, and random field theory. *Neuroimage* **62,** 811–815.

Park, M.-K., Hwang, S.-H., Jung, S., Hong, S.-S., and Kwon, S.-B. (2014). Lesions in the splenium of the corpus callosum: clinical and radiological implications. *Neurology Asia* **19,**.

Smith, S. M., Douaud, G., Chen, W., Hanayik, T., Alfaro-Almagro, F., Sharp, K., and Elliott, L. T. (2021). An expanded set of genome-wide association studies of brain imaging phenotypes in uk biobank. *Nature neuroscience* **24,** 737–745.

Stein, J. L., Hua, X., Lee, S., Ho, A. J., Leow, A. D., Toga, A. W., Saykin, A. J., Shen, L., Foroud, T., Pankratz, N., et al. (2010). Voxelwise genome-wide association study (vgwas). *neuroimage* **53,** 1160–1174.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64,** 479–498.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., Consortium, W.-M. H., et al. (2013). The wu-minn human connectome project: an overview. *Neuroimage* **80,** 62–79.

Verstynen, T. D., Weinstein, A. M., Schneider, W. W., Jakicic, J. M., Rofey, D. L., and Erickson, K. I. (2012). Increased body mass index is associated with a global and distributed decrease in white matter microstructural integrity. *Psychosomatic medicine* **74,** 682.

Vounou, M., Janousova, E., Wolz, R., Stein, J. L., Thompson, P. M., Rueckert, D., Montana, G., Initiative, A. D. N., et al. (2012). Sparse reduced-rank regression detects genetic

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467021; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

associations with voxel-wise longitudinal phenotypes in alzheimer's disease. *Neuroimage* **60,** 700–716.

Vounou, M., Nichols, T. E., Montana, G., Initiative, A. D. N., et al. (2010). Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. *Neuroimage* **53,** 1147–1159.

Wang, H., Nie, F., Huang, H., Kim, S., Nho, K., Risacher, S. L., Saykin, A. J., Shen, L., and Initiative, A. D. N. (2012). Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the adni cohort. *Bioinformatics* **28,** 229–237.

Woo, C.-W., Krishnan, A., and Wager, T. D. (2014). Cluster-extent based thresholding in fmri analyses: pitfalls and recommendations. *Neuroimage* **91,** 412–419.

Zalesky, A., Fornito, A., and Bullmore, E. T. (2010). Network-based statistic: identifying differences in brain networks. *Neuroimage* **53,** 1197–1207.

Zhao, B., Li, T., Yang, Y., Wang, X., Luo, T., Shan, Y., Zhu, Z., Xiong, D., Hauberg, M. E., Bendl, J., et al. (2021). Common genetic variation influencing human white matter microstructure. *Science* **372,**.

Zhao, B., Zhang, J., Ibrahim, J. G., Luo, T., Santelli, R. C., Li, Y., Li, T., Shan, Y., Zhu, Z., Zhou, F., et al. (2019). Large-scale gwas reveals genetic architecture of brain white matter microstructure and genetic overlap with cognitive and mental health traits (n= 17,706). *Molecular psychiatry* pages 1–13.

Zheng, Z., Huang, D., Wang, J., Zhao, K., Zhou, Y., Guo, Z., Zhai, S., Xu, H., Cui, H., Yao, H., et al. (2020). Qtlbase: an integrative resource for quantitative trait loci across multiple human molecular phenotypes. *Nucleic acids research* **48,** D983–D991.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.02.467021; this version posted November 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Zhu, H., Khondker, Z., Lu, Z., and Ibrahim, J. G. (2014). Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers. *Journal of the American Statistical Association* **109,** 977–990.

Zou, H., He, D., and Zhou, Y. (2021). On sure screening with multiple responses. *Statistica Sinica* .