PLOS Computational Biology

RESEARCH ARTICLE

IBAS: Interaction-bridged association studies discovering novel genes underlying complex traits

Dinghao Wang[1☯], **Pathum Kossinna**[2☯], **Karen Ardila**[3], **Senitha Kumarapeli**[1,4], **M. Ethan MacDonald**[3,5,6], **Jingjing Wu**[1], **Qingrun Zhang**[1,2,5,6*]

1 Department of Mathematics & Statistics, University of Calgary, Calgary, Alberta, Canada, **2** Department of Biochemistry & Molecular Biology, University of Calgary, Calgary, Alberta, Canada, **3** Department of Biomedical Engineering, University of Calgary, Calgary, Alberta, Canada, **4** Schulich School of Medicine & Dentistry, University of Western Ontario, London, Ontario, Canada, **5** Alberta Children's Hospital Research Institute, University of Calgary, Calgary, Alberta, Canada, **6** Hotchkiss Brain Institute, University of Calgary, Calgary, Alberta, Canada

☯ joint first authors.
* qingrun.zhang@ucalgary.ca

Citation: Wang D, Kossinna P, Ardila K, Kumarapeli S, MacDonald ME, Wu J, et al. (2026) IBAS: Interaction-bridged association studies discovering novel genes underlying complex traits. PLoS Comput Biol 22(8): e1014640. https://doi.org/10.1371/journal.pcbi.1014640

Editor: Ilya Ioshikhes, CANADA

Received: October 29, 2025

Accepted: July 29, 2026

Published: August 17, 2026

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1014640

Abstract

Genetic contributions to complex traits are often mediated through coordinated gene–gene interaction networks, yet most existing association frameworks focus on marginal single-gene effects and overlook higher-order dependency structures. Direct modeling of interactions remains challenging due to combinatorial complexity and statistical instability. We introduce Interaction-Bridged Association Study (IBAS), a general framework that incorporates pathway-level interaction patterns into genotype–phenotype association analysis without explicitly enumerating interactions. IBAS leverages transcriptomic reference data to construct low-dimensional representations of pathway activity, which guide SNP-weighting and gene-level association testing within a kernel-based framework. In perturbation-based simulations, IBAS demonstrates improved stability and reproducibility compared to conventional TWAS and gene-based methods, while maintaining well-calibrated Type I error under phenotype permutation. Application to the WTCCC datasets identifies both known and novel genes across multiple complex diseases, including candidates with modest marginal effects missed by standard approaches. These findings are supported by replication in an independent cohort, and analyses across multiple reference tissues revealing both shared and tissue-specific signals. Overall, IBAS provides a statistically robust and computationally tractable framework for incorporating interaction effects into association mapping, extending beyond the single-gene paradigm and enabling more comprehensive characterization of complex trait. IBAS is available on GitHub at: https://github.com/QingrunZhangLab/IBAS

Data availability statement: GTEx gene expression data are publicly available from the GTEx Portal (https://gtexportal.org/home/datasets). The individual-level GTEx whole-genome sequencing data and the Type 1 Diabetes Genetics Consortium dataset used in this study are controlled-access data available through dbGaP under accessions phs000424 and phs000911, respectively. Instructions for requesting access to protected GTEx data are provided at https://gtexportal.org/home/protectedDataAccess. WTCCC data are controlled-access data available by application through the European Genome-phenome Archive under dataset accessions EGAD00000000001–EGAD00000000009. The IBAS source code is publicly available at https://github.com/QingrunZhangLab/IBAS.

Funding: Q.Z. is supported by an NSERC Discovery Grant (RGPIN-2025-05344), a New Frontiers in Research Fund (NFRFE-2023-00291), and an NSERC RTI grant (RTI-2021-00675); J.J.W is supported by an NSERC Discovery Grant (RGPIN-2024-06154); D.W. is partly supported by the Alberta Innovates Graduate Student Scholarship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Understanding how genes influence complex diseases remains a major challenge in genetics. While many studies focus on individual genes, biological systems operate through coordinated interactions among groups of genes. However, directly modeling these interactions is difficult due to their complexity and instability. In this study, we introduce Interaction-Bridged Association Studies (IBAS), a new framework that captures gene–gene interaction patterns without explicitly testing every possible interaction. Instead, IBAS uses gene expression data to summarize pathway-level activity and leverages this information to guide genetic association testing. We show that IBAS produces more stable and reproducible results compared to existing methods, particularly when data are noisy or variable. Applying IBAS to multiple disease datasets, we identify both well-known and previously unreported genes, including candidates with subtle effects that are often missed by traditional approaches. Many findings are replicated in independent cohorts and across different tissues, supporting their robustness. Overall, IBAS provides a practical and reliable way to uncover how coordinated gene activity contributes to complex traits, offering new insights into disease mechanisms.

Introduction

Discovering genetic basis of complex traits is a long-standing theme in genetics and genomics [1]. Towards this end, genotype-phenotype association mapping has emerged as an essential tool over the last 20 years [2,3]. In recent times, while identifying genes statistically associated with diseases remains a primary goal, researchers have gradually shifted their focus to their underlying mechanisms through various means of interpretations [4]. Along this line, the availability of multi-scale -omics data especially transcriptome allows researchers to develop sophisticated statistical methods to improve both the discovery of genes and their interpretations [5–10]. Among many biologically sensible hypothetical mechanisms, it is intuitive to expect that the variation of gene-gene interaction networks would play a critical role mediating genetic effects to phenotype. Methods such as iPath [11] take this premise and study it at the transcriptomic level, investigating perturbations of given pathways at the sample level and the effect on clinical phenotypes. However, despite gene-gene interactions having been studied for a century [12,13], no consensus has been reached in understanding the genetic basis of variation of interactions and its implications to phenotypic changes. In this study, the term gene–gene interaction refers to coordinated transcriptional activity within biological pathways, reflected by structured co-expression or covariance patterns among genes, rather than explicit statistical interaction or genetic epistasis between individual genes or variants. Our objective is therefore not to explicitly model pairwise interaction terms, but to learn low-dimensional representations that summarize these coordinated pathway-level expression programs.

In our view, the lack of statistical genetic models studying the genetic basis of such pathway-level interaction structures is largely due to the astronomical number of potential combinations, which leads to serious statistical instability. Indeed, studying the interaction patterns associated with phenotypes is already a $p \gg n$ problem, suffering from statistical instability as well as computational burden [14,15]. By adding an extra layer of the genetic basis of variation of interaction patterns, it only adds substantial risk of instability, potentially leading to many results that may not be robust to perturbation of input data.

Transcriptome-Wide Association Study (TWAS) is a popular model to conduct association mapping utilizing transcriptomes [5,16]. Although it only focuses on single genes, the in-depth characterization of TWAS by our group [17,18] offered a unique insight of developing models incorporating interactions. The mainstream format of TWAS is a two-step protocol: first, one can form an expression prediction model using (usually cis) genotypes in a reference dataset (e.g.,: GTEx [19]). The predicted expression is called Genetically Regulated eXpression, or GReX. Then, in the second step, in the main dataset for association mapping (that does not contain expressions), one can predict expressions using the available genotypes and finally associate the predicted expression to phenotype. It is worth noting that this perspective differs from the conventional objective of TWAS. In mainstream TWAS, gene expression is typically used to construct genetically regulated expression (GReX), and the association between GReX and the phenotype is then tested, often with the goal of identifying causal genes under an instrumental-variable framework. In contrast, the perspective adopted in our previous work and in this study treats gene expression as a data bridge that links genotype information to downstream association testing. Rather than explicitly modeling GReX–phenotype associations, expression-derived information is used to guide SNP-level modeling, including variant selection, weighting, and interaction-aware aggregation. This distinction clarifies that our framework is not a standard TWAS model, but instead an expression-informed variant-set association approach.

Our recent works showed this interpretation of "predicting expressions" was not the essence of TWAS, but rather, misleading. First, theoretical power analysis showed that TWAS could be more powerful than a hypothetical scenario in which expression data is available in the main GWAS dataset [17]. Additionally, TWAS could be underpowered compared to GWAS when the expression heritability is low [17]. Both results question the interpretation of the prediction of expression in TWAS. As such, we proposed to interpret the "prediction" step as a selection of genetic variants directed by expression. From the perspective of Machine Learning, the first step in TWAS is essentially feature selection and the second, feature aggregation. With this interpretation in mind, one may disregard GReX and instead conduct feature selection and aggregation independently. As a result, methodological research focusing on optimal combinations of feature selection and aggregation approaches could be flexibly conducted. Indeed, novel methods splitting these two steps developed by us [18,20] and independently by others [21] showed higher power than standard TWAS.

The above unique insight into TWAS has opened a door of conducting association mapping mediated by any "data-bridge" by replacing the gene expression and its predicted value with another entity for feature selection and a form of feature aggregation [18]. Indeed, another development in our group has provided a useful tool leveraging brain images [22] in a similar manner.

In this work, we designed an unconventional framework by forming a mediator representing interactions computed using gene expressions through statistical learning techniques. We then used it to facilitate the discovery of the genetic basis of variations of interactions and their consequence to phenotypic changes. More specifically, starting from a gene expression matrix of all genes in a prespecified pathway, we extract representative statistics such as PCA [23], t-SNE [24] or UMAP [25] to form the objective function for feature selection. Linear methods such as PCA naturally capture dominant covariance structures, while nonlinear approaches including t-SNE and UMAP can additionally preserve complex local relationships induced by coordinated expression patterns. These learned representations serve as interaction-informed mediators that bridge transcriptomic organization and downstream genotype–phenotype association analysis. Then, following our previous work [18,20,26], a kernel machine [27] is employed for feature aggregation aimed at association test. The outcomes are single genes whose genetic variants alter the variation of gene-gene interactions in the focal pathway,

which also impact the phenotypic changes. This method is named "Interaction Bridged Association Study", or IBAS. Importantly, such representative statistics contributed by many genes in the pathway is more robust to noise than single genes alone, therefore intuitively could be more stable mediators than single genes in a typical TWAS protocol. Indeed, perturbation experiments using GTEx data show that IBAS is more stable than standard TWAS protocols that rely on single genes. We further applied IBAS to the seven diseases of the Wellcome Trust Case Control Consortium, or WTCCC [28] and discovered novel genes underlying pathways and diseases.

Materials and methods

In this section, we first introduce the details of IBAS implementation, followed by the descriptions of methods used to compare to IBAS. Then the methods to generate perturbation expression data to assess stability are presented. Finally, the data sources are described.

The IBAS Framework

Step 1 of IBAS utilizes a reference dataset where gene expression and genotype data are available. Known biological pathways (or even user specified collection of genes known to contain interactions) are then used to extract the expression data. The extracted subset of gene expressions (corresponding to a particular pathway) then undergoes dimensionality reduction (using methods such as PCA, t-SNE or UMAP) to produce interaction components ("data-bridges") **(**Fig 1A). These components are expected to capture interactions at the transcriptome level. In **Step 2**, each component is then associated with individual *cis*-variants corresponding to the genes in the pathway (Fig 1B). The coefficients obtained for each variant indicate their contribution to the interactions captured by the corresponding interaction component. **Step 3** of IBAS involves a GWAS dataset where genotype data is assessed together with a phenotype of interest. Variants in the GWAS dataset are then aggregated to the gene level, weighted by their contribution to interactions (obtained in Step 2 depicted in Fig 1B) and tested for association with phenotypes using the Sequence Kernel Association Test, or SKAT [27] (Fig 1C). This results in the identification of genes associated with the phenotype of interest through the mediation of the pathway being considered.

The natural cascade of genetic variants affecting phenotype may be depicted in Fig 1D**, upper panel**: from variants to gene expression, to pathways and finally phenotype. TWAS captures the effects of variants on gene expression, and then gene expression on phenotype, ignoring the contribution of interactions at the pathway level (Fig 1D**, middle panel**). IBAS on the other hand, captures the effects of cis-variants on the interactions at the pathway level (where it could be argued that the gene-level effect itself is present) and identifies association with the phenotype (Fig 1D**, lower panel**). Therefore, IBAS reveals the genetic basis of complex traits from a compensatory angle.

The implementation of IBAS is structured into the following four components:

IBAS Implementation (I): Biological pathways. Biological pathways refer to known interactions among genes, proteins and metabolites [29]. These interactions are often curated from literature by various projects such as Kyoto Encyclopedia of Genes and Genomes (KEGG) [30–32], Reactome [33] or WikiPathways [34]. These databases provide access to both downloading and visualizing these pathways and their known interactions. This work utilizes the KEGG database through the use of the KEGGREST [35] R package. However, IBAS could be applied in a similar manner to any of the pathway databases or even other databases containing known biological interactions such as the Gene Ontology Resource [36].

IBAS Implementation (II): Dimensionality Reduction. The downloaded pathways provide context for sub-setting the reference data, but we must turn to dimensionality reduction to uncover the complex interactions within the pathways. Here, we take advantage of three methods often used in the analysis of genomics data: PCA [23], t-distributed Stochastic Neighbour Embedding (t-SNE) [24] and Uniform Manifold Approximation and Projection (UMAP) [25]. These methods are based on the concept of identifying a low-dimensional representation of points in a high-dimensional space.

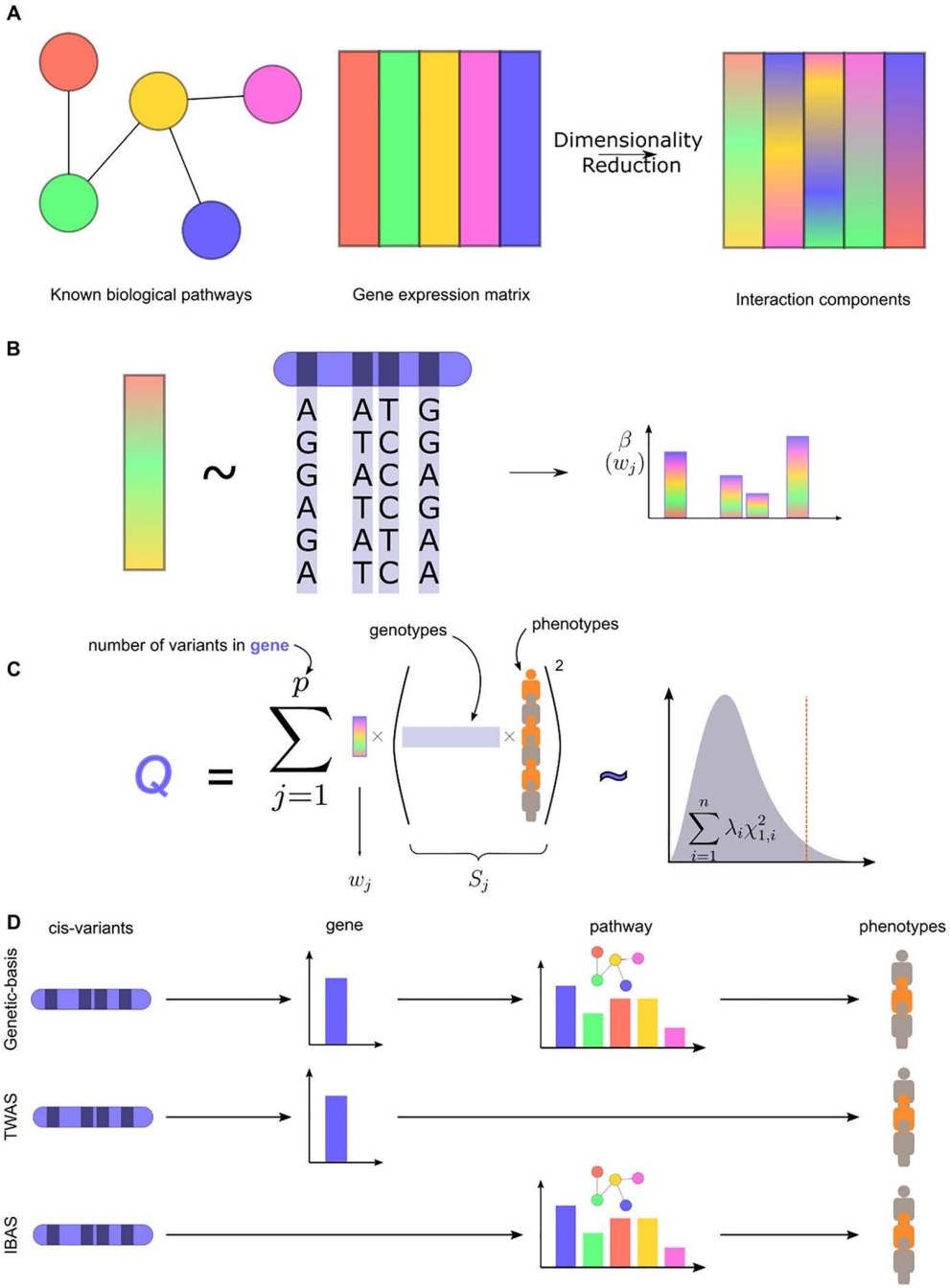

Fig 1. Overview of IBAS Framework. A) Gene expression matrices of a reference dataset (such as GTEx) are subset for known biological pathways (from sources such as KEGG, GO, etc.) and undergo dimensionality reduction (using methods such as t-SNE and UMAP) to produce interaction components ("data-bridges"). **B)** Individual interaction components are then associated with *cis*-variants of the genes in the pathway. The coefficients (or significance) of association are interpreted as the contribution of the variants towards the interaction effects at the transcriptome level. **C)** Genotypes in a GWAS dataset are then tested for association with a phenotype of interest, aggregated at the gene level using the Sequence Kernel Association Test, weighted by the interaction contributions obtained in **(B)**. **D)** The natural cascade of genetic effects affecting phenotype are modelled to various extents by existing methods. TWAS methods capture the effect of variants on individual genes' expression and ignore the contributions of interactions at the pathway level on phenotype while IBAS captures the effect of variants on these interactions (where the individual gene contribution may be also present).

https://doi.org/10.1371/journal.pcbi.1014640.g001

PCA utilizes the principle of iteratively identifying the combinations of components of the higher-dimensional space that maximize captured variance. These components are identified in such a way that they are uncorrelated with all previous PCs resulting in orthogonal combinations of the original data [37]. These components have been identified in various fields to capture variance and identify differences in subpopulations of the data [38,39]. *t-SNE* was originally designed for the purpose of visualizing high-dimensional data in a low-dimensional space (typically 2-dimensional). t-SNE is now often used in the visualization of transcriptomic data, particularly in RNA-Seq [40]. t-SNE is based upon the concept of minimizing the probability distribution of points in a low-dimensional space close to their probability distribution in a high-dimensional space. *UMAP* uses concepts similar to t-SNE, to achieve a low-dimensional embedding. The mathematics behind UMAP are fairly complex [25], and at a high-level, it utilizes "fuzzy simplicial complex" which is a type of weighted graph between the points in the high-dimensional space with the edges denoting the likelihood that two points are connected.

In the current implementation, we retain the first 10 principal components (PCs) as the latent representation of pathway-level expression (or all components when the pathway contains fewer than 10 genes). This choice is motivated by the observation that leading PCs capture the dominant correlation structure among genes, while higher-order components tend to reflect noise. Consistent with PCA loading patterns, the first few PCs are typically driven by groups of highly correlated genes with large variance contributions.

PCA in IBAS serves as a representation learning step to extract low-dimensional features that summarize pathway-level interaction structure and informs SNP weight estimation. Therefore, the number of components acts as a tuning parameter: using too few components may fail to capture relevant structure, while too many may introduce noise and dilute the signal. In practice, a moderate number of components (e.g., 10 PCs) provides a stable balance.

For alternative methods, we reduce pathway dimensionality to a maximum of 10 components for UMAP and 2 components for t-SNE, (or all available components when the pathway contains fewer genes than the specified limit). Similar considerations apply to these methods, where parameters (e.g., number of neighbors, minimum distance, and perplexity) control the learned representation and are chosen based on commonly used settings.

IBAS Implementation (III): Interaction Directed Feature Selection. Once a pathway has undergone dimensionality reduction, it is then in a form suitable for association analysis to "select" relevant genetic variants (here we reference feature selection with regards to feature prioritization). Using the same reference dataset, cis-variants are tested for association with each individual component of the interaction components. Each component is analyzed individually as it is expected that different components would capture a different set of interactions among the gene expression in the pathway. While any number of methods could be used for this association analysis, we use the linear mixed model [41,42] which accounts for population stratification by the use of a kinship matrix (which is also called a Genomic Relationship Matrix (GRM) in some literature).

For a specific pathway, the linear mixed model is defined as:

$$\boldsymbol{v} = \mathbf{X}\beta + \boldsymbol{u}_g + \epsilon \tag{1}$$

where $\boldsymbol{v} \in \mathbb{R}^{N\times 1}$ denotes a vector of interaction components in IBAS (e.g., one of the leading principal components) derived from the gene expression matrix of a pathway, where N is the sample size. $\mathbf{X} \in \mathbb{R}^{N\times(n_c+n_p+1)}$ is the design matrix containing the intercept, the variants to be tested, and other covariates, where n_p is the number of cis-variants in this pathway, n_c is the number of covariates, and the additional 1 corresponds to the intercept. $\beta \in \mathbb{R}^{(n_c+n_p+1)\times 1}$ denotes the vector of coefficients corresponding to the features in $\mathbf{X}$. $\boldsymbol{u}_g \sim \mathcal{N}(\mathbf{0}, \sigma_g^2\mathbf{K})$ denotes the random effects, and $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma_e^2\mathbf{I})$ denotes the residual error. Here $\mathbf{K}$ denotes the kinship matrix which provides a measure of the degree of relatedness of the individuals in the sample. σ_g^2 and σ_e^2 represent the variance components of the genetic random effects and residual error, respectively, and $\mathbf{I}$ denotes the identity matrix.

This association is carried out for all cis-variants of the genes in the pathway. While it is possible that trans-variants also do influence the expression of the genes in the pathway, such testing is extremely resource-intensive and would suffer from a large multiple-testing problem [43]. Typically, cis-variants are considered as the variants within 1Mb of the start and end of the gene and we used this definition for our analysis as well.

IBAS Implementation (IV): Genotype-Phenotype Association Test via Feature Aggregation. Finally, gene-level testing is conducted using the Sequence Kernel Association Test (SKAT). This test combines the effect of multiple variants in a kernel-based score test to test phenotypic association of the aggregate:

$$Q = \boldsymbol{y}^T \mathbf{K_s} \boldsymbol{y} \tag{2}$$

where $\boldsymbol{y} \in \mathbb{R}^{N\times 1}$ is the normalized phenotype vector, and $\mathbf{K_s} \in \mathbb{R}^{N\times N}$ is the kernel matrix constructed from the centralized genotype matrix $\mathbf{G} \in \mathbb{R}^{N\times c}$ in the GWAS dataset, where c is the number of cis-variants of a gene. In the original SKAT framework, the kernel is defined as

$$\mathbf{K_s} = \mathbf{G}\mathbf{W}_{\text{SKAT}}\,\mathbf{G}^{\top},$$

and $\mathbf{W}_{\text{SKAT}} = \text{diag}(w_1, \ldots, w_c)$ is a diagonal c by c weight matrix. The weights w_j are typically assigned based on minor allele frequency using a Beta distribution.

While multiple kernels may be defined for the use in the score test [27], we use the modified kernel previously used in our work [18,20]:

$$\mathbf{K_w} = \mathbf{G}\mathbf{W}_{\text{IBAS}}\,\mathbf{G}^{\top} \tag{3}$$

where $\mathbf{W}_{\text{IBAS}} = \text{diag}(\alpha_1, \ldots, \alpha_i, \ldots, \alpha_c)$. α_i here, denotes the coefficients of association derived from Equation 1 for the reference data. These coefficients of association may be selected as the direct coefficients from the linear mixed model (i.e., $\alpha_i = |\beta_i|$; coefficient-based weighting) or as the p-values associated with each coefficient from the same (i.e., $\alpha_i = -\log p_i$; p-value based weighting). This is then used in Equation 2 to calculate Q which follows a mixture of chi-square distributions [44] (Fig 1C).

When a single interaction component (e.g., the first principal component) is used, SNP weights reflect the association between genetic variants and a single dominant interaction pattern within the pathway. However, complex biological interactions are often distributed across multiple components, and relying on a single component may fail to capture the full interaction structure. To address this, we incorporate multiple interaction components and aggregate their corresponding SNP-level weights.

Specifically, for coefficient-based weighting, the meta-weight for SNP i ($i = 1, 2, ..., c$) is defined as

$$\widetilde{w}_i = \frac{1}{K}\sum_{k=1}^{K} |\beta_{ik}|,$$

where β_{ik} is the regression coefficient of SNP $i$obtained from the k-th interaction component, and K is the number of components (e.g., the top 10 principal components).

For p-value-based weighting, we combine the evidence across components using Fisher's method:

$$X_i = -2\sum_{k=1}^{K} \log(p_{ik}),$$

where p_{ik} is the p-value corresponding to SNP i for the k-th component. Under the null hypothesis, $X_i \sim \chi^2_{2K}$, from which a combined p-value is obtained and subsequently transformed into SNP weights.

These meta-weights summarize the overall contribution of each SNP to the dominant interaction structures within the pathway. Unless otherwise specified, coefficient-based aggregation is used in this study.

IBAS performs association testing at the gene level. For each gene within a pathway, SNP-level weights derived from pathway-informed interaction components are applied to its cis-variants, and a gene-level association statistic is computed using a weighted SKAT framework. Thus, similar to TWAS, the basic unit of association in IBAS is the gene rather than individual SNPs.

The output of IBAS is therefore a set of genes associated with the phenotype of interest, mediated through pathway-level interaction effects. As each pathway may contain a large number of genes, multiple hypothesis testing correction is performed using the Benjamini–Hochberg procedure [45] to control false discoveries.

Theoretical and empirical illustration of PCA for interaction structure

To demonstrate that PCA can recover gene–gene interaction structure and its associated co-expression patterns, we design a simulation with two gene-expression scenarios.

Let $p = 10$ denote the number of genes in a pathway and $N = 100$ the number of samples. For each scenario, we generate a gene expression matrix. For each scenario, we generate a gene expression matrix $X \in \mathbb{R}^{p\times N}$, where the i-th column $x_i \in \mathbb{R}^p$ represents the expression profile of the p genes in sample i, and

$$x_i \sim \mathcal{N}(0, \Sigma), \quad i = 1, ..., N,$$

In the independent scenario, we set the covariance matrix to be isotropic, $\Sigma = \Sigma_0 = \sigma^2 I_p$ with $\sigma^2 = 1.44$, yielding an expression matrix $X^{(0)}$. In the interaction scenario, we instead generate an expression matrix $X^{(1)}$ under a structured covariance $\Sigma = \Sigma_1$, where correlations among a subset of genes are introduced to simulate gene–gene interactions. Specifically, Σ_1 contains two disjoint correlated blocks: $\{gene_1, gene_2\}$ share pairwise covariance τ, and $\{gene_3, gene_4, gene_5\}$ share the same pairwise covariance τ (with $\tau = 1.152$), while all remaining gene pairs are uncorrelated. Each gene has marginal variance σ^2.

We then perform PCA by eigen decomposing the empirical gene–gene covariance matrix $\hat{\Sigma} = \frac{1}{N-1} XX^T \in \mathbb{R}^{p\times p}$. We let $\Lambda = diag\left(\lambda_1 \geq \cdots \geq \lambda_p\right)$ to denote the eigenvalues, which quantify the variance explained by each principal component, and let $V = [v_1, \ldots, v_p]$ to denote the PC loading vectors.

Under Σ_1, the two correlated blocks generate inflated leading eigenvalues $\sigma^2 + 2\tau$ and $\sigma^2 + \tau$ (S1 Notes), implying that for $\tau > 0$, variance concentrates along a small number of dominant eigendirections and the explained-variance ratio $\lambda_1 / \sum_{j-1}^{p} \lambda_j$ (and similarly for the next PCs). Correspondingly, the leading loading vectors align with the correlated gene modules, assigning large-magnitude weights to genes $\{gene_1, gene_2\}$ and $\{gene_3, gene_4, gene_5\}$. In contrast, under Σ_0 (independence), the empirical loadings fluctuate around zero without a reproducible block structure.

In our simulation results, PCA clearly distinguishes the interaction and independent scenarios. Under the interaction model, PC1 shows structured loadings with large magnitudes concentrated on genes from the dominant correlated module (Fig 2A), while PC2 captures the remaining module (Fig 2B). Consistent with this behavior, variance is strongly concentrated in the first few principal components in the interaction scenario (Fig 2C), reflecting the emergence of a low-dimensional dependence structure. Theoretically, under the normalized interaction model with within-block correlation $\rho = 0.8$, the leading eigenvalues are expected to be $1 + 2\rho = 2.6$ and $1 + \rho = 1.8$ for the second eigenvalue. In our simulation, the corresponding empirical eigenvalues are 2.74 and 1.86, respectively, closely matching these theoretical predictions and confirming that the interaction structure inflates the leading eigendirections as expected. In contrast, under the independent model, both PC1 and PC2 exhibit loadings that fluctuate around zero with no reproducible block pattern across genes (Fig 2A and 2B), and the explained variance is more evenly distributed across components (Fig 2C).

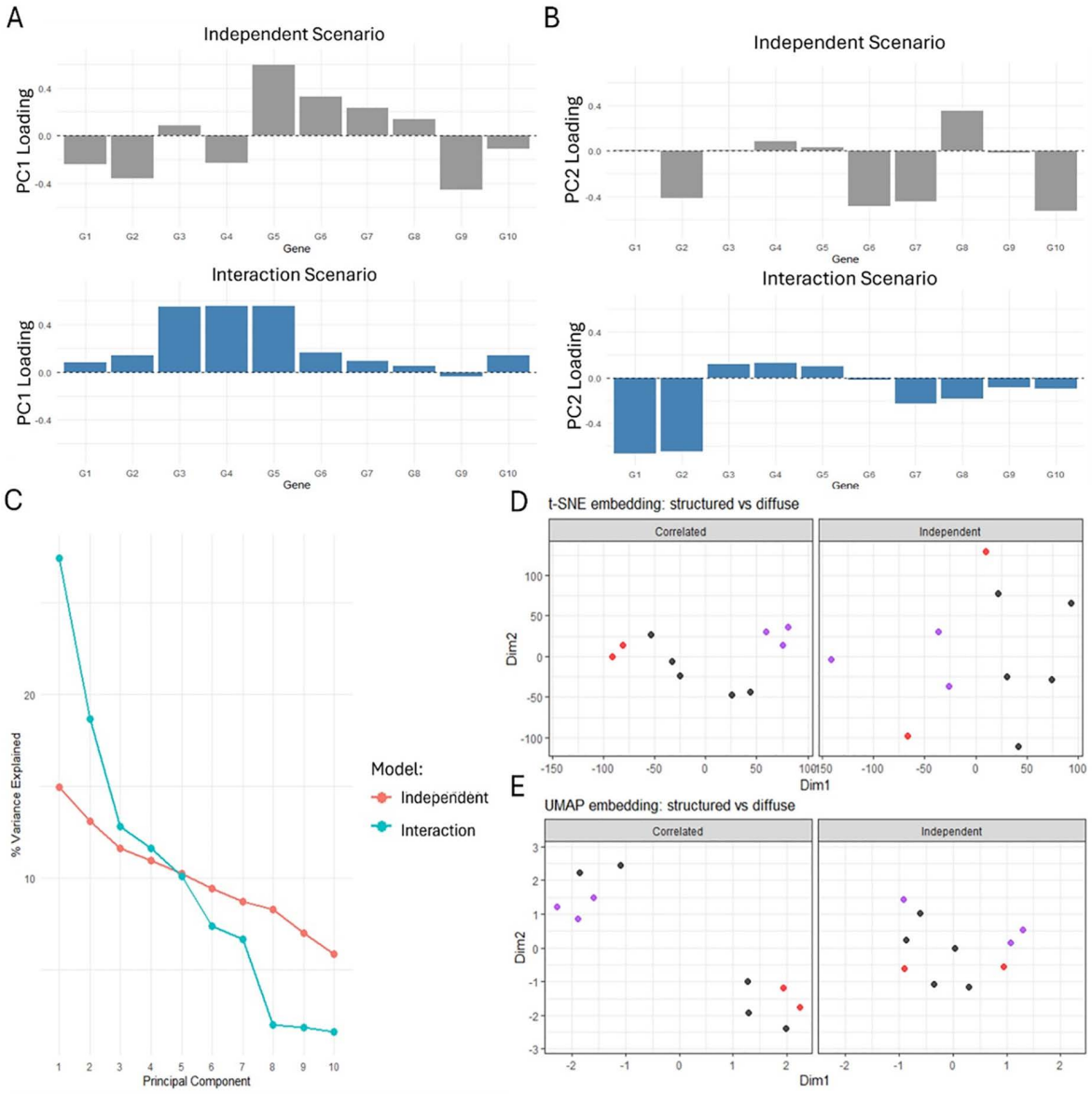

Fig 2. PCA and manifold learning reveal interaction-induced gene modules. A–B) PC1 and PC2 loadings for a simulated pathway under independent and interaction (block-correlated) scenarios. Under interaction, loadings reflect module structure ($\{gene_3, gene_4, gene_5\}$ for PC1; $\{gene_1, gene_2\}$ for PC2), while under independence they appear unstructured. **C)** Variance explained by PCs, showing stronger concentration in leading components under interaction. **D–E)** t-SNE and UMAP embeddings. Correlated genes cluster by module under interaction (Pink and Yellow) but show no clear grouping under independence.

https://doi.org/10.1371/journal.pcbi.1014640.g002

Prior work in data mining has also highlighted PCA as a fundamental tool for detecting correlation structure in high-dimensional data. For example, Böhm et al. [46] note that PCA can "perfectly find" linear correlations between genes, and they build correlation-based clustering methods on top of PCA for applications including gene expression analysis.

Beyond PCA, we also examined whether nonlinear manifold learning methods can recover the same interaction-driven dependence patterns. t-distributed stochastic neighbor embedding (t-SNE) is a nonlinear

dimension-reduction method that learns a low-dimensional embedding by preserving local neighborhood structure, typically by minimizing a divergence between neighborhood similarities in the original space and in the embedding, e.g.,

$$\min \mathrm{KL}(P \parallel Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

As a result, genes with similar expression profiles tend to be mapped close together. Therefore, when gene–gene interaction structure induces strong within-module correlation, t-SNE tends to place genes from the same correlated module in close proximity, whereas under independence such clustering is not expected. Consequently, when gene–gene interaction structure induces strong within-module correlation, genes in the same module share similar local neighborhoods and are mapped nearby in the t-SNE embedding. In Fig 2D, this is reflected by the tight grouping of $\{gene_1, gene_2\}$ (red) and the clustering of $\{gene_3, gene_4, gene_5\}$ (purple) under the interaction scenario, whereas under independence the highlighted genes are intermixed without clear module separation.

Similarly, Uniform manifold approximation and projection (UMAP) is a nonlinear embedding method that builds a weighted k-nearest-neighbor graph to represent local neighborhood structure in the original space and then learns a low-dimensional embedding that best preserves these neighbor relationships by minimizing a cross-entropy objective. A common choice for the embedding-space affinity is

$$\widehat{w}_{ij} = \frac{1}{1 + a \parallel y_i - y_j \parallel^{2b}}.$$

thus, when gene–gene interaction structure induces strong within-module correlation, genes within the same module tend to share local neighborhoods and are mapped close together, whereas under independence this clustering behavior is not expected. Under the interaction scenario, correlated genes receive consistently higher neighbor weights and form stable local neighborhoods, causing UMAP to place them in close proximity and yield module-level clustering. In contrast, under independence, neighbor relationships are weaker and less structured, producing embeddings with less coherent gene grouping (Fig 2E).

Although the above derivation assumes an idealized equi-covariance block structure, it illustrates a general mechanism: dependence among genes induces dominant eigendirections in the covariance (or correlation) matrix, concentrating variance into a small number of principal components. In more general settings with heterogeneous correlations, PCA no longer admits closed-form eigenpairs, but the leading PCs still recover the dominant low-dimensional dependence patterns. Beyond PCA, nonlinear embedding methods such as t-SNE and UMAP can also reflect gene–gene interaction structure by preserving local neighborhood relationships among genes. When interaction-driven co-expression induces within-module similarity, these methods tend to map genes from the same correlated module to nearby locations in the embedding space, producing module-level clustering patterns.

Estimation of type I error rates

To assess whether IBAS properly controls false positive rates, we conducted an empirical Type I error evaluation using phenotype permutation. Let $y_i \in \{0, 1\}$ denote the binary phenotype for individual i, and let n_1 and n_0 denote the numbers of cases and controls in the original dataset. For each replicate, a null phenotype vector

$$\mathbf{y}^{(b)} = (y_1^{(b)}, \ldots, y_n^{(b)})$$

was generated by randomly permuting the case–control labels while preserving the counts n_1 and n_0. This procedure removes any genuine genotype–phenotype association while retaining the original genotype matrix **G**, linkage disequilibrium structure, and gene-level variant composition.

For each permuted phenotype vector, the complete IBAS workflow was applied, including interaction-component construction, variant weighting, and the final gene-level association test using the SKAT statistic Q defined in Equation (2). The empirical Type I error rate was then estimated as

$$\widehat{T1E} = \frac{1}{N_g} \sum_{i}^{N_g} \mathbf{1}\left(p_i < \alpha\right)$$

where p_i denotes the gene-level p-value obtained under the null phenotype, α is the nominal significance level, and N_g is the total number of tested genes.

In addition to this numerical estimate, calibration of the null distribution was visually assessed using quantile–quantile (QQ) plots comparing observed p-values to the theoretical $\text{Uniform}(0, 1)$ distribution. Agreement between observed and expected quantiles indicates appropriate control of false positives under the null hypothesis.

Simulation framework for generating perturbed data

To test whether interaction-based protocol is more stable than single-gene based ones, perturbation experiments were conducted to compare them. Simulations were carried out by adding perturbations to gene expression data from the Whole Blood tissue in GTEx [19] to evaluate robustness to changes in reference data. Each expression value was regenerated using a uniform distribution $U \sim \left(x(1 + r),\ x(1 - r)\right)$ where x is the current expression value and $r = \{0.1,\ 0.25,\ 0.5\}$. This procedure preserves overall pathway-level expression patterns while introducing gene-level noise, mimicking variability arising from factors such as sampling time and sequencing batch effects. Increasing values of rcorrespond to higher noise levels, and 10 replicates were generated at each level, resulting in 30 perturbed expression datasets.

These datasets were used, together with unperturbed genotype data, to construct data bridges for IBAS under different settings (e.g., dimensionality reduction methods and weighting strategies), as well as to train PrediXcan models. In addition, SKAT-O was applied as a baseline method without expression-derived weighting, and MAGMA was included as a reference approach that does not use SNP-level weights. The resulting models were then evaluated through association testing in the WTCCC cohorts to compare the stability and robustness of results across methods.

Simulation framework for empirical power evaluation

To complement the perturbation-based stability analysis and empirical Type I error evaluation, we designed a controlled power simulation with predefined causal genes. The simulation was intended to mimic a pathway-mediated disease architecture, where multiple genes within the same coordinated regulatory program jointly contribute to phenotypic variation. In each replicate, causal genes were planted within pathway-level expression structures, and phenotypes were generated under additive, hybrid, and interaction models across multiple heritability levels, allowing us to evaluate the ability of IBAS and competing methods to recover known causal genes under controlled genetic architectures. As in the main IBAS analysis, GTEx expression data were used as the reference transcriptome to derive pathway-level PCA components and estimate SNP weights, while WTCCC genotype data were used as the target dataset from which simulated phenotypes were generated.

Specifically, in each simulation replicate, one pathway was randomly selected, and the genes within that pathway were ranked according to the absolute values of their PC1 loadings derived from the reference expression data. Because PC1 captures the dominant coordinated expression pattern within a pathway, genes with large absolute loadings represent major contributors to this pathway-level regulatory structure. We therefore randomly selected 10 causal genes from the top 50 ranked genes, mimicking a setting in which disease risk is driven by genetic perturbations of genes central to a

shared pathway program. Then for each causal gene, five cis-variants were randomly selected. When two genes were assigned as an interaction pair, their selected SNPs were matched one-to-one to form five SNP-pair interaction terms.

For each model, a genotype-derived genetic component g was first constructed from the selected causal variants and then combined with environmental noise to achieve the target heritability. Given a prespecified h^2, we generated

$$y = g + \epsilon,$$

where $\epsilon \sim N\left(0, \sigma_\epsilon^2 I\right)$, and σ_ϵ^2 was chosen such that

$$h^2 = \frac{\mathrm{Var}(g)}{\mathrm{Var}(g) + \mathrm{Var}(\epsilon)}.$$

We considered four heritability levels: $h^2 = 0.05$, 0.1, 0.4, and 0.7.

To evaluate method performance across different levels of interaction involvement, we used three phenotype-generating architectures. The additive model served as a baseline setting in which phenotypic variation was driven by marginal SNP effects, the interaction model represented an interaction-enriched setting driven by gene-gene interaction terms, and the hybrid model provided an intermediate architecture containing both marginal and interaction-mediated effects.

In the additive model, phenotypic variation was driven by marginal SNP effects across all 10 causal genes:

$$g_{\mathrm{add}} = \sum_{j=1}^{10} \sum_{k=1}^{5} \beta_{jk} x_{jk},$$

where x_{jk} is the genotype vector of the k-th selected SNP in causal gene j, and $\beta_{jk} \sim N(0, 1)$ is the corresponding additive effect size.

In the hybrid model, four causal genes contributed additive effects, while the remaining six causal genes were assigned into three interaction pairs,

$$g_{\mathrm{hyb}} = \sum_{j=1}^{4} \sum_{k=1}^{5} \beta_{jk} x_{jk} + \sum_{(a,b) \in \mathcal{P}_{\mathrm{hyb}}} \sum_{k=1}^{5} \gamma_{abk} \left(x_{ak} \odot x_{bk}\right).$$

Where $\mathcal{P}_{\mathrm{hyb}} = \left(G_5, G_6\right), \left(G_7, G_8\right), \left(G_9, G_{10}\right)$.

In the interaction model, the 10 causal genes were assigned into five gene pairs, $\mathcal{P}_{\mathrm{int}} = \left(G_1, G_2\right), \left(G_3, G_4\right), \left(G_5, G_6\right), \left(G_7, G_8\right), \left(G_9, G_{10}\right)$, and phenotypic variation was generated only from cross-gene SNP interaction effects:

$$g_{\mathrm{int}} = \sum_{(a,b) \in \mathcal{P}_{\mathrm{int}}} \sum_{k=1}^{5} \gamma_{abk} \left(x_{ak} \odot x_{bk}\right),$$

where $\odot$ denotes elementwise multiplication across individuals, and γ_{abk} is the interaction effect size.

For each combination of genetic architecture and heritability level, 500 simulation replicates were generated. We compared IBAS using PCA-derived weights with SKAT-O, PrediXcan, and MAGMA. For each method, gene-level p-values were adjusted using the Benjamini–Hochberg procedure at FDR 0.05, and empirical power was defined as the proportion of planted causal genes declared significant after correction.

$$\mathrm{Power} = \frac{\left|\hat{G}_{\mathrm{sig}}\right|}{\left|G_{\mathrm{causal}}\right|},$$

where G_{causal} is the set of planted causal genes and $\hat{G}_{sig}$ is the set of causal genes significant after BH correction.

Real data analysis: Reference data

Publicly available, normalized, and filtered expression data from the "Whole Blood" tissue of the GTEx Consortium [19], comprising 670 individuals and 20,315 transcripts, were used as the reference expression dataset. This data was then corrected for PEER factors [47], the top 5 genotype principal components, age and gender as suggested by the GTEx Portal. This expression data was then used for dimensionality reduction by subsetting for individual pathways. For tissue-specific evaluation, GTEx "Brain Cerebellum" and "Pancreas" expression data were processed using the same pipeline as "Whole Blood" and used as alternative reference datasets for IBAS training for Bipolar Disorder and Type 1 Diabetes, respectively.

The GWAS genotype dataset consisted of 838 individuals and 43,066,422 SNPs in the raw VCF data. Since IBAS relies on SNP-level information shared between the GWAS dataset and the reference dataset used for expression modeling, a high degree of overlap in SNP positions is desirable to ensure sufficient variants are available for weighting and association testing. Therefore, imputation was carried out using SHAPEIT4 [48] for phasing and IMPUTE5 [49] for the final imputation based on data from the 1000 Genome Project [50]. This results in 88,863,451 SNPs for the imputed data. The data was then filtered to include only individuals who had expression data available and general QC (MAF < 0.05, deviation from HWE $p < 10^{-8}$ and relatedness > 0.025) was applied, resulting in 6,187,683 SNPs for 553 individuals. This data was converted to the {tped, tfam} file format and JAWAMix5 was used to convert it to hdf5 [51] for out-of-core association analysis with the interaction components obtained from the expression data.

Real data analysis: GWAS data containing genotype and phenotype

The Wellcome Trust Case Control Consortium [28] contains genotype data for 13,241 cases and 2,938 controls (individual sample sizes are provided in S1 Table) across 7 diseases: bipolar disease (BD), coronary artery disease (CAD), Crohn's disease (CD), rheumatoid arthritis (RA), type 1 diabetes (T1D), type 2 diabetes (T2D) and hypertension (HT). The original dataset only contained 393,272 variants since this data was obtained using the Affymetrix GeneChip 500K arrays and thus imputation [49] was conducted as in GTEx genotype data to produce 73,162,888 variants. To further evaluate the performance of IBAS in larger cohorts, we additionally analyzed an independent T1D genome-wide association dataset from dbGaP (Accession: phs000911). This dataset comprises 10,791 samples, substantially exceeding the sample size of the WTCCC T1D cohort (4,901 samples). The inclusion of this dataset enables assessment of the reproducibility and scalability of IBAS findings in a higher-powered setting.

Results

IBAS is more stable than single gene-mediated TWAS

The stability comparison between IBAS and PrediXcan across different reference datasets conveys a consistent and encouraging message. We utilized real data from the GTEx Consortium to construct the data bridge and investigated its effects on the Type 1 Diabetes (T1D) cohort from the WTCCC dataset (Fig 3), along with six additional cohorts presented in S1–S6 Figs to ensure that this stability generalizes across diverse genetic architectures.

All methods considered, including IBAS, perform gene-level association testing using cis-variants within a kernel-based framework. The key distinction lies in how SNP-level weights are constructed. In PrediXcan and kTWAS, weights are derived from gene-specific genotype–expression models, whereas IBAS obtains weights from regression on pathway-level interaction components. Thus, IBAS incorporates pathway-level information while retaining gene-level testing.

To broaden benchmarking beyond TWAS-based approaches, we additionally included the kernel-based gene association method SKAT-O [52] in the perturbation simulations for the T1D cohort. SKAT-O is a widely used gene-level

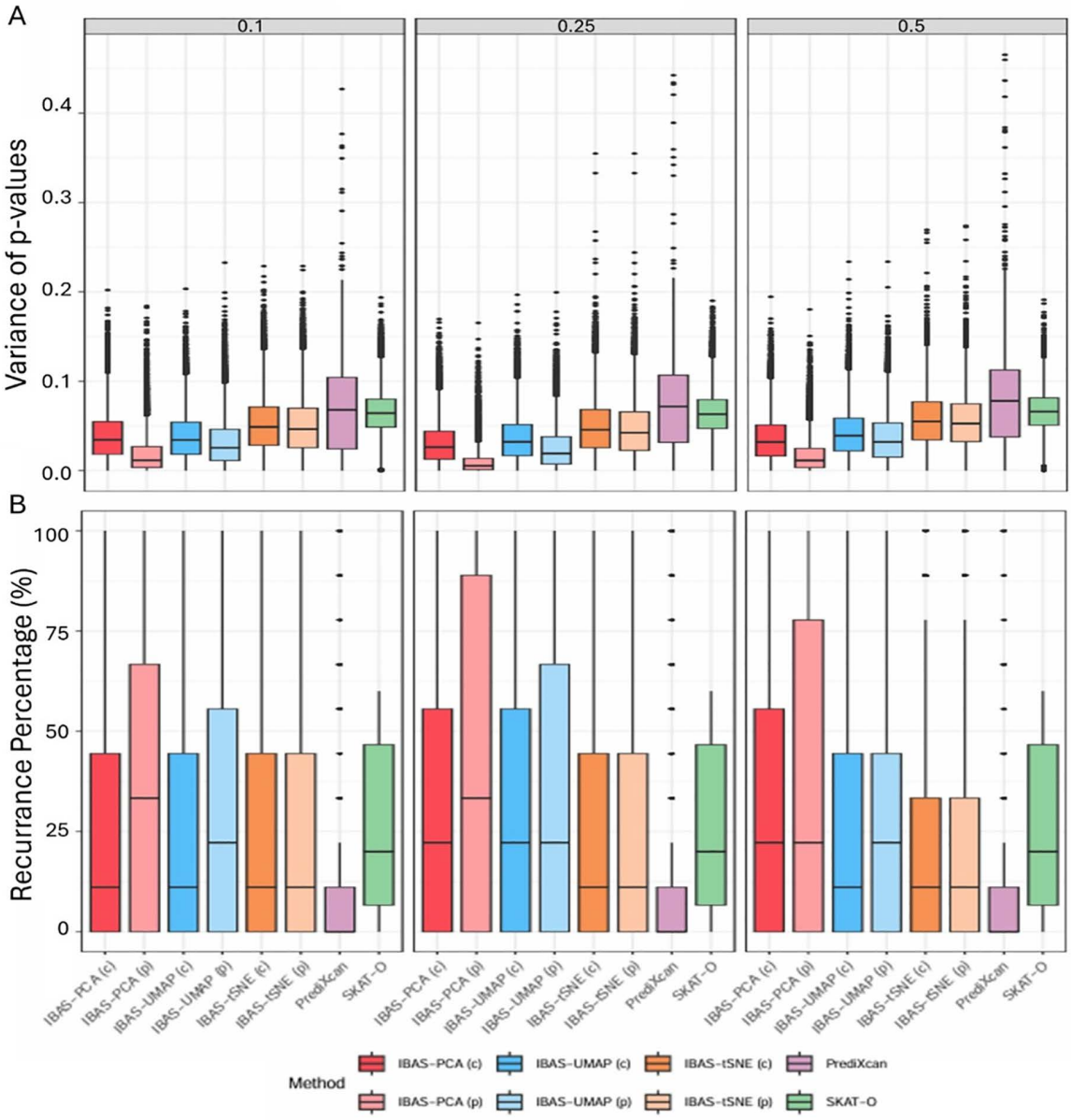

Fig 3. Simulated Perturbations reveal robustness of IBAS to noise in reference data. Simulations were conducted by perturbing gene expression values of each gene by percentage ranges (0.1, 0.25 and 0.5) in the reference data. 10 perturbed expression datasets (replicates) were generated for each of the noise levels and variability of results for case-control association with the Wellcome Trust Case Control Consortium Type 1 Diabetes cohort was evaluated (other cohorts were tested and available in S1-S6 Figs) using IBAS, PrediXcan and SKAT-O. PCA, t-SNE and UMAP dimensionality reduction methods as well as p-value based (p) and coefficient based (c) interaction association weights were tested in the case of IBAS. **A)** Indicates the variance of observed p-values in genes identified as significant ($p < 0.05$) in at least one replicate while **B)** indicates the recurrent significance of genes identified as significant in at least one replicate.

https://doi.org/10.1371/journal.pcbi.1014640.g003

framework that aggregates variant effects without relying on transcriptomic mediators. As an additional reference, we also applied the genotype-based gene analysis method MAGMA [53] to the WTCCC dataset. Because MAGMA does not incorporate gene expression, it is not included in the main figure; its results are instead provided in the S2 Table.

Across all settings, IBAS-based methods showed clear advantages over both PrediXcan and SKAT-O in the stability of discoveries. Fig 3A shows that IBAS achieves substantially lower variance in p-values across a wide range of perturbation levels for genes identified as significant ($p < 0.05$) in at least one replicate. In particular, PCA-based IBAS demonstrates markedly reduced variability compared with competing methods. Weighting by p-value, rather than by coefficient, further improves stability across all dimensionality reduction approaches.

A similar pattern is observed in the recurrence analysis (Fig 3B), which measures the proportion of genes repeatedly identified as significant across perturbations. PrediXcan exhibits low recurrence, indicating that discoveries are highly sensitive to small changes in the reference data. In contrast, IBAS consistently achieves higher recurrence rates, with PCA-based, p-value–weighted IBAS performing best.

Taken together, these results demonstrate that interaction-based modeling in IBAS yields substantially more robust and reproducible gene-level associations than single gene-mediated TWAS approaches, while also outperforming alternative aggregation methods such as SKAT-O. The MAGMA results provide additional context, confirming that IBAS maintains advantages over conventional GWAS-based gene analysis frameworks.

Type I Error is under control in IBAS

To further verify the statistical validity of IBAS, we evaluated its calibration under the null hypothesis using phenotype permutation while preserving the original case–control counts. Across replicated null datasets, the empirical Type I error at the nominal level of 0.05 was estimated to be 0.0524, demonstrating that the IBAS testing procedure maintains appropriate control of false positives and closely matches the expected nominal rate. Quantile–quantile analysis of gene-level p-values additionally showed strong agreement with the theoretical Uniform(0,1) distribution (Fig 4), further confirming that the null distribution is well calibrated and that the observed stability of IBAS is not accompanied by inflation of spurious associations.

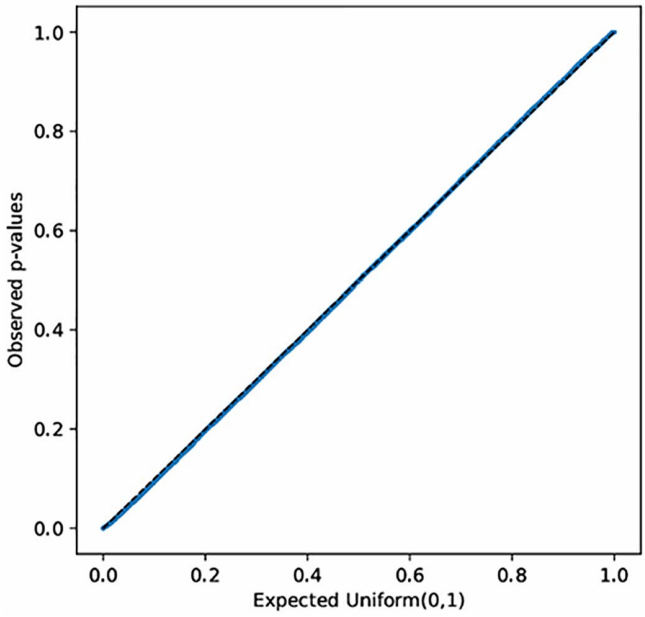

Fig 4. Evaluation of Type I error control in IBAS using QQ plot of null p-values. Quantile–quantile (QQ) plot of gene-level p-values obtained from the IBAS framework under the null hypothesis based on permuted phenotype vectors. The observed p-values closely follow the theoretical **Uniform**$(0, 1)$ distribution (diagonal line), indicating well-calibrated Type I error and appropriate control of false positives.

https://doi.org/10.1371/journal.pcbi.1014640.g004

IBAS improves empirical power in simulation studies

To further evaluate the performance of IBAS, we assessed empirical power in simulations with predefined causal genes. This analysis tested whether the pathway-level interaction bridge used by IBAS improves the recovery of true genetic signals across different genetic architectures. By comparing additive, hybrid, and interaction architectures, the simulation further assessed whether the advantage of IBAS depends on the extent to which phenotypic variation is mediated through coordinated genetic effects within pathways. Across all three simulation models, IBAS using PCA-derived weights showed stronger recovery of causal genes than SKAT-O, PrediXcan, and MAGMA (Fig 5). IBAS achieved comparable or higher power under the additive model, with increasingly clear advantages under the hybrid and interaction models. These results suggest that the pathway-level interaction representations learned by IBAS provide informative genetic prioritization beyond marginal gene effects or unweighted variant aggregation, enabling more effective recovery of coordinated genetic signals.

Although this simulation provides a simplified representation of complex disease biology, it offers a controlled setting for evaluating the recovery of known causal genes. The higher empirical power of IBAS suggests that pathway-informed interaction weights can improve sensitivity beyond conventional gene-based methods, while the real-data analyses further support its practical utility in identifying biologically meaningful genes and pathways across complex diseases.

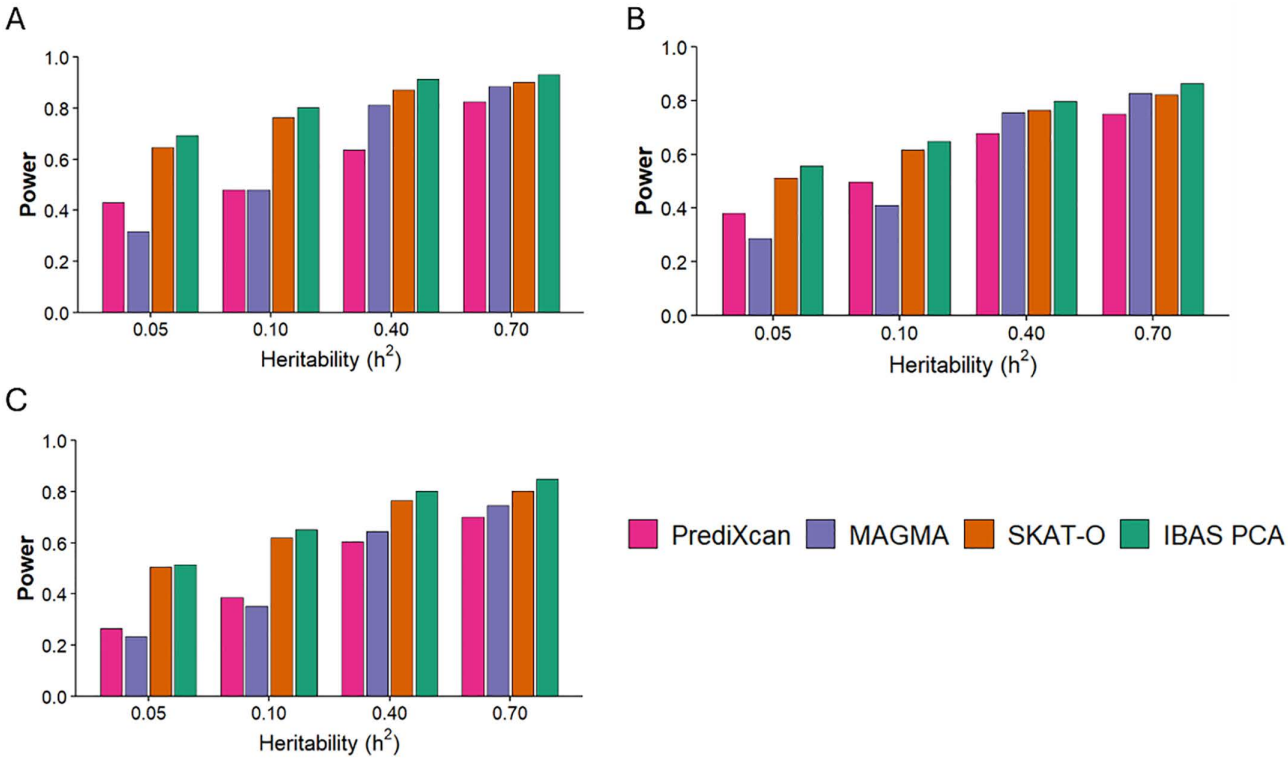

Fig 5. Empirical power evaluation using planted causal gene simulations. Empirical power was evaluated as the proportion of predefined causal genes recovered after Benjamini–Hochberg correction across 500 simulation replicates for each heritability level. IBAS using PCA-derived weights was compared with SKAT-O, PrediXcan, and MAGMA under three phenotype-generating models: **A)** additive model, where causal variants contribute through marginal SNP effects; **B)** hybrid model, where both additive and cross-gene interaction effects contribute to the phenotype; and **C)** interaction-enriched model, where phenotypic variation is driven by cross-gene product terms designed to mimic pathway-level dependency effects. Across the three settings, IBAS showed consistently strong recovery of causal genes, with more pronounced advantages under interaction-enriched architectures.

https://doi.org/10.1371/journal.pcbi.1014640.g005

PCA loadings identify biologically coherent gene modules within pathways

To characterize the biological content of the interaction components generated by dimensionality reduction, we analyzed gene loadings within representative pathways and evaluated whether dominant contributors correspond to known functional modules. We found that genes with the largest absolute loadings consistently formed biologically coherent groups aligned with established pathway functions. In each examined case, the leading principal components were driven by coordinated gene modules reflecting structured pathway-level gene activity. These results demonstrate that the low-dimensional interaction representations derived in IBAS capture biologically organized pathway mechanisms.

For a pathway containing p genes, PCA decomposes the covariance matrix of the expression matrix into eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ and corresponding eigenvectors $v_1, v_2, \ldots, v_p$. The first principal component (PC1), associated with λ_1, captures the maximal variance within the pathway-level expression space. The entries of $v_1 = (v_{11}, \ldots, v_{1p})^\top$ represent gene loadings, quantifying each gene's contribution to the dominant interaction axis. Genes were ranked by absolute loadings $| v_{1j} |$, as magnitude reflects strength of contribution independent of sign. The top five genes were then evaluated for functional consistency with pathway biology.

As an illustrative example, we examined the MAPK signaling pathway (hsa04010), under Environmental Information Processing and Signal Transduction. Ranking genes by absolute loadings of the first principal component identified *HSPA1A*, *HSPA6*, *CSF1*, *HSPA1L*, and *HSPA1B* as the strongest contributors to the dominant variance axis (S3 Table). Four of these genes belong to the HSP70 heat-shock protein family, which are well known to be strongly co-expressed under cellular stress conditions. MAPK cascades regulate stress-responsive transcriptional programs, and coordinated induction of heat-shock genes is a characteristic downstream response. In addition, *CSF1*, a cytokine involved in immune and inflammatory signaling, engages MAPK pathways and participates in coordinated activation programs. The emergence of these genes as top contributors indicates that the principal component captures a structured co-expression pattern consistent with stress and inflammatory signaling within the pathway.

A similar pattern was observed in the Cytokine–cytokine receptor interaction pathway (hsa04060), categorized under Environmental Information Processing and Signaling Molecules and Interaction (S3 Table). The genes with the largest absolute loadings included *CCL4*, *IL10RB*, *TNFRSF1A*, *CCL3*, and *CXCR2*, comprising chemokines and their cognate receptors central to immune signaling. Chemokines such as *CCL3* and *CCL4* are frequently co-induced during inflammatory activation, while receptors including *CXCR2* and *TNFRSF1A* mediate coordinated cytokine responses. The concentration of these immune signaling genes among the top loadings indicates that the principal component reflects a coherent inflammatory signaling program within the pathway, arising from shared covariance structure among functionally related genes.

Together, these examples demonstrate that principal components derived from pathway-level expression matrices capture coordinated gene activity that aligns with established biological functions. Because IBAS leverages these structured interaction representations as mediators in genotype–phenotype association testing, the method effectively links genetically driven variation in coordinated pathway programs to phenotypic outcomes. This provides both statistical stabilization through dimensionality reduction and mechanistic interpretability grounded in known pathway biology.

IBAS discovers novel genes through the mediation of interactions

IBAS uncovered significant discoveries across the 7 diseases of the WTCCC case-control cohort; a number of which were previously identified in the DisGeNET database [54] with an additional number of discoveries that were novel (Fig 6). Most discoveries were for the Rheumatoid Arthritis (RA) and T1D cohorts both of which also contain the most annotations in the DisGeNET database being autoimmune diseases that have been extensively studied. However, IBAS was also able to uncover significantly associated genes for Bipolar Disorder (BD), Coronary Artery Disease (CAD), Hypertension (HT) and Type-2 Diabetes (T2D) (S4 Table). Furthermore, the discovery of numerous novel genes is expected and welcomed since most association studies incorporate purely genetic information (or, in the case of TWAS, transcriptomic information of a single gene).

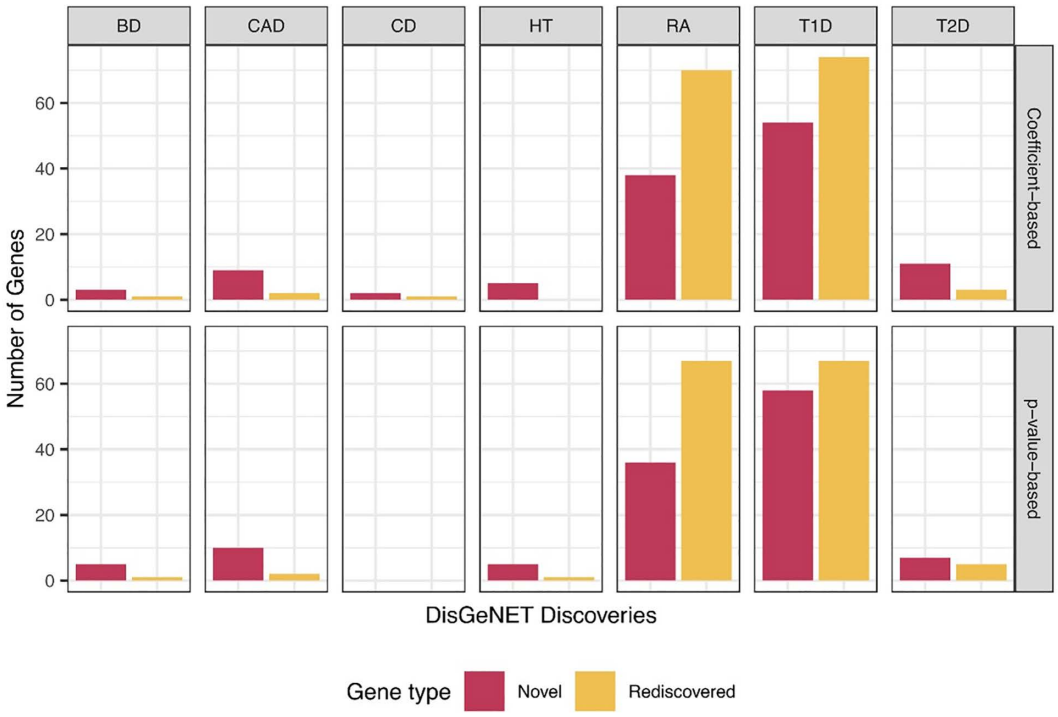

Fig 6. Evaluation of genes associated with disease uncovered by IBAS. Genes identified as associated with each of the 7 diseases (columns) in the WTCCC dataset (with GTEx used as reference) across both coefficient-based and p-value based (rows) for the PCA data-bridge were annotated as "Rediscovered" if found as previously associated with the same disease in the DisGeNET database. Remaining genes are annotated as "Novel". Evaluation of t-SNE and UMAP can be found in S13-S14 Figs.

https://doi.org/10.1371/journal.pcbi.1014640.g006

IBAS shows contrastive results compared to other gene-disease association tools (Fig 7). PrediXcan [5] and kTWAS [20] are used for this comparison since they represent analyses based on the typical TWAS framework and a kernel-based framework similar to IBAS but without the interaction effects, respectively. Results from IBAS are once again based on the meta-weights (mean for coefficient based and Fisher combined p-values for p-value based across all dimensionalities reduced components).

kTWAS and PrediXcan were found to report fewer results in most diseases, but with a level of overlap between them that suggests that while kTWAS does gain more power due to the use of the kernel-based test for feature aggregation as shown previously [20], they extract similar information from transcriptomic data. IBAS on the other hand, incorporates interactions from gene expression and thus is expected to discover genes based on their association with these interactions. The high number of genes discovered in T1D and RA by both kTWAS and PrediXcan further support the theory previously mentioned that there may be a high marginal effect from genes associated with these diseases. It is also noteworthy that IBAS uncovers several genes associated with HT: one of which, *CLOCK*, has known associations with HT in the DisGeNET database (Fig 6). This further strengthens the hypothesis that IBAS is able to effectively uncover genes that may be associated with disease through their interactions at the pathway level.

Biological annotation of results

IBAS distinguishes itself from previous annotations in the WTCCC cohort by identifying genes that may have a small marginal effect but contribute to disease phenotype mediated through relevant pathways (Fig 8). In particular, observing the patterns in diseases where many significant genes are identified such as RA and T1D, previously annotated diseases are

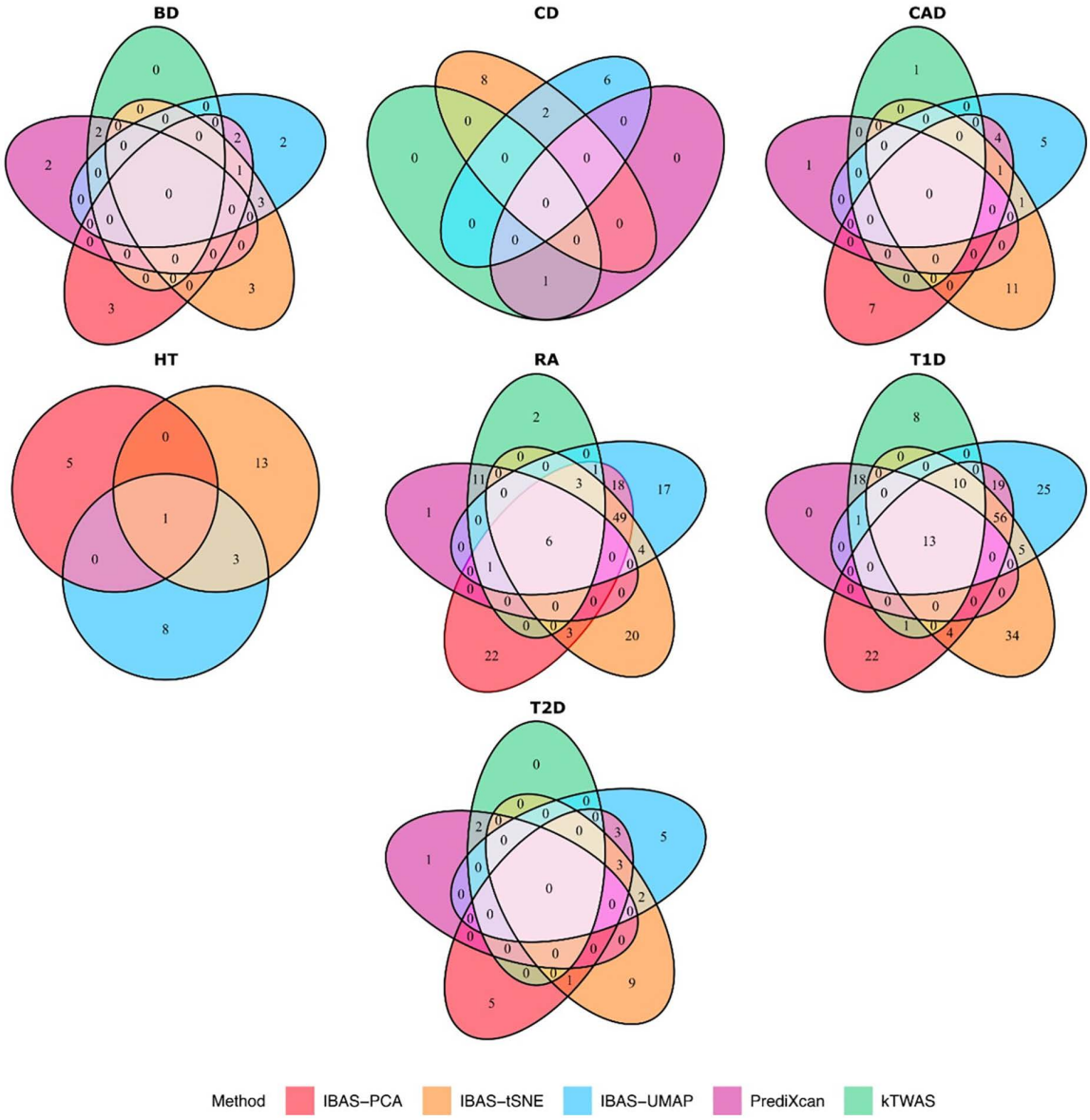

Fig 7. Comparison of IBAS results with kTWAS and PrediXcan. Genes identified by each of the IBAS methods (PCA, t-SNE and UMAP) as well as PrediXcan and kTWAS were compared to identify overlap and uniqueness of discoveries across the 7 diseases of the WTCCC dataset.

https://doi.org/10.1371/journal.pcbi.1014640.g007

identified across multiple pathways – indicating that their strong marginal effect may have contributed to their significance. This could explain the discovery of a large number of genes in the HLA region being identified as significant for these auto-immune disorders [55] as well as *TNF* [55] and *ITPR3* [56]. The weaker marginal effects of genes may explain the lower number of genes identified to be associated with the other diseases since most methods do not account for pathway-level interactions as IBAS does.

IBAS identifies a key mechanism suggesting the mediation of metabolism related pathways and genes in BD. 4 out of the 5 novel genes identified by IBAS are associated with various metabolic processes [57–60]. The role of metabolism in

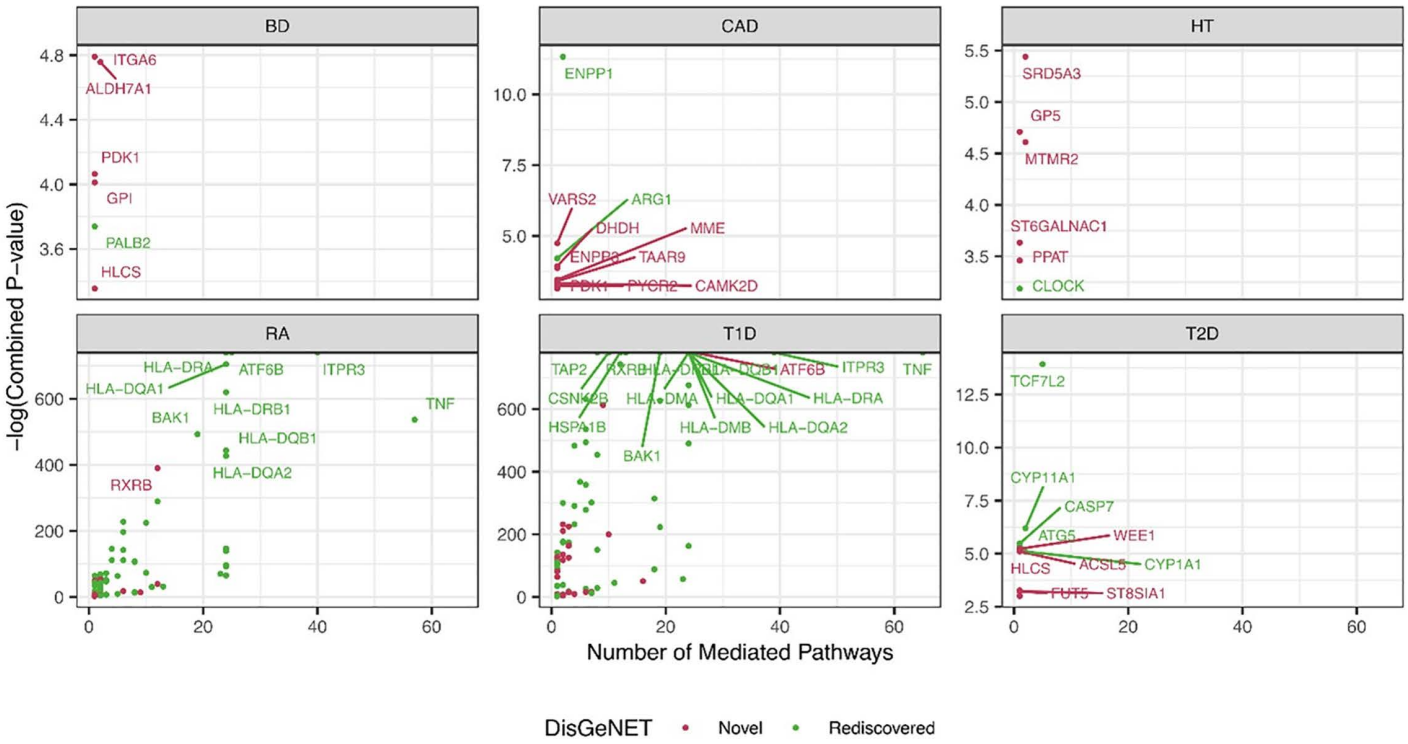

Fig 8. IBAS identifies associations with marginal effects as well as interaction-driven effects. Comparison of significance of previously discovered ("Rediscovered") and "Novel" genes associated with the 6 disease (no significant genes associated with CD using IBAS-PCA) cohorts of the WTCCC dataset. Combined P-values (y-axis) indicate Fisher combined p-values of genes identified across multiple pathways; while the x-axis denotes the number of pathways mediated through which the gene was identified in as associated with the phenotype.

https://doi.org/10.1371/journal.pcbi.1014640.g008

psychiatric disorders has long been investigated [61] with a clear association established between dysfunctional metabolism and psychiatric disorders including BD [62].

HLCS, a gene known for its role in biotin metabolism [63] in the body, is found to be highly associated with BD through the biotin metabolism pathway [32,64] by IBAS indicating a potential pathway that could mediate the development of BD. Biotin or Vitamin B7 is known to be vital to brain function as it is a crucial component in the process of glucose metabolism [65]. Furthermore, *HLCS* is highly specific to neurons (both inhibitory and excitatory) in the brain according to the Human Protein Atlas [66,67]. Changes in the excitatory and inhibitory (E/I) synaptic balance have been previously associated with bipolar disorder in animal models [68] and the specificity of HLCS as well as its importance in metabolism make it a potential mediator in the emergence of the BD phenotype. While no studies have been conducted on the direct association of *HLCS* on BD, one study identified increased methylation at CpG sites in *HLCS* in borderline personality disorder patients in their genome-wide association study but failed to find the same in a validation analysis [69]. Increased methylation could indeed be a potential mechanism of repression of *HLCS* expression, resulting in dysfunctional biotin-dependant metabolic pathways which in turn could create an imbalance in the activation of E/I neurons contributing to manic episodes.

In the case of other significant associations, IBAS hints at potential biologically pivotal candidates in disease etiology. *VARS2*, the most significant gene identified with association to CAD (excluding all previously annotated genes in DisGeNET) has been identified to cause heart failure in zebrafish models when knocked-out [70]. Another significant gene, *ENPP3*, has known cardiac-specific function and is downstream of *NKX2–5*, a transcription factor associated with

congenital heart disease [71]. *PPAT*, identified to be significant in the HT phenotype, was concluded to be an important gene in the control of BP rhythm in rats [72]. A significant gene (*ACSL5*) identified in T2D, has been studied to be the primary target of the most strongly associated variant of T2D in the *TCF7L2* gene [73], and is important in the metabolism of fatty acids.

To further clarify the unit of association in IBAS, we present Manhattan plots of gene-level association results across the genome. Each point represents a gene, with its genomic position defined by gene location and its significance determined by the IBAS gene-level test statistic (Figs 9 and S7–S12).

Sensitivity to the number of principal components

To evaluate the sensitivity of IBAS to the choice of dimensionality, we repeated the WTCCC T1D analysis using K = 5, 10, and 20 retained pathway components. As expected, the exact composition of the top-ranked genes varied across settings, reflecting the distributed covariance structure within biological pathways. The overlap between the top 50 genes identified using K = 10 and K = 20 was 23 genes (46%), and 14 genes were consistently identified across all three analyses. Importantly, these shared genes included several well-established T1D-associated genes within the major histocompatibility complex (MHC) region, including AGER, C2, HLA-B, HLA-DRA, LTA, TNF, and TNXB. Many of these genes were also prioritized in the independent dbGaP T1D cohort, indicating that the principal biological signals recovered by IBAS are reproducible across datasets and are not driven by a particular choice of dimensionality. These results suggest that the strongest disease-relevant signals, particularly MHC-region T1D genes, are stable across reasonable choices of K, whereas the exact ranking and composition of lower-priority genes remain sensitive to dimensionality. Notably, the MHC region represents the most robust and extensively validated genetic locus for T1D, and the persistence of multiple MHC genes across all K values and across independent cohorts supports the stability of the major biological conclusions derived from IBAS.

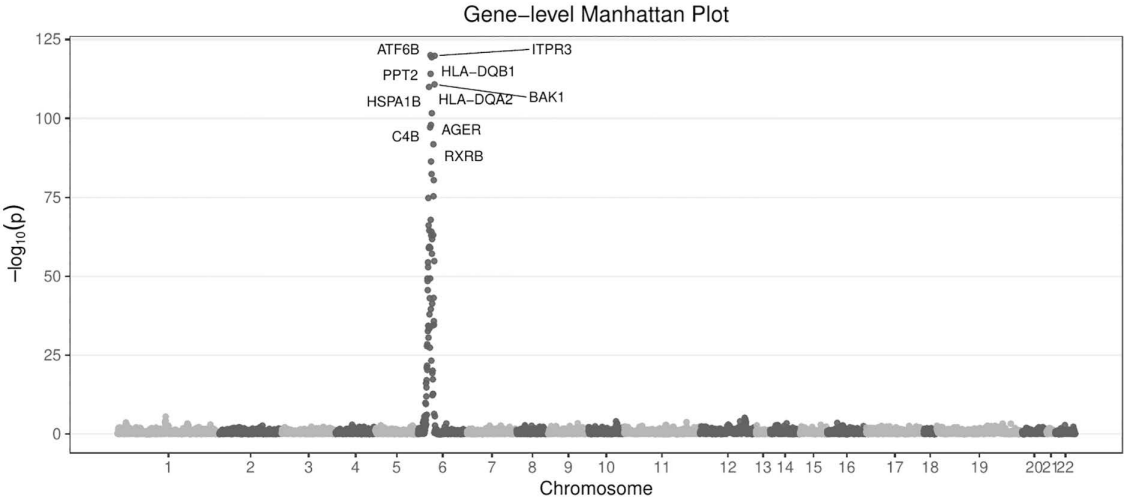

Fig 9. Gene-level Manhattan plot for type 1 diabetes (T1D) in the WTCCC dataset using IBAS. Gene-level Manhattan plot for the WTCCC type 1 diabetes (T1D) dataset using the IBAS framework. Each point represents a gene, plotted by chromosomal position (x-axis) and its association significance measured as $-\log_{10}(p)$ (y-axis). Gene-level statistics were constructed using PCA-based interaction components and coefficient-based variant weighting, with expression weights derived from GTEx Whole Blood tissue. The labeled genes correspond to the top 10 most significant associations.

https://doi.org/10.1371/journal.pcbi.1014640.g009

Validation of IBAS discoveries in a larger independent T1D cohort

To further evaluate whether the discoveries from the WTCCC cohort remain detectable in larger datasets, we applied IBAS to an independent T1D genome-wide association cohort from dbGaP (Accession: phs000911). For comparability with the WTCCC analysis, the same reference data, pathway definitions, and interaction-derived variant weights from GTEx whole blood were used. Genotype preprocessing and quality control followed the same general pipeline as for the WTCCC data.

Applying IBAS to the dbGaP cohort produced a set of significant genes that showed substantial concordance with the WTCCC discoveries. The results reported here are based on the PCA IBAS framework with coefficient-based variant weighting, consistent with the primary configuration used in the WTCCC analysis. Approximately 70% of the genes identified in the WTCCC analysis were also detected in the dbGaP dataset, indicating that the interaction-mediated signals uncovered by IBAS are reproducible across independent cohorts despite differences in sample composition and cohort size (S5 Table). The shared genes are strongly enriched in the extended major histocompatibility complex (MHC) region, including multiple classical HLA genes such as *HLA-DQA1*, *HLA-DQB1*, *HLA-DRB1*, *HLA-A*, *HLA-B*, and *HLA-C*, as well as immune-related genes located in the same region such as *TNF*, *LTA*, *LTB*, *TAP1*, *TAP2*, and *PSMB8*. Additional immune-associated genes, such as *ITPR3*, were also consistently detected across the two cohorts. This pattern is consistent with the well-established role of the MHC region as the dominant genetic contributor to T1D susceptibility.

Importantly, the larger dbGaP cohort also revealed additional genes that were not detected in the WTCCC analysis but have strong biological relevance to T1D. These include several canonical non-HLA T1D susceptibility genes such as *INS*, *IL2RA*, *CTLA4*, and *ERBB3*, which are well known to regulate immune responses and pancreatic β-cell function. The appearance of these established loci in the larger dataset suggests that increasing sample size improves the sensitivity of IBAS to detect genes with moderate effects mediated through interaction patterns. At the same time, the preservation of a substantial subset of WTCCC discoveries demonstrates that IBAS captures stable interaction-mediated genetic signals rather than cohort-specific artifacts.

The partial overlap between cohorts is also expected given the complex genetic architecture of T1D. Association signals are highly concentrated within the extended MHC region, where strong linkage disequilibrium and dense gene clusters can lead to multiple correlated genes appearing significant depending on cohort structure and statistical power. Moreover, because IBAS identifies genes whose variants influence interaction structures within biological pathways, rather than relying solely on marginal single-gene effects, the exact ranking of genes can vary across datasets while still reflecting the same underlying interaction-driven genetic mechanisms.

Together, these results provide additional support that IBAS identifies biologically meaningful genes underlying T1D and that its discoveries remain detectable when applied to larger and independent cohorts.

Evaluation of tissue-specific transcriptome references in IBAS

To assess whether tissue specificity of the reference transcriptome influences IBAS discoveries, we conducted additional analyses using disease-relevant GTEx tissues as alternative reference datasets. Specifically, GTEx Brain Cerebellum expression data were used to train IBAS for the Bipolar Disorder (BD) cohort, and GTEx Pancreas expression data were used for Type 1 Diabetes (T1D). While the primary analyses employed Whole Blood expression due to its broad regulatory coverage, the IBAS framework is not restricted to a single tissue and can naturally incorporate alternative transcriptomic references when biologically appropriate. Across both diseases, tissue-matched references produced gene sets that partially overlapped with Whole Blood discoveries but also revealed additional biologically interpretable candidates, indicating that the interaction structures captured during dimensionality reduction depend on the biological context encoded in the training transcriptome.

In the BD analysis, the Brain-derived reference yielded an expanded set of associated genes, several of which have clear functional relevance to neuronal regulatory processes and psychiatric disease biology (S6 Table). These include

genes involved in neuronal chromatin regulation (*EHMT2*), synaptic adhesion and structural connectivity (*CTNNA2*), and axon guidance and neural signaling (*EFNA2*), all of which represent plausible mediators of transcriptional programs affecting neural circuitry. The Brain-based analysis also identified multiple immune-related loci, including members of the TNF signaling pathway and the HLA region, consistent with increasing evidence supporting neuroimmune contributions to psychiatric disorders. Additional candidates associated with mitochondrial respiration and neuronal energy metabolism (*NDUFS7*, *ATP6V1G2*) further highlight the importance of energy-demanding neuronal processes. Together, these findings indicate that when IBAS is trained on brain-derived transcriptomic interaction patterns, it preferentially highlights genes involved in neural regulatory, immune, and metabolic systems that are biologically consistent with the central nervous system context of BD.

To evaluate whether this tissue-dependent pattern extends beyond psychiatric disease, we performed a parallel analysis for T1D using pancreas-derived transcriptomic data. The pancreas-trained model again retained core immune components detectable from Whole Blood, including genes in the TNF pathway and multiple HLA loci, reflecting the established autoimmune basis of T1D (S7 Table). Beyond these shared immune signals, the pancreas-based reference additionally emphasized regulators of cellular stress and inflammatory signaling such as *ATF6B*, a key mediator of the unfolded protein response implicated in endoplasmic reticulum stress, and *ADAM17*, an important modulator of cytokine activation and inflammatory signaling. Additional candidates including *AGER*, an inflammatory receptor associated with metabolic and immune dysfunction, further suggest that pancreas-derived interaction structures capture regulatory programs related to β-cell stress and immune-mediated damage.

Importantly, these observations do not suggest that Whole Blood references are inappropriate; rather, they demonstrate that different tissues encode distinct but complementary interaction-mediated regulatory programs. Because IBAS uses the reference transcriptome primarily to derive pathway-level interaction weights rather than to predict expression directly, the method can flexibly accommodate tissue-specific transcriptomic resources such as those provided by GTEx. This flexibility enables biologically informed analyses tailored to disease context when tissue-matched expression data are available, while still allowing robust discoveries using broadly sampled tissues when they provide greater statistical power.

Computational complexity and runtime evaluation

IBAS introduces additional computational cost due to its dimensionality reduction step, compared to PrediXcan and MAGMA. For a pathway with N samples, g genes, and m variants, PCA scales on the order of $\mathcal{O}(Ng^2 + g^3)$, while t-SNE and UMAP typically scale between $\mathcal{O}(N\log N)$ and $\mathcal{O}(N^2)$, depending on the implementation. The downstream association and aggregation steps scale approximately linearly with the number of variants, i.e., $\mathcal{O}(Nm)$. In our experiments, MAGMA required 52 seconds, PrediXcan 221.7 seconds, and IBAS 667 seconds, with the additional runtime of IBAS primarily driven by the dimensionality reduction step (~400 seconds). Although IBAS incurs higher computational cost, this additional step enables the extraction of pathway-level interaction patterns through low-dimensional representations, which are not captured by standard TWAS or GWAS approaches. All methods were executed on the same computational node (80 CPU cores; 4 × Intel Xeon Gold 6148 @ 2.40GHz; 3022 GB RAM), ensuring a fair comparison.

Discussion

Dimensionality selection in representation learning

The choice of dimensionality in the representation learning step is an important practical consideration in IBAS. In this work, we adopt a fixed number of leading components (e.g., top 10 PCs), which provides a stable summary of pathway-level interaction structure in our experiments. However, this choice is not unique, and alternative criteria such as cumulative variance explained or data-adaptive selection strategies could also be considered. More broadly, the dimensionality reduction step introduces a tuning parameter that may influence the balance between signal capture and noise incorporation. While our results suggest that IBAS is robust to reasonable choices of dimensionality and to parameter

settings in alternative methods such as UMAP and t-SNE, a systematic investigation of optimal representation learning strategies remains an important direction for future work.

Stability and validations

We expect that the development of IBAS is a significant progress towards the goal of developing statistically stable models involving interactions. This assumption is because IBAS does not aim to explore all combinations of interactions and therefore does not run the risk of overfitting at this stage. Indeed, we have revealed that the performance of IBAS is stable with respect to the alternation of input expression data (Fig 3). Additionally, as t-SNE and UMAP can extract nonlinear components incorporating interactions flexibly, its performance should be also stable to alternations of the prior knowledge of membership of pathways.

More rigorous validations are needed to quantitatively support the above assumptions. First, numerical simulations with known genetic architectures may be conducted to assess the theoretical properties of IBAS. This is not a trivial task as there is very limited literature reporting patterns of gene-gene interactions that involve many genes in a pathway. During the development of IBAS, we have carried out simulations following the three non-linear genetic architectures, i.e., epistatic, compensatory, and heterogenous models. These non-linear models are better accepted by the field and have been used in the previous publications in our group [18,20,74]. However, these architectures are designed to mimic interactions of two genes and are therefore not suitable for the purpose of verifying IBAS. Instead, we opted to use perturbations of real data in our simulations since there was a lack of literature and benchmarking surrounding the simulation of realistic pathway-level interactions.

Association v.s. Causality

It is worth noting that, although IBAS used PCA/t-SNE/UMAP represented gene-gene interactions to mediate the association test, we are not claiming the causality of "genetics ⇒ interaction patterns ⇒ phenotype". Instead, there could be all kinds of causality models, for instance, the pleiotropic model in which genetics causes changes in both interaction patterns and phenotype. Here, the key insight is that we used the interaction patterns to methodologically mediate the discovery, despite the real biological causality relationship is unknown. The same issue applies to many existing works leveraging multi-scale omics to identify association between genetics and phenotype [5–10,18,20,75].

Tissue choice influences IBAS discoveries while providing complementary biological insights

A critical consideration in the IBAS framework is the choice of reference tissue used to construct the interaction bridge. Our primary discovery analyses relied on Whole Blood transcriptomes, which provide large sample sizes and broad regulatory coverage across common biological pathways. Nevertheless, many complex diseases are expected to involve interaction networks that are partly tissue-specific, particularly in organs directly related to disease pathology. To evaluate this possibility, we performed comparative analyses using disease-relevant GTEx tissues as alternative reference transcriptomes. For Bipolar Disorder, training IBAS on Brain Cerebellum expression data revealed additional genes with clear roles in neuronal regulation and circuitry, including *EHMT2* and *CTNNA2*, supporting the relevance of neural interaction programs. Similarly, for Type 1 Diabetes, using pancreas-derived transcriptomic data emphasized immune and cellular stress regulators such as members of the *TNF* and *HLA* pathways together with stress-response genes including *ATF6B* and inflammatory mediators such as *ADAM17*, consistent with autoimmune β-cell damage as a central disease mechanism. Importantly, these observations do not imply that Whole Blood is unsuitable; rather, they indicate that different tissues encode complementary interaction-mediated regulatory structures. This flexibility allows researchers to select reference tissues according to disease biology while still leveraging broadly sampled tissues when they offer greater statistical power, thereby enhancing the practical applicability of IBAS across diverse complex traits.

Strengths and Limitations of IBAS

IBAS offers several methodological advantages for association studies involving gene–gene interactions. First, by leveraging dimensionality reduction, IBAS captures structured interaction patterns at the pathway level without explicitly enumerating combinatorial interactions, thereby avoiding the severe statistical and computational burden associated with high-order interaction modeling. Second, the use of pathway-informed representation learning enables IBAS to incorporate biologically meaningful correlation structures into SNP weighting, which can improve power compared to methods that rely solely on gene-level or SNP-level signals. Third, the modular design of IBAS allows flexibility in choosing representation learning techniques (e.g., PCA, UMAP, t-SNE) and weighting strategies, making the framework adaptable to different data types and biological contexts.

Despite these advantages, IBAS also has several limitations. The method depends on the quality and relevance of the reference transcriptomic data, including tissue selection and sample size, which may influence the extracted interaction patterns. In addition, the dimensionality reduction step introduces tuning parameters (e.g., number of components or embedding parameters), and while our results suggest robustness to reasonable choices, suboptimal settings may affect performance. Furthermore, IBAS relies on pathway annotations, and incomplete or inaccurate pathway definitions may limit its ability to capture relevant interaction structures. Finally, while IBAS improves stability by summarizing interaction patterns, it does not explicitly model specific interaction pairs, which may limit interpretability at the individual interaction level.

Supporting information

S1 Fig. Simulated perturbations reveal robustness of IBAS to noise in reference data (bipolar disorder, BD). Gene expression values were perturbed at levels of 0.1, 0.25, and 0.5 to generate 10 replicate datasets per noise level. Association variability was evaluated using the WTCCC BD cohort for both IBAS and PrediXcan. IBAS was implemented with PCA, t-SNE, and UMAP dimensionality reduction, combined with p-value-based (p) and coefficient-based (c) weighting. (a) Variance of observed p-values among genes identified as significant ($p < 0.05$) in at least one replicate. (b) Recurrence of significant genes across replicates.
(DOCX)

S2 Fig. Simulated perturbations reveal robustness of IBAS to noise in reference data (coronary artery disease, CAD).
(DOCX)

S3 Fig. Simulated perturbations reveal robustness of IBAS to noise in reference data (Crohn's disease, CD).
(DOCX)

S4 Fig. Simulated perturbations reveal robustness of IBAS to noise in reference data (hypertension, HT).
(DOCX)

S5 Fig. Simulated perturbations reveal robustness of IBAS to noise in reference data (rheumatoid arthritis, RA).
(DOCX)

S6 Fig. Simulated perturbations reveal robustness of IBAS to noise in reference data (type 2 diabetes, T2D).
(DOCX)

S7 Fig. Gene-level Manhattan plot for bipolar disorder (BD) in the WTCCC dataset using IBAS. Each point represents a gene, plotted by chromosomal position (x-axis) and association significance as $-\log_{10}(p)$ (y-axis). Gene-level statistics were derived using PCA-based interaction components and coefficient-based variant weighting, with expression weights from GTEx Whole Blood. Labeled genes correspond to the top 10 associations.
(DOCX)

S8 Fig. Gene-level Manhattan plot for coronary artery disease (CAD) in the WTCCC dataset using IBAS.
(DOCX)

S9 Fig. Gene-level Manhattan plot for Crohn's disease (CD) in the WTCCC dataset using IBAS.
(DOCX)

S10 Fig. Gene-level Manhattan plot for rheumatoid arthritis (RA) in the WTCCC dataset using IBAS.
(DOCX)

S11 Fig. Gene-level Manhattan plot for type 2 diabetes (T2D) in the WTCCC dataset using IBAS.
(DOCX)

S12 Fig. Gene-level Manhattan plot for hypertension (HT) in the WTCCC dataset using IBAS.
(DOCX)

S13 Fig. Evaluation of genes associated with disease identified by IBAS using UMAP. Genes identified across seven WTCCC diseases (columns) using coefficient-based and p-value-based weighting (rows) were annotated as "Rediscovered" if previously reported in the DisGeNET database, and "Novel" otherwise.
(DOCX)

S14 Fig. Evaluation of genes associated with disease identified by IBAS using t-SNE.
(DOCX)

S1 Table. Sample sizes of WTCCC cohorts.
(DOCX)

S2 Table. Significant genes identified by MAGMA in the WTCCC type 1 diabetes (T1D) cohort for comparison with IBAS-based results.
(CSV)

S3 Table. Top genes ranked by absolute PC1 loadings in the MAPK signaling pathway (hsa04010).
(CSV)

S4 Table. IBAS-identified genes across WTCCC diseases with DisGeNET annotation support.
(CSV)

S5 Table. IBAS-identified genes in the dbGaP cohort.
(CSV)

S6 Table. IBAS-identified genes in bipolar disorder (BD) using brain-derived reference weights.
(CSV)

S7 Table. IBAS-identified genes in type 1 diabetes (T1D) using pancreas-derived reference weights.
(CSV)

S1 Notes. Spectral justification for PCA-based interaction representation in IBAS
(DOCX)

Author contributions

Conceptualization: Dinghao Wang, Pathum Kossinna, Qingrun Zhang.

Data curation: Karen Ardila.

Formal analysis: Dinghao Wang, Pathum Kossinna, Karen Ardila, Senitha Kumarapeli, Ethan M. MacDonald, Jingjing Wu, Qingrun Zhang.

Methodology: Dinghao Wang, Pathum Kossinna, Qingrun Zhang.

Software: Dinghao Wang, Pathum Kossinna.

Supervision: Qingrun Zhang.

Validation: Dinghao Wang.

Visualization: Dinghao Wang, Pathum Kossinna.

Writing – original draft: Dinghao Wang, Pathum Kossinna, Qingrun Zhang.

Writing – review & editing: Dinghao Wang, Pathum Kossinna, Ethan M. MacDonald, Jingjing Wu, Qingrun Zhang.

References

1. Glazier AM, Nadeau JH, Aitman TJ. Finding genes that underlie complex traits. Science. 2002;298(5602):2345–9. https://doi.org/10.1126/science.1076641 PMID: 12493905
2. Klein RJ, Zeiss C, Chew EY, Tsai J-Y, Sackler RS, Haynes C, et al. Complement factor H polymorphism in age-related macular degeneration. Science. 2005;308(5720):385–9. https://doi.org/10.1126/science.1109557 PMID: 15761122
3. Young AI, Benonisdottir S, Przeworski M, Kong A. Deconstructing the sources of genotype-phenotype associations in humans. Science. 2019;365(6460):1396–400. https://doi.org/10.1126/science.aax3710 PMID: 31604265
4. Sheng X, Guan Y, Ma Z, Wu J, Liu H, Qiu C, et al. Mapping the genetic architecture of human traits to cell types in the kidney identifies mechanisms of disease and potential treatments. Nat Genet. 2021;53(9):1322–33. https://doi.org/10.1038/s41588-021-00909-9 PMID: 34385711
5. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, et al. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet. 2015;47(9):1091–8. https://doi.org/10.1038/ng.3367 PMID: 26258848
6. Zeng P, Zhou X, Huang S. Prediction of gene expression with cis-SNPs using mixed models and regularization methods. BMC Genomics. 2017;18(1):368. https://doi.org/10.1186/s12864-017-3759-6 PMID: 28490319
7. Xie R, Wen J, Quitadamo A, Cheng J, Shi X. A deep auto-encoder model for gene expression prediction. BMC Genomics. 2017;18(Suppl 9):845. https://doi.org/10.1186/s12864-017-4226-0 PMID: 29219072
8. Brandes N, Linial N, Linial M. PWAS: Proteome-Wide Association Study. Lecture Notes in Computer Science. Springer International Publishing. 2020. p. 237–9. https://doi.org/10.1007/978-3-030-45257-5_20
9. Okada H, Ebhardt HA, Vonesch SC, Aebersold R, Hafen E. Proteome-wide association studies identify biochemical modules associated with a wing-size phenotype in Drosophila melanogaster. Nat Commun. 2016;7:12649. https://doi.org/10.1038/ncomms12649 PMID: 27582081
10. Xu Z, Wu C, Pan W, Alzheimer's Disease Neuroimaging Initiative. Imaging-wide association study: Integrating imaging endophenotypes in GWAS. Neuroimage. 2017;159:159–69. https://doi.org/10.1016/j.neuroimage.2017.07.036 PMID: 28736311
11. Su K, Yu Q, Shen R, Sun S-Y, Moreno CS, Li X, et al. Pan-cancer analysis of pathway-based gene expression pattern at the individual level reveals biomarkers of clinical prognosis. Cell Rep Methods. 2021;1(4):100050. https://doi.org/10.1016/j.crmeth.2021.100050 PMID: 34671755
12. William B, Gregor M, Leighton AG. Mendel's principles of heredity, by W. Bateson. Cambridge [Eng.], University Press, 1909; 1909. https://www.biodiversitylibrary.org/item/15713
13. Fisher RA. The correlation between relatives on the supposition of Mendelian inheritance. Earth Environ Sci Trans R Soc Edinb. 1919;52:399–433.
14. Fang G, Wang W, Paunic V, Heydari H, Costanzo M, Liu X, et al. Discovering genetic interactions bridging pathways in genome-wide association studies. Nat Commun. 2019;10(1):4274. https://doi.org/10.1038/s41467-019-12131-7 PMID: 31537791
15. Zhang Q, Long Q, Ott J. AprioriGWAS, a new pattern mining strategy for detecting genetic variants associated with disease through interaction effects. PLoS Comput Biol. 2014;10(6):e1003627. https://doi.org/10.1371/journal.pcbi.1003627 PMID: 24901472
16. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BWJH, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet. 2016;48(3):245–52. https://doi.org/10.1038/ng.3506 PMID: 26854917
17. Cao C, Ding B, Li Q, Kwok D, Wu J, Long Q. Power analysis of transcriptome-wide association study: Implications for practical protocol choice. PLoS Genet. 2021;17(2):e1009405. https://doi.org/10.1371/journal.pgen.1009405 PMID: 33635859
18. Cao C, Kossinna P, Kwok D, Li Q, He J, Su L, et al. Disentangling genetic feature selection and aggregation in transcriptome-wide association studies. Genetics. 2022;220(2):iyab216. https://doi.org/10.1093/genetics/iyab216 PMID: 34849857
19. Consortium G, Ardlie KG, Deluca DS, Segrè AV, Sullivan TJ, Young TR. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015;348:648–60.

20. Cao C, Kwok D, Edie S, Li Q, Ding B, Kossinna P, et al. kTWAS: integrating kernel machine with transcriptome-wide association studies improves statistical power and reveals novel genes. Brief Bioinform. 2021;22(4):bbaa270. https://doi.org/10.1093/bib/bbaa270 PMID: 33200776
21. Tang S, Buchman AS, De Jager PL, Bennett DA, Epstein MP, Yang J. Novel Variance-Component TWAS method for studying complex human diseases with applications to Alzheimer's dementia. PLoS Genet. 2021;17(4):e1009482. https://doi.org/10.1371/journal.pgen.1009482 PMID: 33798195
22. He J, Antonyan L, Zhu H, Li Q, Enoma D, Zhang W. A statistical method for image-mediated association studies discovers genes and pathways associated with four brain disorders. bioRxiv. 2023. https://doi.org/10.1101/2023.06.16.545326
23. Hotelling H. Analysis of a complex of statistical variables into principal components. J Educ Psychol. 1933;24:417.
24. der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research. 2008;9.
25. McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. In: 2018. https://arxiv.org/abs/1802.03426
26. He J, Li Q, Cao C, Zhu H, Shang K, Liu A. IMAS: A novel statistical method for image-mediated association studies – application to the UK Biobank images discovers image and genetic variants associated with four brain disorders. 2022.
27. Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, et al. Powerful SNP-set analysis for case-control genome-wide association studies. Am J Hum Genet. 2010;86(6):929–42. https://doi.org/10.1016/j.ajhg.2010.05.002 PMID: 20560208
28. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nat. 2007;447:661–78.
29. van Iersel MP, Kelder T, Pico AR, Hanspers K, Coort S, Conklin BR, et al. Presenting and exploring biological pathways with PathVisio. BMC Bioinformatics. 2008;9:399. https://doi.org/10.1186/1471-2105-9-399 PMID: 18817533
30. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45(D1):D353–61. https://doi.org/10.1093/nar/gkw1092 PMID: 27899662
31. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res. 2023;51(D1):D587–92. https://doi.org/10.1093/nar/gkac963 PMID: 36300620
32. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30. https://doi.org/10.1093/nar/28.1.27 PMID: 10592173
33. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, et al. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. 2005;33(Database issue):D428-32. https://doi.org/10.1093/nar/gki072 PMID: 15608231
34. Kelder T, van Iersel MP, Hanspers K, Kutmon M, Conklin BR, Evelo CT, et al. WikiPathways: building research communities on biological pathways. Nucleic Acids Res. 2012;40(Database issue):D1301-7. https://doi.org/10.1093/nar/gkr1074 PMID: 22096230
35. Tenenbaum D, Maintainer BP. KEGGREST: Client-side REST access to the Kyoto Encyclopedia of Genes and Genomes (KEGG). 2022.
36. Consortium GO. The gene ontology resource: 20 years and still GOing strong. Nucleic Acids Res. 2019;47: D330–D338.
37. Lever J, Krzywinski M, Altman N. Points of Significance: Principal component analysis. Nat Methods. 2017;14:641–2. https://doi.org/10.1038/NMETH.4346
38. Reich D, Price AL, Patterson N. Principal component analysis of genetic data. Nat Genet. 2008;40(5):491–2. https://doi.org/10.1038/ng0508-491 PMID: 18443580
39. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philos Trans A Math Phys Eng Sci. 2016;374(2065):20150202. https://doi.org/10.1098/rsta.2015.0202 PMID: 26953178
40. Kobak D, Berens P. The art of using t-SNE for single-cell transcriptomics. Nat Commun. 2019;10(1):5416. https://doi.org/10.1038/s41467-019-13056-x PMID: 31780648
41. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, et al. Efficient control of population structure in model organism association mapping. Genetics. 2008;178(3):1709–23. https://doi.org/10.1534/genetics.107.080101 PMID: 18385116
42. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S-Y, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010;42(4):348–54. https://doi.org/10.1038/ng.548 PMID: 20208533
43. Goeman JJ, Solari A. Multiple hypothesis testing in genomics. Stat Med. 2014;33(11):1946–78. https://doi.org/10.1002/sim.6082 PMID: 24399688
44. Davies RB. The distribution of a linear combination of $\chi 2$ random variables. J R Stat Soc Ser C Appl Stat. 1980;29:323–33. https://doi.org/10.2307/2346911
45. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B: Statistical Methodology. 1995;57(1):289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x
46. Böhm C, Kailing K, Kröger P, Zimek A. Computing Clusters of Correlation Connected objects. In: Proceedings of the 2004 ACM SIGMOD international conference on Management of data, 2004. 455–66. https://doi.org/10.1145/1007568.1007620
47. Stegle O, Parts L, Durbin R, Winn J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLoS Comput Biol. 2010;6(5):e1000770. https://doi.org/10.1371/journal.pcbi.1000770 PMID: 20463871

48. Delaneau O, Zagury J-F, Robinson MR, Marchini JL, Dermitzakis ET. Accurate, scalable and integrative haplotype estimation. Nat Commun. 2019;10(1):5436. https://doi.org/10.1038/s41467-019-13225-y PMID: 31780650
49. Rubinacci S, Delaneau O, Marchini J. Genotype imputation using the Positional Burrows Wheeler Transform. PLoS Genet. 2020;16(11):e1009049. https://doi.org/10.1371/journal.pgen.1009049 PMID: 33196638
50. Consortium 1000 Genomes Project, others. A global reference for human genetic variation. Nature. 2015;526:68.
51. Long Q, Zhang Q, Vilhjalmsson BJ, Forai P, Seren Ü, Nordborg M. JAWAMix5: an out-of-core HDF5-based java implementation of whole-genome association studies using mixed models. Bioinformatics. 2013;29(9):1220–2. https://doi.org/10.1093/bioinformatics/btt122 PMID: 23479353
52. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet. 2012;91(2):224–37. https://doi.org/10.1016/j.ajhg.2012.06.007 PMID: 22863193
53. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol. 2015;11(4):e1004219. https://doi.org/10.1371/journal.pcbi.1004219 PMID: 25885710
54. Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. 2017;45(D1):D833–9. https://doi.org/10.1093/nar/gkw943 PMID: 27924018
55. Serrano NC, Millan P, Páez M-C. Non-HLA associations with autoimmune diseases. Autoimmun Rev. 2006;5(3):209–14. https://doi.org/10.1016/j.autrev.2005.06.009 PMID: 16483921
56. Huang Y-C, Lin Y-J, Chang J-S, Chen S-Y, Wan L, Sheu JJ-C, et al. Single nucleotide polymorphism rs2229634 in the ITPR3 gene is associated with the risk of developing coronary artery aneurysm in children with Kawasaki disease. Int J Immunogenet. 2010;37(6):439–43. https://doi.org/10.1111/j.1744-313X.2010.00943.x PMID: 20618519
57. GPI Gene - GeneCards | G6PI Protein | G6PI Antibody. https://www.genecards.org/cgi-bin/carddisp.pl?gene=GPI 2023 July 26.
58. Yang J-S, Hsu J-W, Park S-Y, Lee SY, Li J, Bai M, et al. ALDH7A1 inhibits the intracellular transport pathways during hypoxia and starvation to promote cellular energy homeostasis. Nat Commun. 2019;10(1):4068. https://doi.org/10.1038/s41467-019-11932-0 PMID: 31492851
59. Dupuy F, Tabariès S, Andrzejewski S, Dong Z, Blagih J, Annis MG, et al. PDK1-Dependent Metabolic Reprogramming Dictates Metastatic Potential in Breast Cancer. Cell Metab. 2015;22(4):577–89. https://doi.org/10.1016/j.cmet.2015.08.007 PMID: 26365179
60. Zempleni J, Liu D, Camara DT, Cordonier EL. Novel roles of holocarboxylase synthetase in gene regulation and intermediary metabolism. Nutr Rev. 2014;72(6):369–76. https://doi.org/10.1111/nure.12103 PMID: 24684412
61. Zuccoli GS, Saia-Cereda VM, Nascimento JM, Martins-de-Souza D. The Energy Metabolism Dysfunction in Psychiatric Disorders Postmortem Brains: Focus on Proteomic Evidence. Front Neurosci. 2017;11:493. https://doi.org/10.3389/fnins.2017.00493 PMID: 28936160
62. Rosso G, Cattaneo A, Zanardini R, Gennarelli M, Maina G, Bocchio-Chiavetto L. Glucose metabolism alterations in patients with bipolar disorder. J Affect Disord. 2015;184:293–8. https://doi.org/10.1016/j.jad.2015.06.006 PMID: 26120808
63. Zempleni J, Kuroishi T. Biotin. Advances in Nutrition. 2012;3:213–4. https://doi.org/10.3945/an.111.001305
64. Kanehisa Laboratories. KEGG PATHWAY: hsa00780. https://www.genome.jp/dbget-bin/www_bget?pathway:hsa00780. Accessed 2023 July 26.
65. Kennedy DO. B vitamins and the brain: mechanisms, dose and efficacy—A review. Nutrients. 2016;8:68. https://doi.org/10.3390/NU8020068
66. Karlsson M, Zhang C, Méar L, Zhong W, Digre A, Katona B, et al. A single-cell type transcriptomics map of human tissues. Sci Adv. 2021;7(31):eabh2169. https://doi.org/10.1126/sciadv.abh2169 PMID: 34321199
67. The Human Protein Atlas. Single cell type - HLCS. https://www.proteinatlas.org/ENSG00000159267-HLCS/single%20cell%20type2023 July 28.
68. Lee Y, Zhang Y, Kim S, Han K. Excitatory and inhibitory synaptic dysfunction in mania: an emerging hypothesis from animal model studies. Exp Mol Med. 2018;50(4):1–11. https://doi.org/10.1038/s12276-018-0028-y PMID: 29628501
69. Teschler S, Bartkuhn M, Künzel N, Schmidt C, Kiehl S, Dammann G, et al. Aberrant methylation of gene associated CpG sites occurs in borderline personality disorder. PLoS One. 2013;8(12):e84180. https://doi.org/10.1371/journal.pone.0084180 PMID: 24367640
70. Kayvanpour E, Wisdom M, Lackner MK, Sedaghat-Hamedani F, Boeckel J-N, Müller M, et al. VARS2 Depletion Leads to Activation of the Integrated Stress Response and Disruptions in Mitochondrial Fatty Acid Oxidation. Int J Mol Sci. 2022;23(13):7327. https://doi.org/10.3390/ijms23137327 PMID: 35806332
71. Barth JL, Clark CD, Fresco VM, Knoll EP, Lee B, Argraves WS, et al. Jarid2 is among a set of genes differentially regulated by Nkx2.5 during outflow tract morphogenesis. Dev Dyn. 2010;239(7):2024–33. https://doi.org/10.1002/dvdy.22341 PMID: 20549724
72. Murata Y, Ueno T, Tanaka S, Kobayashi H, Okamura M, Hemmi S, et al. Identification of Clock Genes Related to Hypertension in Kidney From Spontaneously Hypertensive Rats. Am J Hypertens. 2020;33(12):1136–45. https://doi.org/10.1093/ajh/hpaa123 PMID: 33463674
73. Xia Q, Chesi A, Manduchi E, Johnston BT, Lu S, Leonard ME, et al. The type 2 diabetes presumed causal variant within TCF7L2 resides in an element that controls the expression of ACSL5. Diabetologia. 2016;59(11):2360–8. https://doi.org/10.1007/s00125-016-4077-2 PMID: 27539148
74. Li Q, Cao C, Perera D, He J, Chen X, Azeem F, et al. Statistical model integrating interactions into genotype-phenotype association mapping: An application to reveal 3D-genetic basis underlying autism. bioRxiv. 2020;:2020.07.27.222364. https://doi.org/10.1101/2020.07.27.222364
75. Rui Xie, Quitadamo A, Cheng J, Xinghua Shi. A predictive model of gene expression using a deep learning framework. In: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2016. 676–81. https://doi.org/10.1109/bibm.2016.7822599