Autoencoder-transformed transcriptome improves genotype-phenotype association studies

Jiayi Bian[1¶], Qing Li[2¶], Albert Leung[2], Guotao Yang[2], Jun Yan[3,4], Jingjing Wu[1*], Quan Long[1,2,4,5,6*]

[1] Department of Mathematics and Statistics, University of Calgary, Calgary, Alberta, Canada
[2] Department of Biochemistry & Molecular Biology, University of Calgary, Calgary, Alberta, Canada
[3] Department of Physiology and Pharmacology, University of Calgary, Calgary, Alberta, Canada
[4] Mathison Centre for Mental Health Research & Education, Hotchkiss Brain Institute, University of Calgary, Calgary, Alberta, Canada
[5] Department of Medical Genetics, University of Calgary, Calgary, Alberta, Canada
[6] Alberta Children's Hospital Research Institute, University of Calgary, Calgary, Alberta, Canada

[*] Corresponding author
Email: quan.long@ucalgary.ca (QL) and jinwu@ucalgary.ca (JW)

[¶] Joint first authors. These authors contributed equally to this work.

Abstract

Transcriptome-wide association study (TWAS) is an emerging model leveraging gene expressions to direct genotype-phenotype association mapping. A key component in TWAS is the prediction of gene expressions; and many statistical approaches have been developed along this line. However, a problem is that many genes have low expression heritability, limiting the performance of any predictive model. In this work, hypothesizing that appropriate denoising may improve the quality of expression data (including heritability), we propose AE-TWAS, which adds a transformation step before conducting standard TWAS. The transformation is composed of two steps by first splitting the whole transcriptome into co-expression networks (modules) and then using autoencoder (AE) to reconstruct the transcriptome data within each module. This transformation removes noise (including nonlinear ones) from the transcriptome data, paving the path for downstream TWAS. We applied AE-TWAS to the GTEx whole blood transcriptome data and GWAS data of five human diseases, showing two inspiring properties of AE-TWAS: (1) After transformation, the transcriptome data enjoy higher expression heritability at the low-heritability spectrum and possess higher connectivity within the modules. (2) The transferred transcriptome indeed enables better performance of TWAS; and moreover, the newly formed highly connected genes (i.e., hub genes) are more functionally relevant to diseases, evidenced by their functional annotations and overlap with TWAS hits. Taking together, we show that autoencoder transformation produces “better” transcriptome, which in turn enables improved expression-assisted genotype-phenotype association mapping. The impact of this work may be beyond the field of gene mapping: AE can be deemed as a nonlinear extension of principal component analysis (PCA) that is used for removing artifacts in expression data routinely. As such, this work may inspire more expression-based applications to be carried out after an appropriate AE-transformation, unlocking the use of AE-denoised transcriptome in many fields.

Keywords: Autoencoder, De-noising, Heritability, Hub genes, Transcriptome-wide association studies.

Author Summary

We propose to use autoencoder (AE) transformed expression data in transcriptome-wide association studies (TWAS). Currently, TWAS studies have restricted power because of the limiting performance of predictive models. This is largely due to the low expression heritability of many genes in expression data of the reference data set. The quality of expression data (including heritability) is essential, so we aim to develop a method using appropriate denoising. Unlike past studies, which only used linear methods such as principal component analysis (PCA) to remove confounders in expression, we incorporate autoencoder (AE), a nonlinear extension of PCA to remove artifacts in expression data. Our new method, AE-TWAS, is a two-step process. First, we group highly correlated genes in transcriptome into co-expression networks (modules), and then use AE to recreate the transcriptome data within each module. Secondly, the transformed transcriptome is used for downstream TWAS. When we applied AE-TWAS to real diseases, two inspiring discoveries emerged: (1) AE-TWAS boosted the heritability of low-heritability genes in transcriptome data after AE transformation. (2) The transferred transcriptome led to better performance of TWAS and were functionally more relevant to diseases. Our work unlocks the use of AE-denoised transcriptome in more expression-based applications.

Introduction

The transcriptome-wide association study (TWAS) is a popular technique for integrating transcriptome data into genome-wide association studies (GWAS). Unlike GWAS, which associates phenotypic variance directly with genetic variants, TWAS quantitatively aggregates multiple genetic variants into a single test directed by transcriptome data (1,2). The key concept behind TWAS is genetically regulated expression (GReX), which is the component of gene expression attributed to genetic variants. The mainstream format of TWAS is a two-step protocol: First, using a reference dataset containing both genotype and expression data, a linear model is trained to predict GReX as a weighted linear combination of regulatory DNA elements. Techniques including ElasticNet (2), Bayesian sparse linear mixed models (BSLMM) (3) are used to train this genotype-based expression predictor. Then, in the second step, in the GWAS dataset which contains genotype and phenotype (but not expression), the above trained model is used to predict GReX which is subsequently associated with phenotypic traits. TWAS has since achieved popularity and success in identifying the genetic basis of complex traits (4–9), inspiring similar protocols for other endophenotypes such as IWAS (10,11) for images and PWAS (12) for proteins.

Despite of its success in many projects, due to the low heritability of gene expressions, out of around 20,000 genes, usually only 7,000-10,000 genes may be analyzed in a TWAS project (2,5,13). Although many efforts are devoted into the improvement of the predictions, current methods in TWAS (1,2,14–16) do not specifically handle the noise in the input gene expression profiles, including the influence of experimental artifacts, environmental factors, or, importantly, contributions from other interacting genes. Consequently, denoising data may improve expression heritability, which should enhance the genotype-expression model in the reference dataset. Although the current preprocessing methods can remove confounders such as PCs (principal components) in a linear model (17–19), the complex nonlinear confounders within the gene network are not addressed.

Towards this line, Autoencoder (AE) can denoise data and reduce complicated nonlinear confounders. AE is a specific type of neural network to encode input data into a compressed and meaningful representation (i.e., its latent variables), and then decode them back to reconstruct the input (20) (e.g., the configuration in **Fig 1B**). By pursuing minimal discrepancies between the input and reconstructed, an AE effectively filters out biased noise from inputs and retaining essential patterns in the reconstructed, therefore lead to denoised outputs (21). In cases where both the encoder and decoder employ linear operations, a linear AE is achieved, which provides latent variables equivalent to principal component analysis (PCA). Thus, AE can be considered as a nonlinear generalization of PCA, which has been used for processing expression data in a linear way (22).

To test the use of AE in data processing and its impact to the downstream expression-assisted association studies (i.e., TWAS), we implemented a new protocol, AutoEncoder-transformed TWAS (AE-TWAS), which utilizes AE-transformed transcriptome data into a TWAS protocol (**Fig 1**). AE-TWAS first preprocesses transcriptome data by clustering genes into distinct modules

using weighted gene co-expression network analysis (WGCNA) (23,24). Each module contains a set of genes which are linked by substantial correlations (i.e., forming a co-expression network) (**Fig 1A**). Next, AE-TWAS applies a seven-layer conventional AE to the genes in each module (**Fig 1B**). The AE outputs (Z' in **Fig 1B**) are referred to as AE-transformed transcriptome. Finally, AE-TWAS uses the transformed transcriptome to carry out TWAS (**Fig 1C**). More specifically, AE-TWAS employs the widely used Bayesian sparse linear mixed model (BSLMM) (3) to train the genotype-expression model in the reference dataset and then associates the predicted gene expressions to the traits in the target GWAS dataset.

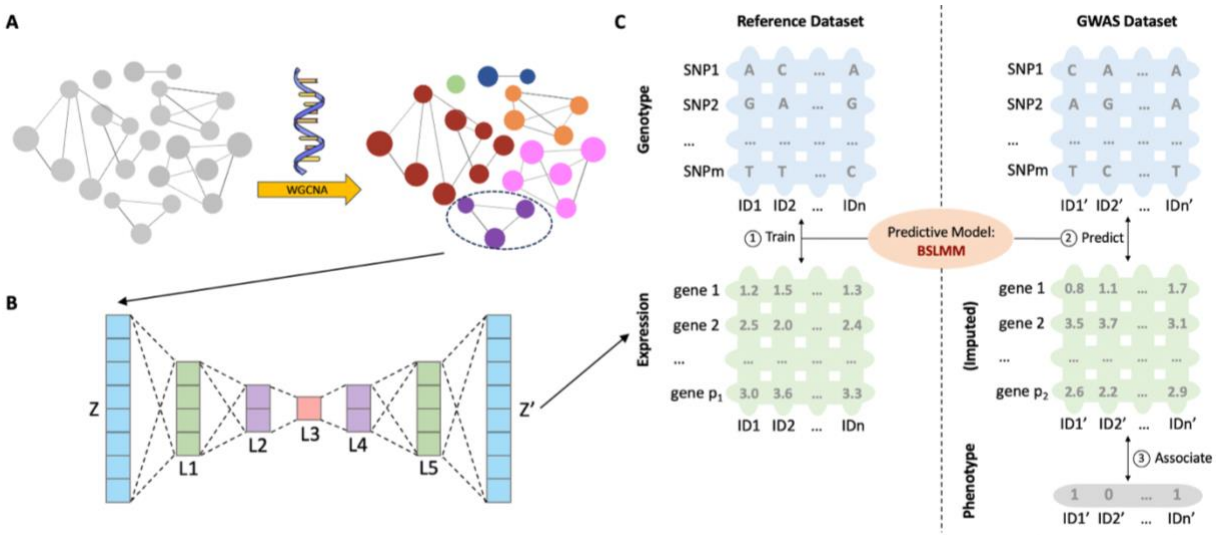

Fig 1: The AE-TWAS protocol. A) AE-TWAS first preprocesses transcriptome data by clustering genes into distinct modules using weighted gene co-expression network analysis (WGCNA). Each module contains a set of genes which are linked by substantial correlations (i.e., forming a co-expression network). B) AE-TWAS then applies a seven-layer conventional AE to the genes in each module. The AE outputs Z' is referred to as AE-transformed transcriptome. C) AE-TWAS uses the transformed transcriptome Z' to carry out TWAS. It employs the BSLMM to train the genotype-expression model in the reference dataset and then associates the predicted gene expressions to the traits in the target GWAS dataset.

We applied AE-TWAS to whole-blood gene expression data on curated genes with 670 subjects in Genotype-Tissue Expression Project version 8 (GTEx v8) (18,19); and re-analyzed gene-trait associations on five diseases in three datasets, namely Wellcome Trust Case Control Consortium (WTCCC) (25), MSSNG (26) and Schizophrenia (27). The applications yielded two unexpected and inspiring discoveries. First, the heritability of most lowly heritable genes ($h^2 < 0.05$) increased after AE transformation, although that for some high-heritability genes decreased (**Fig 2**). Extending single genes to pairs of genes, the correlations between gene-pairs within co-expression networks increased significantly after the AE transformation (**Fig 3**). Second, by comparing AE-TWAS identified genes to the ones identified by a standard TWAS protocol (2), it was observed that the AE-TWAS identified genes had higher proportions of genes that can be validated in DisGeNET, a comprehensive disease-gene database (28–31) in most diseases when compared to the standard TWAS protocol (**Fig 4**). Surprisingly, such enhancement caused by AE transformation was also observed at the level of hub genes in the

co-expression network (**Fig 5**). Taking together, these results suggest that the AE-based denoising should be routinely adapted in expression data processing, just like the use of PCA currently, to facilitate many the downstream analyses including TWAS.

Results

AE-transformation increased expression heritability and correlations between genes

Seven-layer AE recovered transcriptome with high accuracy

The network construction and module detection protocol using WGCNA (23,24) (**Materials & Methods**) on whole-blood gene expression data resulted in 6 modules encompassing 8,191 genes (**S2 Table**). The remaining 4,455 genes not clustered into any modules were excluded from subsequent analysis. We calculated the reconstruction accuracy of the optimal AE model within each module as the average R^2 among all genes calculated from AE-transformed gene expressions and original normalized expressions, leading to high reconstruction accuracy ranging from 0.85 to 0.95 (**S2 Table**). These results indicate that the reconstruction accuracy is high, laying the ground for subsequent analyses.

AE transformation improved expression heritability for lowly heritable genes

We evaluated the change in heritability for the 8,191 genes and visualized them based on their original heritability intervals before AE transformation. Among the seven heritability intervals, the low-spectrum intervals [0, 0.001), [0.001,0.025) and [0.025, 0.05) had a greater number of genes with increased heritability (**Fig 2A; S3 Table**). This suggests that, after removing noises, the previously lowly heritable genes are now with higher heritability, allowing more genes to be TWAS-able. In contrast, the heritability of highly heritable genes reduced (**Fig 2A; S3 Table**). We expected this as might be caused by the reduction of contribution from other genes. Such changes are also evident from the overall distributions of expression heritability before and after the transformation (**Fig 2B; S3 Table**). Nevertheless, the current limitation of TWAS is largely due to that many genes have low expression heritability therefore cannot utilize expressions in the association mapping. The increase of heritability at this spectrum unlocks the potential of broader use of TWAS and other expression-directed association mappings.

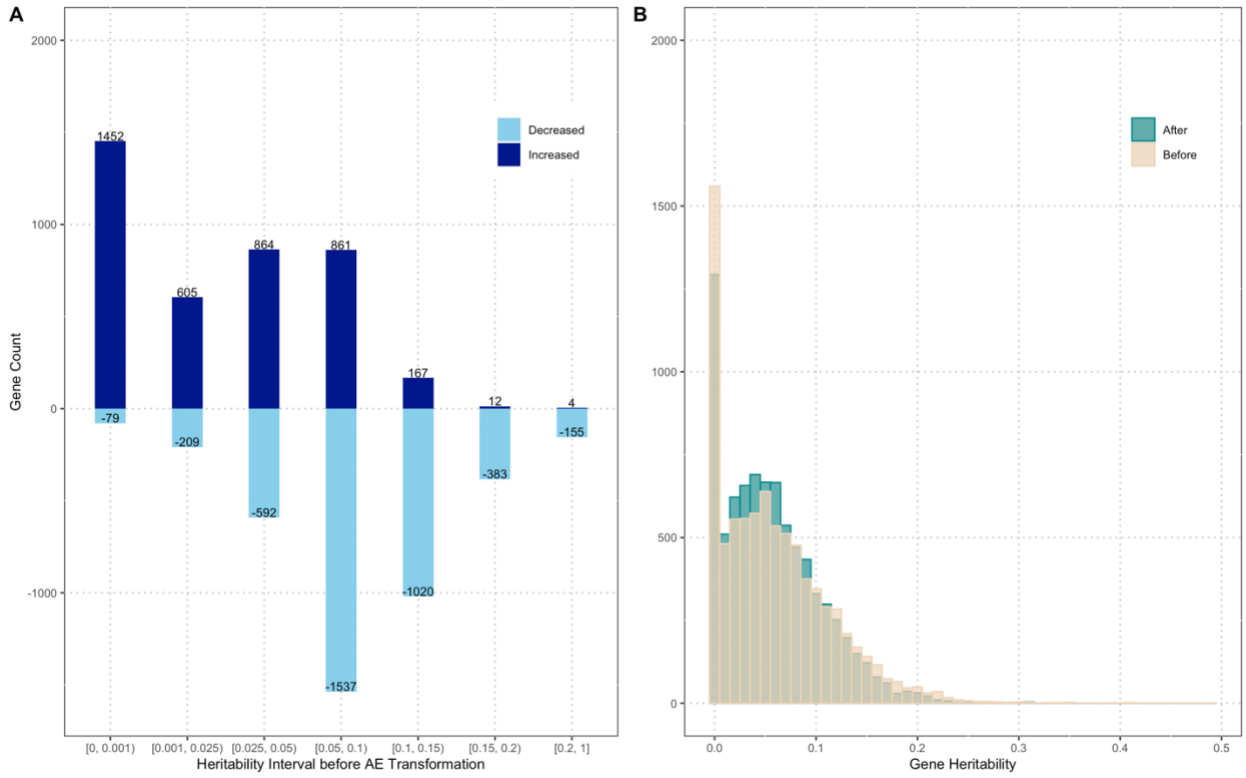

Fig 2: Gene heritability estimation before and after AE transformation. A) Number of genes with increased and decreased heritability in each original heritability interval before AE transformation. B) Distribution of heritability for 8,191 genes before and after AE transformation.

AE transformation increased correlation of gene pairs co-expression networks

Besides the heritability that is for single genes, we calculated the correlation (using absolute value) for all pairs of genes in each module before and after AE transformation (**Materials & Methods**). Evidently, all modules experienced significant increase in absolute correlation among gene pairs following AE transformation (**Fig 3; S4 Table**). It is noteworthy that all gene pairs (100%) within modules 59, 66, 74, and 75 showed a significant improvement in absolute correlation (> 0.1). Additionally, the majority of gene pairs in modules 34 (87.3%) and 72 (97.3%) exhibited an increase, although with a limited improvement in absolute correlation (0.01 for module 34 and 0.03 for module 72).

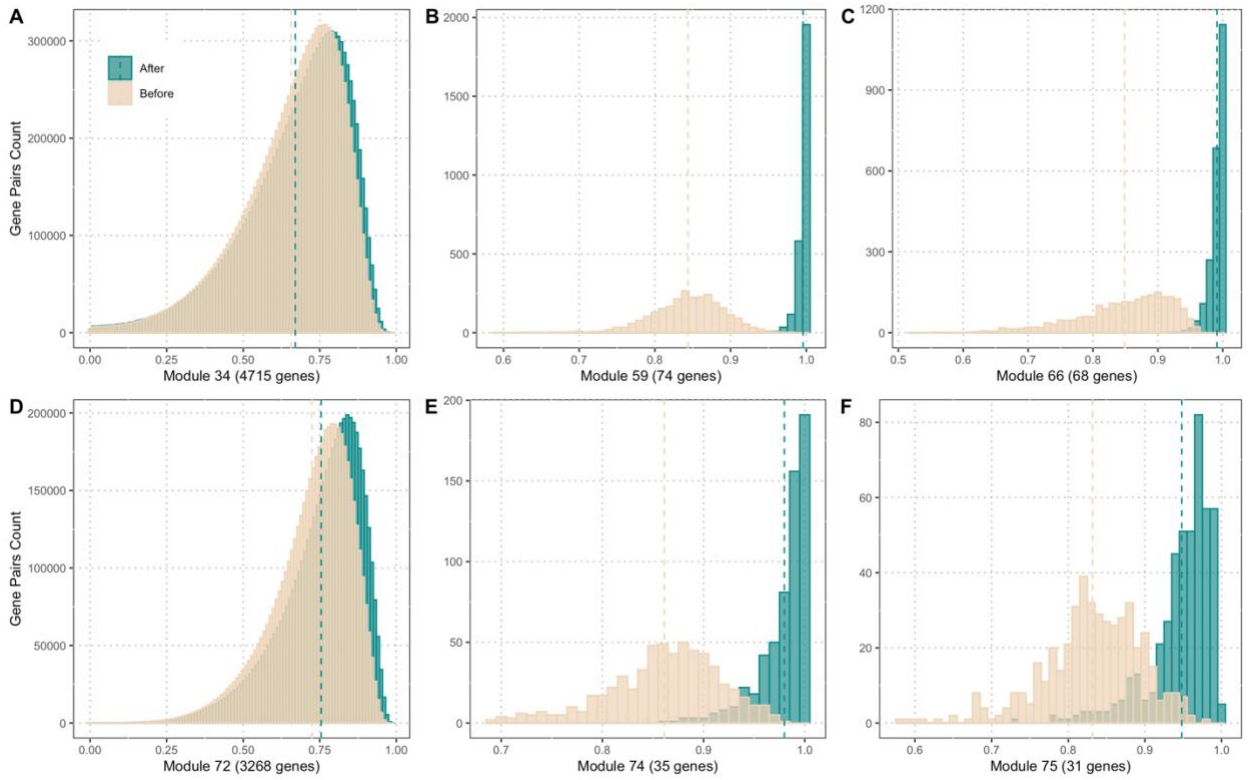

Figure 3: Histogram of absolute correlation among gene pairs within each module before and after AE transformation. Absolute correlation for each gene pair before AE transformation (yellow) and after AE transformation (green) are depicted. Each panel represents histogram for each module.

AE-transformed expressions improved identification of gene-trait associations

To reveal the improvement of AE-transformed expressions, we carried out TWAS using expressions before and after transformation on five diseases including the MSSNG database for autism spectrum disorder (ASD) (26), a Schizophrenia dataset (SCZ) (27), and three WTCCC datasets (25), namely type 1 diabetes (T1D), Crohn's disease (CD) and rheumatoid arthritis (RA) (**Materials & Methods**). The full outcomes are elaborated in (**S5-9 Tables**). We then examined the role of identified genes in each of the five diseases using DisGeNET (28–31), a well-established repository of validated gene-disease associations (**Materials & Methods**).

AE-transformed transcriptomes lead to TWAS outcomes with higher functional relevance

Using DisGeNET as the gold standard, we compared the number of validated associations and the validation rate (number of validated associations divided by the total number of discovered associations) (**Materials & Methods**). After removing genes discovered by both methods, we also calculated the number of unique validated associations reported exclusively by one of the two methods and the unique validation rate. Looking at the numbers only, AE-TWAS model

identified slightly more genes than the standard TWAS for T1D, CD, and RA in the WTCCC dataset, although did not outperform the standard TWAS on ASD (673 vs. 823) and SCZ (324 vs. 332) (**Fig 4A; S10 Table**). The same trend was observed for unique validated associations (**Fig 4B; S10 Table**). However, when considering the validation rate and unique validation rate, AE-TWAS outperformed the standard TWAS model in four of the five diseases with a large margin, with only RA being the exception. (**Fig 4C,D; S10 Table**). Aggregately, the outcomes suggest that AE-TWAS outperformed the standard TWAS method in identifying gene-trait associations.

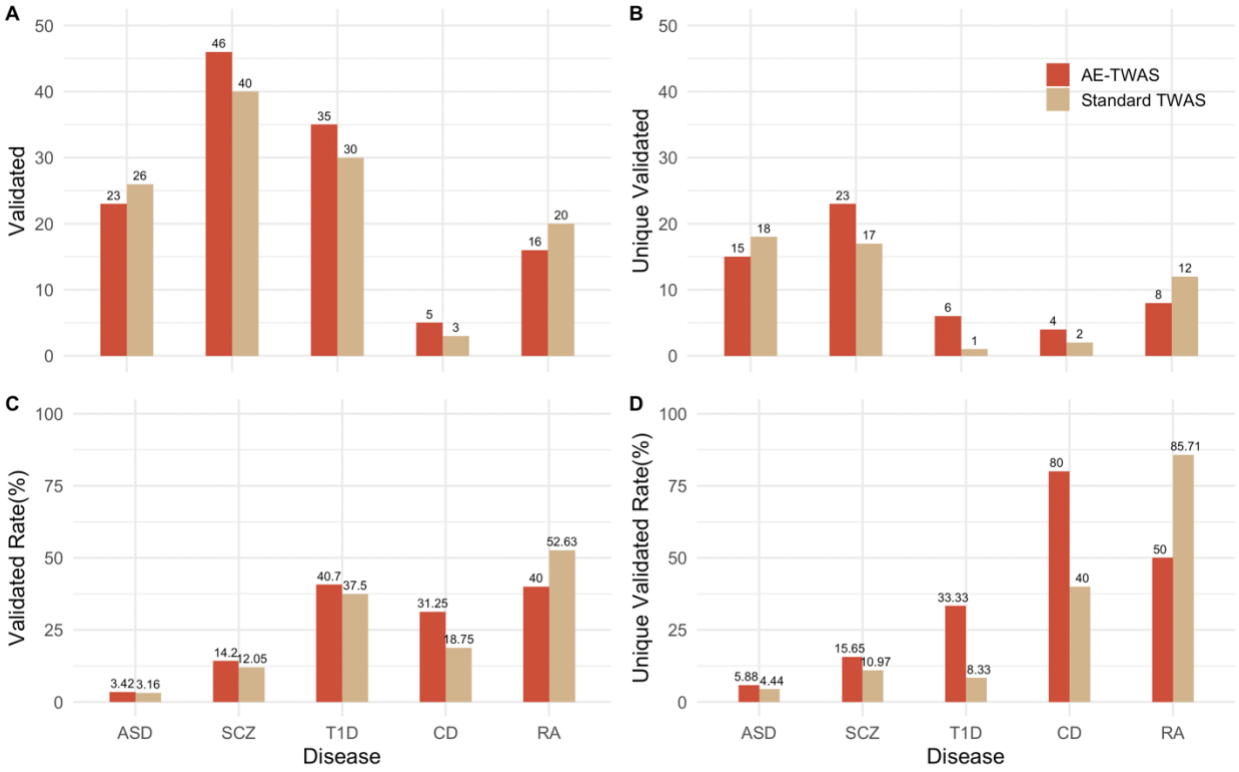

Fig 4: The DisGeNET functional annotations of genes identified by AE-TWAS and standard TWAS in five diseases. A) Number of validated genes (validated) reported in the DisGeNET database. B) Number of validated genes (unique validated) reported exclusively by one of the two protocols in the DisGeNET database. C) Proportion (validated rate) of all reported genes which are validated in the DisGeNET database. D) Proportion (unique validated rate) of genes reported exclusively by one of the two protocols in the DisGeNET database.

AE-transformed genes enjoy higher co-expression network connectivity, which overlap with TWAS discoveries

We first checked the total connectivity of genes and learned that, except for module 34, the rest five modules enjoyed a significant increase of connectivity after AE transformation (**Fig 5A; S11 Table**). The distributions of connectivity of uniquely discovered genes by AE-TWAS and standard TWAS protocols in five diseases also showed interesting properties: for all 5 diseases, the genes uniquely discovered by each protocol enjoyed increased connectivity after transformation (**Fig 5B ; S12 Table**).

Hub genes in AE-transformed networks are more functionally relevant to diseases

To verify the functional relevance of the genes with high connectivity, we selected "hub genes" within each module using the cut-off of the 3rd quantile of total connectivity within each module before AE transformation. This resulted in 2049 and 2350 hub genes, including 554 and 855 unique hub genes before and after AE transformation. For these hub genes, we then examined their functional roles in DisGeNET and in TWAS for the five diseases. Functional verification using DisGeNET showed that all the 5 diseases enjoyed increased number of validated and unique validated associations, as well as the validation rate and unique validation rate after AE transformation (**Fig 5C and 5E; S13 Table)**. We then further checked whether these hub genes are discovered by TWAS before transformation (or AE-TWAS after transformation, respectively) and verified by DisGeNET. Consistent to the above messages, all the 5 diseases enjoyed increased number of unique validated associations and unique validation rate after transformation (using AE-TWAS protocol) (**Fig 5D and 5F; S14 Table**).

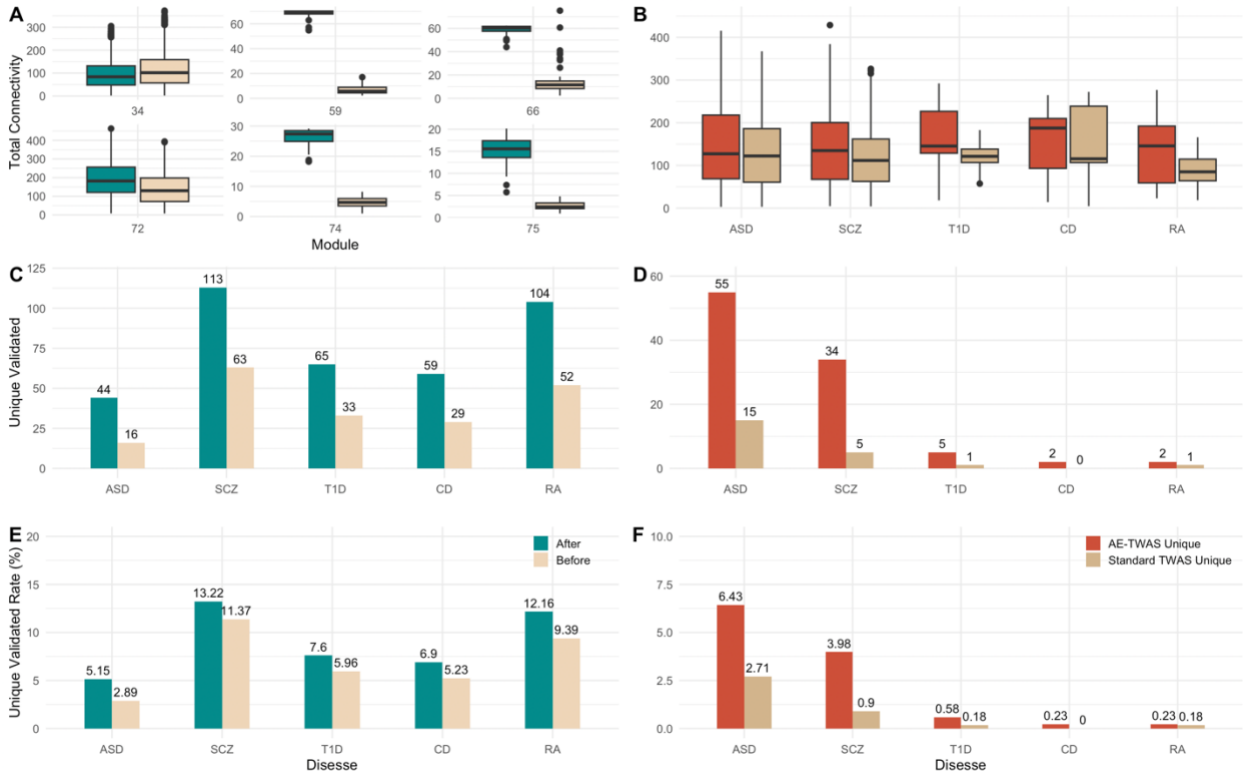

Fig 5: Network connectivity analysis of 8,191 genes clustered in 6 modules by WGCNA. A) Boxplot of total connectivity of 8191 genes within each module before and after AE transformation. B) Boxplot of total connectivity of uniquely discovered genes by AE-TWAS and standard TWAS in five diseases. C) Number of significant hub genes uniquely validated in DisGeNET for the five diseases before and after AE transformation. D) Number of significant hub genes uniquely validated in TWAS for the five diseases by the two protocols. E) Unique validation rate of significant hub genes validated in DisGeNET for the five diseases before and after AE transformation. F) Unique validation rate of significant hub genes validated in TWAS for the five diseases by the two protocols.

Literature support of hub genes uniquely identified by AE-TWAS and validated in DisGeNET

To further investigate biological mechanisms supporting the newly discovered genes, we conducted a literature search on those genes that became hub genes after AE transformation and were both discovered uniquely by AE-TWAS and validated in DisGeNET. The gene names and associated p-values and scores are presented in (**S15 Tables**). We annotated one gene for each of the five diseases that has the highest GDA score in DisGeNet (28–31). We were able to identify multiple lines of evidence in literature for these genes (**S1 Text**). Full investigation of their mechanisms may be left for future experimental efforts; however, the present literature provides preliminary evidence of the effectiveness of AE transformation and AE-TWAS.

DISCUSSION

In summary, we presented the AE-TWAS framework, which integrates gene expressions after AE transformation into TWAS studies. We trained AE models on GTEx v8 using gene networks (modules) generated by WGCNA and conducted TWAS with BSLMM using gene expressions before and after transformation. We evaluated the change in gene heritability and correlation between gene pairs within each module before and after AE transformation. Furthermore, we compared our framework to standard TWAS in identifying genes associated with five different diseases and characterized the functional relevance in relation to the transformation-altered connectivity and hub genes.

The impact of our study is beyond just to elucidate the genetic architecture of complex human diseases. It reveals that by harnessing modern machine learning techniques, one can denoise data to further genuine biological discoveries. Specifically, in this work, we employed autoencoders (AE) to investigate nonlinear interactions among gene expressions and eliminate noise. To our knowledge, our research provides the first evidence that AE-transformed gene expressions improve expression heritability estimation for low-heritability genes and enhance the correlation between genes in clustered co-expression networks (modules). These results suggest that AE models can learn patterns among genes and produce reconstructed outputs with essential information. Therefore, such AE-based data processing may be routinely used in daily expression analysis, similar to the current standard protocol of using PCA (the linear special case of an AE).

We acknowledge that our study has certain limitations. First, our evaluation focused solely on the performance using the BSLMM method for GReX prediction. However, our AE-TWAS framework can also be adapted to other TWAS methods, such as ElasticNet implemented by PrediXcan and kernel methods articulated by ourselves (11,15,16,32). We chose BSLMM because our previous experiences show its supremacy to ElasticNet (16) and popularity in the field. Second, due to the complicated correlation structure within the transcriptome and the black-box and nonlinear nature of AE models, it is challenging to perform a simulation (in which the gold-standard is known) that can quantify the outperformance of AE-TWAS framework. We also haven't provided closed-form mathematical derivations to quantify the improvement theoretically. To ease these limitations, future research could involve applying the AE-TWAS

framework to various gene expression datasets and evaluating its performance across a range of TWAS methods. Third, in our real data analysis of five disease datasets, we demonstrated that our AE-TWAS framework outperformed the standard TWAS in terms of both the validation rate and the unique validation rate for four diseases: T1D, CD, ASD, and SCZ. However, this was not the case for RA, which may be attributable to potentially under-explored genetic basis. Finally, The AE-TWAS framework can be extended to other omics data and TWAS methods, making it a versatile tool in association studies for uncovering the genetic basis of complex human diseases. We again leave this to the future work.

Materials & Methods

Data sources and quality control

Gene expressions

We obtained RNA-seq transcriptome data and genome-wide genotyping data from the GTEx v8 (18,19). We used whole-blood samples with 670 subjects, as this is the most representative general dataset when the disease-causing tissue is unknown. The target tissue initially contains 56,200 genes. After removing low-expression genes with average transcripts per million (TPM) below 1 and genes located outside of autosomes as we focused on human complex diseases unrelated to sex, 12,646 genes remained. Covariates, including genotyping principal components (PCs), were obtained from the GTEx portal. For each gene, we adjusted the gene expression for the top five genotyping PCs, age, sex, sequencing platform, PCR protocol, and 60 confounding factors using a probabilistic estimation of expression residuals (PEER) analysis (17–19), following the GTEx best practices. In our analysis , we used log transformed gene expression data after covariate adjustment ($log_2(1 + gene\ expression)$) to maintain gene expression values in a reasonable scale.

a probabilistic estimation of expression residuals (PEER) analysis

GWAS datasets

We used five diseases datasets (**S1 Table**). The WTCCC dataset includes approximately 2,000 individuals for each of seven diseases and a shared set of around 3,000 controls: type 2 diabetes (T2D), type 1 diabetes (T1D), hypertension (HT), rheumatoid arthritis (RA), coronary artery disease (CAD), Crohn's disease (CD), and bipolar disorder (BD) (25). We focused our analysis on T1D, CD, and RA, as few genes have been identified in association with the other diseases. The MSSNG database provides whole genome sequencing (WGS) and their phenotype data for over 11,500 individuals with autism spectrum disorder (ASD) (26). The Schizophrenia dataset (dbGaP phs000473.v2.p2) provides whole exome sequences (WES) and their phenotype data for over 12,000 individuals from a population-based Swedish case-control cohort with schizophrenia (SCZ) (27). For quality control, we first excluded genetic variants with a minor allele frequency (MAF) lower than 1%, as well as variants that deviated from Hardy-Weinberg equilibrium (HWE). We then mapped all genotype data in our GWAS datasets to the Genome Reference Consortium Human Build 38 (GRCh38) using the UCSC LiftOver tool (33) to ensure

consistent genomic coordinates with genotypes in GTEx v8. This process resulted in a pruned set of 392,937 shared variants with GTEx v8 in the WTCCC dataset, 5,057,872 shared variants in the MSSNG dataset, and 401,515 shared variants in the Schizophrenia dataset carried forward in our analysis.

The AE-TWAS protocol and its applications

Weighted gene co-expression network analysis (WGCNA)

WGCNA is an R package consisting of a comprehensive collection of R functions for performing various aspects of weighted correlation network analysis (23,24). In our study, WGCNA is used to form co-expression networks (modules). It follows a step-by-step protocol for network construction and module detection. WGCNA first defines the weighed network by its adjacency matrix A, a symmetric matrix with entries ranging from 0 to 1, where the component a_{ij} indicates the connection strength between gene i and gene j. This adjacency matrix is calculated by raising the co-expression similarity component s_{ij} to a power: $a_{ij} = s_{ij}^{\beta}, \beta \geq 1$, where s_{ij} is the absolute value of the correlation between gene i and gene j. The power β is selected using the function pickSoftThreshold in WGCNA based on the criterion of approximate scale-free topology.

Supported by WGCNA, AE-TWAS analyzed network topology on a set of candidate powers and chose the optimal power as the lowest at which the scale-free topology fit index reached 0.8. Using this criterion, the optimal power was set as 16. Afterward, AE-TWAS calculated the adjacency matrix A using the optimal power on gene expression data and transformed it into a topological overlap matrix (TOM) (34). Modules were then detected through hierarchical clustering on the TOM using the dynamic tree cut method (35). Finally, AE-TWAS merged modules with highly similar expression profiles based on their eigengenes, clustering them based on their correlation using a 0.75 cut-off. Genes that were not clustered into any module were removed for our subsequent analysis.

Autoencoder transformation

AE-TWAS applied a seven-layer conventional AE to the genes in each module clustered by WGCNA in GTEx whole-blood samples. This conventional AE consists of one input layer, one output layer, and five hidden layers. Within each module, let Z represents the original input and Z' is the reconstructed output. Let q be the number of nodes in the input and output layers, then the number of nodes in each of the five hidden layers are $q/2, q/4, q/8, q/4$, and $q/2$ respectively. Both the encoder and decoder in our seven-layer conventional AE are defined as fully connected multi-layer perceptron (MLP), and sigmoid function were consistently use as activation functions. The optimal AE model is trained to minimize the reconstruction error with the least-square optimization: $L = \frac{1}{q}\sum_{i=1}^{q} ||Z_i - Z_i'||_2^2$ (20).

To compare models' performance, within each module, we divided the whole-blood samples randomly into training and testing sets using a 9:1 ratio, resulting in 603 subjects in the training dataset and 67 subjects in the testing dataset. We then normalized the data in both sets,

scaling gene expression values for each gene to a range of (0,1) across all subjects. These normalized gene expressions values served as inputs of the AE models. We updated the weights using the ADAM optimization algorithm (36). The batch size was set equal to the size of the training set. We set the learning rate at 1×10^{-3} and the decay rate at 0.5 after 2,000 epochs, with a total of 20,000 epochs. The AE models were implemented using the PyTorch package.

TWAS analysis based on Autoencoder-transformed transcriptomes

Assuming that the expression matrix Z is an $n \times p_1$ matrix where n and p_1 represent the number of individuals and genes respectively. The transformed Z' has the same dimensions of Z, and is utilized to train the predictive model in the reference dataset using BSLMM (3):

$$Z' \sim 1_n \mu + Xw + \nu + \epsilon \quad (1)$$
$$w \sim \pi N(0, \sigma_a^2 + \sigma_b^2) + (1 - \pi) N(0, \sigma_b^2) \quad (2)$$

Here X is an $n \times m$ genotype matrix measured on the same n individuals at m cis-genetic variants (within 1 Mb of the gene), w is the corresponding effects of variants which come from a mixture of two normal distributions, ν is the random effect term, and ϵ is the error term.

AE-TWAS then predicted gene expressions as linear combinations of the effect estimates obtained from BSLMM and the genotypes derived from the GWAS dataset, which solely provides genotype and phenotype data. Finally, the imputed gene expressions, $\hat{Z}$, were associated with disease trait Y (an $n' \times 1$ vector where n' represents the number of individuals in GWAS dataset) to identify significant gene-trait associations: $Y \sim \hat{Z} + \varepsilon$.

We applied AE-TWAS on five diseases in three datasets (WTCCC, MSSNG and Schizophrenia datasets) (25) and compared their results in identifying significant gene-trait associations to a standard TWAS protocol (2). Significant genes are defined as genes with p-values less or equal than 0.05 after false discovery rate (FDR) correction (37).

Calculation of expression heritability

We estimated the gene expression heritability, which represents the proportion of expression variance explained by genotype, for genes in whole-blood samples using a variance-component model with a genetic relationship matrix (GRM) derived from genotype data. This estimation was conducted using the Genome-wide Complex Trait Analysis (GCTA) tool (38). For each gene, variants located within 1 Mb of the gene's start or end were used to construct the GRM. GCTA then calculates the proportion of the variance in gene expression attributed to these local variants using the following linear mixed model (39):

$$Z = Xb + G_{local} + e \quad (3)$$
$$var(Z) = A_{local}\sigma_{local}^2 + I\sigma_e^2 \quad (4)$$

In this equation, Z represents gene expression and b is a vector of fixed effects. A_{local} refers to the GRM calculated from local variants, while the random effect G_{local} denotes the genetic

effect attributable to the set of local variants, with $var(G_{local}) = A_{local}\sigma^2_{local}$. Variance components were calculated using restricted maximum likelihood (REML) (40).

Calculation of gene pairs' correlation and gene's connectivity

We investigated the correlation among gene pairs within each module before and after AE transformation. The correlation between gene i and gene j is defined as $cor(x_i, x_j)$ where x_i and x_j represent the expression profiles of the two genes. We considered the absolute value of the correlation. For 8,191 genes clustered in 6 modules, we calculated their total connectivity before and after AE transformation using *intramodularConnectivity* function in the WGCNA package (23,24). For each gene, the total connectivity is calculated as the sum of intramodular and extra-modular connectivity.

Functional verification of discovered genes

To verify the functional relevance of gene-trait associations identified in the three GWAS datasets, we conducted a gene-disease search using the DisGeNET repository (v7.0), a platform containing 1,134,942 gene-disease associations (GDAs), between 21,671 genes and 30,170 traits (28–31). For each disease, DisGeNET provides a specific summary of gene-disease associations, containing information on identified significant genes and their corresponding GDAs. These association summaries can be accessed using the Concept Unique Identifier (CUI) number assigned to each disease (**S1 Table**).

Acknowledgments

New Frontiers in Research Fund, Exploration NFRFE-2018-00748 (QL)
Alberta Innovates LevMax-Health Program Bridge Funds 222300769 (QL)
Canada Foundation for Innovation 36605 (QL)
NSERC Discovery Grant (RGPIN-2018-04328 J.W.; RGPIN-2017-04860: QL)
Alberta Innovates Graduate Student Scholarships (JB)

Author contributions (QL= Quan Long; QL2 = Qing Li)

Conceptualization: JB, QL2, JW, QL.
Funding, Resources & Supervision: JY, JW, QL2.
Methodology: JB, QL2, JW, QL.
Formal analysis and investigation: JB, QL2, AL, GY, QL.
Manuscript writing: JB, QL2, and QL, with contributions from all authors.

Competing interests

Authors declare that they have no competing interests.

Data and materials availability

AE-TWAS source code and users' manual: https://github.com/theLongLab/AE-TWAS
GTEx gene expression: https://gtexportal.org/home/datasets
GTEx whole genome sequencing data: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v9.p2
The gene model file: https://www.gtexportal.org/home/datasets/gencode.v26.GRCh38.genes.gtf
WGCNA package in R: https://cran.r-project.org/web/packages/WGCNA/WGCNA.pdf
The PyTorch package used to train AE models: https://pytorch.org/
WTCCC dataset: https://www.wtccc.org.uk/
MSSNG database: https://research.mss.ng/
Schizophrenia dataset: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000473.v2.p2
BSLMM model used to train the genotype-expression model: https://github.com/genetics-statistics/GEMMA
Elastic net model used to perform gene-trait association: https://github.com/hakyimlab/MetaXcan
GCTA used to calculate gene expression heritability: https://yanglab.westlake.edu.cn/software/gcta/#Overview
DisGeNET repository used to perform functional gene-disease analysis: https://www.disgenet.org/

References

1. 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 [Internet]. 2016 Mar 1 [cited 2023 May 23];48(3):245–52. Available from: https://pubmed.ncbi.nlm.nih.gov/26854917/
2. Gamazon ER, Wheeler HE, Shah KP, Mozaffari S V., Aquino-Michaels K, Carroll RJ, et al. A gene-based association method for mapping traits using reference transcriptome data. Nature Genetics 2015 47:9 [Internet]. 2015 Aug 10 [cited 2023 May 23];47(9):1091–8. Available from: https://www.nature.com/articles/ng.3367
3. Zhou X, Carbonetto P, Stephens M. Polygenic Modeling with Bayesian Sparse Linear Mixed Models. PLoS Genet [Internet]. 2013 Feb [cited 2023 May 24];9(2):e1003264. Available from: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003264
4. Thériault S, Gaudreault N, Lamontagne M, Rosa M, Boulanger MC, Messika-Zeitoun D, et al. A transcriptome-wide association study identifies PALMD as a susceptibility gene for calcific aortic valve stenosis. Nature Communications 2018 9:1 [Internet]. 2018 Mar 7 [cited 2023 May 24];9(1):1–8. Available from: https://www.nature.com/articles/s41467-018-03260-6
5. Mancuso N, Gayther S, Gusev A, Zheng W, Penney KL, Kote-Jarai Z, et al. Large-scale transcriptome-wide association study identifies new prostate cancer risk regions. Nature Communications 2018 9:1 [Internet]. 2018 Oct 4 [cited 2023 May 24];9(1):1–11. Available from: https://www.nature.com/articles/s41467-018-06302-1
6. Gusev A, Mancuso N, Won H, Kousi M, Finucane HK, Reshef Y, et al. Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. Nature Genetics 2018 50:4 [Internet]. 2018 Apr 9 [cited 2023 May 24];50(4):538–48. Available from: https://www.nature.com/articles/s41588-018-0092-1
7. Wu L, Shi W, Long J, Guo X, Michailidou K, Beesley J, et al. A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. Nature Genetics 2018 50:7 [Internet]. 2018 Jun 18 [cited 2023 May 24];50(7):968–78. Available from: https://www.nature.com/articles/s41588-018-0132-x
8. Ratnapriya R, Sosina OA, Starostik MR, Kwicklis M, Kapphahn RJ, Fritsche LG, et al. Retinal transcriptome and eQTL analyses identify genes associated with age-related macular degeneration. Nature Genetics 2019 51:4 [Internet]. 2019 Feb 11 [cited 2023 May 24];51(4):606–10. Available from: https://www.nature.com/articles/s41588-019-0351-9
9. Chen H, Wang T, Huang S, Zeng P. New novel non-MHC genes were identified for cervical cancer with an integrative analysis approach of transcriptome-wide association study. J Cancer [Internet]. 2021 Jan 1 [cited 2023 May 24];12(3):840–8. Available from: http://www.jcancer.org://creativecommons.org/licenses/by/4.0/

10. Xu Z, Wu C, Pan W. Imaging-wide association study: Integrating imaging endophenotypes in GWAS. Neuroimage. 2017 Oct 1;159:159–69.
11. He J, Antonyan L, Zhu H, Li Q, Enoma D, Zhang W, et al. A statistical method for image-mediated association studies 1 discovers genes and pathways associated with four brain disorders 2 3. Available from: https://doi.org/10.1101/2023.06.16.545326
12. Brandes N, Linial N, Linial M. PWAS: Proteome-wide association study - Linking genes and phenotypes by functional variation in proteins. Genome Biol [Internet]. 2020 Jul 14 [cited 2023 May 24];21(1):1–22. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02089-x
13. Li B, Verma SS, Veturi YC, Verma A, Bradford Y, Haas DW, et al. Evaluation of PrediXcan for prioritizing GWAS associations and predicting gene expression. Pac Symp Biocomput [Internet]. 2018 [cited 2023 May 24];23(212669):448. Available from: /pmc/articles/PMC5749400/
14. Rodriguez-Fontenla C, Carracedo A. UTMOST, a single and cross-tissue TWAS (Transcriptome Wide Association Study), reveals new ASD (Autism Spectrum Disorder) associated genes. Translational Psychiatry 2021 11:1 [Internet]. 2021 Apr 30 [cited 2023 May 24];11(1):1–11. Available from: https://www.nature.com/articles/s41398-021-01378-8
15. 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. Vol. 22, Briefings in Bioinformatics. Oxford University Press; 2021.
16. 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 Feb 1;220(2).
17. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nature Protocols 2012 7:3 [Internet]. 2012 Feb 16 [cited 2023 May 24];7(3):500–7. Available from: https://www.nature.com/articles/nprot.2011.457
18. Ardlie KG, DeLuca DS, Segrè A V., Sullivan TJ, Young TR, Gelfand ET, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science (1979) [Internet]. 2015 May 8 [cited 2023 May 24];348(6235):648–60. Available from: https://www.science.org/doi/10.1126/science.1262110
19. Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, et al. Genetic effects on gene expression across human tissues. Nature 2017 550:7675 [Internet]. 2017 Oct 12 [cited 2023 May 25];550(7675):204–13. Available from: https://www.nature.com/articles/nature24277
20. Bank D, Koenigstein N, Giryes R. Autoencoders. 2020 Mar 12 [cited 2023 May 25]; Available from: http://arxiv.org/abs/2003.05991
21. Wang J, Xie X, Shi J, He W, Chen Q, Chen L, et al. Denoising Autoencoder, A Deep Learning Algorithm, Aids the Identification of A Novel Molecular Signature of Lung Adenocarcinoma. Genomics Proteomics Bioinformatics. 2020 Aug 1;18(4):468–80.

22. Casella M, Dolce P, Ponticorvo M, Marocco D. From Principal Component Analysis to Autoencoders: a comparison on simulated data from psychometric models. In: 2022 IEEE International Workshop on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering, MetroXRAINE 2022 - Proceedings. Institute of Electrical and Electronics Engineers Inc.; 2022. p. 377–81.
23. Langfelder P, Horvath S. WGCNA: An R package for weighted correlation network analysis. BMC Bioinformatics [Internet]. 2008 Dec 29 [cited 2023 May 25];9(1):1–13. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-559
24. Langfelder P, Horvath S. Fast R Functions for Robust Correlations and Hierarchical Clustering.
25. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007 447:7145 [Internet]. 2007 Jun 7 [cited 2023 May 25];447(7145):661–78. Available from: https://www.nature.com/articles/nature05911
26. Yuen RKC, Merico D, Bookman M, Howe JL, Thiruvahindrapuram B, Patel R V., et al. Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. Nat Neurosci. 2017 Mar 29;20(4):602–11.
27. Kiezun A, Garimella K, Do R, Stitziel NO, Neale BM, McLaren PJ, et al. Exome sequencing and the genetic basis of complex traits. Nature Genetics 2012 44:6 [Internet]. 2012 May 29 [cited 2023 May 25];44(6):623–30. Available from: https://www.nature.com/articles/ng.2303
28. Piñero J, Queralt-Rosinach N, Bravo À, Deu-Pons J, Bauer-Mehren A, Baron M, et al. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database (Oxford) [Internet]. 2015 [cited 2023 May 25];2015. Available from: https://pubmed.ncbi.nlm.nih.gov/25877637/
29. Piñero J, Bravo Á, 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 [Internet]. 2017 Jan 4 [cited 2023 May 25];45(D1):D833–9. Available from: https://academic.oup.com/nar/article/45/D1/D833/2290909
30. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res [Internet]. 2020 Jan 8 [cited 2023 May 25];48(D1):D845–55. Available from: https://academic.oup.com/nar/article/48/D1/D845/5611674
31. Piñero J, Saüch J, Sanz F, Furlong LI. The DisGeNET cytoscape app: Exploring and visualizing disease genomics data. Comput Struct Biotechnol J. 2021 Jan 1;19:2960–7.
32. He J, Li Q, Zhang Q. rvTWAS: identifying gene-trait association using sequences by utilizing transcriptome-1 directed feature selection 2 3. Available from: https://doi.org/10.1101/2023.07.16.549227
33. Karolchik D, Hinrichs AS, Kent WJ. The UCSC Genome Browser. Curr Protoc Bioinformatics [Internet]. 2009 Dec 1 [cited 2023 May 25];28(1):1.4.1-1.4.26.

Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/0471250953.bi0104s28

34. Yip AM, Horvath S. Gene network interconnectedness and the generalized topological overlap measure. BMC Bioinformatics. 2007;8.
35. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: The Dynamic Tree Cut package for R. Bioinformatics. 2008;24(5):719–20.
36. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. 2014 Dec 22; Available from: http://arxiv.org/abs/1412.6980
37. Sesia M, Bates S, Candès E, Candès C, Marchini J, Sabatti C. False discovery rate control in genome-wide association studies with population structure. 2023;29.
38. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A tool for genome-wide complex trait analysis. Am J Hum Genet. 2011 Jan 7;88(1):76–82.
39. Torres JM, Gamazon ER, Parra EJ, Below JE, Valladares-Salgado A, Wacher N, et al. Cross-tissue and tissue-specific eQTLs: Partitioning the heritability of a complex trait. Am J Hum Genet. 2014;95(5):521–34.
40. Gilmour AR, Thompson R, Cullis BR. Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models [Internet]. Vol. 51. 1995. Available from: https://about.jstor.org/terms

Supporting information

S1 Table: Description of the five diseases in three datasets and their CUI in the DisGeNET database.

S2 Table: WGCNA generated modules and the AE-reconstruction accuracy of the optimal AE model within each module (average R^2 among all genes) in whole-blood samples.

S3 Table: Gene heritability estimation before and after AE transformation.

S4 Table: Average absolute correlation among gene pairs within each module before and after AE transformation.

S5 Table: All significant genes identified by one of the two protocols in ASD.

S6 Table: All significant genes identified by one of the two protocols in SCZ.

S7 Table: All significant genes identified by one of the two protocols in T1D.

S8 Table: All significant genes identified by one of the two protocols in CD.

S9 Table: All significant genes identified by one of the two protocols in RA.

S10 Table: Number of significant genes validated in DisGeNET for the five diseases identified by the two protocols.

S11 Table: Summary statistics of total connectivity of 8191 genes within each module before and after AE transformation.

S12 Table: Summary statistics of total connectivity of discovered and uniquely discovered genes by AE-TWAS and standard TWAS in five diseases.

S13 Table: Number of significant hub genes validated in DisGeNET for the five diseases before and after AE transformation.

S14 Table: Number of significant hub genes discovered in TWAS and validated by DisGeNET for the five diseases by the two protocols.

S15 Table: Description of the selected hub genes in each of the five diseases in literature search.

S1 Text: Literature support of hub genes uniquely identified by AE-TWAS and validated in DisGeNET.