Epigenetic Element-Based Transcriptome-Wide Association Study Identifies Novel Genes for Bipolar Disorder

Shi Yao[1], Jing-Miao Ding[1], Hao Wu[1], Ruo-Han Hao[1], Yu Rong[1], Xin Ke[1], Jing Guo[1], Shan-Shan Dong[1], Yan Guo[1]

[1]Key Laboratory of Biomedical Information Engineering of Ministry of Education, Biomedical Informatics & Genomics Center, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an, Shaanxi, P. R. China, 710049

[*]Corresponding authors: Yan Guo, Ph.D.

PHONE: 86-29-62818386

E-MAIL: guoyan253@xjtu.edu.cn

Running title: ETWAS for Bipolar Disorder

Abstract

Bipolar disorder (BD) is a highly heritable neuropsychiatric disorder characterized by recurrent episodes of depression and mania. Since the signals identified by genome-wide association study (GWAS) often reside in the non-coding regions, understanding the biological relevance of these genetic loci has proven to be complicated. Transcriptome-wide association studies (TWAS) providing a powerful approach to identify novel disease risk genes and uncover possible causal genes at loci identified previously by GWAS. However, these methods did not consider the importance of epigenetic regulation in gene expression. Here, we developed a novel epigenetic element-based transcriptome-wide association study (ETWAS) that tests the effects of genetic variants on gene expression levels with the epigenetic features as prior and further mediates the association between predicted expression and BD. We conducted an ETWAS consisting of 20,352 cases and 31,358 controls and identified 44 transcriptome-wide significant hits. We found 14 conditionally independent genes, and 11 did not previously implicate with BD, which regarded as novel candidate genes, such as *ASB16* in the cerebellar hemisphere ($P = 9.29\times10^{-8}$). We demonstrated that several genome-wide significant signals from the BD GWAS driven by genetically regulated expression, and conditioning of *NEK4* explaining 90.1% of the GWAS signal. Additionally, ETWAS identified genes could explain heritability beyond that explained by GWAS-associated SNPs ($P = 0.019$). In conclusion, ETWAS is a powerful method, and we identified several novel candidate genes associated with BD.

Introduction

Bipolar disorder (BD) is a severe neuropsychiatric disorder characterized by recurrent episodes of depression and mania that affect thought, perception, emotion, and social behavior. BD has a lifetime prevalence of 1-2%, and the World Health Organization ranks BD among the most significant contributors to the global burden of disease [1]. Based on twin studies, the narrow-sense heritability of BD has estimated to be over 70% [2, 3], suggesting a substantial involvement of genetic factors is driving the phenotypic variance. Genome-wide association study (GWAS) has seen great strides and invaluable utilities in revealing initial insights into BD's genetic architecture. Recently, a large-scale GWAS identified 30 loci that were significantly associated with BD [4]. Despite the significant success of GWAS in delineating elements that contribute to the genetic architecture of psychiatric disorders, only a small fraction of this heritability is explained by associated loci [4], leaving a substantial proportion of genetic risk factors uncharacterized.

Most of the identified variants mapped through GWAS reside in non-coding regions of the genome [5], that may be involved in modulating gene regulatory programs [5-9]. Recent mechanistic studies have demonstrated that GWAS-identified variants located in the active chromatin regions more frequently and highly-enriched with expression quantitative trait loci (eQTL) [10, 11]. Moreover, most common risk variants identified to date are only associated with diseases with modest effect sizes, and many risk variants have not identified via a typical GWAS, even with a large

sample size [12]. Transcriptome-wide association study (TWAS) that systematically investigates the association of genetically predicted gene expression with disease risk, providing a powerful approach to identify novel disease risk genes and uncover possible causal genes at loci identified previously by GWAS [13-17].

Nevertheless, gene expression highly regulated in many steps, including transcriptional regulation, splicing, end modification, export, and degradation. Transcriptional regulation of DNA into mRNA can occur on both genetic and epigenetic levels. The epigenetic regulation alters the accessibility of DNA to transcription factors by chemical modification of chromatin. For example, several post-translational modifications occur on the histones that could change chromatin structure and function [18], making it accessible or vice versa to transcription factors. Functional class quantification in 11 diseases from the Wellcome Trust Case Control Consortium, including BD, showed that 80% of the common variants that contribute to phenotype variability could attribute to DNase I hypersensitivity sites [19]. They are likely to regulate chromatin accessibility and transcription, further highlighting the importance of transcript regulation in the epigenetic level. Histone modifications are involved in both activation and repression of transcription [20] and further linked to diseases [21]. Researchers have developed a growing body of computational methods to predict gene expression from histone modification signals of chromatin structure [22-25]. Thus, integrating epigenetic features is essential for the prediction of gene expression besides genetic variants.

In this study, we set out to develop a four-step quantitative pipeline named epigenetic element-based transcriptome-wide association studies (ETWAS), based on the interpretation of epigenetic element, genotype, gene expression, and phenotype. We used ETWAS to investigate the association between gene expression and BD risk using the largest BD cohort currently available (as of 2020); the cohort consists of 20,352 BD cases and 31,358 controls from Europe. We found that ETWAS outperformed than original methods, and we identified 14 conditionally independent genes associated with BD risk in 13 brain tissues. We additionally identified 11 genes that not previously implicated with BD.

Materials and Methods

Data Resources

GTEx RNA sequencing data sets

We used transcriptome and high-density genotyping data of European decedent from the Genotype-Tissue Expression (GTEx) study Pilot Project V8 (dbGap accession: phs000424.v8.p2) [11] to establish gene expression prediction models. In brief, genomic DNA samples obtained from study participants genotyped using Illumina OMNI 2.5M or 5M SNP Array, and RNA samples from 51 tissue sites were sequenced to generate transcriptome profiling data. In this project, we used genotyping and transcriptome data of whole blood and 13 brain tissues to build epigenetic-based gene expression prediction models. The tissue abbreviations, sample

sizes are listed in Supplementary Table 1.

Geuvadis RNA sequencing data set

We obtained freely available RNA-seq data from 358 European lymphoblastoid cell lines produced by the Genetic European Variation in Health and Disease [26] (Geuvadis) at https://www.ebi.ac.uk/. The genotype data were generated by the 1000 Genomes Project. We used Geuvadis as the validation data set to test the prediction models generated in the GTEx whole blood as the result of cross-study prediction evaluation.

Epigenetic elements

The chromatin states of relevant tissues were downloaded from the Roadmap Epigenome Project [27]. We briefly utilized the 15-state model and grouped them into four categories: *promoter*, *enhancer*, *transcription*, and *others*. Individually, *TssA* and *TssAFlnk* are considered to be a *promoter*; *TxFlnk*, *Tx*, and *TxWk* are considered to be a *transcription*; *EnhG* and *Enh* are considered to be an *enhancer*, and the rest are classified as the *others*. Freely available transcription factor binding sites (TFBS) were obtained from the Encyclopedia of DNA Elements (ENCODE) [28], and the DNase I hypersensitive sites (DHS) were obtained from the Roadmap Epigenome Project. The URLs of all epigenetic data are listed in Supplementary Table 2.

GWAS summary statistics

We used the most recent summary statistics of the Psychiatric Genomics Consortium (PGC) Bipolar Disorder Working Group, comprising 20,352 BD cases and 31,358 controls of European descent (Supplementary Table 3). Details on participant ascertainment and quality control were previously reported by Eli et al. [4]. Cases were required to meet international consensus criteria (DSM-IV or ICD-10) for a lifetime diagnosis of BD established using structured diagnostic instruments. We evaluated ETWAS's performance by identifying significantly associated genes in an earlier reported summary statistic that did not overlap a genome-wide significant SNP, and looking for newly genome-wide significant SNPs in the expanded study. The earlier reported summary data is the first large collaborative BD GWAS by the same group in PGC [29], comprised of 7,481 BD patients and 9,250 controls of European descent (Supplementary Table 3).

ETWAS Framework

Our current study based on the premise that gene expression is heritable [30]. Considering the high heritability genes typically enriched for trait associations [15], we only focused on the highly heritable genes in further analyses. We estimated the *cis* heritability (1 Mb window around each gene) for each gene using restricted maximum likelihood analysis, a variance-component model with a genetic relationship matrix (GRM) estimated from genotype data in GCTA software [31]. A significant nonzero heritability gene defined as genes with 95% confidence not overlapping with zero. Finally, we removed genes in the MHC region and on

non-autosomal chromosomes.

Our prediction framework includes four main sections which act sequentially (Figure 1). For each gene, we assessed predictive performance using tenfold cross-validation R^2 between predicted and true expression in the initial reference data. We trained and evaluated the models for gene expression prediction in each round of tenfold cross-validation using the following steps. (i) We performed eQTL analyses within 1 Mb of the transcription start/end sites of genes using only the training data to avoid overfitting. (ii) We annotated the eQTL with epigenetic annotations. (iii) We selected multiple SNP sets according to the eQTL *P*-value threshold, chromatin state, TFBS annotation, and DHS annotation. (iv) For each SNP set, we built an expression prediction model in the training dataset by using the lasso and the elastic net ($\alpha = 0.5$) methods as implemented in the *glmnet* R package.

Second, for each model, we evaluated its prediction performance by the Pearson's correlation coefficient R^2 between the predicted gene expression and the observed gene expression of the testing set in fold *y* and averaged all the cross-validation data. For each gene *x*, the model with the highest mean correlation coefficient was selected as the best model. Based on the parameters of the best model and statistics of eQTL analyses using all samples, we constructed each gene's final prediction model. Finally, after getting the best model for gene *x*, we could predict expression directly for genotyped samples using the effect sizes from the reference panels and measure the

association between predicted expression and a trait. On the other hand, the ImpG-Summary algorithm [32] has been used to extend to train on the genetic component of expression based on GWAS summary data [15]. Thus, we could indirectly estimate the association between predicted expression and a trait as the weighted linear combination of SNP-trait standardized effect sizes while accounting for linkage disequilibrium (LD) among SNPs.

Evaluating Prediction Accuracy

The within-study prediction was evaluated by tenfold cross-validation in the whole blood and 13 brain tissues of the highly heritable genes. The accuracy was evaluated using tenfold cross-validation R^2 between predicted and true expression. We then evaluated the prediction models on an independent cohort. Cross-study prediction accuracy was measured by using weights derived from the best models with all whole blood samples to predict gene expression levels in the Geuvadis dataset.

We compared the cross-validation R^2 between ETWAS and recent work in parallel to ours, which imputed expression using only genetic variants with different models, such as lasso, and elastic net [13, 16]. Briefly, lasso uses an L1 penalty as a variable selection method to select a sparse of predictors [33]. In contrast, the elastic net linearly combines the L1 and L2 penalties of lasso and ridge regression to perform variable selection [34]. We used the R package *glmnet* to implement lasso and elastic net ($\alpha = 0.5$).

Transcriptomic Imputation

ETWAS was done using gene expression prediction models derived above in 13 brain tissues. A strict Bonferroni-corrected study-wise threshold was used: $P = 2.55\times10^{-6}$ (0.05/19,632, the total number of highly heritable genes across tissues). FUSION (http://gusevlab.org/projects/fusion/) was used to conduct the transcriptome-wide association testing. The 1000 Genomes v3 LD panel was used for the ETWAS.

Conditionally Testing GWAS Signals

To determine how much GWAS signal remains after the expression association from ETWAS is removed, conditional and joint testing was done for genome-wide Bonferroni-corrected ETWAS genes via FUSION. Each BD GWAS SNP associations were conditioned on the joint gene model one SNP at a time. The overlapping genes were set to be in the range of 500 KB around each SNP, and the defined regions included only the transcribed region of the genes. The FUSION tool was used to perform the conditional and joint analyses with the *cis*-genetic component of expression we generated.

Functional Relevance of Conditionally Independent Genes

Differential expression analyses

We examined whether the gene or transcript expression of the identified genes were differentially expressed between BD patients and controls. The data generated for this

analysis represent Freeze 1 and 2 of the PsychENCODE consortium dataset. Details about data generation, RNA-Seq quality control, and normalization were previously reported by Michael et al. [35]. In brief, several levels of the transcriptomic organization—gene-level, transcript isoform, and local splicing, were generated in brain samples from 1,695 individuals with an autism spectrum disorder, schizophrenia, and bipolar disorder, as well as controls.

Gene's function analysis based on MGI data

We used the Mouse Genome Informatics (MGI) [36] to check whether the conditionally independent genes are relevant to BD. MGI is the international database resource for the laboratory mouse, providing integrated genetic, genomic, and biological data to assist the study of human health and disease. Knockout mice can be used to identify genes whose deletion results in the disease of mental health.

Partitioned Heritability

To investigate whether the ETWAS identified genes could explain additional heritability, partitioned heritability of BD was estimated using LD score regression (LDSC) following previously described methodology [37]. The goal of the partitioned analysis is to quantify the heritability directly explained by SNPs in each functional category using the summary statistics. We partitioned the heritability explained by conditionally independent genes and GWAS loci by SNPs within 500 kb of the genes and GWAS loci. LD score files were made for the two categories separately and

simultaneously using the open-source software available at https://github.com/bulik/ldsc/wiki/Partitioned-Heritability.

GAD, GWAS Catalog and Drug Target Enrichment Analyses

To demonstrate ETWAS's ability to identify BD-related genes, we performed the Genetic Association Database (GAD) disease enrichment analyses, GWAS catalog enrichment analyses, and drug targets enrichment analyses. A relaxed threshold of 0.01, rather than Bonferroni-correction, was used for GAD disease enrichment analyses since Bonferroni-correction assumes independence and genes tend to be correlated due to co-expression. More genes will allow for better recapitulation and prioritization of appropriate pathways. GAD disease enrichment analyses were carried out in DAVID (https://david.ncifcrf.gov/home.jsp) using default settings. For the GWAS catalog enrichment analyses, reported BD genes were derived from the comprehensive NHGRI catalog of disease-associated variants identified using GWAS [38]. We excluded studies listed in the GWAS catalog that included the PGC samples to make sure our known gene list was independent of the current analysis. We then counted the number of known disease-associated genes that had an ETWAS *P*-value below 0.01. We compared this count to the null expectation based on 10,000 randomly drawn gene sets of similar size to the known disease gene set to derive an enrichment *P*-value. Similarly, we carried out the drug targets enrichment analyses by using the drug targets reported in the Open Targets Platform (https://www.targetvalidation.org/) [39]. The Open Targets Platform integrates

evidence from genetics, genomics, transcriptomics, drugs, animal models, and scientific literature to score and rank target-disease associations for drug target identification. We consider only targets whose overall association score is higher than 0.1. Similarly, we counted the number of drug targets for BD that had an ETWAS *P*-value below 0.01 and compared this count to the null expectation based on 10,000 randomly drawn gene sets.

Phenome-wide Association Studies

To identify phenotypes that may be associated or co-morbid with BD, we conducted a phenome-wide association study (pheWAS) for each SNP in the final model of the identified genes. The top five phenotypes (excluding BD) were reported via public data provided by GWASAtlas (https://atlas.ctglab.nl). To further determine the relationship between BD and the phenotypes identified from pheWAS, genetic correlation of the traits was done for currently available GWAS data via LDSC (https://github.com/bulik/ldsc). The latest GWAS summary statistics were used for the correlation.

Results

Model Generation and Evaluation

Heritability of Gene Expression

We estimated the narrow-sense heritability of genes and found that the mean heritability of gene expression is 0.016 for all protein-coding genes in whole blood,

and 0.04 in brain tissues range from 0.031 to 0.049 in different tissues (Supplementary Table 1). We identified 2,239 high heritability genes in whole blood and 19,632 highly heritable gene-tissue pairs of 9,492 genes in 13 brain tissues. Among the heritable genes in the brain, we found that almost half of them (5,155/9,492) are significant only in one brain tissue (Supplementary Figure 1). Therefore, it is necessary to apply all brain-related tissues in GTEx to generate gene expression models. We observed a high correlation of gene expression heritability in different brain tissues, with a correlation coefficient ranging from 0.40 to 0.79 (Supplementary Figure 2).

Epigenetic Annotation Improve the Performance of Gene Expression Prediction

The prediction performance of gene expression prediction is better in highly heritable genes (Supplementary Figure 3), which can be supplementary to prove the appropriateness of focusing on high heritability genes. We evaluated whether the expression of highly heritable genes could be accurately imputed in 13 brain tissues from genotype with the epigenetic elements as prior. We noted that the cross-validation performance significantly increased with the number of active annotations increasing in 11 tissues, and the genes with at least two active annotations 1.01× to 1.19× outperformed than the genes with 0 active annotations (Figure 2A, Supplementary Figure 4). We identified a negative correlation between cross-validation R^2 and SNP number in the best models (Figure 2B, Supplementary Table 4), which consistent with the sparsity of the local architecture of gene

expression and a handful of genetic variants seem to contribute to the variability in gene expression [40]. We demonstrated that the best SNP sets for predicting gene expression significantly enriched in active epigenetic elements by comparing the epigenetic annotation frequency between the best SNP sets and all variants used (Pearson's chi-squared test, $P < 0.01$) (Figure 2C, Supplementary Figure 5). Specifically, the best SNP sets are distributed in the active HMM annotation at a frequency of 1.32-1.45 times that of all variants, while distributed in the TFBS region and DHS region at a frequency of 2.12-2.43 and 1.70-2.40 times that of all variants. We further evaluated ETWAS's performance and compared them with recent work in parallel to ours via cross-validation. On average, our pipeline attained slightly better performance than using lasso, elastic net, and top SNP (Figure 2D, Supplementary Figure 6). For ETWAS, the average tenfold cross-validated prediction R^2 value ranging from 0.11 to 0.16 in different brain tissues (Supplementary Table 4).

ETWAS Performance in a Separate Cohort

We also tested the prediction models on separate cohorts. We used prediction models trained in GTEx blood to compare predicted and observed expression in Geuvadis. The average prediction R^2 is 0.034. The top six genes with the highest performance from this analysis are illustrated in Supplementary Figure 7, proving a comparison of the predicted and observed expression. Among these genes, the *CHURC1* trained in GTEx performed best, and the squared correlation, R^2, between predicted and observed expression levels in Geuvadis is 0.80. A quantile-quantile plot showed

expected (under the null, the correlation between two independent vectors with the same sample size) and observed (between observed and predicted) R^2 from ETWAS is given (Supplementary Figure 8). We found a substantial departure from the null distribution, indicating that the ETWAS captured a substantial proportion of the transcriptome variability.

ETWAS Identifies New BD Associations

ETWAS Significant Hits

We applied ETWAS to identify genes associated with BD using summary data comprising 20,352 BD cases and 31,358 controls of European descent. ETWAS identified 34 susceptibility genes associated with BD, comprising 44 total associations after Bonferroni corrections ($P < 2.55\times10^{-6}$, Figure 3A, Supplementary Table 5). For example, the most significant gene, *NEK4*, associated with BD in three tissues ($P_{CBG} = 1.66\times10^{-9}$, $P_{ACC} = 7.38\times10^{-9}$, $P_{NAB} = 1.88\times10^{-6}$).

Expression signals explain several BD loci

Since several of the ETWAS hits overlapped with significant BD loci, conditional and joint analyses were performed to establish whether these signals were due to multiple-associated features or conditionally independent. We identified 14 conditionally independent genes through conditional analyses. We observed that *NEK4* explains most of the signal at its loci, a region contained 27 significantly ETWAS hits (rs2071044 lead SNP_{GWAS} $P = 9.10\times10^{-9}$, conditioned on *NEK4* lead

SNP_{GWAS} P = 0.57, explaining 0.901 of the variances) (Figure 3B). Similarly, conditioning on *PBX4* explains most of the loci's variance on chromosome 19 (rs1064395 lead SNP_{GWAS} $P = 9.70\times10^{-10}$, conditioned on *PBX4* lead SNP_{GWAS} P = 0.011, explaining 0.583 of the variances) (Supplementary Figure 9A). We also found that *LMAN2L* explains most of the signal at its loci (rs6746896 lead SNP_{GWAS} $P = 7.00\times10^{-9}$, conditioned on *LMAN2L* lead SNP_{GWAS} P = 0.059, explaining 0.674 of the variances) (Supplementary Figure 9B).

Expression signals drive BD ETWAS loci

Among the conditionally independent genes, we identified 11 genes that were not implicated in the original BD GWAS, which regarded as novel candidate genes for BD (Table 1). Conditioning on the expression of *BRF2* explains 0.352 of the variances (rs12677998 lead SNP_{GWAS} $P = 1.10\times10^{-6}$, conditioned on *BRF2* lead SNP_{GWAS} $P = 1.60\times10^{-3}$) (Figure 3C). Similarly, conditioning on the expression of *ADD3* in chromosome 10, *FADS1* in chromosome 11, *CDAN1* in chromosome 15, *HDAC5* and *ASB16* in chromosome 17 explains most of the variances of the lead SNP (Supplementary Figure 10).

ETWAS increase power to find BD associations

Differential expression analyses

We found 11 novel genes significantly associated with BD, indicating the expression of these genes may be associated with BD risk independent of the GWAS-identified

risk variants. We examined whether the gene or transcript expression of the conditionally independent genes were differentially expressed between BD patients and controls using the mRNA expression profiles from PsychENCODE consortium dataset. We identified 10 of the 14 conditionally independent genes associated with BD at the gene expression level or the transcript level, including eight novel genes (Supplementary Table 6).

Functional validation in MGI

Five of the novel genes can be found annotated phenotypes in the knock out mice model. Since the phenotypes MGI arranged did not include BD, we listed all the phenotypes of conditionally independent novel genes reported in MGI (Supplementary Table 7). We found three genes (*HDAC5*, *ASB16*, and *CDAN1*) associated with cardiovascular disease (CVD) relevant phenotypes, such as cardiac hypertrophy, abnormal heart morphology, abnormal heart shape, and an enlarged heart. There is an integration of the various factors that putatively underlie the association of BD with CVD [41], which indirectly suggested that the genes we found may be related to BD.

Partitioned Heritability of ETWAS identified BD genes

We partitioned the heritability explained by SNPs around conditionally independent genes and found that conditionally independent genes explained 3.00% (se = 0.46%) of the estimated heritability, an 8.5× enrichment ($P = 1.31\times10^{-8}$) compared to the

percentage of SNPs. The GWAS loci explained 4.58% (se = 0.45%) of the estimated heritability, a 7.9× enrichment ($P = 5.46\times10^{-17}$) compared to the percentage of SNPs. Combining the SNPs from ETWAS and GWAS explained a much larger percentage of heritability (6.46%, se = 0.65%), a 7.8× enrichment ($P = 9.29\times10^{-17}$) compared to the percentage of SNPs in either category. We performed t-tests to compare the partitioned heritability from GWAS loci with the partitioned heritability from GWAS loci and the ETWAS identified conditionally independent genes. As shown in Figure 4A, ETWAS identified genes significantly increased the proportion of explained heritability ($P = 0.019$).

GAD, GWAS Catalog and Drug Target Enrichment Analyses

The GAD disease enrichment analyses detected four significantly enriched diseases (Figure 4B), including bipolar disorder ($P = 1.7\times10^{-3}$), and attention deficit disorder with hyperactivity ($P = 0.017$). We found that BD had a significant enrichment ($P < 10^{-4}$) of GWAS catalog reported genes (Figure 4C) and Open Targets Platform reported drug targets (Figure 4D) in the ETWAS results, which suggested that there are likely to be true disease associations among the genes that fail to meet strict genome-wide significance.

Validation of novel loci in subsequent BD study

We further employed ETWAS to identify new expression-trait associations using an early released summary association data for BD in 2012, comprising 7,481 BD cases

and 9,250 controls of European descent [29]. We identified four conditionally independent novel genes using the 2012 BD GWAS summary, that is, associations more than 500 kb away from any genome-wide significant SNPs in that study. We then looked for genome-wide significant SNPs at these loci in the larger 2019 BD GWAS [4] (expanded to 20,352 BD cases and 31,358 controls). We identified all the four novel BD-associated genes contained genome-wide significant SNPs in the 2019 GWAS summary data ($P < 5\times10^{-8}$, Supplementary Table 8). Thus, ETWAS is highly predictive of robust phenotypic associations.

Phenome-wide association study

A total of 158 phenotypes, such as cognitive, immunological, metabolic, neurological, psychiatric, were significantly associated with the SNPs in the final model of the BD genes (Figure 5A). To determine whether the pheWAS traits were genetically correlated with BD, a genetic correlation was done between BD and the most recent GWAS for each of the phenotypes. We found 38 phenotypes, including nine psychiatric phenotypes, genetically correlated with BD ($P < 0.05$) (Figure 5B). After Bonferroni corrections, there were nine phenotypes (schizophrenia, depression, frequency of depressed mood in last 2 weeks, gray matter, educational attainment, worrier/anxious feelings, neuroticism, reaction time, mean time to correctly identify matches) positive correlated with BD and one phenotype (number of symbol digit matches made correctly) negative correlated with BD. These phenotypes have

previously been implicated as risk factors for BD, reaffirming the SNPs' relevance in the final models.

Discussion

In this study, we developed a four-step quantitative pipeline named epigenetic element-based transcriptome-wide association studies (ETWAS). We identified 44 genes with genetically-predicted expression levels associated with BD risk. Additionally, through conditional and joint analyses, we identified 14 conditionally independent genes associated with BD risk in 13 brain tissues. Eleven of these genes that not previously implicated with BD, suggesting they are potential novel candidate genes.

We applied conditional and joint association methods to identify genes with significant ETWAS associations when analyzed jointly to account for multiple associations in the same locus. Importantly, the ETWAS expression signals were driving the significance for several previously implicated BD loci when conditioned on the ETWAS genes. For example, the conditioning of *NEK4* led to explain 90.1% of the GWAS signal, which suggests that after considering the predicted expression signal of *NEK4*, there is little residual association signal from the genetic variant in the GWAS locus. We identified 14 conditionally independent genes through conditional analyses, three (*NEK4*, *LMAN2L*, and *PBX4*) implicated in the original BD GWAS, and the rest eleven genes regarded as novel candidate genes for BD.

Zhihui et al. [42] have reported the association between psychiatric risk alleles and mRNA expression of *NEK4*, and the overexpression of *NEK4* could reduce mushroom density spines in rat primary cortical neurons, the most mature form of all spines that responsible for long-term memory. Future studies could interrogate whether expression differences of other candidate genes are consistent with our findings.

Since gene's heritability providing an upper bound of the predictive accuracy, expression of genes that are not significantly heritable at current sample sizes are not included in this project. Some of the strongly implicated genes in BD risk were not assayed, such as *ITIH1*, *FADS2*, and *NCAN* [4], due to non-significant heritability estimates in any of the brain tissue. We detected the overlap between the conditionally independent genes identified via ETWAS and GWAS reported genes with significantly heritable in at least one brain tissue. We identified only one gene (*LMAN2L*) reported in the GWAS discovery dataset (20,352 cases and 31,358 controls). However, another three genes (*ADD3*, *HDAC5*, and *PACS1*) were implicated in a more massive combined data with 29,764 BD cases and 169,118 controls (Supplementary Figure 11). Furthermore, two genes (*NEK4* and *FADS1*) were Bonferroni-corrected significant in several brain tissues, while others only significant in specific tissues. Since expression regulation may be common across tissue types, it was refreshing not to see consistency across panels. For instance, *PBX4* had a *P*-value of 2.13×10^{-8} in the cerebellum but a *P*-value of 0.069 in the cortex. Similarly, *HDAC5* had a *P*-value of 2.41×10^{-8} in the cerebellar hemisphere but a *P*-value of 4.90×10^{-3} in

the substantia nigra. Although it may be due to tissue-specificity, it is essential to note that it may also be due to specific effects and the quality of the RNA data and panel size of different tissue types from GTEx.

The GAD disease enrichment analyses detected four GAD disease terms, including bipolar disorder, attention deficit disorder with hyperactivity (ADHD), magnesium, and arterial stiffness. Sandra et al. [43] conducted a nationwide follow-up study with a total of 2,409,236 individuals investigating the relationship between ADHD and anxiety with the onset of BD. They found that prior diagnoses of ADHD and anxiety disorders are associate with an increased risk of bipolar disorder. By using a Mendelian randomization analysis, Wenwen et al. [44] found that magnesium supplementation would increase BD risk. Lithium, the most common treatment for bipolar disorder, could inhibit glycogen synthase kinase-3 by competition for magnesium [45]. Jess et al. [46] performed a cross-sectional metabolic and vascular function evaluation on a sub-sample near completion after a mean follow-up of 27 years and identified chronicity of mood symptoms contribute to vasculopathy in a dose-dependent fashion. GWAS catalog enrichment analyses and drug target enrichment analyses further suggest that ETWAS identified genes are likely to be true disease associations even those fail to meet strict genome-wide significance.

Although we are convinced the method we used to predict gene expression with the combination of genetic variants and regulatory elements has significant potential to

delineate further the biological mechanisms for human complex diseases, our current study's limitations also need to be addressed. First, although gene expression is amenable to genetic prediction with relatively modest sample sizes because of the sparse genetic architecture of gene expression [40], recent evidence also suggests that larger expression reference panels will help increase the total number of significant cis-heritable genes available for prediction [15]. However, due to the limitation of the published transcriptome data sets, the sample size of our reference data was not large enough, as mentioned above. Second, in part due to the historical paucity of eQTL in populations of non-European ancestry, all subjects from the two reference panels were limited to be European ancestry, and the results may not apply to other populations. Since the genetic predictors of gene expression are more accurate in populations of similar ancestry [47], further study with a larger sample size of different races with both genotype and gene expression levels is needed. Third, although ultrarare variants have been reported to drive substantial cis heritability of human gene expression [48], it is unrealistic to include singletons in expression prediction models at present. Finally, identification of BD-associated genes by ETWAS does not imply causality, and functional studies are needed to determine underlying mechanisms of risk comprehensively. Larger transcriptome and GWAS datasets for BD are likely to improve statistical power for gene identification in the future. Likewise, transcriptome datasets from specific ancestry could also improve future BD ETWAS approaches.

In summary, we demonstrated our ETWAS method to investigate the association between gene expression and BD risk using the largest BD cohort currently available with 20,352 BD cases and 31,358 controls. We found that ETWAS outperformed than original methods and epigenetic annotation could improve gene expression prediction performance. Here, we show that 44 genes reach Bonferroni-corrected significance. We identified 14 conditionally independent genes through conditional analyses, and 11 of them were not previously implicated with BD. We demonstrated that several of the genome-wide significant signals from the BD GWAS are driven by genetically regulated expression. Additionally, ETWAS identified genes could explain heritability beyond that explained by GWAS-associated SNPs. GAD and GWAS catalog enrichment analyses highlight the potential function of the identified genes that fail to meet strict genome-wide significance. We further identified several phenotypes previously associated with BD by querying the SNPs in the final model of identified genes in phenome databases. In conclusion, ETWAS is a powerful method that increases statistical power to identify genes associated with BD. We hope ETWAS could provide novel insights into the identification of additional susceptibility genes and further delineate the biological mechanisms for other human complex diseases.

Conflict of Interest statement

The authors have no financial conflicts of interest.

Acknowledgments

This study is supported by the National Natural Science Foundation of China (81872490), Innovative Talent Promotion Plan of Shaanxi Province for Young Sci-Tech New Star (2018KJXX-010), and the Fundamental Research Funds for the Central Universities. This study is also supported by the High-Performance Computing Platform of Xi’an Jiaotong University.

Reference

1. Merikangas KR, Jin R, He JP, Kessler RC, Lee S, Sampson NA, et al. Prevalence and correlates of bipolar spectrum disorder in the world mental health survey initiative. Arch Gen Psychiatry. 2011;68(3):241-51. doi: 10.1001/archgenpsychiatry.2011.12. PubMed PMID: 21383262; PubMed Central PMCID: PMCPMC3486639.

2. Edvardsen J, Torgersen S, Roysamb E, Lygren S, Skre I, Onstad S, et al. Heritability of bipolar spectrum disorders. Unity or heterogeneity? J Affect Disord. 2008;106(3):229-40. doi: 10.1016/j.jad.2007.07.001. PubMed PMID: 17692389.

3. Smoller JW, Finn CT. Family, twin, and adoption studies of bipolar disorder. Am J Med Genet C Semin Med Genet. 2003;123C(1):48-58. doi: 10.1002/ajmg.c.20013. PubMed PMID: 14601036.

4. Stahl EA, Breen G, Forstner AJ, McQuillin A, Ripke S, Trubetskoy V, et al. Genome-wide association study identifies 30 loci associated with bipolar disorder. Nat Genet. 2019;51(5):793-803. doi: 10.1038/s41588-019-0397-8. PubMed PMID: 31043756; PubMed Central PMCID: PMCPMC6956732.

5. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science. 2012;337(6099):1190-5. doi: 10.1126/science.1222794. PubMed PMID: WOS:000308414000028.

6. Soldner F, Stelzer Y, Shivalila CS, Abraham BJ, Latourelle JC, Barrasa MI, et al. Parkinson-associated risk variant in distal enhancer of alpha-synuclein modulates target gene expression. Nature. 2016;533(7601):95-9. doi: 10.1038/nature17939. PubMed PMID:

27096366; PubMed Central PMCID: PMCPMC5042324.

7. Sekar A, Bialas AR, de Rivera H, Davis A, Hammond TR, Kamitaki N, et al. Schizophrenia risk from complex variation of complement component 4. Nature. 2016;530(7589):177-+. doi: 10.1038/nature16549. PubMed PMID: WOS:000369916700030.

8. Claussnitzer M, Dankel SN, Kim KH, Quon G, Meuleman W, Haugen C, et al. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. New Engl J Med. 2015;373(10):895-907. doi: 10.1056/NEJMoa1502214. PubMed PMID: WOS:000360439500005.

9. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. P Natl Acad Sci USA. 2009;106(23):9362-7. doi: 10.1073/pnas.0903103106. PubMed PMID: WOS:000266817500047.

10. Nicolae DL, Gamazon E, Zhang W, Duan SW, Dolan ME, Cox NJ. Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. Plos Genet. 2010;6(4). doi: ARTN e1000888

10.1371/journal.pgen.1000888. PubMed PMID: WOS:000277354200007.

11. Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015;348(6235):648-60. doi: 10.1126/science.1262110. PubMed PMID: 25954001; PubMed Central PMCID: PMCPMC4547484.

12. Park JH, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, et al. Estimation of effect size distribution from genome-wide association studies and implications for future

discoveries. Nat Genet. 2010;42(7):570-5. doi: 10.1038/ng.610. PubMed PMID: 20562874; PubMed Central PMCID: PMCPMC4615599.

13. 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-+. doi: 10.1038/ng.3367. PubMed PMID: WOS:000360394100022.

14. 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. Nat Genet. 2018;50(4):538-+. doi: 10.1038/s41588-018-0092-1. PubMed PMID: WOS:000429529300014.

15. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BW, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet. 2016;48(3):245-52. doi: 10.1038/ng.3506. PubMed PMID: 26854917; PubMed Central PMCID: PMCPMC4767558.

16. Barbeira AN, Dickinson SP, Bonazzola R, Zheng JM, Wheeler HE, Torres JM, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat Commun. 2018;9. doi: ARTN 1825 10.1038/s41467-018-03621-1. PubMed PMID: WOS:000431630400007.

17. Manor O, Segal E. Robust Prediction of Expression Differences among Human Individuals Using Only Genotype Information. Plos Genet. 2013;9(3). doi: ARTN e1003396 10.1371/journal.pgen.1003396. PubMed PMID: WOS:000316866700067.

18. Kouzarides T. Chromatin modifications and their function. Cell. 2007;128(4):693-705. doi: 10.1016/j.cell.2007.02.005. PubMed PMID: WOS:000245098600012.

19. Gusev A, Lee SH, Trynka G, Finucane H, Vilhjalmsson BJ, Xu H, et al. Partitioning

Heritability of Regulatory and Cell-Type-Specific Variants across 11 Common Diseases. Am J Hum Genet. 2014;95(5):535-52. doi: 10.1016/j.ajhg.2014.10.004. PubMed PMID: WOS:000344845000005.

20. Dong XJ, Weng ZP. The correlation between histone modifications and gene expression. Epigenomics-Uk. 2013;5(2):113-6. doi: 10.2217/Epi.13.13. PubMed PMID: WOS:000317346400003.

21. Bannister AJ, Kouzarides T. Regulation of chromatin by histone modifications. Cell Res. 2011;21(3):381-95. doi: 10.1038/cr.2011.22. PubMed PMID: WOS:000288064900003.

22. Singh R, Lanchantin J, Robins G, Qi YJ. DeepChrome: deep-learning for predicting gene expression from histone modifications. Bioinformatics. 2016;32(17):639-48. doi: 10.1093/bioinformatics/btw427. PubMed PMID: WOS:000384666800031.

23. Dong XJ, Greven MC, Kundaje A, Djebali S, Brown JB, Cheng C, et al. Modeling gene expression using chromatin features in various cellular contexts. Genome Biol. 2012;13(9). doi: ARTN R53

10.1186/gb-2012-13-9-r53. PubMed PMID: WOS:000313182600006.

24. Frasca M, Pavesi G. A Neural Network Based Algorithm for Gene Expression Prediction from Chromatin Structure. Ieee Ijcnn. 2013. PubMed PMID: WOS:000349557200247.

25. Karlic R, Chung HR, Lasserre J, Vlahovicek K, Vingron M. Histone modification levels are predictive for gene expression. P Natl Acad Sci USA. 2010;107(7):2926-31. doi: 10.1073/pnas.0909344107. PubMed PMID: WOS:000274599500041.

26. Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature.

2013;501(7468):506-11. doi: 10.1038/nature12531. PubMed PMID: 24037378; PubMed Central PMCID: PMCPMC3918453.

27. Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518(7539):317-30. doi: 10.1038/nature14248. PubMed PMID: 25693563; PubMed Central PMCID: PMCPMC4530010.

28. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis C, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57-74. doi: 10.1038/nature11247. PubMed PMID: WOS:000308347000039.

29. Sklar P, Ripke S, Scott LJ, Andreassen OA, Cichon S, Craddock N, et al. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. Nat Genet. 2011;43(10):977-U162. doi: 10.1038/ng.943. PubMed PMID: WOS:000295316200012.

30. Monks SA, Leonardson A, Zhu H, Cundiff P, Pietrusiak P, Edwards S, et al. Genetic inheritance of gene expression in human cell lines. Am J Hum Genet. 2004;75(6):1094-105. doi: Doi 10.1086/426461. PubMed PMID: WOS:000224866400013.

31. Yang JA, Lee SH, Goddard ME, Visscher PM. GCTA: A Tool for Genome-wide Complex Trait Analysis. Am J Hum Genet. 2011;88(1):76-82. doi: 10.1016/j.ajhg.2010.11.011. PubMed PMID: WOS:000286501500007.

32. Pasaniuc B, Zaitlen N, Shi H, Bhatia G, Gusev A, Pickrell J, et al. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. Bioinformatics. 2014;30(20):2906-14. doi: 10.1093/bioinformatics/btu416. PubMed PMID:

WOS:000343083600009.

33. Tibshirani R. Regression shrinkage and selection via the Lasso. J Roy Stat Soc B Met. 1996;58(1):267-88. PubMed PMID: WOS:A1996TU31400017.

34. Zou H, Hastie T. Regularization and variable selection via the elastic net. J Roy Stat Soc B. 2005;67:301-20. doi: DOI 10.1111/j.1467-9868.2005.00503.x. PubMed PMID: WOS:000227498200007.

35. Gandal MJ, Zhang P, Hadjimichael E, Walker RL, Chen C, Liu S, et al. Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. Science. 2018;362(6420):1265-+. doi: ARTN eaat8127

10.1126/science.aat8127. PubMed PMID: WOS:000452994400042.

36. Blake JA, Eppig JT, Kadin JA, Richardson JE, Smith CL, Bult CJ. Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse. Nucleic Acids Research. 2017;45(Database issue):D723-D9.

37. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 2015;47(11):1228-+. doi: 10.1038/ng.3404. PubMed PMID: WOS:000363988200005.

38. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47(D1):D1005-D12. doi: 10.1093/nar/gky1120. PubMed PMID: WOS:000462587400139.

39. Carvalho-Silva D, Pierleoni A, Pignatelli M, Ong C, Fumis L, Karamanis N, et al. Open Targets Platform: new developments and updates two years on. Nucleic Acids Res.

2019;47(D1):D1056-D65. doi: 10.1093/nar/gky1133. PubMed PMID: 30462303; PubMed Central PMCID: PMCPMC6324073.

40. Wheeler HE, Shah KP, Brenner J, Garcia T, Aquino-Michaels K, Cox NJ, et al. Survey of the Heritability and Sparse Architecture of Gene Expression Traits across Human Tissues. Plos Genetics. 2016;12(11). doi: ARTN e1006423

10.1371/journal.pgen.1006423. PubMed PMID: WOS:000392129600028.

41. Goldstein BI, Carnethon MR, Matthews KA, McIntyre RS, Miller GE, Raghuveer G, et al. Major Depressive Disorder and Bipolar Disorder Predispose Youth to Accelerated Atherosclerosis and Early Cardiovascular Disease A Scientific Statement From the American Heart Association. Circulation. 2015;132(10):965-86. doi: 10.1161/Cir.0000000000000229. PubMed PMID: WOS:000361110300010.

42. Yang ZH, Zhou DY, Li HJ, Cai X, Liu WP, Wang L, et al. The genome-wide risk alleles for psychiatric disorders at 3p21.1 show convergent effects on mRNA expression, cognitive function, and mushroom dendritic spine. Mol Psychiatr. 2020;25(1):48-66. doi: 10.1038/s41380-019-0592-0. PubMed PMID: WOS:000502329400007.

43. Meier SM, Pavlova B, Dalsgaard S, Nordentoft M, Mors O, Mortensen PB, et al. Attention-deficit hyperactivity disorder and anxiety disorders as precursors of bipolar disorder onset in adulthood. Brit J Psychiat. 2018;213(3):555-60. doi: 10.1192/bjp.2018.111. PubMed PMID: WOS:000441903100009.

44. Cheng WW, Zhu Q, Zhang HY. Mineral Nutrition and the Risk of Chronic Diseases: A Mendelian Randomization Study. Nutrients. 2019;11(2). doi: ARTN 378

10.3390/nu11020378. PubMed PMID: WOS:000460829700169.

45. Gould TD, Quiroz JA, Singh J, Zarate CA, Manji HK. Emerging experimental therapeutics for bipolar disorder: insights from the molecular and cellular actions of current mood stabilizers. Mol Psychiatr. 2004;9(8):734-55. doi: 10.1038/sj.mp.4001518. PubMed PMID: WOS:000222851700003.

46. Fiedorowicz JG, Coryell WH, Rice JP, Warren LL, Haynes WG. Vasculopathy Related to Manic/Hypomanic Symptom Burden and First-Generation Antipsychotics in a Sub-Sample from the Collaborative Depression Study. Psychother Psychosom. 2012;81(4):235-43. doi: 10.1159/000334779. PubMed PMID: WOS:000305099500005.

47. Mogil LS, Andaleon A, Badalamenti A, Dickinson SP, Guo XQ, Rotter JI, et al. Genetic architecture of gene expression traits across diverse populations. Plos Genetics. 2018;14(8). doi: ARTN e1007586

10.1371/journal.pgen.1007586. PubMed PMID: WOS:000443389100038.

48. Hernandez RD, Uricchio LH, Hartman K, Ye C, Dahl A, Zaitlen N. Ultrarare variants drive substantial cis heritability of human gene expression. Nat Genet. 2019;51(9):1349-55. doi: 10.1038/s41588-019-0487-7. PubMed PMID: 31477931; PubMed Central PMCID: PMCPMC6730564.

Figure Legends

Figure 1. Schematic of ETWAS approach. The prediction framework includes four sections. First, for each gene, we divide the SNPs within 1 Mb of the transcription start/end site of the gene into multiple SNP sets according to the eQTL *P*-value and epigenetic annotations. We construct multiple elastic net and lasso models with the SNPs in each SNP set using tenfold cross-validation R^2 in the initial reference data. Second, we evaluate the prediction performance of each SNP set by Pearson's correlation coefficient between the predicted and observed expression, and the SNP set with the highest mean correlation coefficient is selected as the best model. Third, we construct the final prediction model with the parameters of the best SNP set using all the samples in the reference data. We estimate the associations between predicted expressions and traits with the combination of SNP-trait effect sizes while accounting for linkage disequilibrium among SNPs.

Figure 2. Epigenetic data improved the performance of gene expression prediction in brain tissues. (A) The correlation coefficient R^2 of the best prediction models in cross-validation were sorted according to the active annotation number and grouped into three categories: 0, 1, ≥2. One asterisk (*) indicates *P*-value smaller than 0.05 ($P < 0.05$), two asterisks (**) indicates *P*-value smaller than 0.01 ($P < 0.01$). (B) The correlation between the prediction performance and SNP number used in the best model. The x-axis represents cross validation r^2 and the y-axis represents the SNP number of the best model. The prediction performance significantly negative

associated with SNP number in the best model. (C) The epigenetic annotation distributions of the SNPs used for all genes and the best ETWAS models in two brain tissues. (D) Accuracy of individual-level expression imputation models. Accuracy was estimated using cross-validation R^2 between predicted and true expression. Bars show the mean estimate across 2 brain tissues and five methods: Top, Lasso, Enet, ETWAS.lasso and ETWAS.enet.

Figure 3. ETWAS identifies new BD associations. (A) Manhattan plot of the ETWAS (upper) and GWAS (lower) results for BD (n = 20,352 cases and n = 31,358 controls). The blue line represents the Bonferroni-corrected significant thresholds, $P = 2.55\times 10^{-6}$ for ETWAS and $P = 5\times10^{-8}$ for GWAS. (B-C) Regional association of ETWAS hits. Chromosome 3 (B) and chromosome 8 (C) regional association plot. The marginally associated ETWAS genes are shown in blue and the conditionally significant genes are shown in green. The bottom panel shows a regional Manhattan plot of the GWAS data before (gray) and after (blue) conditioning on the predicted expression of the green genes.

Figure 4. (A) Partitioned heritability of conditionally independent genes and GWAS loci. The null expectation, equal to the percentage of SNPs in each category (gray), and *P* values report the difference from this expectation. Fold enrichment relative to the null expectation is shown in parentheses below each category. Error bars show 1.96× SE. (B) GAD disease enrichment analyses of ETWAS genes. (C) GWAS

catalog enrichment analyses of ETWAS genes. The two histograms show the expected number of genes with $P < 0.01$ based on 10,000 random permutations. The large red points show the observed number of previously known BD genes/targets that fall below this threshold. (D) Drug targets enrichment analyses of ETWAS genes.

Figure 5. (A) Heatmap of the associations between each SNP in the final models of the 14 identified genes and the top five phenotypes (excluding BD) identified via pheWAS. (B) Genetic correlation between BD and phenotypes associated with top BD eQTLs. The asterisk in the box indicates the correlation passes the Bonferroni significance threshold (0.05/158).

Table 1. Conditionally independent ETWAS genes for BD.

Gene	Tissue	Best eQTL	[a]Cross-validation R^2	TWAS Z-score	[b]TWAS *P*-value	[c]Implicated in 2018 BD GWAS
NEK4	CBG	rs2019065	0.108	6.03	1.66×10^{-9}	Yes
NEK4	ACC	rs2255107	0.073	5.78	7.38×10^{-9}	Yes
NEK4	NAB	rs731831	0.131	4.77	1.88×10^{-6}	Yes
LMAN2L	SUB	rs11891926	0.146	-5.48	4.36×10^{-8}	Yes
PBX4	CER	rs2288865	0.072	5.60	2.13×10^{-8}	Yes
RP11-382A20.3	CEH	rs8034801	0.094	5.59	2.32×10^{-8}	No
HDAC5	CEH	rs7207464	0.093	5.58	2.41×10^{-8}	No
PACS1	FRO	rs7114014	0.107	5.55	2.84×10^{-8}	No
FTCD	NAB	rs2839258	0.195	-5.39	7.08×10^{-8}	No
ASB16	CEH	rs9910055	0.185	5.34	9.29×10^{-8}	No
BRF2	HYP	rs12549353	0.141	5.33	1.01×10^{-7}	No
ADD3	CER	rs4918489	0.155	5.09	3.68×10^{-7}	No
FADS1	CEH	rs174568	0.197	-5.07	3.90×10^{-7}	No
FADS1	CER	rs174535	0.174	-4.82	1.43×10^{-6}	No
HIST2H2AA3	COR	rs2039800	0.034	-4.83	1.39×10^{-6}	No
ZNF584	SCC	rs1550813	0.168	4.82	1.45×10^{-6}	No
CDAN1	CER	rs1359003	0.073	4.78	1.73×10^{-6}	No

[a]The cross-validation R^2 between predicted and observed gene expression is based on tenfold cross-validation within training data. [b]To account for multiple testing, we used a significance threshold of 2.55×10^{-6} for all phenotypes. [c]Whether there are any genome-wide significant SNPs within 500 kb away from the gene. Note: CBG, Brain Caudate basal ganglia; ACC, Brain Anterior cingulate cortex BA24; NAB, Brain Nucleus accumbens basal ganglia; CER, Brain Cerebellum; CEH, Brain Cerebellar Hemisphere; FRO, Brain Frontal Cortex BA9; SUB, Brain Substantia nigra; HYP, Brain Hypothalamus; SCC, Brain Spinal cord cervical c-1; COR, Brain Cortex.

1. Train models for the *x*-th gene expression in *y*-th fold

1.1 Various SNP sets for prediction model

SNP sets	eQTL *P*-value	Chromatin state	TFBS	DHS	Number of SNPs
1	0.05	Transcription	Yes	Yes	5
2	0.05	Transcription	Yes	No	4
3	0.05	Enhancer	No	Yes	4
…		…			…
z	0.001	Other	Yes	No	12

eQTL *P*-value in the **training set**: 0.05, 1×10^{-2}, 1×10^{-3}, 1×10^{-4}, 1×10^{-5}, 1×10^{-6}

Chromatin states: State model from Roadmap

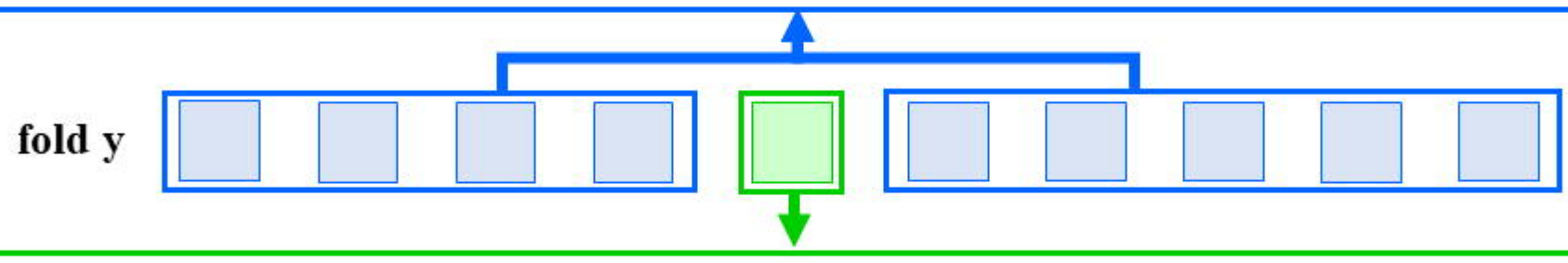
TFBS: Transcription factor binding sites

DHS: DNase I hypersensitive sites

1.2 Effect size of *j*-th variant for gene expression in the training set:

$$E_x \sim \beta_0 + \beta_1 * SNP_1 + \beta_2 * SNP_2 + \cdots + \beta_j * SNP_j$$

E_x: the gene expression for the *x*-th gene; β_j: effect size of *j*-th variant in *z*-th model for *x*-th gene in lasso and elastic net; SNP_j: number of reference alleles of *j*-th variant; β_0: the contribution of other factors that determine the expression trait.



2. Test for the best model for *x*-th gene

2.1 Evaluation of gene expression prediction performance in testing set ($P_{x,y,z}$)

$P_{x,y,z}$: Pearson's correlation coefficient R^2 between the predicted gene expression and observed gene expression of testing set in fold y.

2.2 Choose the best model for *x*-th gene through 10-fold cross validation

SNP sets	eQTL *P*-value	Chromatin state	TFBS	DHS	P1	P2	…	Py	…	P10	Mean
1	0.05	Transcription	Yes	Yes	0.1	0.3		$P_{x,y,1}$		0.2	0.2
2	0.05	Transcription	Yes	No	0.2	0.2		$P_{x,y,2}$		0.4	0.3
3	0.05	Enhancer	No	Yes	0.6	0.4		$P_{x,y,3}$		0.5	0.5
…		…			…			…			…
z	0.001	Other	Yes	No	$P_{x,1,z}$	$P_{x,2,z}$	…	$P_{x,y,z}$	…	$P_{x,10,z}$	$P_{x,z}$

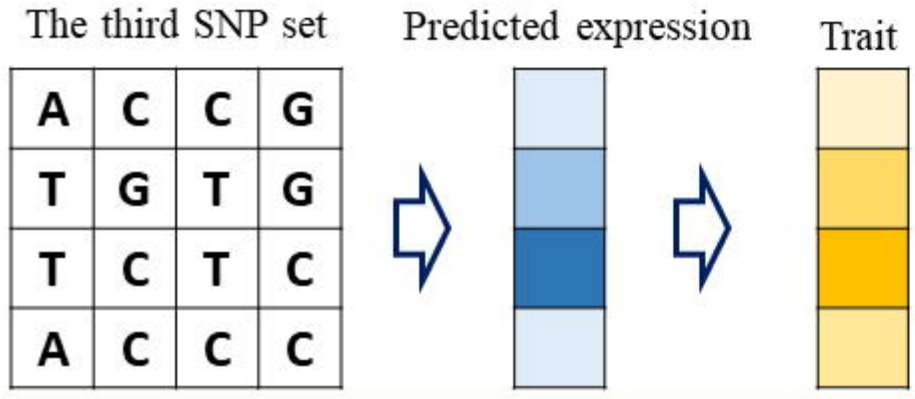
The best model is defined as the SNP sets with the highest performance ($P_{x,z}$) in testing set.

3. Construct the final model

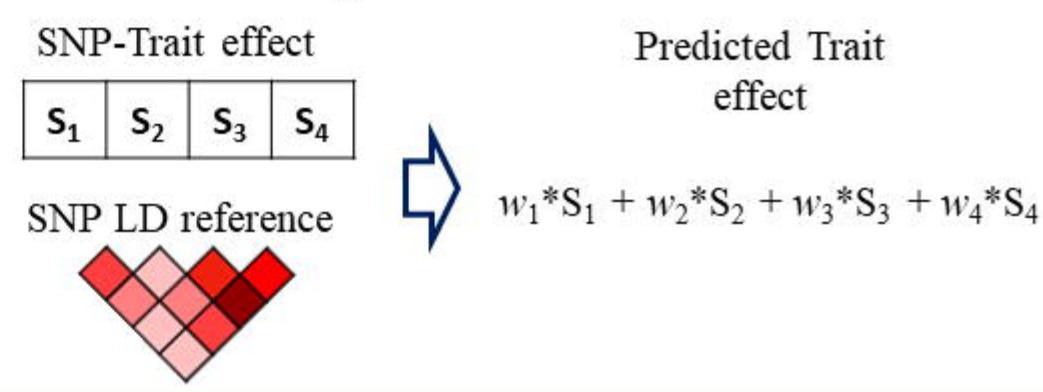
Use the SNP set of best model to construct the final model with all samples

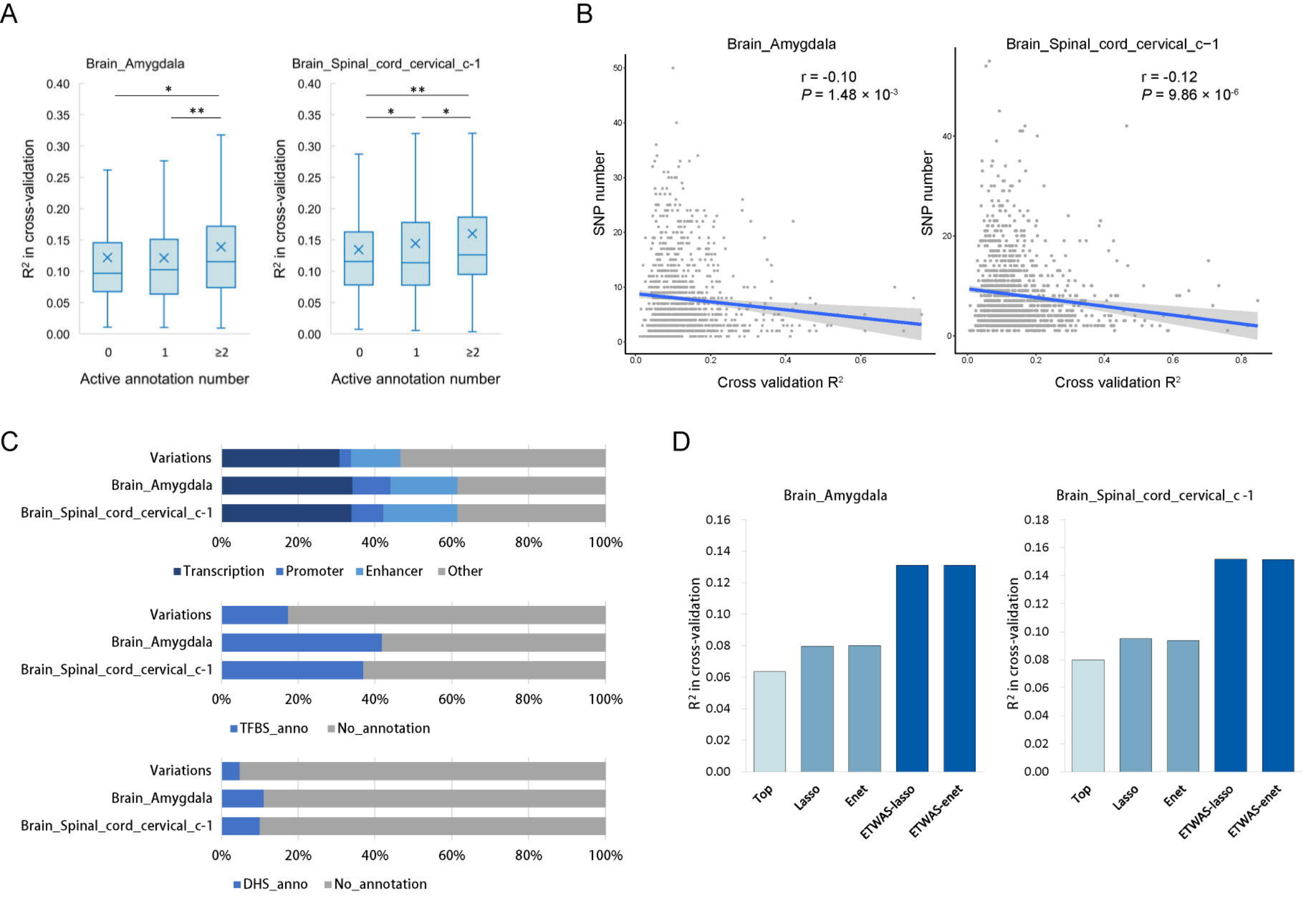
In this example, the third SNP set enjoy the highest performance: **eQTL *P*-value**: 0.05 **Chromatin state**: Enhancer **TFBS**: No **DHS**: Yes

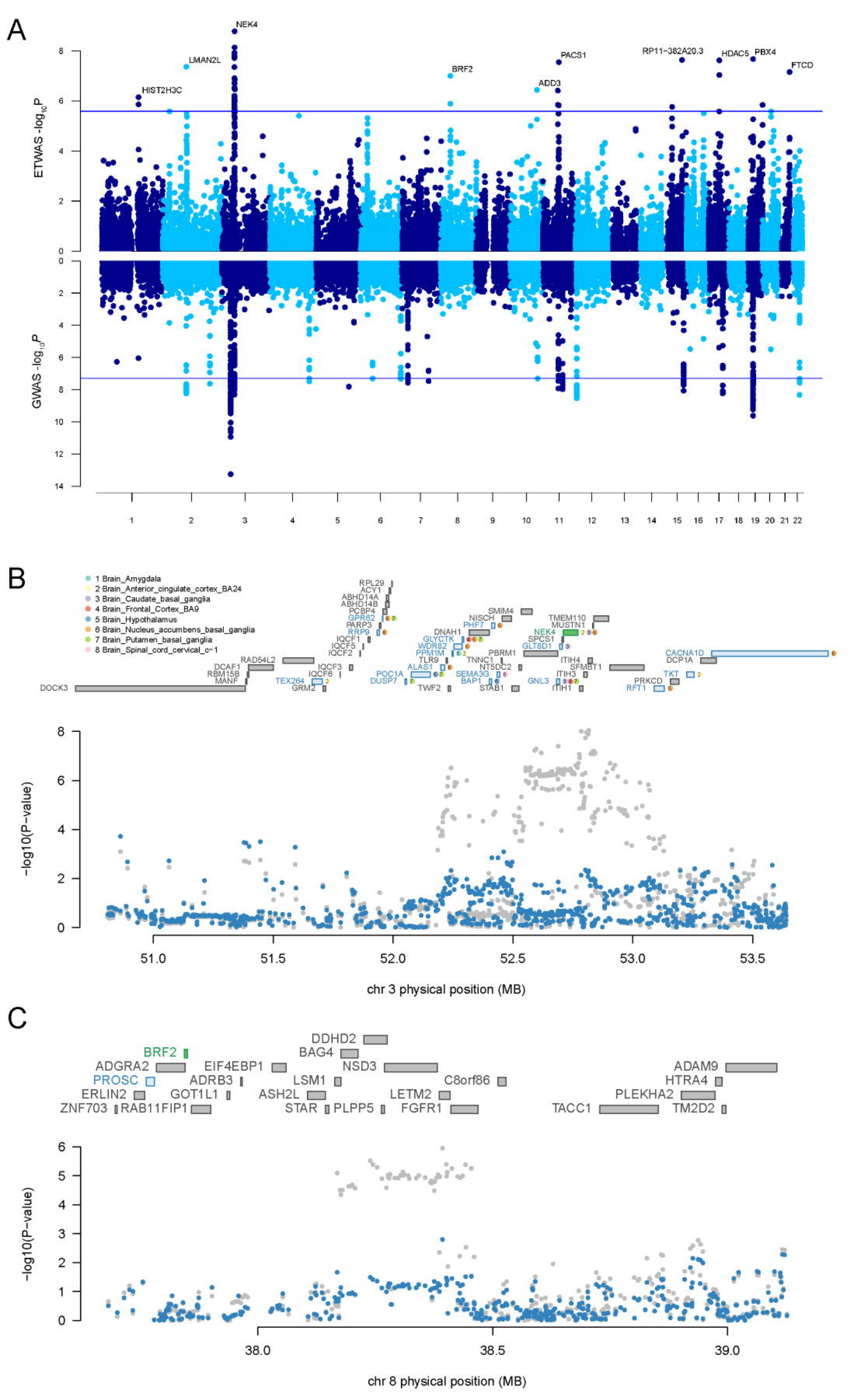
4.1 Individual TWAS

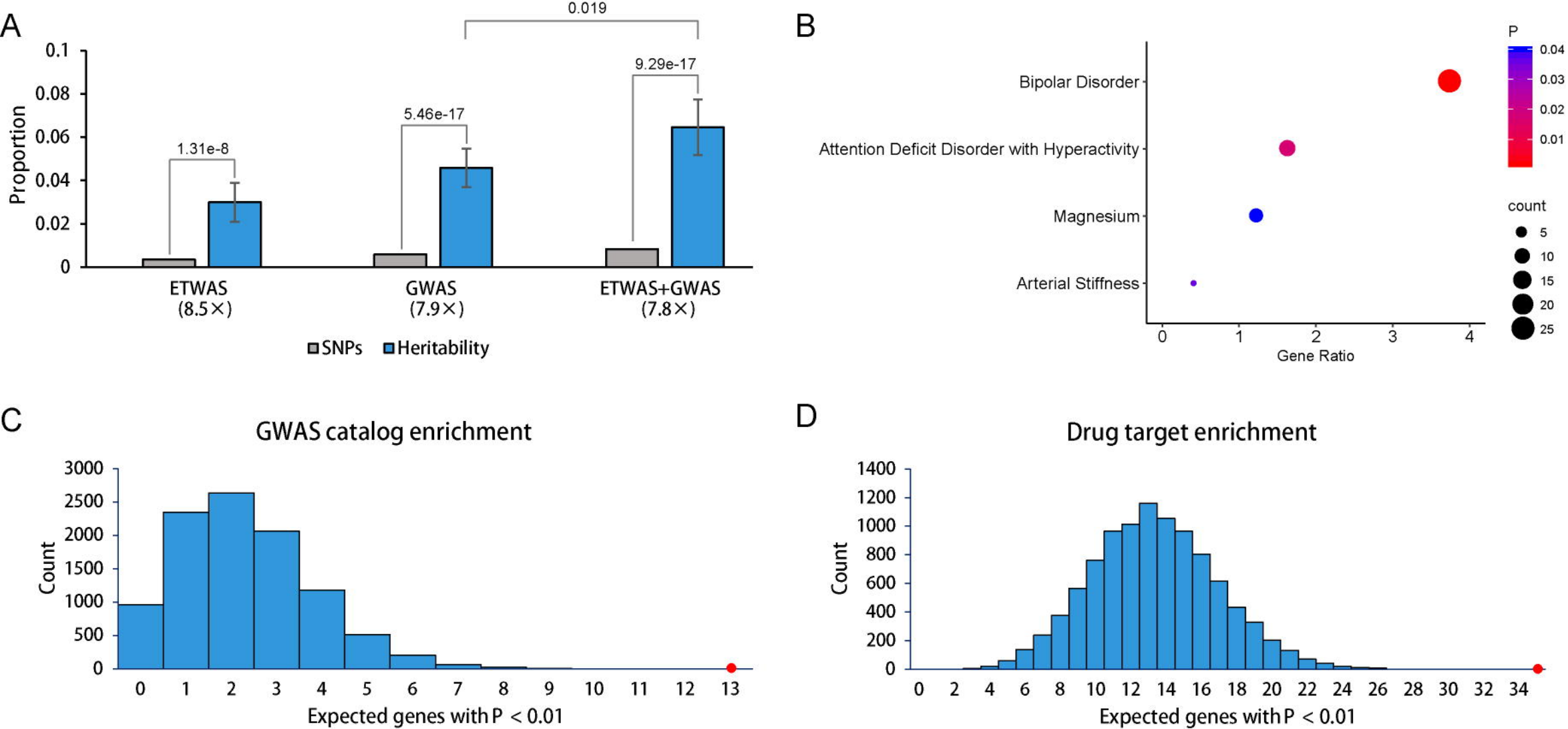


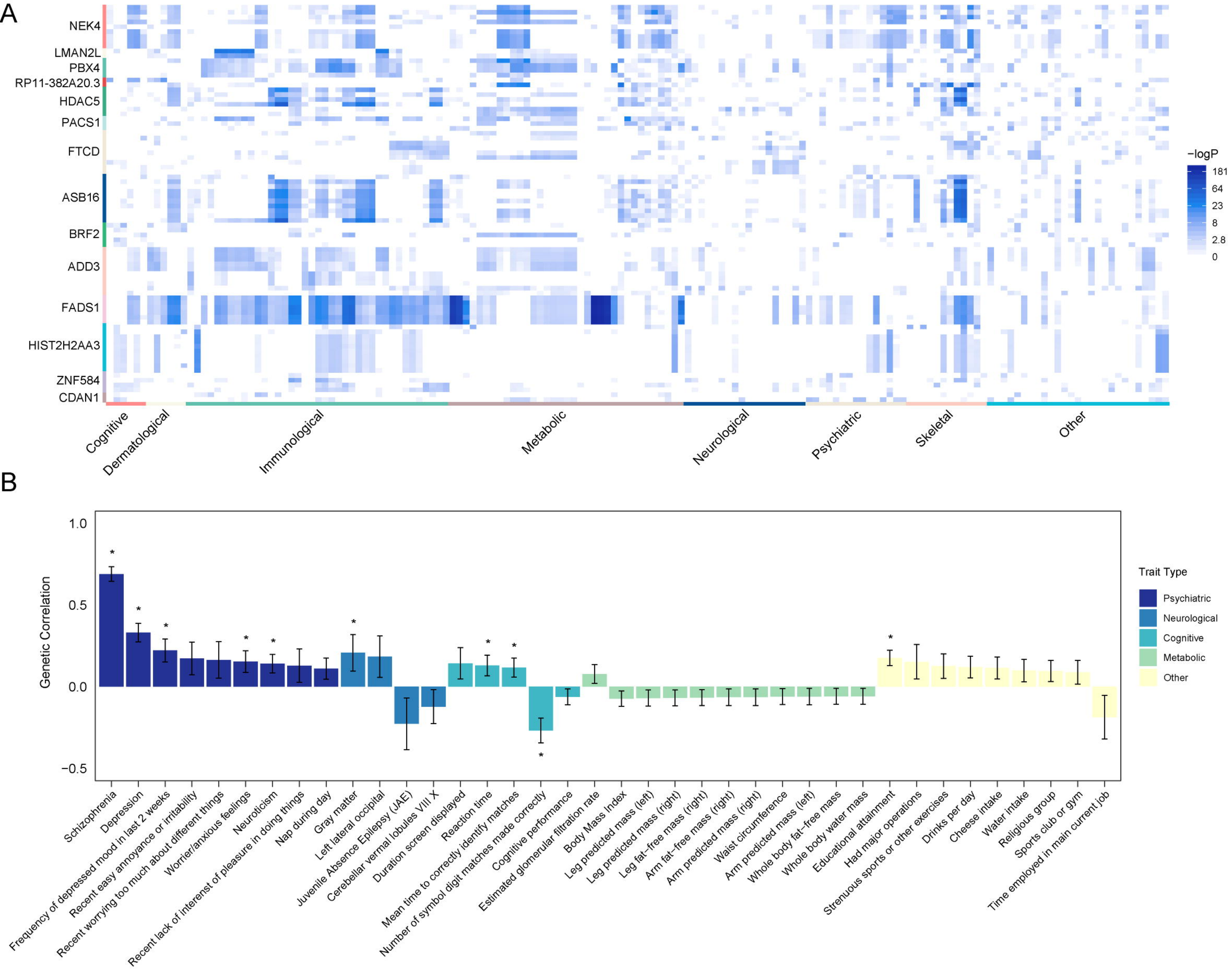
4.2 Summary-based TWAS




A
Brain_Amygdala
Brain_Spinal_cord_cervical_c-1
R² in cross-validation
Active annotation number
0
1
≥2
B
Brain_Amygdala
r = -0.10
P = 1.48 × 10-3
Brain_Spinal_cord_cervical_c−1
r = -0.12
P = 9.86 × 10-6
SNP number
Cross validation R²
C
Variations
Brain_Amygdala
Brain_Spinal_cord_cervical_c-1
Transcription
Promoter
Enhancer
Other
TFBS_anno
No_annotation
DHS_anno
D
Brain_Amygdala
Brain_Spinal_cord_cervical_c -1
R² in cross-validation
Top
Lasso
Enet
ETWAS-lasso
ETWAS-enet


A
NEK4
LMAN2L
HIST2H3C
BRF2
PACS1
ADD3
RP11-382A20.3
HDAC5
PBX4
FTCD
ETWAS -log10P
GWAS -log10P
B
1 Brain_Amygdala
2 Brain_Anterior_cingulate_cortex_BA24
3 Brain_Caudate_basal_ganglia
4 Brain_Frontal_Cortex_BA9
5 Brain_Hypothalamus
6 Brain_Nucleus_accumbens_basal_ganglia
7 Brain_Putamen_basal_ganglia
8 Brain_Spinal_cord_cervical_c-1
DOCK3
RAD54L2
DCAF1
RBM15B
MANF
IQCF3
IQCF6
TEX264
GRM2
IQCF1
IQCF5
IQCF2
RPL29
ACY1
ABHD14A
ABHD14B
PCBP4
GPR62
PARP3
RRP9
POC1A
DUSP7
ALAS1
TWF2
DNAH1
GLYCTK
WDR82
PPM1M
TLR9
SEMA3G
BAP1
STAB1
SMIM4
NISCH
PHF7
PBRM1
TNNC1
NT5DC2
NEK4
SPCS1
GLT8D1
GNL3
TMEM110
MUSTN1
ITIH4
SFMBT1
ITIH3
ITIH1
CACNA1D
DCP1A
TKT
PRKCD
RFT1
-log10(P-value)
chr 3 physical position (MB)
C
DDHD2
BRF2
BAG4
ADGRA2
EIF4EBP1
NSD3
PROSC
ADRB3
LSM1
C8orf86
ERLIN2
GOT1L1
ASH2L
LETM2
ZNF703
RAB11FIP1
STAR
PLPP5
FGFR1
ADAM9
HTRA4
PLEKHA2
TACC1
TM2D2
-log10(P-value)
chr 8 physical position (MB)


A
0.019
9.29e-17
5.46e-17
1.31e-8
Proportion
0.1
0.08
0.06
0.04
0.02
0
ETWAS
(8.5×)
GWAS
(7.9×)
ETWAS+GWAS
(7.8×)
SNPs
Heritability
B
Bipolar Disorder
Attention Deficit Disorder with Hyperactivity
Magnesium
Arterial Stiffness
0
1
2
3
4
Gene Ratio
P
0.04
0.03
0.02
0.01
count
5
10
15
20
25
C
GWAS catalog enrichment
Count
3000
2500
2000
1500
1000
500
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13
Expected genes with P < 0.01
D
Drug target enrichment
Count
1400
1200
1000
800
600
400
200
0
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34
Expected genes with P < 0.01


A
NEK4
LMAN2L
PBX4
RP11-382A20.3
HDAC5
PACS1
FTCD
ASB16
BRF2
ADD3
FADS1
HIST2H2AA3
ZNF584
CDAN1
Cognitive
Dermatological
Immunological
Metabolic
Neurological
Psychiatric
Skeletal
Other
−logP
181
64
23
8
2.8
0
B
Genetic Correlation
1.0
0.5
0.0
−0.5
Schizophrenia
Depression
Frequency of depressed mood in last 2 weeks
Recent easy annoyance or irritability
Recent worrying too much about different things
Worrier/anxious feelings
Neuroticism
Recent lack of interenst of pleasure in doing things
Nap during day
Gray matter
Left lateral occipital
Juvenile Absence Epilepsy (JAE)
Cerebellar vermal lobules VIII X
Duration screen displayed
Reaction time
Mean time to correctly identify matches
Number of symbol digit matches made correctly
Cognitive performance
Estimated glomerular filtration rate
Body Mass Index
Leg predicted mass (left)
Leg predicted mass (right)
Leg fat−free mass (right)
Arm fat−free mass (right)
Arm predicted mass (right)
Waist circumference
Arm predicted mass (left)
Whole body fat−free mass
Whole body water mass
Educational attainment
Had major operations
Strenuous sports or other exercises
Drinks per day
Cheese intake
Water intake
Religious group
Sports club or gym
Time employed in main current job
Trait Type
Psychiatric
Neurological
Cognitive
Metabolic
Other