Complex Psychiatry

Review Article

Complex Psychiatry 2023;9:130–144
DOI: 10.1159/000530223

Received: August 23, 2022
Accepted: March 9, 2023
Published online: April 11, 2023

Integrative Post-Genome-Wide Association Study Analyses Relevant to Psychiatric Disorders: Imputing Transcriptome and Proteome Signals

Huseyin Gedik[a] Roseann E. Peterson[b] Brien P. Riley[b]
Vladimir I. Vladimirov[c] Silviu-Alin Bacanu[b]

[a]Integrative Life Sciences, Virginia Institute of Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA, USA; [b]Institute for Genomics in Health, SUNY Downstate Health Sciences University, Brooklyn, NY, USA; [c]Department of Psychiatry, College of Medicine-Phoenix, University of Arizona, Phoenix, AZ, USA

Keywords
Expression quantitative trait locus · Protein quantitative trait locus · Psychiatry · Proteome-wide association study · Transcriptome-wide association study

Abstract
Background: The genome-wide association study (GWAS) is a common tool to identify genetic variants associated with complex traits, including psychiatric disorders (PDs). However, post-GWAS analyses are needed to extend the statistical inference to biologically relevant entities, e.g., genes, proteins, and pathways. To achieve this goal, researchers developed methods that incorporate biologically relevant intermediate molecular phenotypes, such as gene expression and protein abundance, which are posited to mediate the variant-trait association. Transcriptome-wide association study (TWAS) and proteome-wide association study (PWAS) are commonly used methods to test the association between these molecular mediators and the trait. ***Summary:*** In this review, we discuss the most recent developments in TWAS and PWAS. These methods integrate existing "omic" information with the GWAS summary statistics for trait(s) of interest. Specifically, they impute transcript/protein data and test the association between imputed gene expression/protein level with phenotype of interest by using (i) GWAS summary statistics and (ii) reference transcriptomic/proteomic/genomic datasets. TWAS and PWAS are suitable as analysis tools for (i) primary association scan and (ii) fine-mapping to identify potentially causal genes for PDs. ***Key Messages:*** As post-GWAS analyses, TWAS and PWAS have the potential to highlight causal genes for PDs. These prioritized genes could indicate targets for the development of novel drug therapies. For researchers attempting such analyses, we recommend Mendelian randomization tools that use GWAS statistics for both trait and reference datasets, e.g., summary Mendelian randomization (SMR). We base our recommendation on (i) being able to use the same tool for both TWAS and PWAS, (ii) not requiring the pre-computed weights (and thus easier to update for larger reference datasets), and (iii) most larger transcriptome reference datasets are publicly available and easy to transform into a compatible format for SMR analysis.

Introduction

Genome-wide association study (GWAS) is a statistical framework for univariate testing of the association between a trait outcome and genetic variants, such as single

karger@karger.com
www.karger.com/cxp

Correspondence to:
Huseyin Gedik, gedikh@vcu.edu

Karger

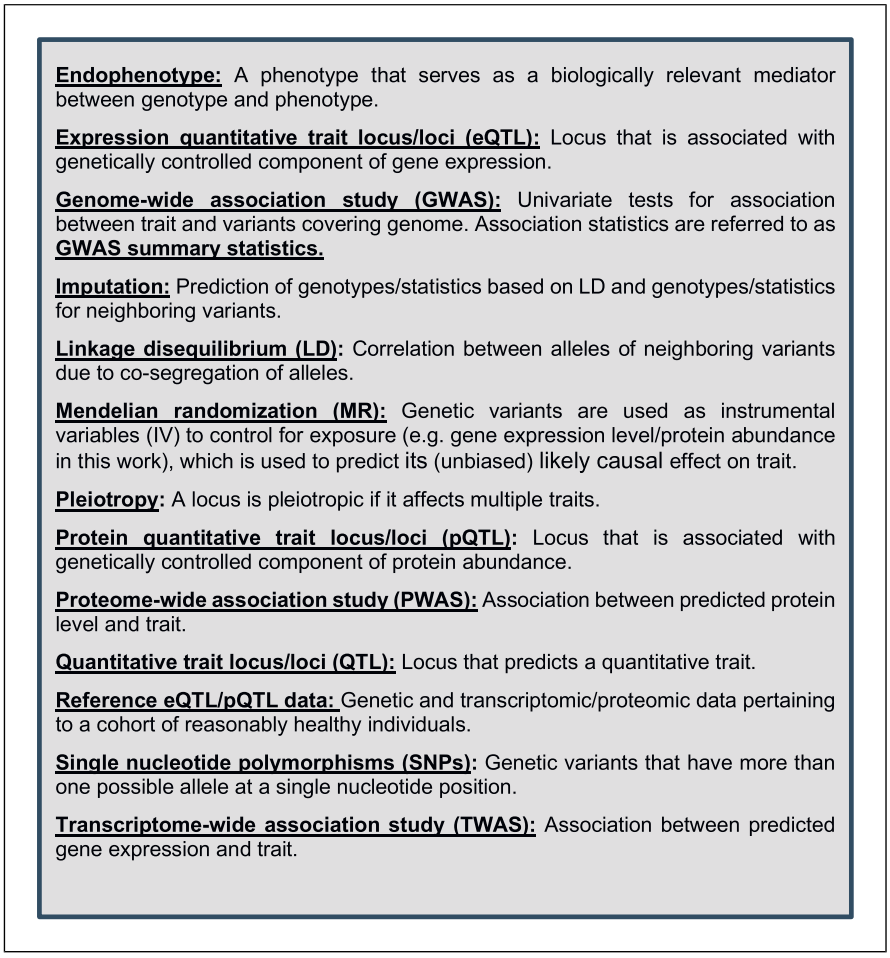

Fig. 1. Key terms.

nucleotide polymorphisms (SNPs), among the thousands or millions of genotyped variants in a genome-wide scan (see Fig. 1 for definitions of key terms). GWAS signals are often found in non-coding segments of the human genome [1]. In certain genomic regions with these signals, large correlations are detected between variants (due to neighboring allele co-segregation). This is commonly known as linkage disequilibrium (LD) (Fig. 2). As a result, significant GWAS signals are observed across multiple genes (or loci). Therefore, it is not straightforward to infer a biologically informative association between these signals and the trait [2]. Consequently, post-GWAS methods are needed to connect variant signals to genes using biologically relevant mediators like gene transcript and protein levels. The enrichment of GWAS signals among variants that influence transcript/protein levels provides evidence that these molecular measures mediate the association of variant with a trait [3, 4]. Researchers can better understand the underlying etiology of disorders by integrating transcriptome and genome data. This enables them to prioritize biologically relevant GWAS signals associated with the trait. Prioritized genes would ultimately indicate molecular targets for future therapeutic intervention or clinical diagnosis/prognosis in common diseases [5, 6].

Loci associated with transcript levels are expression quantitative trait loci (eQTLs) and those associated with protein levels are protein QTLs (pQTLs). Genes with at least one eQTL are expression genes (eGenes). Large consortia have discovered eQTLs and pQTLs for various tissues/fluids by testing the correlation between genetic variants and transcript [7, 8] and protein levels [9–12]. Henceforth, we refer to this association data from eQTL/pQTL studies as reference transcriptome/proteome data.

By combining GWAS results with information on eQTLs (variant weights derived from reference data), it is possible to impute gene expression [13–15] and thus researchers can perform a gene-based association scan between trait and predicted transcript level. Such an approach is known as a transcriptome-wide association

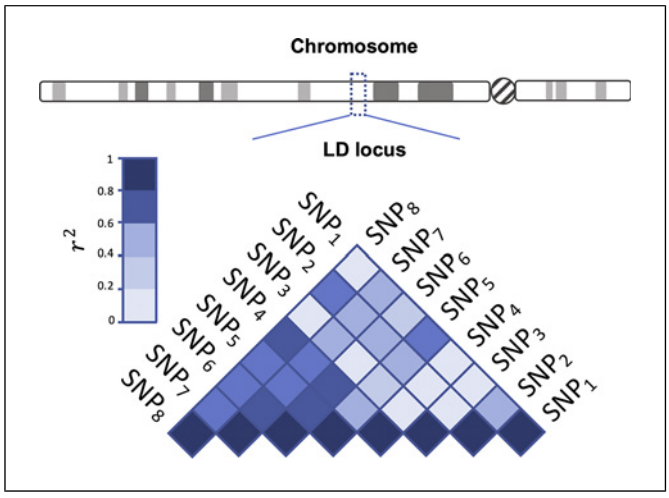

Fig. 2. Triangular matrix with color shades showing the strength of the LD (r^2, square of pair-wise Pearson's correlation coefficient). Darker colors denote higher LD between SNP pairs in the designated locus.

study (TWAS). It is a widely used method to uncover potentially causal loci and genes that were not previously identified as significant by GWAS [16, 17].

Although profiling the human proteome is not yet as advanced as DNA and RNA sequencing, it has recently made notable progress in terms of scaling up the number of analytes [18]. Reference proteome data in population-based cohorts have enabled the discovery of pQTLs, e.g., in human blood plasma [4] and brain tissues [19]. Similar to the TWAS paradigm, proteome-wide association studies (PWASs) are conducted by integrating GWAS summary statistics and reference pQTL information [20–22] to test the association between trait and protein abundance (Fig. 3a). These methods simply test the association between trait and predicted gene expression/protein levels (Fig. 3b).

To help researchers uncover the molecular basis of GWAS signals, we discuss the TWAS/PWAS methods and their associated transcriptome/proteome reference data. The main text of the review is divided into three sections: GWAS, TWAS, and PWAS. First, we briefly review GWAS, its shortcomings, and the need for post-GWAS analyses. Second, we lay out the need for biologically informative mediators (transcriptome and proteome data) that are empirically supported in the literature. We provide details on various transcriptome and proteome reference datasets as well as TWAS methods, which are also used in PWAS. Finally, we discuss the limitations and possible future improvements of the TWAS and PWAS approaches.

Methods to Link Genome, Transcriptome, and Proteome Data to Phenotypes

Genome-Wide Association Studies

Advances in genotyping technology allowed for the development of GWAS - a genome-wide scanning for the association signals between the genotype and the phenotype using a regression model (Fig. 4). The ultimate aim of GWAS is to identify common risk variants (minor allele frequency >0.01). However, as common variants have generally small effects on phenotypes, for signal detection in GWAS, large sample sizes are essential. With adequate sample sizes, GWASs provided numerous statistically significant association signals for many disorders with complex etiologies, including metabolic, neurodegenerative, and neuropsychiatric disorders [23]. Although some GWAS analyses employ mixed effect models or logistic regression, we give details for only the simple linear model (Fig. 4).

Limitations of GWAS in Identifying Biologically Informative Loci

GWAS signals are limited in explaining the biological mechanisms involved in complex traits. While there are many reasons for this, we discuss two of the most important ones. First, most signals are found in the non-coding regions of the human genome [1, 24], i.e., genomic regions that do not code for proteins. Second, observing large correlations between alleles of SNPs in close proximity, which is generally known as LD (Fig. 2) [25]. A genomic region with elevated pairwise correlations between alleles at two or more loci is commonly denoted as an "LD block" [26, 27]. This phenomenon is a direct consequence of the lack of recombination at the LD block in meiosis through many consecutive generations [25]. Because any causal signal within an LD block induces associations with numerous other SNPs within the same block, further fine-mapping of the region is necessary in order to identify the causal variant(s).

Support for Biologically Relevant Mediators between Genes and Traits

While few GWAS signals are directly informative for the biological underpinnings of the disease, researchers showed that these genomic regions were enriched in regulatory effects on gene expression [2, 3] and protein abundance levels [28]. GWASs provided empirical support for transcript levels as a variant-trait mediator by showing enrichment of eQTL signals in association signals of many common disorders [3, 29, 30]. Similarly, pQTLs were shown to co-localize with GWAS signals in

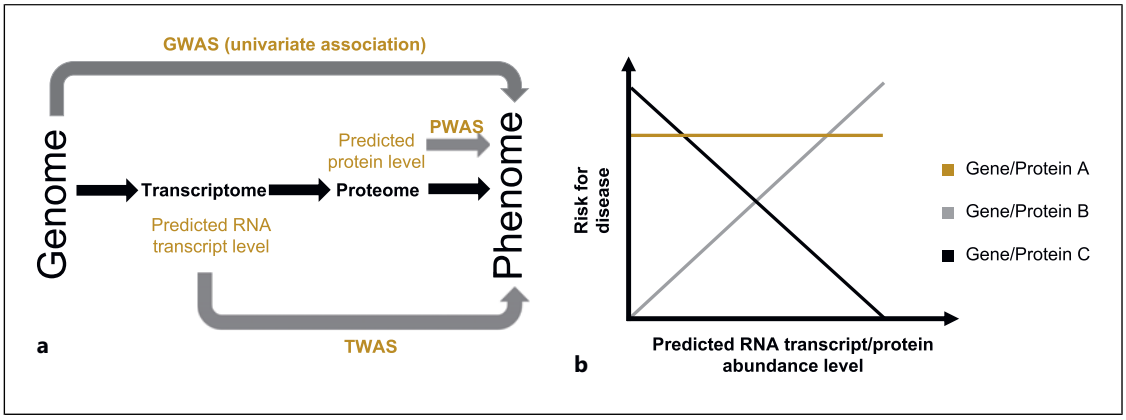

Fig. 3. a Associations among omics of biological complexity at molecular level. **b** TWAS and PWAS aim to infer a potential causal path between the predicted transcript/protein level and the risk for disease.

A common association model for outcome in GWAS is an ordinary linear regression model:

$$Y \sim \beta_0 + \boldsymbol{G}\beta_G + \boldsymbol{X}\beta_X + \boldsymbol{\varepsilon} \qquad \varepsilon \sim N(0, \sigma^2) \qquad (\mathbf{1})$$

where the Y is the phenotype vector, $\boldsymbol{G}$ is the additively coded genotype, $\boldsymbol{X}$ is the matrix of covariates, and $\boldsymbol{\varepsilon}$ is the error vector of independent and identically distributed normal variates having mean 0 and variance σ^2 **(1)**. Covariates may include age, sex, continental genetic ancestry, ancestral principal components and other confounding factors, e.g., batch/platform/site effects.

Fig. 4. Statistical model testing for the association between genetic variants and the trait of interest.

traits with complex etiologies [31]. Thus, the empirical evidence supports that RNA transcript levels and protein abundance serve as molecular mediators of variant-trait association.

Transcriptomics: TWASs

Currently, there are a large number of TWAS methods (Table 1). PrediXcan, TWAS, and S-PrediXcan were the first among these methods [13, 14, 32]. To summarize the analysis steps, TWAS methods first use a reference cohort that has genetic and transcriptomic data from the same subjects (Fig. 5a). Second, it selects the best prediction model for imputing the gene expression from the individual-level genotype (Fig. 5b). Third, it uses the best fitting prediction model to predict genetically regulated gene expression (GReX) (Fig. 5c). The fourth step is to compute the trait-gene expression association statistic by regressing outcome on GReX (and relevant covariates) (Fig. 5d). The prediction model in TWAS requires precomputed weights for each gene and tissue under investigation. The weights fall into three main categories. The first is based on single tissue models including peripheral blood, whole blood, adipose tissue, and brain tissues. The second one is the multiple tissue-based model, whose weights are estimated from GTEX v8. The third one is the cross-tissue model weights [33] which are based on GTEx v6 and GTEx v8.

Subject-level genotype data for cohorts are not easy to access due to privacy concerns. Subsequently, S-PrediXcan and TWAS-FUSION do not require individual level genetic data by employing only GWAS summary statistics for traits under investigation. Because these methods infer only association and are not directly concerned with the causality from GReX to trait, we refer to them as primary TWAS methods.

Primary TWAS Methods

PrediXcan. Among primary TWAS methods, one of the first was PrediXcan [13]. It was followed by its GWAS summary statistics-based version, S-PrediXcan [37]. These

Table 1. TWAS methods for inferring gene-trait association and eQTL colocalization

Method	GWAS input	References	Prediction model or statistical test	Tissue
Gene-level TWAS methods				
JEPEG	GSS	Lee et al., [34] 2015	Multivariate gene based	Single
PrediXcan	GT	Gamazon et al., [13] 2015	Elastic net, LASSO	Single
TWAS/TWAS-FUSION	GT/GSS	Gusev et al., [14] 2016	BLUP, BSLMM/Elastic net, LASSO	Single
JEPEGMIX	GSS	Lee et al., [35] 2016	Mix ancestry multivariate gene based	Single
SMR	GSS	Zhu et al., [36] 2016	MR	Single
S-PrediXcan	GSS	Barbeira et al., [37] 2018	Elastic net, LASSO	Single
COMM	GT	Yang et al., [38] 2019	Mixed joint model	Single
MultiXcan	GSS	Barbeira et al., [32] 2019	Cross tissue PCA regression	Multiple
UTMOST	GSS	Hu et al., [39] 2019	Cross tissue Group LASSO and GBJ	Multiple
TIGAR	GT/GSS	Nagpal et al., [40] 2019	Dirichlet process regression	Single
COMM-S2	GSS	Yang et al., [41] 2020	Mixed joint model	Single
TisCoMM/ TisCoMM-S^2	GT/GSS	Shi et al., [42] 2020	Mixed joint model	Multiple
PMR-Egger	GT/GSS	Yuan et al., [43] 2020	Probabilistic two-sample MR	Single
BGW-TWAS	GSS	Luningham et al., [44] 2020	Bayesian variable selection regression*	Single
MR-JTI	GT	Zhou et al., [45] 2020	MR	Multiple
InTACT	GT/GSS	Bae et al., [46] 2021	BLUP and Cauchy distribution	Multiple
VC-TWAS	GT/GSS	Tang et al., [47] 2021	SKAT**	Single
TWAS fine-mapping methods				
FOCUS	GSS	Mancuso et al., [48] 2019	BSLMM, GBLUP	Single
FOGS	GSS	Wu and Pan [49] 2020	SPU***	Single
HEIDI	GSS	Zhu et al., [36] 2016	MR	Single
TWAS pathway analysis				
JEPEGMIX2-P	GSS	Chatzinakos et al., [50] 2020	Pathway analysis in multiple ancestries	Single
TWAS-GSEA	GSS	Pain et al., [51] 2019	Linear mixed model	Single
eQTL/GWAS signal-based colocalization methods				
Coloc	GSS	Giambartolomei et al., [52] 2014	Bayesian test (colocalization)	Single
eCAVIAR	GSS	Hormozdiari et al., [53] 2016	Bayesian test (posterior colocalization probability)	Single
Enloc	GSS	Wen et al., [54] 2017	Bayesian hierarchical model	Single

GT, genotype data as input from GWAS; GSS, GWAS summary statistics; JEPEG, joint effect on phenotype of eQTL/functional SNPs associated with a gene; TWAS, transcriptome-wide association, JEPEGMIX, JEPEG for mixed ethnicity cohorts; SMR, summary data-based Mendelian randomization, COMM, collaborative mixed model; UTMOST, uunified test for molecular signature; GBJ; generalized Berk-Jones; TIGAR, transcriptome integrated genetic association resource; COMM-S2, collaborative mixed models for GWAS summary statistics; TisCoMM, tissue-specific collaborative mixed model; PMR-Egger, probabilistic two-sample Mendelian randomization-Egger regression; BGW-TWAS, Bayesian genome-wide TWAS; MR-JTI, Mendelian randomization joint tissue imputation; InTACT, integrative test of associations via Cauchy transformation; VC-TWAS, novel variance-component TWAS; FOCUS, fine-mapping of causal genes; FOGS, fine-mapping of gene sets; JEPEGMIX2-P, JEPEGMIX2 pathway; TWAS-GSEA, TWAS-based gene set enrichment analysis; HEIDI, heterogeneity in dependent instruments; coloc, colocalization; eCAVIAR, eQTL causal variants identification in associated regions; enloc, enrichment estimation-aided colocalization analysis; BLUP, best linear unbiased predictor; LASSO, least absolute shrinkage and selection operator; PCA, principal component analysis; SUP, sum of powered score; BSLMM, Bayesian sparse linear mixed model . *The method for statistical inference was from another study [55]. **The statistical model is published elsewhere in detail apart from the main research article [56]. ***The primary method relies on another study [57].

methods select the optimal prediction model using regularization methods, e.g., LASSO [58] or elastic net [59]. The prediction models used in PrediXcan are available in PredictDB (see data resources listed in Data Availability Statement). Some of the most recent updates on transcriptome prediction models are based on PsychENCODE [60], GTEx v8 [8] and its cross-tissue models [39]. While for GTEx v8, they used elastic net for the model selection method, the improved versions employ a multivariate adaptive shrinkage (mash) [61]. This method also has an R implementation called mashr (see data resources listed in Data Availability Statement).

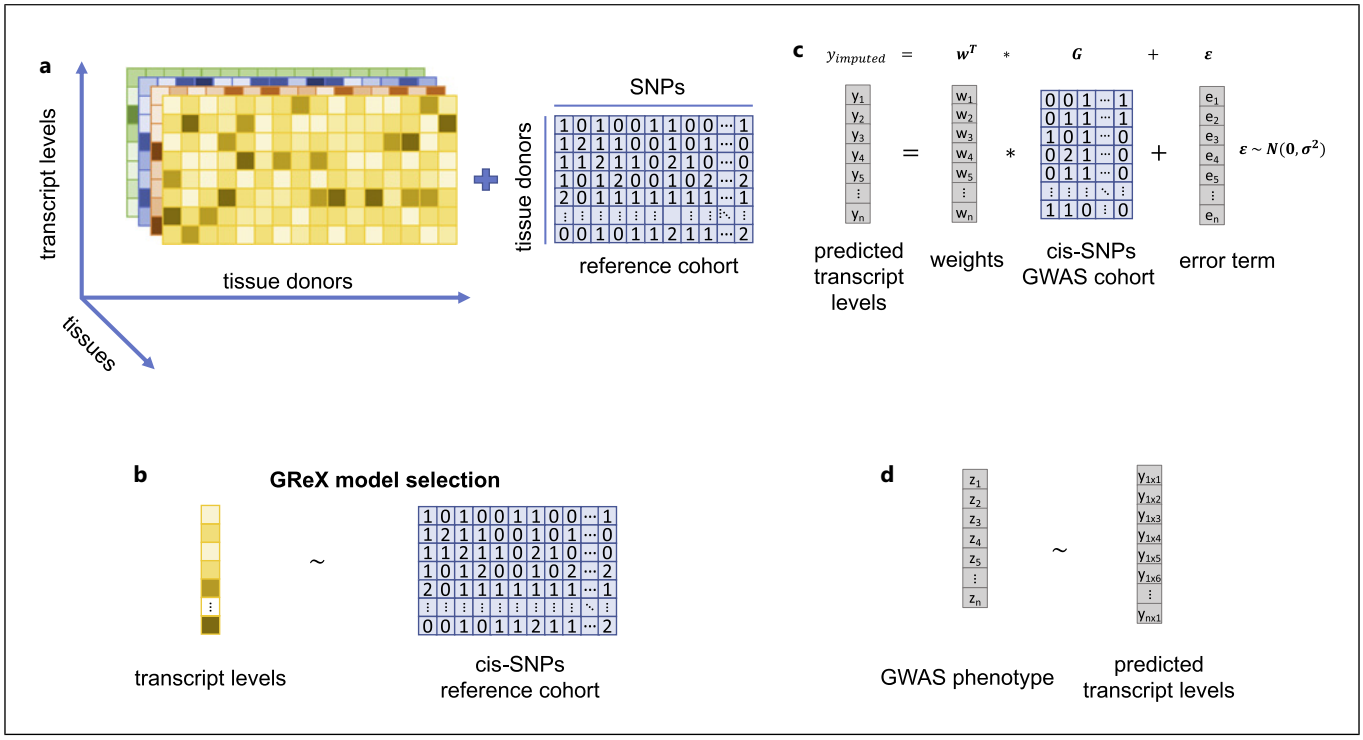

Fig. 5. a Gene expression matrices have transcript-level measurement for each gene by rows. Columns are associated with individual donors for different tissues shown as color coded. The intensity of the color shows the amount of the transcript levels. **b** Transcriptome imputation relies on the eGene statistical prediction model to assign weights to each cis-eQTL. **c** These weights from GReX model will be used for imputing transcript levels. **d** TWAS is association between trait and predicted transcript level.

TWAS/TWAS-FUSION. TWAS [14] and its updated version, TWAS-FUSION, employ LD between SNP pairs and their GWAS Z-scores to impute the transcriptomic Z-score for a gene. To generate these weights for each SNP from the best GReX prediction model, TWAS-FUSION uses the best linear unbiased predictor, Bayesian sparse linear mixed model [62], elastic net, LASSO, or top SNP. The pre-computed weights for prediction models are publicly available (see data resources listed in Data Availability Statement).

JEPEG/JEPEGMIX

JEPEG [34] is another TWAS-like gene-level method that modeled several classes of functional effects of SNP (eQTL, transcription factor binding, or miRNA associated) on the phenotype. Its extension to cosmopolitan cohorts, JEPEGMIX, became a pure gene-level TWAS tool later [35]. The latest iteration of the tool, JEPEGMIX2-P, added the much-needed pathway-level inference for TWAS analysis. To the best of our knowledge, this method is still the only tool that makes pathway inferences by directly estimating the LD of gene-level TWAS statistics (even in multiple ancestry cohorts).

SMR

Since the causal path from genetic variants to trait is posited to be mediated by gene expression, Mendelian randomization (MR) methods are very effective in estimating the effect of gene expression on the trait of interest. For statistical inference, MR tools need to estimate the effect of genetic variants (instrumental variables) on (i) gene expression (exposure) and (ii) trait (outcome). Summary Mendelian randomization (SMR) extended the MR framework to infer gene expression-trait association using only GWAS summary statistics [36]. Its main advantage is that it does not require pre-computed predictive weights for transcripts/protein abundances, unlike many other TWAS methods (S-PrediXcan, TWAS-FUSION, JEPEGMIX, etc.). SMR requires inputs of summary statistics from GWAS and eQTL studies, which are assumed to be independent cohorts.

Multiple Tissue TWAS Methods

Multiple tissue integration TWAS methods aggregate TWAS effect estimates from different tissues. Some of these methods are updated versions of primary TWAS methods. It is well established that different types of tissues share eQTLs, especially the cis-eQTLs that most TWAS methods use [63]. Thus, prediction models can be improved by combining eQTL information across tissues. Because reference eQTL data were limited at the beginning, multi-tissue approaches were developed to improve prediction power for under-sampled tissues [32]. MultiXcan [32], UTMOST [39], TisCoMM/TisCoMM-S2 [42], MR-JTI [45], and InTACT [46] are examples of this group of TWAS methods. We briefly present some of these.

Among these methods, UTMOST [39] uses a generalized Berk Johns method to combine the joint tissue covariance structure and the single tissue TWAS effects. MultiXcan jointly analyzes PrediXcan effect estimates from all or multiple tissues [32]. MR-JTI is a multi-tissue integration TWAS that uses MR for causal inference. MR-JTI has been shown to outperform UTMOST and PrediXcan [45] (see online suppl. Table 1 for more details on the TWAS methods; for all online suppl. material, see https://doi.org/10.1159/000530223).

TWAS Fine-Mapping Methods

Similar to overlapping significant GWAS signals at the variant level, TWAS yields multiple gene signals at a single locus due to LD. To get closer to the true causal signals, and better prioritize genes, researchers use estimated LD between signals to probabilistically fine-map TWAS findings. Two such methods are TWAS-FOCUS [48] and FOGS [49] (Table 1). There is an updated version of TWAS-FOCUS, which implements multi-ancestry in TWAS fine-mapping (MA-FOCUS) [64].

Also, SMR provides the heterogeneity in dependent instruments (HEIDI) test. It is a goodness of fit for the causal model (smaller HEIDI p values indicate a poor fit for the causality assumption). Thus, unlike commonly used primary TWAS methods, which do not test for causality, SMR can be directly used to fine-map TWAS signals.

Fine-Mapping Based on the Colocalization of eQTL and GWAS Signals

To eliminate the confounding due to LD and improve the detection of more likely causal loci, researchers have also developed methods for colocalization [53]. They use eQTL and trait GWASs to probabilistically assess whether variants are likely causal for both the molecular mediator (expression of the gene under investigation for TWAS) and the outcome. This approach can also be applied to proteomics by substituting pQTL/protein abundance for eQTL/gene expression. Among the most well-known colocalization methods, we can list Coloc [52], eCAVIAR [53], and enloc [54].

Transcriptome Reference Repositories with Human Tissues and Cell Lines

For computing the weights for SNPs used in gene expression prediction models in various tissues, analytical tools require a reference transcriptome from public repositories (Table 2). These repositories have both genomic (DNA sequence or genotype) and transcriptomic (RNA sequence or array) information either at the bulk tissue [8] or single-cell level [65] from the same cohort (Fig. 6).

Among various repositories, the most comprehensive is GTEx [8]. In version 8, GTEx contains data from 52 different human tissues and 2 cell lines from 948 individuals. Among all, 67.15% of these individuals are female, 84.6% are of European and 12.9% are of African ancestry [8, 72–74]. Genotype data were obtained from 838 subjects. Other relatively larger eQTL repositories are PsychENCODE [60], eQTLGen [21], CAGE [66], Geuvadis [7], ALSPAC [67], MuTHER [68], DGN (case-control study) [70] and a meta-analysis of brain tissue eQTLs (Brain-eMeta) [71].

PsychENCODE combines data from participating research groups under one data repository for ease of data sharing [60, 69]. Their web portal includes RNA sequencing data from postmortem bulk brain tissue and single cells, as well as genotype data.

eQTLgen is the largest meta-analysis of 37 (peripheral) blood eQTL studies [21]. Cis-eQTL detection yielded 16,987 eGenes, which were replicated in an independent eQTL mapping study on single cells from whole blood (monocytes, natural killer cells, CD8+ T cells, and lymphocytes). The majority of cis-eQTL (92%) were not distant (<100 KB) from the transcription start or end site of eGenes. Some of the more distal cis-eQTLs were located in Hi-C regions.

The Consortium for the Architecture of Gene Expression (CAGE) study includes highly related individuals in their analysis [66]. They meta-analyzed the gene expression data on whole blood from five different study cohorts (BSGS, coronary artery disease [CAD] [75], Emory-Georgia tech center for health discovery and well-being [CHDWB] [76], Estonian genome center of the University of Tartu [EGCUT] [77], Morocco [78]). These cohorts were predominantly of European ancestry, except the Moroccan cohort. The CAGE study is included in the eQTLgen consortia study.

Table 2. Biorepositories have human tissue level or single-cell RNA transcript levels and genotype data

Repository or study	Samples	No. of donors	Reference	Genotype	Transcript	No. of distinct eGenes or probes
Population based						
eQTLGen	Peripheral blood	31,684	Võsa et al., [21] 2021	SNP array	Expression array and RNA-seq	19,942
CAGE	Peripheral blood	2,765	Lloyd-Jones et al., [66] 2017	SNP array	Expression array	36,778
ALSPAC	159 placenta tissues	869	Bryois et al., [67] 2014	SNP array	Bulk tissue RNA-seq	3,534
GTEx version 8	47 different tissues and 2 cell lines	838	Aguet et al., [8] 2020	WGS	Bulk tissue RNAseq	143 trans/ 23,268 cis
Geuvadis	Lymphoblastoid cell lines	462*	Lappalainen et al., 2013	SNP array	Bulk cell RNAseq	13,703
MuTHER	Subcutaneous adipose/ lymphoblastoid cells/skin	166/156/160	Nica et al., [68] 2011	SNP array	Whole genome expression array	1,822
Case-control design						
PEC	Bulk and single cell	1,387**	Wang et al., [69] 2018	SNP array and WGS	Single cell and bulk tissue RNA-seq	15,626
DGN	Peripheral blood	463 cases with MDD/459 controls	Battle et al., [70] 2014	SNP array	Bulk tissue RNAseq	8,208
Brain-eMeta version 1	Brain tissues	1,194***	Qi et al., [71] 2018	SNP array/ sequencing	Expression array and RNA-seq	28,538

These bio-specimens are sometimes obtained from multiple discovery cohorts of donors. CAGE, Consortium for the Architecture of Gene Expression; GTEx, genotype tissue expression; GEUVADIS, Genetic European in Health and Disease; ALSPAC, Avon Longitudinal Study of Parents and Children; MuTHER, Multiple tissue Human expression resource; PEC, PsychENCODE; DGN, Depression Genes and Networks study. *462 samples in eQTL discovery. **eQTL analysis is a subset of donors from GTEx, BrainSpan, CommonMind, CommonMind-HBCC, BrainGVEX, Lieberman schizophrenia control, BipSeq, UCLA-ASD, and Yale-ASD studies. ***Effective sample size of eQTL meta-analysis based on a subset of donors from GTEx brain, CMC, and ROSMAP.

One of the earliest large-scale blood eQTL mappings in humans was conducted by the Geuvadis consortium [7]. It contains RNA-seq data for lymphoblastoid cell lines. Geuvadis includes subjects of Yoruba and European descent who participated in the 1000 Genomes project phase 1.

For the brain, brain-eMeta (BrainMeta version 1) meta-analyzed several eQTL studies (GTEx brain tissues, Common Mind Consortium [CMC] [79], GTEx [74], and ROSMAP [80]). (Meta-analysis of cis-eQTL in correlated samples is the method for combining effect estimates [71]). Aside from the study by Siebert et al. [81], it is one of the largest publicly available brain eQTL datasets. In its latest update (BrainMeta version 2), the sample size increased to 2,865 [82] (see online suppl. Table 2 for more detail).

The remaining eQTL repositories, which we briefly describe below, are smaller in terms of the number of donors. Unlike most other studies, the eQTL discovery cohort from ALSPAC reported CNVs as eQTLs [67].

In the multiple tissue human expression resource (MuTHER) study, authors conducted eQTL mapping on cis-regulated gene expression in two different tissues (skin, fat) and 1 cell line (lymphoblastoid) sampled from female twins (age above 40 years) with European descent [68]. They also showed that eQTLs for 12.5% of eGenes are shared among all tissues and for 71.44% of eGenes are relevant to a single tissue.

The Depression Genes and Networks study (DGN) cohort includes case-control individuals with major depression of European ancestry [70]. The age range is between 21 and 60 years. The DGN study also identified the splicing QTLs and allelic expression.

Proteomics: PWASs

DNA is transcribed into mRNA, and then mRNA is translated into protein [83]. Therefore, proteins could be considered molecular mediators between genes and traits.

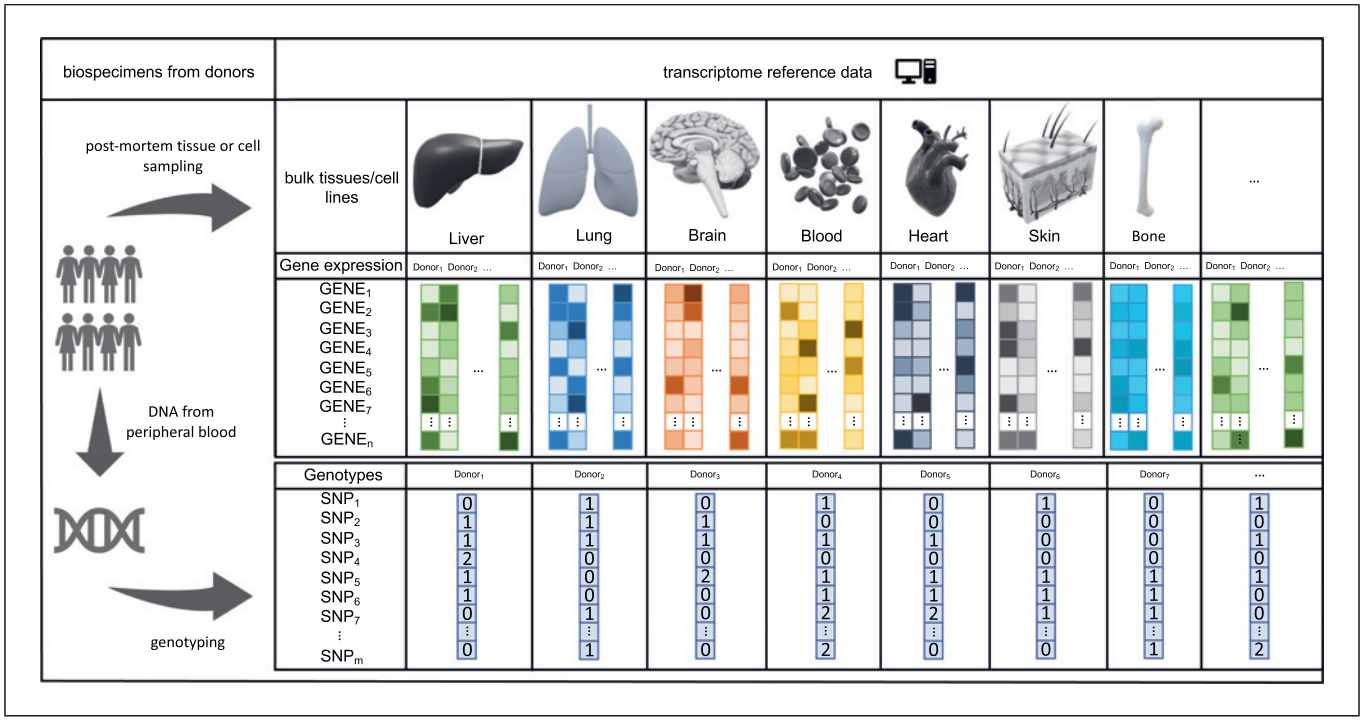

Fig. 6. Reference transcriptome biorepositories have large data on both genotype and the transcript-level measurements from the postmortem tissue donors.

Thus, identifying a group of fluid/tissue-specific proteins in a molecular pathway associated with the risk for psychiatric disorders (PDs) could help researchers better understand the etiology and identify possible drug targets for treatment. The levels of these proteins in the blood serum would also serve as possible biomarkers for the prognosis or diagnosis of the PDs.

Although the biochemical pathway from genes to proteins is causal, the correlation between RNA and protein abundance is not particularly strong in humans [84] and yeast [85]. For instance, only 40% of the variance in the transcriptome was explained by the variance in the proteome of mammalian cells [86]. In a study of human cell lines, 67% of the explained variance in protein abundance was attributable to transcript levels and sequence elements [87]. Similarly, analysis of protein abundance in human lymphoblastoid cell lines showed that many pQTLs are not eQTLs for the same gene [88]. These findings suggest that the regulatory mechanisms in transcript levels and protein abundance are different. Thus, proteome studies provide additional information to the findings at transcriptome level.

Testing the mediator effect of pQTL on traits in PWAS is plausible because (i) it has been shown that the human proteome is under genetic control [89] and (ii) GWAS of protein abundance reported that the common variation has heritable effects on the plasma protein [90]. So, PWAS could employ the same methodological framework used for TWAS. Recent PWAS analyses suggest that 70% of the significant gene-level associations do not yield TWAS signals [91]. It is reasonable to conclude that PWAS complements TWAS findings.

PWAS Methods

Most PWAS methods mimic TWAS methods [13, 14, 36]. The main difference between the TWAS and PWAS paradigms is that the estimation of SNP weights in PWAS is now based on a reference proteome study. With more blood and brain pQTL reference data available, researchers conducted PWAS on the PD GWAS [91, 92]. Unfortunately, there are still very few pre-computed weights to conduct PWAS. Predictive weights for PWAS analysis are available for the brain [92] and blood [22]. Researchers might apply MR methods, such as SMR, that do not require computing weights.

Reference Proteome Repositories

Reference proteome datasets are large-scale proteome studies applying various protein assay methods that detect and quantify proteins in human solid tissues or

Table 3. Selected large-scale pQTL discovery studies to identify associations between protein abundance and genotype

Cohorts	Samples	No. of donors	Reference	Genotype	Proteome technology	No. of proteins
Population-based study						
KORA/QMDiab studies	Blood plasma	997/338	Suhre et al., [96] 2017	SNP array	SOMAscan	1,124
INTERVAL study	Blood plasma	3,301	Sun et al., [4] 2018	SNP array	SOMAscan	2,994
AGES cohort 1	Blood serum	5,457	Emilsson et al., [30] 2018	SNP array	SOMAscan	4,137
ROS/MAP and Banner Sun Health	dPFC	330 and 149	Robins et al., [11] 2021	WGS	TMT isobaring labeling MS	7,376
AGES cohort 2	Blood serum	5,368	Gudjonsson et al., [97] 2021	SNP array	SOMAscan	4,782
AGES Reykjavik cohort	Blood serum	5,343	Emilsson et al., [98] 2022	SNP array	SOMAscan	1,942
Fenland cohort	Blood plasma	10,708	Pietzner et al., [20] 2021	SNP array	SOMAscan	4,775
ARIC	Blood serum	9,084*	Zhang et al., [22] 2022	SNP array	SOMAscan	871**
deCODE	Blood serum	35,559	Ferkingstad et al., [99] 2021	SNP array/WGS	SOMAscan	4,719
Case-control studies						
Cases with PDs and control***	CSF	133	Sasayama et al., [12] 2017	SNP array	SOMAscan	1,126
Individuals with neurological disorders	CSF/blood plasma/ parietal lobe cortex	835/ 529/380	Yang et al., [100] 2021	SNP array	SOMAscan	713/931/1,079

We grouped the studies in two categories: population based and case-control studies. ARIC, Atherosclerosis Risk in Communities; SOMAScan, slow off-rate modified aptamer scan; KORA, Cooperative Health Research in the Augsburg Region; QMDiab, The Qatar Metabolomics Study on Diabetes; CSF, cerebro-spinal fluid; ROS, Religious Orders Study; MAP, Memory and Aging Project; dPFC, dorsolateral prefrontal cortex; TMT, tandem mass tag; MS, mass spectrometry; AGES, The Age, Gene/ Environment Susceptibility-Reykjavik Study. *After QC steps, 1,871 and 7,213 individuals with African American and European ancestry, respectively. **Number of proteins assayed with both Olink and SOMAscan v4. ***Sixty-six percent of participants have psychiatric condition, 28 schizophrenia, 15 bipolar disorder, and 45 major depressive disorder.

plasma. Mass spectrophotometry methods are considered the gold standard for measuring protein abundance and identifying proteins in humans [93]. Other methods include slow off-rate modified aptamer [94] and Olink [95] assay technology. The majority of the large-scale pQTL mapping studies implement the aptamer-based technology to assay proteins (Table 3) (see online suppl. Table 3 for more detail).

The development in proteome technology allowed high-throughput analysis of proteins in large cohorts [4, 18]. Initially, the majority of studies used blood serum to identify pQTL [4, 17, 28, 97]. However, assays on blood specimens might not be fully representative of other organ tissues, such as the brain. So, after the initial pQTL discoveries in blood, researchers also investigated cerebrospinal fluid (CSF) [12, 100], postmortem tissues [101], and the brain [11].

In addition to the abovementioned studies, there are other smaller pQTL studies on different tissues, such as the liver [9] and human induced pluripotent cell lines [10]. For example, the GTEx consortium analyzed the human proteome in 32 different tissues [101]. However, the sample size (n = 14) is rather low compared to the large-scale serum proteome analysis [4, 97].

The mapping of pQTLs in three different tissues showed that tissue-shared pQTLs are more likely to be cis-pQTLs [100]. Thus, it is likely that some blood pQTLs, especially cis-pQTLs, are shared with brain tissues, supporting the idea of searching for brain biomarkers of PDs in plasma. Also, it is known that microglia in the central nervous system are immune cells and there are some shared pQTLs between the CNS and CSF [102]. For these reasons and due to their physical proximity, CSF might be viewed as a better pQTL proxy than blood for the under-sampled brain tissue.

Discussion

GWAS identifies a large number of loci associated with the risk for PDs in well-powered studies. However, these findings alone do not provide enough information on the molecular underpinnings of PDs. We better understand this molecular etiology by integrating GWAS trait information with molecular endophenotypes [103], which mediate the association between the genome and traits. Here, we focus on two well-known mediators, i.e., transcript and protein levels. Although proteomes and transcriptomes are not investigated in most GWAS cohorts, transcript and protein abundance can be predicted from genotype or GWAS summary statistics using the TWAS and PWAS paradigms. In this review, we summarized (i) available TWAS/PWAS methods and (ii) large publicly available eQTL and pQTL repositories that serve as the reference transcriptome/proteome for the TWAS/PWAS analysis. As a guide for researchers, we further categorized TWAS/PWAS methods by their intended end use, i.e., (i) primary, (ii) fine-mapping, and (iii) pathway-aware tools (Table 1). This work would help researchers choose the TWAS/PWAS methods that are better suited for their analyses.

TWAS/PWAS are powerful paradigms because they model molecular mediators and add biological context to the GWAS findings. In addition to this, TWAS/PWAS also greatly reduce the burden of multiple testing. For instance, TWAS/PWAS only pay the multiple testing penalty for the number (~20 K) of expressed genes/proteins and not the number (>5 M) of variants [14]. Reducing the number of tests would likely increase the power of detecting signals.

Researchers could uncover a larger number of TWAS/PWAS signals whenever large transcriptome and proteome reference datasets are available for numerous (and relevant) tissues. When such datasets are smaller, as in brain-related tissues, a meta-analysis of TWAS signals across tissues might help increase the discovery of signals. This approach yielded both previously identified and novel loci for nicotine addiction when meta-analyzing TWAS statistics from 13 different brain regions [104].

There are many examples of genes that were prioritized by TWAS as likely causal for PDs. For instance, *ADH1B,* which codes for alcohol metabolizing enzyme, was found to be significant in problematic alcohol use S-MultiXcan analysis using all GTEx tissues [105]. In the recent PGC3 SCZ GWAS, one of the genes prioritized by brain TWAS was *RERE*. It was discussed as a putatively causal gene that is biologically relevant to the SCZ pathology [106]. Functional genetic studies assessing effects of SCZ risk variants on chromatin accessibility have shown that some of these variants have a cis-eQTL effect on *RERE* [107]. For BIP, twenty-two genes were included in the credible gene set based on TWAS-FUSION and FOCUS results. The most significant gene was *DCLK3* in brain TWAS (using reference eQTL PsychENCODE brain as transcriptomic reference) [108]. A recent major depression study conducted brain PWAS [109] using TWAS-FUSION and SMR [110, 111]. This study identified 24 proteins that were reported as significant by SMR, HEIDI, and FUSION analyses. The strongest signal was for *B3GLCT* that was thought to play a role in synaptogenesis [109].

Although reference transcriptome and proteome datasets are still underpowered for certain tissues, e.g., the brain, we expect a significant increase in their sample size and cell type specificity over time. For instance, consecutive GTEx iterations substantially increased the reference transcriptome size [8, 72–74]. There are efforts to discover eQTL at the single cell level that serve as transcriptome reference data [65, 112].

Ancestrally diverse reference cohorts will help scientists (i) identify population-specific eQTLs in transcriptome data [8] and (ii) increase cross-population generalizability of eQTLs [113]. Also, ancestral diversity will improve the detection of TWAS/PWAS signals because this results in better accuracy and specificity when imputing the transcriptome and proteome.

Progress in mapping eQTL and pQTL in non-European ancestral populations is still in its early phases. In Atherosclerosis Risk in Communities (ARIC), the African American cohort was part of the human serum pQTL mapping study [22]. In another study, researchers developed a method to map eQTL in a cohort of individuals with non-European ancestry, including individuals of African American and Hispanic ancestry [114]. The same research group also mapped pQTL in cohorts from populations of African American, Chinese, and Hispanic ancestry [115]. These studies are still limited in terms of sample size.

Limitations and Issues Associated with the TWAS and PWAS Methods

1. Primary TWAS/PWAS methods only test for association, but not causality. For improved fine-mapping, such methods need to be followed by fine-mapping/colocalization analyses.
2. Primary non-MR TWAS methods may yield false-positive rates [116].
3. The brain eQTL and brain pQTL datasets include some individuals with certain neurological and PDs.
4. Since the sample size for brain pQTL reference datasets is smaller than that for the blood pQTL, PWAS signal detection and resolution suffers from low sample size.

5. TWAS methods, including SMR, cannot eliminate horizontal pleiotropy.
6. Reference eQTL/pQTL LD data that are used in TWAS/PWAS might not be representative of the LD from the GWAS cohort.
7. The overlap between GWAS and variants in TWAS prediction model can be limited due to the fact that the two approaches target different types of variants. This might limit TWAS's detection power [117].

Conclusion

We believe that applying TWAS and PWAS tools to GWAS summary statistics greatly helps researchers identify risk genes for PDs. Applied researchers might find it more effective to use SMR as the default method for TWAS and PWAS because it eliminates the need for precomputing SNP weights every time the reference cohort or prediction method updates. Moreover, SMR can assist in the process of signal fine-mapping by providing the HEIDI test to identify likely causal risk loci.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Funding Sources

Research in this work was funded by AA022537 (Huseyin Gedik, Brian P. Riley, and Silviu-Alin Bacanu), R01MH118239, and R01DA052453 (Vladimir I. Vladimirov and Silviu-Alin Bacanu) and MH125938, MH126358, NARSAD grant 28632 (Roseann E. Peterson).

Author Contributions

The review was conceptualized and written by Huseyin Gedik, Roseann E. Peterson, Brien P. Riley, Vladimir I. Vladimirov, and Silviu-Alin Bacanu. Throughout all the steps, Vladimir I. Vladimirov and Silviu-Alin Bacanu supervised the investigation.

Data Availability Statement

Some of the datasets referred to in this review are publicly available. The URLs can be found in the supplementary tables and they can be downloaded from HG's GitHub: https://github.com/huseyingedik/A-review-of-integrative-post-GWAS-analyses-relevant-to-psychiatric-disorders-imputing-transcriptome/raw/main/Supplementary_tables.xlsx. PrediXcan predictive weight database: https://predictdb.org/ (accessed on 10/20/2022). TWAS-FUSION prediction model database: http://gusevlab.org/projects/fusion/ (accessed on 11/02/2022). mashr: https://github.com/stephenslab/mashr (accessed on 11/05/2022). Further inquiries can be directed to the corresponding author.

References

1 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 Sep 7;337(6099):1190–5.
2 Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. Genome Res. 2012 Sep 1;22(9):1748–59.
3 Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to Be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. 2010 Apr 1;6(4):e1000888.
4 Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, et al. Genomic atlas of the human plasma proteome. Nature. 2018 Jun;558(7708):73–9.
5 Baird DA, Liu JZ, Zheng J, Sieberts SK, Perumal T, Elsworth B, et al. Identifying drug targets for neurological and psychiatric disease via genetics and the brain transcriptome. PLoS Genet. 2021 Jan 8;17(1):e1009224.
6 Folkersen L, Gustafsson S, Wang Q, Hansen DH, Hedman ÅK, Schork A, et al. Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. Nat Metab. 2020 Oct;2(10):1135–48.
7 Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PAC, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013 Sep;501(7468):506–11.
8 Aguet F, Anand S, Ardlie KG, Gabriel S, Getz GA, Graubert A, et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020 Sep 11;369(6509):1318–30.
9 He B, Shi J, Wang X, Jiang H, Zhu HJ. Genome-wide pQTL analysis of protein expression regulatory networks in the human liver. BMC Biol. 2020 Aug 10;18(1):97.
10 Mirauta BA, Seaton DD, Bensaddek D, Brenes A, Bonder MJ, Kilpinen H, et al. Population-scale proteome variation in human induced pluripotent stem cells. Elife. 2020 Aug 10;9:e57390.
11 Robins C, Liu Y, Fan W, Duong DM, Meigs J, Harerimana NV, et al. Genetic control of the human brain proteome. Am J Hum Genet. 2021 Mar 4;108(3):400–10.
12 Sasayama D, Hattori K, Ogawa S, Yokota Y, Matsumura R, Teraishi T, et al. Genome-wide quantitative trait loci mapping of the human cerebrospinal fluid proteome. Hum Mol Genet. 2017 Jan 1;26(1):44–51.
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 Sep;47(9):1091–8.
14 Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BWJH, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet. 2016 Mar;48(3):245–52.
15 Mancuso N, Shi H, Goddard P, Kichaev G, Gusev A, Pasaniuc B. Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. Am J Hum Genet. 2017 Mar 2;100(3):473–87.

16 Wainberg M, Sinnott-Armstrong N, Mancuso N, Barbeira AN, Knowles DA, Golan D, et al. Opportunities and challenges for transcriptome-wide association studies. Nat Genet. 2019 Apr;51(4):592–9.
17 Yao C, Chen G, Song C, Keefe J, Mendelson M, Huan T, et al. Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. Nat Commun. 2018 Aug 15;9(1):3268.
18 Gold L, Walker JJ, Wilcox SK, Williams S. Advances in human proteomics at high scale with the SOMAscan proteomics platform. N Biotechnol. 2012 Jun 15;29(5):543–9.
19 Wingo AP, Liu Y, Gerasimov ES, Gockley J, Logsdon BA, Duong DM, et al. Integrating human brain proteomes with genome-wide association data implicates new proteins in Alzheimer's disease pathogenesis. Nat Genet. 2021 Feb;53(2):143–6.
20 Pietzner M, Wheeler E, Carrasco-Zanini J, Cortes A, Koprulu M, Wörheide MA, et al. Mapping the proteo-genomic convergence of human diseases. Science. 2021 Oct 14;374(6569):eabj1541.
21 Võsa U, Claringbould A, Westra HJ, Bonder MJ, Deelen P, Zeng B, et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat Genet. 2021 Sep;53(9):1300–10.
22 Zhang J, Dutta D, Köttgen A, Tin A, Schlosser P, Grams ME, et al. Plasma proteome analyses in individuals of European and African ancestry identify cis-pQTLs and models for proteome-wide association studies. Nat Genet. 2022 May 2;54(5):593–602.
23 MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2017 Jan 4;45(D1):D896–901.
24 Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS discovery: biology, function, and translation. Am J Hum Genet. 2017 Jul 6;101(1):5–22.
25 Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, et al. Linkage disequilibrium in the human genome. Nature. 2001 May;411(6834):199–204.
26 Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. High-resolution haplotype structure in the human genome. Nat Genet. 2001 Oct;29(2):229–32.
27 Ardlie KG, Kruglyak L, Seielstad M. Patterns of linkage disequilibrium in the human genome. Nat Rev Genet. 2002 Apr;3(4):299–309.
28 Emilsson V, Ilkov M, Lamb JR, Finkel N, Gudmundsson EF, Pitts R, et al. Co-regulatory networks of human serum proteins link genetics to disease. Science. 2018 Aug 24;361(6404):769–73.
29 Dermitzakis ET. From gene expression to disease risk. Nat Genet. 2008 May 1;40(5):492–3.
30 Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, et al. Genetics of gene expression and its effect on disease. Nature. 2008 Mar 1;452(7186):423–8.
31 Zheng J, Haberland V, Baird D, Walker V, Haycock PC, Hurle MR, et al. Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. Nat Genet. 2020 Oct;52(10):1122–31.
32 Barbeira AN, Pividori M, Zheng J, Wheeler HE, Nicolae DL, Im HK. Integrating predicted transcriptome from multiple tissues improves association detection. PLoS Genet. 2019 Jan 22;15(1):e1007889.
33 Feng H, Mancuso N, Gusev A, Majumdar A, Major M, Pasaniuc B, et al. Leveraging expression from multiple tissues using sparse canonical correlation analysis and aggregate tests improves the power of transcriptome-wide association studies. PLoS Genet. 2021 Apr 8;17(4):e1008973.
34 Lee D, Williamson VS, Bigdeli TB, Riley BP, Fanous AH, Vladimirov VI, et al. JEPEG: a summary statistics based tool for gene-level joint testing of functional variants. Bioinformatics. 2015 Apr 15;31(8):1176–82.
35 Lee D, Williamson VS, Bigdeli TB, Riley BP, Webb BT, Fanous AH, et al. JEPEGMIX: gene-level joint analysis of functional SNPs in cosmopolitan cohorts. Bioinformatics. 2016 Jan 15;32(2):295–7.
36 Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet. 2016 May;48(5):481–7.
37 Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, Torres JM, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat Commun. 2018 May 8;9(1):1825.
38 Yang C, Wan X, Lin X, Chen M, Zhou X, Liu J. CoMM: a collaborative mixed model to dissecting genetic contributions to complex traits by leveraging regulatory information. Bioinformatics. 2019 May 15;35(10):1644–52.
39 Hu Y, Li M, Lu Q, Weng H, Wang J, Zekavat SM, et al. A statistical framework for cross-tissue transcriptome-wide association analysis. Nat Genet. 2019 Mar;51(3):568–76.
40 Nagpal S, Meng X, Epstein MP, Tsoi LC, Patrick M, Gibson G, et al. TIGAR: an improved bayesian tool for transcriptomic data imputation enhances gene mapping of complex traits. Am J Hum Genet. 2019 Aug 1;105(2):258–66.
41 Yang Y, Shi X, Jiao Y, Huang J, Chen M, Zhou X, et al. CoMM-S2: a collaborative mixed model using summary statistics in transcriptome-wide association studies. Bioinformatics. 2020 Apr 1;36(7):2009–16.
42 Shi X, Chai X, Yang Y, Cheng Q, Jiao Y, Chen H, et al. A tissue-specific collaborative mixed model for jointly analyzing multiple tissues in transcriptome-wide association studies. Nucleic Acids Res. 2020 Nov 4;48(19):e109.
43 Yuan Z, Zhu H, Zeng P, Yang S, Sun S, Yang C, et al. Testing and controlling for horizontal pleiotropy with probabilistic Mendelian randomization in transcriptome-wide association studies. Nat Commun. 2020 Jul 31;11(1):3861.
44 Luningham JM, Chen J, Tang S, De Jager PL, Bennett DA, Buchman AS, et al. Bayesian genome-wide TWAS method to leverage both cis- and trans-eQTL information through summary statistics. Am J Hum Genet. 2020 Oct 1;107(4):714–26.
45 Zhou D, Jiang Y, Zhong X, Cox NJ, Liu C, Gamazon ER. A unified framework for joint-tissue transcriptome-wide association and Mendelian randomization analysis. Nat Genet. 2020 Nov;52(11):1239–46.
46 Bae YE, Wu L, Wu C. InTACT: an adaptive and powerful framework for joint-tissue transcriptome-wide association studies. Genet Epidemiol. 2021;45(8):848–59.
47 Tang S, Buchman AS, De Jager PL, Bennett DA, Epstein MP, Yang J. Novel Variance-Component TWAS method for studying complex human diseases with applications to Alzheimer's dementia. PLoS Genet. 2021 Apr 2;17(4):e1009482.
48 Mancuso N, Freund MK, Johnson R, Shi H, Kichaev G, Gusev A, et al. Probabilistic fine-mapping of transcriptome-wide association studies. Nat Genet. 2019 Apr;51(4):675–82.
49 Wu C, Pan W. A powerful fine-mapping method for transcriptome-wide association studies. Hum Genet. 2020 Feb;139(2):199–213.
50 Chatzinakos C, Georgiadis F, Lee D, Cai N, Vladimirov VI, Docherty A, et al. TWAS pathway method greatly enhances the number of leads for uncovering the molecular underpinnings of psychiatric disorders. Am J Med Genet B Neuropsychiatr Genet. 2020;183(8):454–63.
51 Pain O, Pocklington AJ, Holmans PA, Bray NJ, O'Brien HE, Hall LS, et al. Novel insight into the etiology of autism spectrum disorder gained by integrating expression data with genome-wide association statistics. Biol Psychiatry. 2019 Aug 15;86(4):265–73.
52 Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 2014 May 15;10(5):e1004383.
53 Hormozdiari F, van de Bunt M, Segrè AV, Li X, Joo JWJ, Bilow M, et al. Colocalization of GWAS and eQTL signals detects target genes. Am J Hum Genet. 2016 Dec 1;99(6):1245–60.
54 Wen X, Pique-Regi R, Luca F. Integrating molecular QTL data into genome-wide genetic association analysis: probabilistic assessment of enrichment and colocalization. PLoS Genet. 2017 Mar 9;13(3):e1006646.

55 Guan Y, Stephens M. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. Ann Appl Stat. 2011 Sep;5(3):1780–815.
56 Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011 Jul 15;89(1):82–93.
57 Pan W, Kim J, Zhang Y, Shen X, Wei P. A powerful and adaptive association test for rare variants. Genetics. 2014 Aug 1;197(4):1081–95.
58 Tibshirani R. Regression shrinkage and selection via the lasso. J Roy Stat Soc B. 1996;58(1):267–88.
59 Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol. 2005;67(2):301–20.
60 Akbarian S, Liu C, Knowles JA, Vaccarino FM, Farnham PJ, et al. PsychENCODE Consortium. The PsychENCODE project. Nat Neurosci. 2015 Dec;18(12):1707–12.
61 Urbut SM, Wang G, Carbonetto P, Stephens M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. Nat Genet. 2019 Jan;51(1):187–95.
62 Zhou X, Carbonetto P, Stephens M. Polygenic modeling with bayesian sparse linear mixed models. PLoS Genet. 2013 Feb 7;9(2):e1003264.
63 Gamazon ER, Segrè AV, van de Bunt M, Wen X, Xi HS, Hormozdiari F, et al. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. Nat Genet. 2018 Jul;50(7):956–67.
64 Lu Z, Gopalan S, Yuan D, Conti DV, Pasaniuc B, Gusev A, et al. Multi-ancestry fine-mapping improves precision to identify causal genes in transcriptome-wide association studies. Am J Hum Genet. 2022 Aug 4;109(8):1388–404.
65 van der Wijst M, de Vries D, Groot H, Trynka G, Hon C, Bonder M, et al. The single-cell eQTLGen consortium. Elife. 2020 Mar 9;9:e52155.
66 Lloyd-Jones LR, Holloway A, McRae A, Yang J, Small K, Zhao J, et al. The genetic architecture of gene expression in peripheral blood. Am J Hum Genet. 2017 Feb 2;100(2):371–37.
67 Bryois J, Buil A, Evans DM, Kemp JP, Montgomery SB, Conrad DF, et al. Cis and trans effects of human genomic variants on gene expression. PLoS Genet. 2014 Jul 10;10(7):e1004461.
68 Nica AC, Parts L, Glass D, Nisbet J, Barrett A, Sekowska M, et al. The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. PLoS Genet. 2011 Feb 3;7(2):e1002003.
69 Wang D, Liu S, Warrell J, Won H, Shi X, Navarro FCP, et al. Comprehensive functional genomic resource and integrative model for the human brain. Science. 2018 Dec 14;362(6420):eaat8464.
70 Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. 2014 Jan 1;24(1):14–24.
71 Qi T, Wu Y, Zeng J, Zhang F, Xue A, Jiang L, et al. Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. Nat Commun. 2018 Jun 11;9(1):2282.
72 Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The genotype-tissue expression (GTEx) project. Nat Genet. 2013 Jun;45(6):580–5.
73 Ardlie KG, Deluca DS, Segrè AV, Sullivan TJ, Young TR, GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015 May 8;348(6235):648–60.
74 Brown AA, Castel SE, Davis JR, He Y, Jo B, GTEx Consortium. Genetic effects on gene expression across human tissues. Nature. 2017 Oct;550(7675):204–13.
75 Kim J, Ghasemzadeh N, Eapen DJ, Chung NC, Storey JD, Quyyumi AA, et al. Gene expression profiles associated with acute myocardial infarction and risk of cardiovascular death. Genome Med. 2014 Dec;6(5):40–13.
76 Preininger M, Arafat D, Kim J, Nath AP, Idaghdour Y, Brigham KL, et al. Blood-informative transcripts define nine common axes of peripheral blood gene expression. PLoS Genet. 2013 Mar 14;9(3):e1003362.
77 Leitsalu L, Haller T, Esko T, Tammesoo ML, Alavere H, Snieder H, et al. Cohort profile: Estonian biobank of the Estonian genome center, university of Tartu. Int J Epidemiol. 2015 Aug 1;44(4):1137–47.
78 Idaghdour Y, Czika W, Shianna KV, Lee SH, Visscher PM, Martin HC, et al. Geographical genomics of human leukocyte gene expression variation in southern Morocco. Nat Genet. 2010 Jan;42(1):62–7.
79 Fromer M, Roussos P, Sieberts SK, Johnson JS, Kavanagh DH, Perumal TM, et al. Gene expression elucidates functional impact of polygenic risk for schizophrenia. Nat Neurosci. 2016 Nov;19(11):1442–53.
80 Ng B, White CC, Klein HU, Sieberts SK, McCabe C, Patrick E, et al. An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. Nat Neurosci. 2017 Oct;20(10):1418–26.
81 Sieberts SK, Perumal TM, Carrasquillo MM, Allen M, Reddy JS, Hoffman GE, et al. Large eQTL meta-analysis reveals differing patterns between cerebral cortical and cerebellar brain regions. Sci Data. 2020 Oct 12;7(1):340.
82 Qi T, Wu Y, Fang H, Zhang F, Liu S, Zeng J, et al. Genetic control of RNA splicing and its distinct role in complex trait variation. Nat Genet. 2022 Aug 18;54:1355–63.
83 Crick F. Central dogma of molecular biology. Nature. 1970 Aug;227(5258):561–3.
84 de Sousa Abreu R, Penalva LO, Marcotte EM, Vogel C. Global signatures of protein and mRNA expression levels. Mol Biosyst. 2009 Nov 12;5(12):1512–26.
85 Maier T, Güell M, Serrano L. Correlation of mRNA and protein in complex biological samples. FEBS Lett. 2009 Dec 1;583(24):3966–73.
86 Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, et al. Global quantification of mammalian gene expression control. Nature. 2011 May;473(7347):337–42.
87 Vogel C, de Sousa Abreu R, Ko D, Le SY, Shapiro BA, Burns SC, et al. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. Mol Syst Biol. 2010 Jan 1;6(1):400.
88 Wu L, Candille SI, Choi Y, Xie D, Jiang L, Li-Pook-Than J, et al. Variation and genetic control of protein abundance in humans. Nature. 2013 Jul;499(7456):79–82.
89 Melzer D, Perry JRB, Hernandez D, Corsi AM, Stevens K, Rafferty I, et al. A genome-wide association study identifies protein Quantitative Trait Loci (pQTLs). PLoS Genet. 2008 May 9;4(5):e1000072.
90 Lourdusamy A, Newhouse S, Lunnon K, Proitsi P, Powell J, Hodges A, et al. Identification of cis-regulatory variation influencing protein abundance levels in human plasma. Hum Mol Genet. 2012 Aug 15;21(16):3719–26.
91 Liu J, Li X, Luo XJ. Proteome-wide association study provides insights into the genetic component of protein abundance in psychiatric disorders. Biol Psychiatry. 2021 Dec 1;90(11):781–9.
92 Wingo TS, Liu Y, Gerasimov ES, Vattathil SM, Wynne ME, Liu J, et al. Shared mechanisms across the major psychiatric and neurodegenerative diseases. Nat Commun. 2022 Jul 26;13(1):4314.
93 Wilhelm M, Schlegl J, Hahne H, Gholami AM, Lieberenz M, Savitski MM, et al. Mass-spectrometry-based draft of the human proteome. Nature. 2014 May;509(7502):582–7.
94 Gold L, Ayers D, Bertino J, Bock C, Bock A, Brody EN, et al. Aptamer-based multiplexed proteomic technology for biomarker discovery. PLoS One. 2010 Dec 7;5(12):e15004.
95 Assarsson E, Lundberg M, Holmquist G, Björkesten J, Thorsen SB, Ekman D, et al. Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability. PLoS One. 2014 Apr 22;9(4):e95192.
96 Suhre K, Arnold M, Bhagwat AM, Cotton RJ, Engelke R, Raffler J, et al. Connecting genetic risk to disease end points through the human blood plasma proteome. Nat Commun. 2017 Feb 27;8(1):14357.

97 Gudjonsson A, Gudmundsdottir V, Axelsson GT, Gudmundsson EF, Jonsson BG, Launer LJ, et al. A genome-wide association study of serum proteins reveals shared loci with common diseases. Nat Commun. 2022 Jan 25;13(1):480.
98 Emilsson V, Gudmundsdottir V, Gudjonsson A, Jonmundsson T, Jonsson BG, Karim MA, et al. Coding and regulatory variants are associated with serum protein levels and disease. Nat Commun. 2022 Jan 25;13(1):481.
99 Ferkingstad E, Sulem P, Atlason BA, Sveinbjornsson G, Magnusson MI, Styrmisdottir EL, et al. Large-scale integration of the plasma proteome with genetics and disease. Nat Genet. 2021 Dec 1;53(12):1712–21.
100 Yang C, Farias FHG, Ibanez L, Suhy A, Sadler B, Fernandez MV, et al. Genomic atlas of the proteome from brain, CSF and plasma prioritizes proteins implicated in neurological disorders. Nat Neurosci. 2021 Sep;24(9):1302–12.
101 Jiang L, Wang M, Lin S, Jian R, Li X, Chan J, et al. A quantitative proteome map of the human body. Cell. 2020 Oct 1;183(1):269–83. e19.
102 Johnson ECB, Dammer EB, Duong DM, Ping L, Zhou M, Yin L, et al. Large-scale proteomic analysis of Alzheimer's disease brain and cerebrospinal fluid reveals early changes in energy metabolism associated with microglia and astrocyte activation. Nat Med. 2020 May;26(5):769–80.
103 Gottesman II, ShieldsJ. Schizophrenia and genetics; a twin study vantage point [Internet]. Academic Press; 1972. Available from: https://books.google.com/books?id=_7A5zQEACAAJ.
104 Ye Z, Mo C, Ke H, Yan Q, Chen C, Kochunov P, et al. Meta-analysis of transcriptome-wide association studies across 13 brain tissues identified novel clusters of genes associated with nicotine addiction. Genes. 2021;13(1):37.
105 Zhou H, Sealock JM, Sanchez-Roige S, Clarke TK, Levey DF, Cheng Z, et al. Genome-wide meta-analysis of problematic alcohol use in 435,563 individuals yields insights into biology and relationships with other traits. Nat Neurosci. 2020 Jul;23(7):809–18.
106 Trubetskoy V, Pardiñas AF, Qi T, Panagiotaropoulou G, Awasthi S, Bigdeli TB, et al. Mapping genomic loci implicates genes and synaptic biology in schizophrenia. Nature. 2022 Apr;604(7906):502–8.
107 Zhang S, Zhang H, Zhou Y, Qiao M, Zhao S, Kozlova A, et al. Allele-specific open chromatin in human iPSC neurons elucidates functional disease variants. Science. 2020 Jul 31;369(6503):561–5.
108 Mullins N, Forstner AJ, O'Connell KS, Coombes B, Coleman JRI, Qiao Z, et al. Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. Nat Genet. 2021 Jun;53(6):817–29.
109 Wingo TS, Liu Y, Gerasimov ES, Gockley J, Logsdon BA, Duong DM, et al. Brain proteome-wide association study implicates novel proteins in depression pathogenesis. Nat Neurosci. 2021 Jun;24(6):810–7.
110 Howard DM, Adams MJ, Clarke TK, Hafferty JD, Gibson J, Shirali M, et al. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. Nat Neurosci. 2019 Mar;22(3):343–52.
111 Wray NR, Ripke S, Mattheisen M, Trzaskowski M, Byrne EM, Abdellaoui A, et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nat Genet. 2018 May;50(5):668–81.
112 Eraslan G, Drokhlyansky E, Anand S, Fiskin E, Subramanian A, Slyper M, et al. Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function. Science. 2022 May 13;376(6594):eabl4290.
113 Wen X, Luca F, Pique-Regi R. Cross-population joint analysis of eQTLs: fine mapping and functional annotation. PLoS Genet. 2015 Apr 23;11(4):e1005176.
114 Mogil LS, Andaleon A, Badalamenti A, Dickinson SP, Guo X, Rotter JI, et al. Genetic architecture of gene expression traits across diverse populations. PLoS Genet. 2018 Aug 10;14(8):e1007586.
115 Schubert R, Geoffroy E, Gregga I, Mulford AJ, Aguet F, Ardlie K, et al. Protein prediction for trait mapping in diverse populations. PLoS One. 2022 Feb 24;17(2):e0264341.
116 de Leeuw C, Werme J, Savage JE, Peyrot WJ, Posthuma D. Reconsidering the validity of transcriptome-wide association studies. bioRxiv. 2022 Jan 1:2021.
117 Mostafavi H, Spence JP, Naqvi S, Pritchard JK. Limited overlap of eQTLs and GWAS hits due to systematic differences in discovery [Internet]. bioRxiv. 2022;2022.05.07.491045 [cited 2023 Jan 30]. Available from: https://www.biorxiv.org/content/10.1101/2022.05.07.491045v1.