Quantitative Biology 2021, 9(2): 107–121
https://doi.org/10.1007/s40484-020-0207-4

REVIEW

Transcriptome-wide association studies: a view from Mendelian randomization

Huanhuan Zhu[1], Xiang Zhou[1,2,*]

[1] Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA
[2] Center for Statistical Genetics, University of Michigan, Ann Arbor, MI 48109, USA
* Correspondence: xzhousph@umich.edu

Received January 28, 2020; Revised February 23, 2020; Accepted March 2, 2020

***Background*: Genome-wide association studies (GWASs) have identified thousands of genetic variants that are associated with many complex traits. However, their biological mechanisms remain largely unknown. Transcriptome-wide association studies (TWAS) have been recently proposed as an invaluable tool for investigating the potential gene regulatory mechanisms underlying variant-trait associations. Specifically, TWAS integrate GWAS with expression mapping studies based on a common set of variants and aim to identify genes whose GReX is associated with the phenotype. Various methods have been developed for performing TWAS and/or similar integrative analysis. Each such method has a different modeling assumption and many were initially developed to answer different biological questions. Consequently, it is not straightforward to understand their modeling property from a theoretical perspective.**

***Results*: We present a technical review on thirteen TWAS methods. Importantly, we show that these methods can all be viewed as two-sample Mendelian randomization (MR) analysis, which has been widely applied in GWASs for examining the causal effects of exposure on outcome. Viewing different TWAS methods from an MR perspective provides us a unique angle for understanding their benefits and pitfalls. We systematically introduce the MR analysis framework, explain how features of the GWAS and expression data influence the adaptation of MR for TWAS, and re-interpret the modeling assumptions made in different TWAS methods from an MR angle. We finally describe future directions for TWAS methodology development.**

***Conclusions*: We hope that this review would serve as a useful reference for both methodologists who develop TWAS methods and practitioners who perform TWAS analysis.**

Keywords: transcriptome-wide association studies; genome-wide association studies; expression mapping studies

Author summary: Transcriptome wide association studies (TWAS) integrate expression mapping studies and GWAS studies and aim to identify candidate genes whose genetically regulated expression is associated with trait of interest. We present a comprehensive review on a broad category of recently developed and commonly used TWAS methods. Our review covers different modeling assumptions, different inference procedures, modeling of horizontal pleiotropic effects, and extensions of TWAS towards multivariate MR analysis and summary statistics. Our review also aims to provide a unified view of various TWAS methods from the perspective of Mendelian randomization (MR).

INTRODUCTION

Genome-wide association studies (GWASs) have identified thousands of genetic variants that are associated with many common diseases and disease related complex traits. However, most of these identified genetic variants reside outside protein-coding regions, making it challenging to understand the biological mechanism underlying these identified associations. One possible mechanism that a genetic variant may influence the associated trait is through regulating the gene expression level of its neighborhood gene [1]. To investigate such potential

mechanism, many gene expression mapping studies are performed to in parallel to GWASs to characterize the transcriptome landscape and investigate the genetic architecture underlying gene expression variation. These gene expression mapping studies collect both gene expression data and genotype data on the same set of individuals and aim to identify genetic variants associated with gene expression levels. Exemplary expression mapping studies include the Genotype-Tissue Expression (GTEx) project [2], the Genetic European Variation in Disease (GEUVADIS) project [3] and many others [4–13] (a summary of recently large-scale transcriptome datasets is shown in Table 1). With the availability of both GWASs and expression mapping studies, there is a strong recent interest in developing methods to integrate these two data types together. Integrating GWASs and expression mapping studies is commonly referred to as the transcriptome-wide association study (TWAS), which can facilitate our understanding of the molecular and causal mechanisms underlying variant-trait associations.

Several statistical methods have been recently proposed to perform TWAS. For example, PrediXcan [14] performs a weighted SNP-set-based test in GWAS using SNP weights inferred from the expression mapping study based on elastic net [15]. TWAS [1] infers the association between an outcome phenotype and the predicted gene expression level, where the predicted gene expression levels is built upon the Bayesian sparse linear mixed model (BSLMM) [16]. Zeng and Zhou [17] proposed a non-parametric latent Dirichlet process regression (DPR) model that can flexibly model the underlying complex genetic architecture of expression data for TWAS. TIGAR (Transcriptome-Integrated Genetic Association Resource) further implements DPR in a user friendly software for convenient TWAS analysis [18]. SMR (summary data–based Mendelian randomization) [19] and GSMR (generalized SMR) [20] directly tests the causal relationship between gene expression and disease trait under a Mendelian randomization (MR) framework through selecting a single instrumental variable (IV) or multiple independent IVs. The probabilistic Mendelian randomization (PMR) further uses likelihood-based inference framework to both model all cis-SNPs jointly that are in high linkage disequilibrium (LD) with each other and account for horizontal pleiotropic effects, thus substantially enhancing the power of MR analysis in TWAS settings [21]. While these integrative methods were originally proposed to solve different problems, as we will show below, all of them can be viewed as a two-sample MR method with different modeling assumptions and different inference algorithms (more details below). MR is a causal inference method that uses genetic variants as instrumental variables (IVs) to estimate causal effect of an exposure variable (*e.g.*, gene expression) on an outcome of interest in observational studies. Because of their relationship to MR, these methods effectively attempt to identify causal genes associated with diseases or disease related complex traits in the context of TWAS. Besides the aforementioned methods that perform univariate MR analyses where the exposure variables are examined one at a time, several recent methodological extensions have enabled multivariate MR analysis that models many exposure variables jointly [22–27]. For TWAS applications in particular, multivariate MR attempts to either model the same gene across multiple tissues [28–32] or model multiple genes in the same locus [33].

It has been five years since the first TWAS method,

Table 1 A summary of commonly used gene expression database with sample size over 50

Data sets	RNAseq	Sample size	Ref.
ABRP	Blood (Baboons)	63	[7]
GSE19480	Lymphoblastoid cell lines	69	[8]
Braineac	Ten brain regions	134	[5]
NABEC	Four brain regions	150	[6]
CommonMind	Dorsolateral prefrontal cortex	452	[11]
GEUVADIS	Lymphoblastoid cell lines	465	[3]
TCGA	Prostate adenocarcinoma	483	[10]
METSIM	Adipose	563	[9]
GTEx (v8)	54 tissues	838	[2]
DGN	Whole blood	922	[4]
NTR	Blood	1247	[12]
YFS	Blood	1264	[13]

PrediXcan [14], was proposed. Since then, many TWAS analysis methods and software have been developed for uncovering gene-trait associations. However, there has been a lack of systematic review from a technical perspective summarizing the advantages and shortcomings of these existing methods. Most existing reviews on TWAS often covers a limited number of methods and often aims for experimental biologists. For example, Wainberg *et al*. [34] published a work to point out the opportunities and challenges of TWAS. They affirmed the accomplishments of TWAS in prioritizing candidate genes while also expressed their concerns about the causality of these identified genes. However, Wainberg *et al*. only focused on conducting analysis to evaluate the performance of TWAS [1] and S-PrediXcan [35] and also briefly mentioned UTMOST [29] and MultiXcan [28] in one sentence. To complement these existing review works, here, we present a technical review on thirteen recently developed statistical methods for TWAS. We organize the present review from the perspective of MR framework and gear the presentation towards computational biologists and applied statisticians. In particular, our review is organized as follows. In Section "Mendelian randomization analysis" we describe the MR analysis framework, how it is adapted for TWAS, and the modeling assumptions necessarily for causality interpretation. In Section "Different modeling assumptions on the SNP-gene effect sizes $\boldsymbol{\beta}$" we describe different TWAS methods along with their detailed modeling specifications and show how they are interconnected with each other under the MR framework. In Sections "Extensions of TWAS towards multivariate MR analysis" and "Use of summary statistics" we describe several current extensions of TWAS methods towards using multiple tissues, multiple genes and summary statistics, and explain how such extensions can also be included into the MR framework. In the last Discussion section, we provide our view of future development for TWAS methods. We hope that our review can serve as a useful reference for statistical geneticists and computational biologists.

MENDELIAN RANDOMIZATION ANALYSIS

Both MR and TWAS have become popular in the past decade with increased popularity and availability of GWASs (Fig. 1A, B). These two approaches are mathematically interconnected with each other. In this section, we provide a technical review of MR and illustrate how different TWAS methods can be viewed in the MR framework. MR is a causal inference method that uses genetic variants as IVs to infer the presence or absence of a causal effect of an exposure variable (*e.g*., gene expression) on an outcome of interest in observational studies. MR methods have been widely applied to estimate and test the causal relationship among various complex traits [36–39], and, through a two-sample design, can be easily adapted to settings where the exposure variable and outcome are measured on two independent samples of individuals [40,41].

Two-sample MR considers two separate studies in the setting of TWAS: the gene expression study that measures both the expression data and the genotype data on n_1 individuals; and the GWAS that measures both the outcome variable of interest and the genotype data on n_2 individuals. The two studies are often separate from each other with no individual overlap. MR analysis examines one gene at a time and aims to infer the causal effect of gene expression on the outcome trait. For the given gene, we denote $\boldsymbol{z}$ as an n_1-vector of the gene expression measurements in the first sample (*i.e*., the gene expression study). We denote $\boldsymbol{X}$ as an $n_1 \times p$ genotype matrix for the p cis-SNPs that are selected for the gene in the first sample. Note that, while standard MR methods select one or multiple independent IVs, TWAS methods often take advantage of all SNPs that reside in the cis-region of the gene. These cis-SNPs are often in LD with each other and using all cis-SNPs for TWAS can ensure optimal power (more details in the next section below). We denote $\boldsymbol{y}$ as the n_2 vector of the outcome variable (*i.e*., trait) in the second sample (*i.e.*, the GWAS study). For simplicity, we only consider $\boldsymbol{y}$ to be a quantitative trait, although extensions to a binary trait is straightforward, requiring replacing certain linear regression models with logistic regression models. We also denote $\widetilde{\boldsymbol{X}}$ as an $n_2 \times p$ genotype matrix for the same p cis-SNPs in the second sample. We assume $\boldsymbol{z}$, $\boldsymbol{y}$ and each column of $\boldsymbol{X}$ and $\widetilde{\boldsymbol{X}}$ have all been standardized to have a mean of zero and a standard deviation of one. MR analysis incorporates three linear models to link the two studies jointly:

$$\boldsymbol{z} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{\boldsymbol{z}}, \tag{1}$$

$$\widetilde{\boldsymbol{z}} = \widetilde{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{\widetilde{\boldsymbol{z}}}, \tag{2}$$

$$\boldsymbol{y} = a\widetilde{\boldsymbol{z}} + \boldsymbol{\varepsilon}_{\boldsymbol{y}}, \tag{3}$$

where Eq. (1) is for the first sample and Eqs. (2) and (3) are for the second sample. Here, $\widetilde{\boldsymbol{z}}$ is the unobserved gene expression measurements for the n_2 individuals in the second sample; $\boldsymbol{\beta} = (\beta_1,...,\beta_j,...,\beta_p)^{\mathrm{T}}$ is the p-length effect sizes of the cis-SNPs on the exposure; $\boldsymbol{\varepsilon}_{\boldsymbol{z}}$, $\boldsymbol{\varepsilon}_{\widetilde{\boldsymbol{z}}}$, and $\boldsymbol{\varepsilon}_{\boldsymbol{y}}$ are error terms in the three models, and follow multivariate normal distributions $N_{n_1}(\boldsymbol{0},\sigma_z^2\boldsymbol{I}_{n_1})$, $N_{n_2}(\boldsymbol{0},\sigma_z^2\boldsymbol{I}_{n_2})$, and $N_{n_2}(\boldsymbol{0},\sigma_y^2\boldsymbol{I}_{n_2})$, respectively. Note that these three models are joined together with the common variable $\boldsymbol{\beta}$ and the unobserved gene expression $\widetilde{\boldsymbol{z}}$. The goal of MR methods

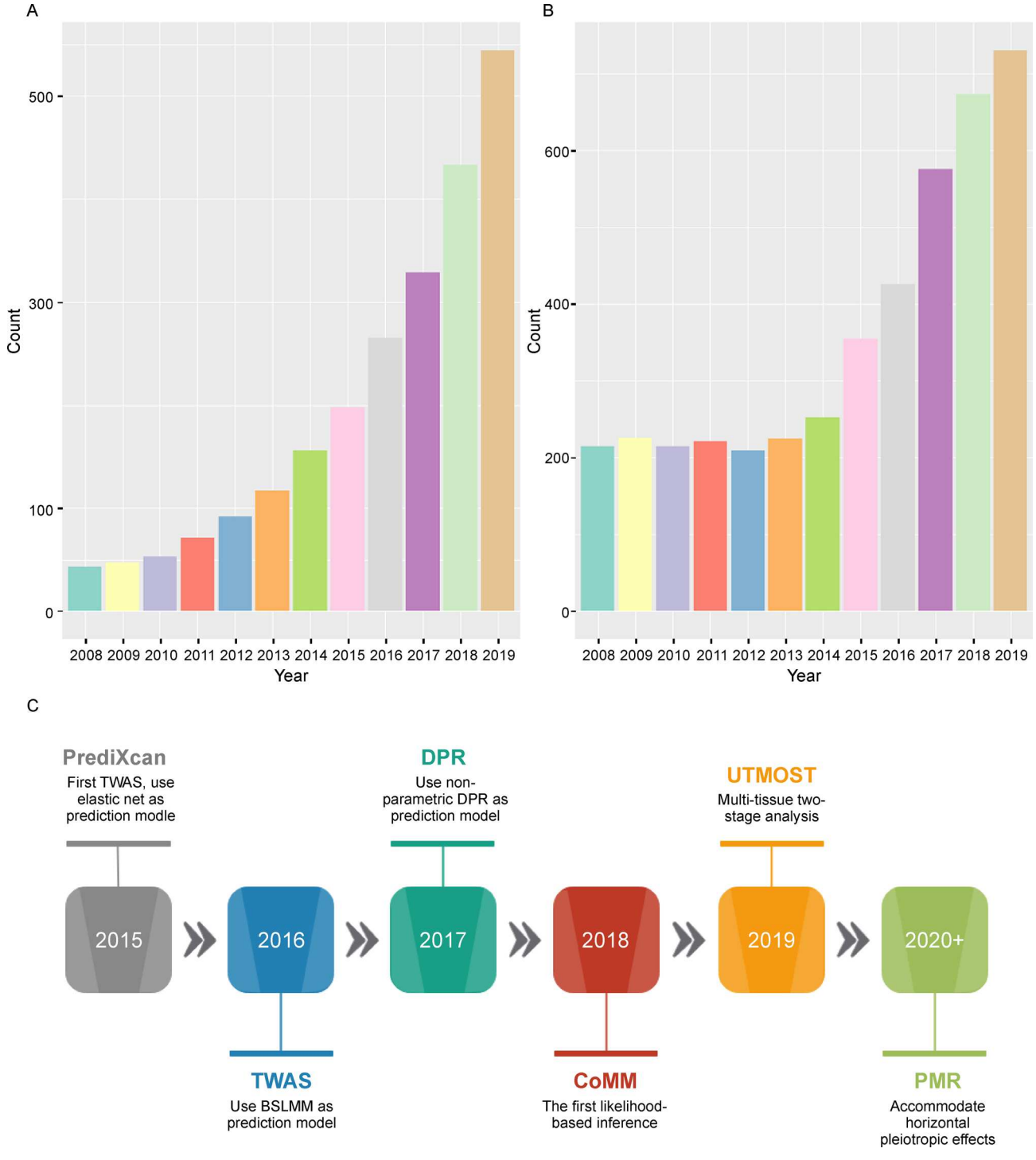

Figure 1. Importance of TWAS and MR methods. (A) The number of publications on PubMed on MR and TWAS over recent years. The generated URL is https://www.ncbi.nlm.nih.gov/pubmed/?term = Mendelian + randomization + or + transcriptome-wide + association + studies. (B) The number of hits on MR based on Google Trends (https://trends.google.com/trends/?geo = US). The search on "transcriptome-wide association studies" was not large enough to generate statistics. (C) Timeline of various TWAS landmark methods throughout the years.

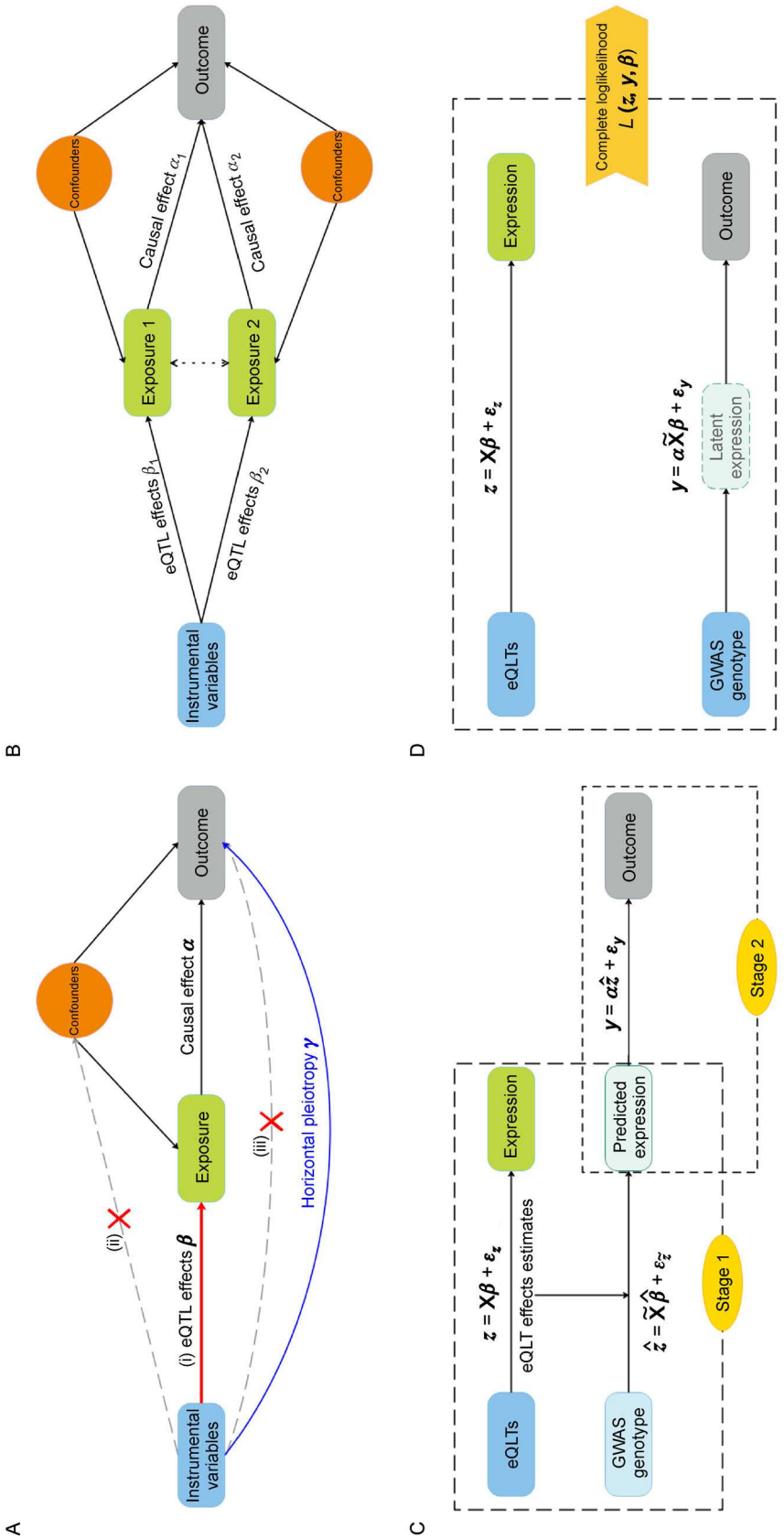

Figure 2. The Mendelian randomization framework for understanding TWAS. (A) The standard MR model assumes three assumptions on instrumental variable: (i) it must be associated with the exposure, (ii) it must not be associated with confounders, and (iii) it is associated with the outcome only through the exposure. The blue solid line represents the horizontal pleiotropic effects. (B) Multivariate MR analysis with two exposures as an example. (C) Scheme of the two-stage TWAS analysis. (D) Scheme of maximum likelihood-based TWAS analysis.

is to make inference on the causal effect a. The causal interpretation of a requires each of the selected IVs to satisfy three main assumptions (Fig. 2A): (i) it must be associated with the exposure, (ii) it must not be associated with confounders, and (iii) it is associated with the outcome only through the exposure.

DIFFERENT MODELING ASSUMPTIONS ON THE SNP-GENE EFFECT SIZES $\boldsymbol{\beta}$

Almost all TWAS methods can be viewed in the above MR framework and different TWAS methods often differ in their modeling assumptions on $\boldsymbol{\beta}$.

Assumptions in traditional MR methods

Perhaps the easiest assumption on $\boldsymbol{\beta}$ is that made in the traditional MR methods, such as SMR [19], GSMR [20], MR-Egger [42], and median-based regression [43]. Both SMR [19] and GSMR [20] have been applied to perform integrative analysis of gene expression study and GWAS in the setting of TWAS. Specifically, SMR [19] selects one SNP in the cis-region of the gene to serve as the IV. To do so, SMR first performs a marginal association analysis for each SNP in turn and selects the one that has the smallest p-value association evidence with the gene expression level. Afterwards, SMR estimates the SNP effect on the outcome trait, the SNP effect on the gene expression, and uses the standard MR ratio method [44] to express the causal effect a as the ratio of the previous two effect estimates. Because the MR ratio method computes p-value based on asymptotic normality, which is often unsatisfied in TWAS settings, the p-values from SMR are often conservative under the null [21,35]. Different from SMR that uses only one IV, GSMR [20] selects multiple independent SNPs in the cis-region to serve as IVs. In particular, GSMR uses the pruning strategy implemented in PLINK to select IVs, estimates the causal effect of each IV in turn using the standard MR ratio method, and eventually combines these causal effect estimates together using the standard inverse-variance weighting (IVW) approach. Importantly, both SMR and GSMR often select a small set of SNPs into Eq. (1). Modeling only a small set of independent SNPs can be restrictive in the setting of TWAS, since this approach neglects the fact that most exposure variables/molecular traits are polygenic/omnigenic and are influenced by numerous SNPs that are in potential LD with each other. Consequently, incorporating multiple correlated SNPs can help explain a larger proportion of variance in the exposure variable than using independent SNPs only, thus helping boost statistical power and improve estimation accuracy of MR analysis [45–48]. Indeed, almost all other TWAS methods include all cis-SNPs of a gene into modeling gene expression in Eq. (1). Note that the number of individuals in the gene expression study is often in the scale of a few hundred while the number of cis-SNPs for a gene is often in the range of a few hundreds to a few thousands, with the detailed number depending on the cis-region size and SNP density in the expression data (Table 1). Consequently, TWAS methods that accommodate all cis-SNPs will often need to make certain modeling assumptions on the SNP effect sizes $\boldsymbol{\beta}$ to ensure model identifiability. Various modeling assumptions on $\boldsymbol{\beta}$ have been proposed.

Elastic net

The first modeling assumption on $\boldsymbol{\beta}$ is the elastic net modeling assumption made in PrediXcan [14]. The elastic net modeling assumption assumes that each element of $\boldsymbol{\beta}$ *a priori* follows a linear combination of LASSO [49] (L_1 penalty) and ridge regression [50] (L_2 penalty) on the cis-SNP effect sizes. In particular, it assumes that

$$\boldsymbol{\beta} \propto \exp(\lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2), \tag{4}$$

where $\|\cdot\|_1$ and $\|\cdot\|_2$ denote the L_1 and L_2 norms, respectively. The elastic net assumption is equivalent to a mixture of normal and Laplace prior. With the above modeling assumption on $\boldsymbol{\beta}$, PrediXcan obtains the estimates of $\boldsymbol{\beta}$ in Eq. (1), plugs in the $\boldsymbol{\beta}$ estimates in Eq. (2) to obtain the genetically predicted gene expression (a.k.a. genetically regulated expression, or GReX), and finally perform analysis in Eq. (3) to obtain the causal effect estimates. Note that the elastic net modeling assumption made in PrediXcan itself is polygenic in nature, as it assumes that all elements of $\boldsymbol{\beta}$ are non-zero *a priori*. However, PrediXcan relies on an optimization algorithm to obtain the maximum a posterior (MAP) estimates for $\boldsymbol{\beta}$ and the MAP estimates is sparse. Therefore, PrediXcan effectively relies on elastic net that combines L_1 and L_2 penalty as a variable selection method to select a sparse set of cis-SNPs with non-zero effects on the gene expression. Similar strategy, pairing a polygenic modeling assumption and a sparse MAP estimation solution, is also used in several other TWAS methods, in particular those applied to multiple-tissue TWAS analysis, for example, UTMOST [29] (more details in the multivariate TWAS section).

Bayesian sparse linear mixed model

The second modeling assumption on $\boldsymbol{\beta}$ is the Bayesian sparse linear mixed model [16] used in the method TWAS [1]. The BSLMM represents a hybrid modeling assumption between a sparse modeling assumption such as the Bayesian variable sparse regression (BVSR, more details below) [51] and the standard polygenic modeling

assumption. Consequently, BSLMM is able to take advantages of LMM and sparse regression models and can adaptively infer the genetic architecture underlying the gene expression variation from the data. Technically, BSLMM assumes that the effect size of each cis-SNP on the gene expression follows a mixture of two normal distributions

$$\beta_j \sim \pi N(0,\ \sigma_a^2 + \sigma_b^2) + (1-\pi)N(0,\ \sigma_b^2). \quad (5)$$

In particular, with probability $1-\pi$, β_j tends to be small and follows a normal distribution with a small background variance of σ_b^2. With probability π (which is often small), β_j tends to be large and follows a normal distribution with a large variance that equals the summation of the background variance and an additional variance. The BSLMM modeling assumption represents a direct attempt for modeling the omnigenic hypothesis that was proposed recently [52]. Specifically, the BSLMM assumption categorizes SNPs into two groups: a small group of SNPs with large effect sizes and a large group of SNPs with small effect sizes. Such SNP categorization is equivalent to assuming that all SNPs have non-zero effects, while a small proportion of them have additional effects. The assumption that all SNPs have non-zero effects attempts to model the omnigenic hypothesis that all genes/SNPs have non-zero effects. The assumption that a small subset of SNPs has additional effects also attempts to model the omnigenic hypothesis that a small subset of genes, termed as core genes, have additional effects. The set of core genes was hypothesized in the omnigenic model to directly underlie disease etiology and contribute disproportionally to disease and disease related complex traits. With the BSLMM modeling assumption on $\boldsymbol{\beta}$, TWAS [1] obtains the estimates of $\boldsymbol{\beta}$ in Eq. (1), plugs in the $\boldsymbol{\beta}$ estimates in Eq. (2) to obtain the genetically predicted gene expression, and finally perform analysis in Eq. (3) to obtain the causal effect estimates. Because of the relatively robust and flexible assumption made in BSLMM, the TWAS method often performs well across a range of TWAS applications.

Dirichlet process regression

The third modeling assumption on $\boldsymbol{\beta}$ is the latent Dirichlet process regression (DPR) [17] implemented in the TWAS methods DPR [17] and TIGAR [18]. DPR relies on a Bayesian non-parametric modeling assumption on the genetic effects on the gene expression. In particular, it assumes that each element of $\boldsymbol{\beta}$ follows a normal distribution, with a further unknown distribution G placed upon the variance parameter. DPR actively infers such unknown distribution G by placing a non-parametric Dirichlet process (DP) prior on the distribution itself:

$$\beta_j \sim N(0,\ \sigma_j^2),\ \ \sigma_j^2 \sim G,\ \ G \sim \mathrm{DP}(\mathrm{IG}(a,\ b),\ \lambda), \quad (6)$$

where the inverse gamma (IG) distribution is the base distribution while the concentration parameter λ determines how the distribution of G differs from the base distribution. By inferring the distribution G based on the data at hand, DPR becomes flexible and is adaptive to a wide range of genetic architectures, leading to accurate gene expression prediction and subsequent power increase for TWAS. Note that the above modeling assumption is also equivalent to assuming each element of $\boldsymbol{\beta}$ follows a mixture of infinitely many normal distributions *a priori*,

$$\beta_j \sim \sum_{\varphi=1}^{+\infty} \pi_\varphi N(0,\ \sigma_\varphi^2),\ \pi_\varphi = v_\varphi \prod_{l=1}^{\varphi-1}(1-v_l),\ v_\varphi \sim \mathrm{Beta}(1,\ \lambda). \quad (7)$$

Here, π_φ is the weight corresponding to the φ-th normal distribution; it is generated from a stick breaking process and determined by v_l that each follows a Beta prior. With the DPR modeling assumption on $\boldsymbol{\beta}$, one can obtain the estimates of $\boldsymbol{\beta}$ in Eq. (1) via two algorithms: either the Monte Carlo Markov Chain or variational Bayesian algorithm. Both these two algorithms are implemented in the DPR software [17] while the second algorithm is also conveniently implemented in the TIGAR software [18]. With the estimated $\boldsymbol{\beta}$ from Eq. (1), one can use Eq. (2) to obtain the genetically predicted gene expression and finally perform analysis in Eq. (3) to obtain the causal effect estimates. Because of the relatively robust and flexible assumption made in DPR, DPR and TIGAR often performs well in TWAS applications.

Linear mixed model

The fourth modeling assumption is the normality assumption on the effect sizes that is used in CoMM [53] and PMR [21]. The normality assumption assumes that each element of $\boldsymbol{\beta}$ follows a normal distribution

$$\beta_j \sim N(0,\ \sigma_\beta^2). \quad (8)$$

The above model effectively assumes that all SNPs have non-zero effects on gene expression and their effect sizes follow a normal distribution. The normality modeling assumption is often referred to as the ridge regression assumption or L2 assumption in statistics literature and is also often referred to as the polygenic modeling assumption or the linear mixed model (LMM) assumption in various GWAS applications. For TWAS applications, CoMM [53] and some of its extensions [32,54], as well as PMR [21], all use this modeling assumption. The

normality modeling assumption is known to be less flexible than the BSLMM and DPR modeling assumption. However, it is also a simple modeling assumption that allows model inference based on a likelihood framework that is known to be more powerful than the two stage inference procedures used in common TWAS methods. Consequently, CoMM and PMR often enjoy substantial power gain over existing TWAS approaches including PrediXcan and TWAS.

Bayesian variable selection regression

Finally, the Bayesian variable selection regression modeling assumption (BVSR) [51] is also recently adapted by the Factored QTL (fQTL) [31] into TWAS settings. In contrast to the above polygenic modeling assumptions (*e.g.*, elastic net, BSLMM, DPR and LMM), BVSR places a sparse modeling assumption on the genetic effects on the gene expression. In particular, BVSR assumes that each element of $\boldsymbol{\beta}$ follows a point-normal distribution

$$\beta_j \sim \pi(0,\sigma_\beta^2) + (1-\pi)\delta_0, \tag{9}$$

where with a small proportion π, β_j is non-zero and follows a normal distribution; and with proportion $1-\pi$, β_j is zero with δ_0 indicating a point mass at zero. The point normal is also commonly referred to as a spike and slab prior. More details about fQTL are given in the multivariate TWAS section.

Overall, different TWAS methods make different modeling assumptions on $\boldsymbol{\beta}$. While the sparse modeling assumptions used in SMR [19], GSMR [20], and fQTL [31] are the easiest to understand, they often do not perform well for TWAS applications as compared to the polygenic modeling assumptions made in most existing TWAS methods such as PrediXcan [14], TWAS [1], DPR [17], TIGAR [18], CoMM [53] and PMR [21]. Indeed, polygenic models (*e.g.*, LMM, BSLMM, DPR) often outperform sparse models (elastic net, LASSO, etc.) in predicting gene expression and TWAS applications [1,18,21]. The superior performance of polygenic modeling assumptions in TWAS is consistent with gene expression heritability studies that reveal a polygenic architecture underlying gene expression level [7]. In terms of models with polygenic assumptions, both BSLMM and DPR are flexible and include some other polygenic models as special cases. Due to the flexible modeling assumption in BSLMM and DPR, TWAS methods using these assumptions often perform well across genes with varying genetic architectures, which is often unknown *a priori*. However, these flexible modeling assumptions also have the shortcomings of being computationally difficulty to fit. Consequently, TWAS methods using these flexible polygenic modeling assumptions often have to rely on a two-step estimation procedure, by constructing the predicted genetic component of gene expression and subsequently estimate its association with the outcome trait. In contrast, simple polygenic modeling assumptions such as the normality assumption allows MR analysis to be carried out in a likelihood framework, thus leading to substantial power gain (more details below).

INFERENCE PRECEDURES AND MODELING OF HORIZONTAL PLEIOTROPIC EFFECTS

In terms of the inference procedure, as briefly explained in the above section, while most TWAS methods perform causal inference in a two-stage regression-based framework (Fig. 2C), several recently developed TWAS methods attempt to perform inference in a maximum likelihood-based framework (Fig. 2D). Specifically, the two-stage regression-based inference algorithm attempts to construct a predictor of gene expression data using the IVs and then perform an association between the predicted gene expression levels and the outcome phenotype. The majority of existing TWAS methods, such as PrediXcan, TWAS, DPR, and TIGAR, rely on a two-stage MR inference procedure: they estimate SNP effect sizes in the reference transcriptome data and pass these estimates to the GWAS study for causal effect inference. In other words, these methods perform gene expression "imputation" and subsequent "association" between imputed expression and outcome phenotype as two separate steps. In contrast, the maximum likelihood-based inference procedure, as used in CoMM and PMR, jointly model all the three equations together and perform inference through maximizing the likelihood function. The two-stage inference procedure in MR has the benefits of simplicity and yields approximately unbiased causal effect size estimates. However, the two-stage inference procedure may also fail to account for the uncertainty in parameter estimates in the transcriptome study and thus resulting in power loss, especially in the presence of weak IVs [45,47]. Indeed, similar to what have been observed in the MR framework, recent TWAS studies also suggest that likelihood-based inference can substantially improve power for TWAS [53].

Beside the difference in inference procedure, different TWAS methods also differ in their ways of modeling horizontal pleiotropic effects. In particular, while most TWAS methods do not account for horizontal pleiotropy, some recently developed TWAS attempt to directly model horizontal pleiotropy. In the TWAS setting, horizontal pleiotropy occurs when an IV affects the outcome through

pathways other than the middle exposure variable [55]. It has been recently observed that pervasive horizontal pleiotropy occurs for both complex traits analysis [56] and for TWAS applications [21]. The horizontal pleiotropy is widely distributed across the genome and is important for our understanding of the genetic architecture of human diseases and disease related complex traits (Fig. 2A). Failing to account for horizontal pleiotropic effects in MR or TWAS analysis can be overly restrictive and can lead to a substantial inflation of test statistics and subsequently false discoveries [21]. Because of the importance and wide presence of horizontal pleiotropy, several MR methods have been developed to test and account for horizontal pleiotropic effects [20,43,56–63] in GWAS and for TWAS applications [21]. For example, MR-PRESSO [56] is proposed to test for horizontal pleiotropic effects without directly controlling for them. CaMMEL [57] controls for horizontal pleiotropic effects without directly testing them. Egger regression [43,58], GLIDE [59], GSMR [20], MR-median method [43], MRMix [60] and Bayesian MR [61,62] test and control for horizontal pleiotropic effects with independent instruments. LDA MR-Egger [63] is developed for testing and controlling for pleiotropic effects in the presence of correlated instruments. More recently, PMR [21] builds upon these previous studies and relies on a jointly integrative TWAS analysis to accommodate the presence of both correlated instruments and horizontal pleiotropy. Specifically, PMR replaces Eq. (3) with the following extended version

$$\boldsymbol{y} = \alpha\widetilde{\boldsymbol{z}} + \widetilde{\boldsymbol{X}}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_{\boldsymbol{y}}, \tag{10}$$

where $\boldsymbol{\gamma}$ is a p-length vector representing the horizontal pleiotropic effects. PMR [21] explored two different modeling assumptions on the horizontal pleiotropic effects $\boldsymbol{\gamma}$. The first modeling assumption is the normality modeling assumption $\gamma_j \sim N(0,\ \sigma_\gamma^2)$, which assumes that all elements of $\boldsymbol{\gamma}$ is non-zero and they all follow a normal distribution *a priori*. The second modeling assumption is the Egger modeling assumption $\gamma_1 = ... = \gamma_p = \gamma$, which assumes that all elements of $\boldsymbol{\gamma}$ equal to each other and all equal to a common scalar value of γ. The first modeling assumption is analogous to the SKAT [64] modeling assumption commonly used in the rare variant test setting while the second modeling assumption is analogous to the burden [65–67] modeling assumption also used in rare variant test setting. PMR when paired with the first modeling assumption is often referred to as PMR-VC while PMR paired with the second modeling assumption is often referred to as PMR-Egger. Both versions of PMR test the causal effect $H_0: \alpha = 0$ while properly controlling for horizontal pleiotropic effects, resulting in a substantial reduction of false positives. Importantly, while PMR no longer requires the third assumption of standard MR model (*i.e.*, instruments are associated with the outcome only through the exposure), it still requires the InSIDE assumption that the instrument-exposure effects and instrument-outcome effects are independent of each other, which is sometimes refered to as the weak exclusion restriction condition [42]. Besides testing for causal effects, both PMR-VC and PMR-Egger can also directly test for horizontal pleiotropic effects by testing the corresponding null hypothesis: $H_0: \sigma_\gamma^2 = 0$ in PMR-VC and $H_0: \gamma = 0$ in PMR-Egger. By testing for horizontal pleiotropic effects, widespread horizontal pleiotropy has been revealed across the transcriptome.

EXTENSIONS OF TWAS TOWARDS MULTIVARIATE MR ANALYSIS

Traditional TWAS methods are univariate in nature and focus on analyzing one exposure and one gene at a time. However, many recently developed TWAS methods are gradually extending from the univariate TWAS analysis to multivariate TWAS by using either multiple exposures or multiple genes (Fig. 2B). For example, TisCoMM [32] is an extension of CoMM [53] and can leverage the co-regulation of cis-SNPs on multiple tissues via a likelihood-based inference. Specifically, TisCoMM regresses expression data across multiple tissues on genotype by the following multiple regression model:

$$\boldsymbol{Z}_{n_1 \times m} = \boldsymbol{X}_{n_1 \times p}\boldsymbol{B}_{p \times m} + \boldsymbol{\varepsilon}_{z, n_1 \times m} \tag{11}$$

where $m\ (k = 1,...,m)$ denotes the number of tissues, $\boldsymbol{Z}$ is the expression matrix with each column representing a tissue measured from n_1 samples, $\boldsymbol{B}$ is the genetic effect matrix with dimension $p \times m$, and $\boldsymbol{\varepsilon}_{\boldsymbol{z}}$ is the error term with dimension $n_1 \times m$. TisCoMM assumes that the genetic effect matrix $\boldsymbol{B} = \mathrm{diag}(\boldsymbol{b})\boldsymbol{W}$, where $\boldsymbol{b} = (b_1,...,b_p) \sim N_p(\boldsymbol{0}, \sigma_b^2\boldsymbol{I}_p)$ is the SNP-dependent component and $\boldsymbol{W}_{p \times m} = (w_{jk})$ is the tissue-dependent component, where w_{jk} is estimated using the marginal regression of gene expression on the j-th SNP in the k-th tissue. The GWAS model is an extension of Eqs. (2) and (3):

$$\boldsymbol{y} = \widetilde{\boldsymbol{X}}_{n_2 \times p}\boldsymbol{B}_{p \times m}\boldsymbol{\alpha}_{m \times 1} + \boldsymbol{\varepsilon}_{y, n_2 \times 1}, \tag{12}$$

where $\boldsymbol{\alpha}_{m \times 1} = (\alpha_1,...,\alpha_k,...,\alpha_m)$ is a vector of causal effects with each element indicating the effect of gene expression in each tissue on the phenotype. TisCoMM uses the PX-EM algorithm to estimate parameters and likelihood ratio tests to make inference on $\boldsymbol{\alpha}_{m \times 1}$ [53].

Similarly, UTMOST [29] uses the same expression model as in Eq. (11). Different from TisCoMM that requires a complete expression matrix $\boldsymbol{Z}$ in order to complete the likelihood-based analysis, UTMOST allows incomplete $\boldsymbol{Z}$ meaning only a subset of tissues is collected from each sample. Denote $\boldsymbol{Z}_k$ as an N_k -length vector of

expression data in the k-th tissue, which is a subset of the k-th column of matrix $\boldsymbol{Z}$ that contain non-missing expression data; $\boldsymbol{X}_k$ is an $N_k \times p$ genotype matrix for the same N_k samples; $\boldsymbol{B}_{\cdot k}$ is the k-th column of matrix $\boldsymbol{B}$ and represents the genetic effect sizes of p SNPs in the k-th tissue; and $\boldsymbol{B}_{j\cdot}$ is the j-th row of $\boldsymbol{B}$ and represents the genetic effect sizes of the j-th SNP across all m tissues. UTMOST estimates $\boldsymbol{B}$ by minimizing the squared loss function with a LASSO penalty on the columns (within-tissue) and a ridge penalty on the rows (cross-tissue):

$$\boldsymbol{B} \propto \exp\left(\lambda_1 \sum_{k=1}^{m} \frac{1}{N_k} \|\boldsymbol{B}_{\cdot k}\|_1 + \lambda_2 \sum_{j=1}^{p} \|\boldsymbol{B}_{j\cdot}\|_2\right). \quad (13)$$

A third method, MultiXcan [28], leverages the substantial sharing of eQTLs across multiple tissues using multivariate regression. Specifically, MultiXcan regresses the phenotype of interest on the predicted expression from multiple tissues:

$$\boldsymbol{y} = \sum_{k=1}^{m} \hat{\boldsymbol{z}}_k \boldsymbol{\alpha}_k + \boldsymbol{\varepsilon}_y, \quad (14)$$

where $\hat{\boldsymbol{z}}_k$ is a vector of the standardized version (zero mean and unit standard deviation) of the predicted expressions in tissue k, *i.e.*, $\sum_j \boldsymbol{B}_{jk}\boldsymbol{X}_j$, where $\boldsymbol{B}_{jk}$ is estimated based on elastic net regression; and $\boldsymbol{\alpha}_k$ is the causal effect of gene expression in the k-th tissue on phenotype $\boldsymbol{y}$. MultiXcan then uses an F-test to jointly infer the significance of gene effects across multiple tissues.

A fourth multivariate TWAS approach is fQTL [31], which decomposes the SNP effect of the j-th SNP in the k-th tissue, $\boldsymbol{B}_{jk}$, into a SNP-dependent component and a tissue-dependent component. That is, fQTL assumes $\boldsymbol{B}_{jk} = \sum_{r=1}^{t} b_{jr}^{snp} (b^{tis})_{rk}^{\mathrm{T}}$, where $t \leqslant$ m. fQTL assumes the BVSR [51] prior on each column of $\boldsymbol{b}^{snp}$ (SNP-dependent genetic effect component) and $\boldsymbol{b}^{tis}$ (tissue-dependent genetic effect component) and estimates the posterior distribution of $\boldsymbol{b}^{snp}$ and $\boldsymbol{b}^{tis}$ based on stochastic variational inference (SVI), which finds the best mean-field approximating distribution to the posterior by optimizing the variational objective function. Afterwards, the posterior distributions of $\boldsymbol{b}^{snp}$ and $\boldsymbol{b}^{tis}$ can be obtained respectively, together with the mean and variance of $\boldsymbol{B}_{jk}$. Finally, fQTL characterizes the distribution of the tissue-specific gene expression from a Gaussian distribution where the mean and variance are related to the mean and variance of $\boldsymbol{B}_{jk}$.

A fifth method, multi-tissue TWAS [30] identifies susceptibility genes by using gene expression panels measured in various tissues from multiple expression consortiums. Specifically, multi-tissue TWAS conducts the univariate TWAS [1] for each tissue using the FUSION software and is able to quantify the tissue-trait relevance by the mean TWAS association statistics from all genes.

Finally, in addition to extending TWAS from single tissue to multiple tissues, a recently developed method FOCUS [33] (Fine-mapping Of CaUsal gene Sets) also attempts to extend TWAS from modeling one gene at a time towards modeling multiple genes simultaneously. FOCUS takes as input GWAS summary data, expression prediction weights, and LD among all SNPs, and estimates the probability of any given set of genes containing the causal genes.

USE OF SUMMARY STATISTICS

Because of consent and privacy concerns, as well as logistic limitations (*e.g.*, large-scale data transfer and storage often require high-end computing infrastructure), it is now becoming increasingly difficult to access complete individual-level data from large-scale association studies. Indeed, using summary statistics across multiple studies and then releasing results in terms of summary statistics has become a standard practice in most studies and it has several advantages over using individual phenotype and genotype data. Using summary statistics in TWAS settings has several important benefits. First, GWAS summary statistics are often stored in the datasets with open access, and it becomes incredibly easy to obtain summary statistics than individual-level data which requires a lengthy process for data approval. Second, many GWAS summary statistics are often obtained through meta-analysis of multiple sub-studies where hundreds of thousands of individuals in total are collected. Since sample size is the most important factor in determining statistical power, using summary statistics can lead to substantial benefits for TWAS. Third, summary statistics-based analysis often offers advantages in computational cost and computing memory storage as compared to individual data-based approaches. Consequently, many existing individual-level TWAS methods can either directly accommodate summary statistics or have corresponding extensions that can accommodate them. For example, the summary statistics version of PrediXcan (S-PrediXcan) [35] and CoMM (CoMM-S^2) [53] are presented as follow-up extensions of the original individual-level data based version. Other methods are proposed to directly use summary statistics without an initial individual-level data model; such examples include UTMOST [29], fQTL [31], and FOCUS [33]. Yet some other methods are presented to work on both individual-level data and summary statistics and such examples include PMR [21], TIGAR [18], TWAS (STWAS) [1], MultiXcan (S-MultiXcan) [28], and TisCoMM (TisCoMM-S^2) [32]. Regardless how the summary statistics version of different TWAS methods were proposed, these

methods often require two important input information from the GWAS study: the marginal association statistics in terms of marginal z-scores obtained in the GWAS, and the SNP correlation/LD matrix obtained from either a sub-sample of the original GWAS or from a reference panel. Certainly, the sample sizes used in these two input data are often orders of magnitude different from each other: while the marginal z-scores are calculated based on often hundreds of thousands of individuals, the SNP correlation matrix is often calculated based on a few thousands of individuals. Even though the SNP correlation matrix is often calculated based on a much smaller set of individuals, the overall parameter estimation accuracy remains high, especially when a polygenic modeling assumption on SNP effect sizes are made [68]. Certainly, besides the two input information from GWAS, the summary statistics version of TWAS methods also requires the input information from the gene expression data. Because the current gene expression studies often contain samples only in the scale of tenths to hundredths (Table 1), many researchers can still use individual-level genotype-expression data, although summary statistics version of some TWAS methods can also make use of summary level data from the gene expression study as input.

We summarized the thirteen TWAS approaches examined above in Table 2 from the following aspects: model designs (two-stage or likelihood-based), number of tissues from the expression mapping study (single or multiple), data types a method is applicable for (individual-level, summary statistics, or both), whether controlling for horizontal pleiotropic effects (yes or no), modelling assumptions on genetic effects (LMM, BSLMM, DPR, etc.), and the URL link of implemented software for each method.

DISCUSSION

Transcriptome-wide association studies have been proposed for five years and have been widely applied for prioritizing candidate genes whose genetically regulated expression is associated with common diseases and disease related complex traits. As we have presented here, almost all TWAS methods can be viewed as a two-sample Mendelian randomization analysis with different modeling assumptions. We have comprehensively review the existing TWAS methods from the perspective of MR.

Most TWAS methods and applications have been focused on using common cis-SNPs that have a reasonably high minor allele frequency (MAF) and that reside in a small cis-region of a gene (*e.g.*, 1 Mb surrounding the transcription factor starting site). In recent years, many GWAS studies have shown that rare genetic variants can play a crucial role in explaining missing heritability and some of them are identified to be associated with many diseases and traits [64,69,70]. Therefore, including rare

Table 2 A summary of thirteen TWAS approaches examined in the present review

Methods	Design	Tissue	Data type	Pleiotropy	Model assumptions	URLs
PrediXcan	Two-stage	Single	Individual	No	Elastic net	https://github.com/hakyimlab/PrediXcan
S-PrediXcan	Two-stage	Single	Summary	No	Elastic net	https://github.com/hakyimlab/MetaXcan
TWAS	Two-stage	Single	Individual /Summary	No	BSLMM	https://bogdan.dgsom.ucla.edu/pages/twas/
DPR	Two-stage	Single	Individual	No	DPR	http://www.xzlab.org/software.html
TIGAR	Two-stage	Single	Individual /Summary	No	DPR	https://github.com/yanglab-emory/TIGAR
CoMM	Likelihood-based	Single	Individual	No	LMM	https://github.com/gordonliu810822/CoMM
CoMM-S^2	Likelihood-based	Single	Summary	No	LMM	https://github.com/gordonliu810822/CoMM
PMR	Likelihood-based	Single	Individual /Summary	Yes	LMM	https://github.com/yuanzhongshang/PMR
UTMOST	Two-stage	Multiple	Summary	No	LASSO & Ridge	https://github.com/Joker-Jerome/UTMOST
MultiXcan	Two-stage	Multiple	Individual /Summary	No	Elastic net	https://github.com/hakyimlab/MetaXcan
TisCoMM	Likelihood-based	Multiple	Individual /Summary	No	LMM	https://github.com/XingjieShi/TisCoMM
fQTL	Two-stage	Multiple	Summary	No	BVSR	https://github.com/ypark/fqtl
FOCUS	Gene-mapping	Multiple genes	Summary	No		https://github.com/bogdanlab/focus

variants into TWAS applications model may have added benefits. In addition, while cis-SNPs explain a substantial fraction of gene expression heritability, the explained expression heritability is nevertheless small. For example, it is estimated that 70%–90% of gene expression heritability is determined by trans-acting factors [7,71]. Therefore, incorporating trans-SNPs into TWAS applications may help improve association power. Finally, while several multivariate TWAS methods have been developed to accommodate multiple tissues, important statistical and computational challenges remain in multivariate TWAS modeling. For example, current multivariate TWAS methods are either combining single-tissue association results together in a relatively simple fashion (*e.g.*, UTMOST) or are only capable to using a small subset of tissues with overlapping samples (*e.g.*, TisCoMM). Modeling more tissues (*e.g.*, 54 tissues in GTEx) simultaneously may help us better understand the transcriptomic mechanism underlying disease etiology. Besides the extensions towards multiple tissues and genes, multivariate TWAS analysis can also be extended towards multiple phenotypes. Multiple phenotypes analyses are widely employed in GWASs and have been proven to be more powerful than testing each phenotype at a time by considering the correlation across phenotypes. Incorporating multiple correlated phenotypes into TWAS may become a potential way to discover genes that are associated with multiple phenotypes. However, this practice needs more investigations due to the complexity of phenotypic structures.

Finally, we caution that, while we have followed the previous MR literature and use "causal effect" through the text, the effect is causal only when certain MR modeling assumptions hold. These MR assumptions are often not straightforward to prove. For example, without measuring all potential confounders, it is not straightforward to argue that the SNP instruments are not associated with any other confounders that may be associated with both exposure and outcome. Therefore, we caution against the over-interpretation of causal inference in observation studies such as TWAS applications. However, we do believe MR is an important step that allows us to move beyond standard linear regressions and is an important analysis that can provide potentially more trustworthy evidence with regard to causality compared to simpler approaches. In summary, we hope that our review could serve as a useful reference for understanding TWAS from the MR perspective and provide researchers useful information for the future development of TWAS methods.

ACKNOWLEDGEMENTS

This study was supported by the National Institutes of Health (NIH) Grants R01HG009124 and the National Science Foundation (NSF) Grant DMS1712933.

COMPLIANCE WITH ETHICS GUIDELINES

The authors Huanhuan Zhu and Xiang Zhou declare that they have no conflict of interests.

All procedures performed in studies were in accordance with the ethical standards of the institution.

REFERENCES

1. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W. J. H., Jansen, R., de Geus, E. J., Boomsma, D. I., Wright, F. A., *et al.* (2016) Integrative approaches for large-scale transcriptome-wide association studies. Nat. Genet., 48, 245–252
2. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., *et al.* (2013) The genotype-tissue expression (GTEx) project. Nat. Genet., 45, 580–585
3. Lappalainen, T., Sammeth, M., Friedländer, M. R., 't Hoen, P. A., Monlong, J., Rivas, M. A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., *et al.* (2013) Transcriptome and genome sequencing uncovers functional variation in humans. Nature, 501, 506–511
4. Battle, A., Mostafavi, S., Zhu, X., Potash, J. B., Weissman, M. M., McCormick, C., Haudenschild, C. D., Beckman, K. B., Shi, J., Mei, R., *et al.* (2014) Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res., 24, 14–24
5. Ramasamy, A., Trabzuni, D., Guelfi, S., Varghese, V., Smith, C., Walker, R., De, T., Coin, L., de Silva, R., Cookson, M. R., *et al.* (2014) Genetic variability in the regulation of gene expression in ten regions of the human brain. Nat. Neurosci., 17, 1418–1428
6. Gibbs, J. R., van der Brug, M. P., Hernandez, D. G., Traynor, B. J., Nalls, M. A., Lai, S.-L., Arepalli, S., Dillman, A., Rafferty, I. P., Troncoso, J., *et al.* (2010) Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. PLoS Genet., 6, e1000952
7. Tung, J., Zhou, X., Alberts, S. C., Stephens, M. and Gilad, Y. (2015) The genetic architecture of gene expression levels in wild baboons. eLife, 4, e04729
8. Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J. B., Stephens, M., Gilad, Y. and Pritchard, J. K. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature, 464, 768–772
9. Stancáková, A., Civelek, M., Saleem, N. K., Soininen, P., Kangas, A. J., Cederberg, H., Paananen, J., Pihlajamäki, J., Bonnycastle, L. L., Morken, M. A., *et al.* (2012) Hyperglycemia and a common variant of GCKR are associated with the levels of eight amino acids in 9,369 Finnish men. Diabetes, 61, 1895–1902
10. Abeshouse, A., Ahn, J., Akbani, R., Ally, A., Amin, S., Andry, C. D., Annala, M., Aprikian, A., Armenia, J., Arora, A., *et al.* (2015) The molecular taxonomy of primary prostate cancer. Cell, 163, 1011–1025
11. Fromer, M., Roussos, P., Sieberts, S. K., Johnson, J. S., Kavanagh, D. H., Perumal, T. M., Ruderfer, D. M., Oh, E. C., Topol, A., Shah,

H. R., *et al.* (2016) Gene expression elucidates functional impact of polygenic risk for schizophrenia. Nat. Neurosci., 19, 1442–1453
12. Wright, F. A., Sullivan, P. F., Brooks, A. I., Zou, F., Sun, W., Xia, K., Madar, V., Jansen, R., Chung, W., Zhou, Y. H., *et al.* (2014) Heritability and genomics of gene expression in peripheral blood. Nat. Genet., 46, 430–437
13. Raitakari, O. T., Juonala, M., Rönnemaa, T., Keltikangas-Järvinen, L., Räsänen, L., Pietikäinen, M., Hutri-Kähönen, N., Taittonen, L., Jokinen, E., Marniemi, J., *et al.* (2008) Cohort profile: the cardiovascular risk in Young Finns Study. Int. J. Epidemiol., 37, 1220–1226
14. Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., Nicolae, D. L., Cox, N. J., *et al.* (2015) A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet., 47, 1091–1098
15. Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. J. R. Stat. Soc. B, 67, 301–320
16. Zhou, X., Carbonetto, P. and Stephens, M. (2013) Polygenic modeling with bayesian sparse linear mixed models. PLoS Genet., 9, e1003264
17. Zeng, P. and Zhou, X. (2017) Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. Nat. Commun., 8, 456
18. Nagpal, S., Meng, X., Epstein, M. P., Tsoi, L. C., Patrick, M., Gibson, G., De Jager, P. L., Bennett, D. A., Wingo, A. P., Wingo, T. S., *et al.* (2019) TIGAR: an improved Bayesian tool for transcriptomic data imputation enhances gene mapping of complex traits. Am. J. Hum. Genet., 105, 258–266
19. Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., Montgomery, G. W., Goddard, M. E., Wray, N. R., Visscher, P. M., *et al.* (2016) Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat. Genet., 48, 481–487
20. Zhu, Z., Zheng, Z., Zhang, F., Wu, Y., Trzaskowski, M., Maier, R., Robinson, M. R., McGrath, J. J., Visscher, P. M., Wray, N. R., *et al.* (2018) Causal associations between risk factors and common diseases inferred from GWAS summary data. Nat. Commun., 9, 224
21. Yuan, Z., Zhu, H., Zeng, P., Yang, S., Sun, S., Yang, C., Liu, J., Zhou, X. (2019) Testing and controlling for horizontal pleiotropy with the probabilistic Mendelian randomization in transcriptome-wide association studies. bioRxiv, 691014
22. Sanderson, E., Davey Smith, G., Windmeijer, F. and Bowden, J. (2019) An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings. Int. J. Epidemiol., 48, 713–727
23. Burgess, S. and Thompson, S. G. (2015) Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. Am. J. Epidemiol., 181, 251–260
24. Rees, J. M. B., Foley, C. N. and Burgess, S. (2019) Factorial Mendelian randomization: using genetic variants to assess interactions. Int. J. Epidemiol., dyz161
25. Burgess, S., Daniel, R. M., Butterworth, A. S. and Thompson, S. G., and the EPIC-InterAct Consortium. (2015) Network Mendelian randomization: using genetic variants as instrumental variables to investigate mediation in causal pathways. Int. J. Epidemiol., 44, 484–495
26. Porcu, E., Rüeger, S., Lepik, K., the eQTLGen Consortium, the BIOS Consortium, Santoni, F. A., Reymond, A. and Kutalik, Z. (2019) Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. Nat. Commun., 10, 3300
27. Zuber, V., Colijn, J. M., Klaver, C. and Burgess, S. (2020) Selecting causal risk factors from high-throughput experiments using multivariable Mendelian randomization. Nat. Commun. 11, 29
28. Barbeira, A. N., Pividori, M., Zheng, J., Wheeler, H. E., Nicolae, D. L. and Im, H. K. (2019) Integrating predicted transcriptome from multiple tissues improves association detection. PLoS Genet., 15, e1007889
29. Hu, Y., Li, M., Lu, Q., Weng, H., Wang, J., Zekavat, S. M., Yu, Z., Li, B., Gu, J., Muchnik, S., *et al.* (2019) A statistical framework for cross-tissue transcriptome-wide association analysis. Nat. Genet., 51, 568–576
30. Mancuso, N., Gayther, S., Gusev, A., Zheng, W., Penney, K. L., Kote-Jarai, Z., Eeles, R., Freedman, M., Haiman, C.Pasaniuc, B., *et al.* (2018) Large-scale transcriptome-wide association study identifies new prostate cancer risk regions. Nat. Commun., 9, 4079
31. Park, Y., Sarkar, A. K., Bhutani, K. and Kellis, M. (2017) Multi-tissue polygenic models for transcriptome-wide association studies. bioRxiv, 107623
32. Shi, X., Chai, X., Yang, Y., Cheng, Q., Jiao, Y., Huang, J., Yang, C. and Liu, J. (2019) A tissue-specific collaborative mixed model for jointly analyzing multiple tissues in transcriptome-wide association studies. bioRxiv, 789396
33. Mancuso, N., Freund, M. K., Johnson, R., Shi, H., Kichaev, G., Gusev, A. and Pasaniuc, B. (2019) Probabilistic fine-mapping of transcriptome-wide association studies. Nat. Genet., 51, 675–682
34. Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A. N., Knowles, D. A., Golan, D., Ermel, R., Ruusalepp, A., Quertermous, T., Hao, K., *et al.* (2019) Opportunities and challenges for transcriptome-wide association studies. Nat. Genet., 51, 592–599
35. Barbeira, A. N., Dickinson, S. P., Bonazzola, R., Zheng, J., Wheeler, H. E., Torres, J. M., Torstenson, E. S., Shah, K. P., Garcia, T., Edwards, T. L., *et al.* (2018) Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat. Commun., 9, 1825
36. Ference, B. A., Robinson, J. G., Brook, R. D., Catapano, A. L., Chapman, M. J., Neff, D. R., Voros, S., Giugliano, R. P., Davey Smith, G., Fazio, S., *et al.* (2016) Variation in PCSK9 and HMGCR and risk of cardiovascular disease and diabetes. N. Engl. J. Med., 375, 2144–2153
37. Helgadottir, A., Gretarsdottir, S., Thorleifsson, G., Hjartarson, E., Sigurdsson, A., Magnusdottir, A., Jonasdottir, A., Kristjansson, H., Sulem, P., Oddsson, A., *et al.* (2016) Variants with large effects on blood lipids and the role of cholesterol and triglycerides in

coronary disease. Nat. Genet., 48, 634–639
38. Pingault, J.-B., O'Reilly, P. F., Schoeler, T., Ploubidis, G. B., Rijsdijk, F. and Dudbridge, F. (2018) Using genetic data to strengthen causal inference in observational research. Nat. Rev. Genet., 19, 566–580
39. Zheng, J., Baird, D., Borges, M.-C., Bowden, J., Hemani, G., Haycock, P., Evans, D. M. and Smith, G. D. (2017) Recent developments in Mendelian randomization studies. Curr. Epidemiol. Rep., 4, 330–345
40. Haycock, P. C., Burgess, S., Wade, K. H., Bowden, J., Relton, C. and Davey Smith, G. (2016) Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. Am. J. Clin. Nutr., 103, 965–978
41. Lawlor, D. A. (2016) Commentary: Two-sample Mendelian randomization: opportunities and challenges. Int. J. Epidemiol., 45, 908–915
42. Bowden, J., Davey Smith, G. and Burgess, S. (2015) Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. Int. J. Epidemiol., 44, 512–525
43. Bowden, J., Davey Smith, G., Haycock, P. C. and Burgess, S. (2016) Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. Genet. Epidemiol., 40, 304–314
44. Smith, G. D. and Ebrahim, S. (2003) 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? Int. J. Epidemiol., 32, 1–22
45. Burgess, S., Small, D. S. and Thompson, S. G. (2017) A review of instrumental variable estimators for Mendelian randomization. Stat. Methods Med. Res., 26, 2333–2355
46. Burgess, S., Butterworth, A. and Thompson, S. G. (2013) Mendelian randomization analysis with multiple genetic variants using summarized data. Genet. Epidemiol., 37, 658–665
47. Burgess, S., Dudbridge, F. and Thompson, S. G. (2016) Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. Stat. Med., 35, 1880–1906
48. Burgess, S. and Thompson, S. G. (2011) Bias in causal estimates from Mendelian randomization studies with weak instruments. Stat. Med., 30, 1312–1323
49. Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. J. R. Stat. Soc. B, 58, 267–288
50. Hoerl, A. E. and Kennard, R. W. (2000) Ridge regression: biased estimation for nonorthogonal problems. Technometrics, 42, 80–86
51. Guan, Y. and Stephens, M. (2011) Bayesian variable selection regression for genome-wide association studies and other large-scale problems. Ann. Appl. Stat., 5, 1780–1815
52. Boyle, E. A., Li, Y. I. and Pritchard, J. K. (2017) An expanded view of complex traits: from polygenic to omnigenic. Cell, 169, 1177–1186
53. Yang, C., Wan, X., Lin, X., Chen, M., Zhou, X. and Liu, J. (2019) CoMM: a collaborative mixed model to dissecting genetic contributions to complex traits by leveraging regulatory information. Bioinformatics, 35, 1644–1652
54. Yang, Y., Shi, X., Jiao, Y., Huang, J., Chen, M., Zhou, X., Sun, L., Lin, X., Yang, C., Liu, J. (2020) CoMM-S^2: a collaborative mixed model using summary statistics in transcriptome-wide association studies. Bioinformatics, 36, 2009–2016
55. Hemani, G., Bowden, J. and Davey Smith, G. (2018) Evaluating the potential role of pleiotropy in Mendelian randomization studies. Hum. Mol. Genet., 27, R195–R208
56. Verbanck, M., Chen, C.-Y., Neale, B. and Do, R. (2018) Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. Nat. Genet., 50, 693–698
57. Park, Y., Sarkar, A. K., He, L., Davila-Velderrain, J., De Jager, P. L. and Kellis, M. (2017) A Bayesian approach to mediation analysis predicts 206 causal target genes in Alzheimer's disease. bioRxiv, 219428
58. Burgess, S. and Thompson, S. G. (2017) Interpreting findings from Mendelian randomization using the MR-Egger method. Eur. J. Epidemiol., 32, 377–389
59. Dai, J. Y., Peters, U., Wang, X., Kocarnik, J., Chang-Claude, J., Slattery, M. L., Chan, A., Lemire, M., Berndt, S. I., Casey, G., *et al.* (2018) Diagnostics for pleiotropy in Mendelian randomization studies: global and individual tests for direct effects. Am. J. Epidemiol., 187, 2672–2680
60. Qi, G. and Chatterjee, N. (2019) Mendelian randomization analysis using mixture models for robust and efficient estimation of causal effects. Nat. Commun., 10, 1941
61. Berzuini C, Guo H, Burgess S, Bernardinelli L. (2020) A Bayesian approach to Mendelian randomization with multiple pleiotropic variants. 2018. Biostatistics, 21, 86–101
62. Li, S. (2017) Mendelian randomization when many instruments are invalid: hierarchical empirical Bayes estimation. ArXiv, 170601389
63. Barfield, R., Feng, H., Gusev, A., Wu, L., Zheng, W., Pasaniuc, B. and Kraft, P. (2018) Transcriptome-wide association studies accounting for colocalization using Egger regression. Genet. Epidemiol., 42, 418–433
64. Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M. and Lin, X. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet., 89, 82–93
65. Li, B. and Leal, S. M. (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am. J. Hum. Genet., 83, 311–321
66. Madsen, B. E. and Browning, S. R. (2009) A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet., 5, e1000384
67. Price, A. L., Kryukov, G. V., de Bakker, P. I., Purcell, S. M., Staples, J., Wei, L.-J. and Sunyaev, S. R. (2010) Pooled association tests for rare variants in exon-resequencing studies. Am. J. Hum. Genet., 86, 832–838
68. Zhou, X. (2017) A unified framework for variance component estimation with summary statistics in genome-wide association studies. Ann. Appl. Stat., 11, 2027–2051
69. Schork, N. J., Murray, S. S., Frazer, K. A. and Topol, E. J. (2009) Common vs. rare allele hypotheses for complex diseases. Curr. Opin. Genet. Dev., 19, 212–219

70. Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H. and Nadeau, J. H. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nat. Rev. Genet., 11, 446–450

71. Price, A. L., Helgason, A., Thorleifsson, G., McCarroll, S. A., Kong, A. and Stefansson, K. (2011) Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. PLoS Genet., 7, e1001317