Quantitative Biology 2021, 9(2): 141–150
https://doi.org/10.15302/J-QB-020-0228

REVIEW

Transcriptome wide association studies: general framework and methods

Yuhan Xie[1], Nayang Shan[2,3], Hongyu Zhao[1], Lin Hou[2,3,4,*]

[1] Department of Biostatistics, School of Public Health, Yale University, New Haven, CT 06510, USA
[2] Center for Statistical Science, Tsinghua University, Beijing 100084, China
[3] Department of Industrial Engineering, Tsinghua University, Beijing 100084, China
[4] MOE Key Laboratory of Bioinformatics, School of Life Sciences, Tsinghua University, Beijing 100084, China
* Correspondence: houl@tsinghua.edu.cn

Received April 25, 2020; Revised June 8, 2020; Accepted July 9, 2020

***Background*: Genome-wide association studies (GWAS) have succeeded in identifying tens of thousands of genetic variants associated with complex human traits during the past decade, however, they are still hampered by limited statistical power and difficulties in biological interpretation. With the recent progress in expression quantitative trait loci (eQTL) studies, transcriptome-wide association studies (TWAS) provide a framework to test for gene-trait associations by integrating information from GWAS and eQTL studies.**

***Results*: In this review, we will introduce the general framework of TWAS, the relevant resources, and the computational tools. Extensions of the original TWAS methods will also be discussed. Furthermore, we will briefly introduce methods that are closely related to TWAS, including MR-based methods and colocalization approaches. Connection and difference between these approaches will be discussed.**

***Conclusion*: Finally, we will summarize strengths, limitations, and potential directions for TWAS.**

Keywords: TWAS; gene imputation; gene-trait association test; eQTL studies; GWAS

Author summary: Transcriptome-wide association studies (TWAS) provide an important framework to test for gene-trait associations by integrating information from GWAS and eQTL studies. In this review, we systematically review the general framework and methods of transcriptome-wide association studies, and discuss their strengths, limitations, and potential future directions.

INTRODUCTION

Genome-wide association studies (GWAS) have been very successful in identifying genetic variants associated with complex human traits. However, these studies exhibit limited statistical power due to the polygenic effect of genetic variants as well as small effect sizes. Moreover, the underlying molecular mechanisms between genetic variation and these traits are not well understood [1,2]. One way that genetic variants can affect traits is through modulating gene expression [3]. The importance of gene expression regulation has been shown by expression quantitative trait loci (eQTL) studies. Such relationships between gene expression and the traits can be studied by measuring both genetic variation and transcriptome data in the same study subjects. Unfortunately, such designs might not be cost effective. For some traits, it is hard to retrieve tissue samples to measure transcriptome levels [3–5]. Thus, we have a limited number of datasets with matched GWAS and transcriptome data.

Transcriptome-wide association studies (TWAS) shed novel insights into complex disease mechanisms when transcriptome data are not available in GWAS samples. These studies integrate information from GWAS and eQTL studies to test for gene-trait associations and prioritize potential causal genes that mediate variants effects on the traits. TWAS can reduce the multiple testing

burden (~number of genes) compared to testing traits with genetic variations (~number of SNPs), and is a powerful tool to identify susceptible risk genes in complex human traits [1,4].

Briefly, there are three steps in a TWAS framework (see Fig. 1). In the first step, an imputation model is trained in a reference panel, for which matched genotype and transcriptome data are available. Usually, the imputation model is a linear model that takes individual level genotype as input, and outputs the predicted gene expression level, which corresponds to the genetically regulated component of gene expression [1]. The number of predictable genes is limited by the power of the reference panel. Publicly available resources of reference panels are summarized in Table 1. In particular, GTEx provides the most comprehensive database, with adequate power to predict gene expression in 54 tissue sites (The number is based on GTEx v8, the number of tissues used is method-dependent). The imputation models can be trained either by using single tissue or by leveraging information from multiple tissues.

In the second step the single-tissue/cross-tissue model is applied to a GWAS cohort of interest to predict (or impute) gene expression. In the third step a gene-trait association test is conducted. The association can be tested either in a relevant tissue or in a cross-tissue manner. The connections between different gene-based tests will be further illustrated in the discussion section.

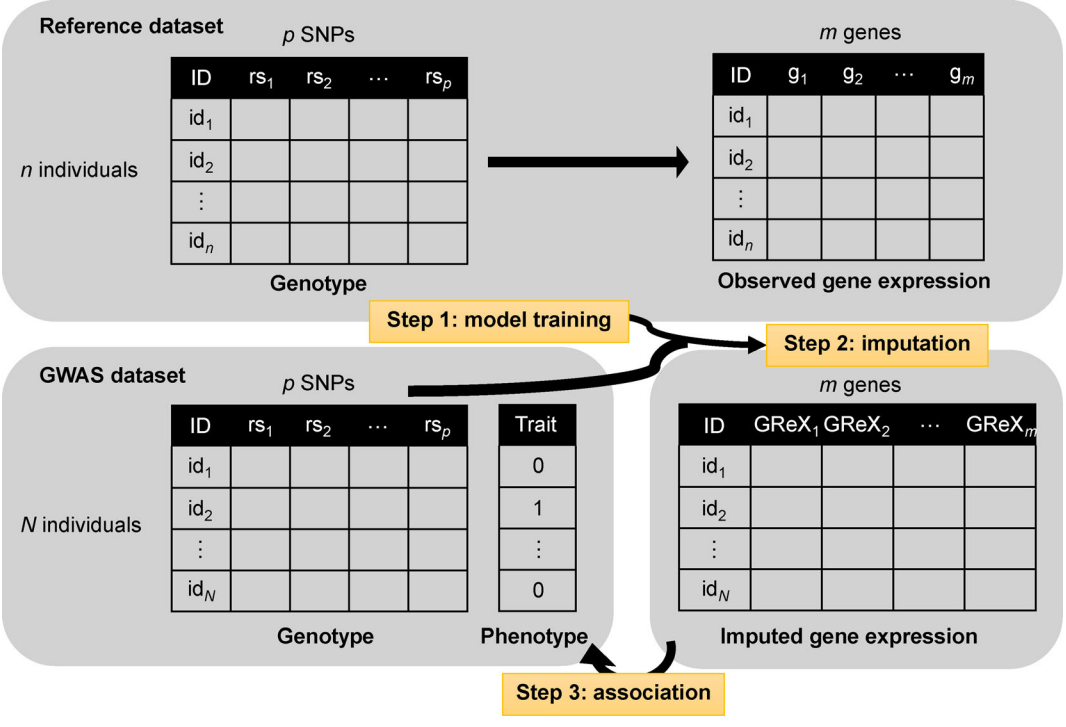

Figure 1. General framework for transcriptome wide association studies.

Table 1 Reference panels for eQTL studies

Database	Abbreviation	Tissue/Cell types	Sample size	Ref.
Genotype-Tissue Expression V8	GTEx v8	54 human tissues	948 individuals	[6]
Genetic European Variation in Health and Disease	GEUVADIS	Lymphoblastoid cell lines	462 individuals	[7]
Depression Genes and Network	DGN	Whole blood	922 individuals	[8]
Netherlands Twin Register	NTR	Peripheral blood	2,752 individual twins	[9]
Metabolic Syndrome in Men	METSIM	Adipose tissue	795 individuals	[10]
North American Brain Expression Consortium	NABEC	Cortex	69 individuals	dbGAP: phs001300.v1.p1
CommonMind Consortium	TCGA	Tumors	980 individuals	[11]

There are two types of cross-tissue tests. The first one constructs a test statistic that combines information from single-tissue tests [12]. The other one directly tests trait of interest with predicted expression in multiple tissues [13].

Alternatively, some approaches [2,4,12–16] combined the second step and the third step, which made use of summary statistics from large GWAS cohorts instead of individual-level data. Such approaches leveraged the publicly available GWAS summary statistics from large GWAS repositories such as the GWAS Catalog [17]. However, these summary-based methods are lack of information on covariance structures of genetic variants (*i.e.*, linkage disequilibrium (LD)), thus we need to use the training set or a population reference set such as the 1000 Genomes Project [18] to estimate this information in the study population. Furthermore, the GWAS summary statistics and the reference transcriptome data should come from independent cohorts. We underscore the importance of independence here as GWAS summary statistics oftentimes include multiple cohorts that may have overlaps with the reference panels.

In 2019, Wainberg *et al*. [19] gave a perspective review on TWAS. They highlighted two challenges in interpreting GWAS results, co-regulation and tissue bias, suggested best practices among current studies, and discussed future opportunities for TWAS. In contrast, our review concentrates in the methodological aspect of TWAS framework, and summarized and discussed the extensions of TWAS methods.

METHODS AND TOOLS

In this section, we review the methods and tools to conduct TWAS analysis (see Table 2). To avoid ambiguity, we refer to the studies under the three-step framework (see introduction) as TWAS approaches, and to distinguish the work of Gusev *et al*. as FUSION [4]. The TWAS framework originates from PrediXcan [1] and FUSION [4]. The major difference between the two methods is that PrediXcan uses individual-level data as input while FUSION can take both individual-level data and summary-level GWAS results.

Gamazon *et al.* firstly introduced the concept of genetically regulated expression (GReX), which excluded trait-altered components as well as non-genetic components such as those affected by environmental factors [1]. Based on this concept, the PrediXcan framework focuses on estimating GReX and correlating them with a phenotype of interest. Specifically, PrediXcan adopts a penalized regression model, Elastic-Net [30], to train weights between observed gene expressions and cis-SNPs using a reference transcriptome dataset. Next, GReX is imputed with individual-level genotype data and the trained weights. Then, a gene-trait association test is performed with imputed GReX and a phenotype of interest.

FUSION adopts a Bayesian sparse linear mixed model (BSLMM) [31] to train weights between observed gene expressions and cis-SNPs with a reference dataset and tests for the association between predicted gene expres-

Table 2 A list of tools for integration of eQTL and GWAS

Name	Website	Requirement	Ref.
FUSION	http://gusevlab.org/projects/fusion/	R	[4]
PrediXcan	https://github.com/hakyimlab/PrediXcan	Python, R	[1]
S-PrediXcan	https://github.com/ hakyimlab/MetaXcan	Python, R	[2]
MultiXcan/S-MultiXcan	https://github.com/ hakyimlab/MetaXcan	Python, R	[13]
UTMOST	https://github.com/Joker-Jerome/ UTMOST/	Python, R	[12]
fQTL	https://ypark.github.io/fqtl/	R(support C++ 14)	[14]
CoMM/CoMM-S2	https://github.com/gordonliu810822/CoMM	R	[15,20]
TF-TWAS	https://github.com/TangYiChing/TF-TWAS	GNU bash, R, Python	[21]
EpiXcan	https://bitbucket.org/roussoslab/epixcan/src/master/	R, Python	[22]
FOCUS	https://github.com/bogdanlab/focus/	Python	[23]
TIGAR	https://github.com/yanglab-emory/TIGAR	BGZIP, TABIX, Python	[16]
moloc	https://github.com/clagiamba/moloc	R	[24]
QTLMatch	https://cran.r-project.org/web/packages/coloc/	R	[25]
COLOC	https://cran.r-project.org/web/packages/coloc/	R	[26]
fastENLOC	https://github.com/xqwen/fastenloc	C++	[27]
eCAVIAR	http://genetics.cs.ucla.edu/caviar/	C++	[28]
Sherlock	http://sherlock.ucsf.edu	Web page	[29]

sion and phenotypes of interest. BLSMM combines Bayesian variable selection (BVSR) [32] and linear mixed model (LMM) [33] with the assumption of a normal mixture prior [16,31]. The summary-based model is implemented in the publicly available software FUSION.

Following PrediXcan and FUSION, other methods extend TWAS by making modifications on each step of the TWAS framework. Here, we discuss these methods in details.

EXTENSIONS IN IMPUTATION MODELS

Cross-tissue model

Given the challenge of training a robust and accurate gene expression model in a single tissue, fQTL [14] and UTMOST [12] model gene expressions in multiple tissues jointly. fQTL adopts a cross-tissue Bayesian multivariate linear regression model that hierarchically decomposes the effect size of cis-SNPs into SNP-specific and tissue-specific effects. In addition, fQTL used the spike-and-slab prior for variable selection and solved the model using an optimization method with fast convergence called stochastic variation inference (SVI) [34]. UTMOST builds a cross-tissue prediction model by using a multivariate regression that penalizes on both within-tissue effects and cross-tissue effects. Compared with the Elastic-Net model on single tissue, UTMOST achieved improved prediction accuracy.

Nonparametric model

To avoid limitations of parametric prediction models, Nagpal *et al.* utilized a nonparametric Bayesian model, latent Dirichlet (DPR) process regression [35] to model the distribution of cis-SNP effect sizes [16]. Compared to traditional parametric methods, this model is data driven and models the complex genetic architecture of gene expression. Actually, Elastic-Net [30] and BSLMM [31] can be considered as special cases of DPR. Further, Nagpal *et al.* employed a variational Bayesian algorithm, which approximates MCMC with less computational burden [36], to obtain posterior estimates of the effect sizes. In real application of ROS/MAP data [37–39], Nagpal *et al*. showed the DPR model can predict more genes than PrediXcan. Their method is implemented in a publicly available software tool called TIGAR.

EXTENSIONS IN GENE-TRAIT ASSOCIATION TEST

Cross-tissue test

In addition to training a cross-tissue imputation model, another idea to integrate information across different tissues is to modify the testing step in the TWAS framework. MultiXcan adopted multivariate regression to test the association between phenotypes of interest and predicted gene expression in multiple tissues [13]. They used the first k principal components of the predicted gene expressions in multiple tissues to address the multi-collinearity problem. Applied to 222 traits studied in UK Biobank, MultiXcan detected more associations than the single-tissue based PrediXcan in 103 traits. In their simulation study, PrediXcan exhibited more benefits under a setting in which the single causal tissue is known and the predicted gene expression captures the regulatory mechanism well in that tissue.

Rather than performing the test jointly in MultiXcan, UTMOST combines the testing results in each tissue to cross-tissue z-scores by using a generalized Berk-Jones (GBJ) test [40]. Compared with single-tissue methods including PrediXcan and FUSION, UTMOST showed a remarkable improvement in the accuracy of gene expression imputation and the number of identified significant genes in 50 GWAS of complex human traits. Li *et al.* investigated the influence of tissue context on gene prioritization in TWAS approaches, using PrediXcan as a representative of single-tissue TWAS and UTMOST as a representative of cross-tissue TWAS [41]. Excluding the overlaps of significant genes identified, they found PrediXcan typically identified genes that are more unique in a single tissue, while UTMOST tended to prioritize gene-trait association in multiple tissues as well as in disease-related tissues.

Probabilistic model

The uncertainties imposed in the imputation step might have an impact in the test of gene-trait association. Bhutani *et al.* discussed possible sources of the uncertainties, including model misspecification and biased parameter estimation, and developed a two-stage Bayesian regression model, BAY-TS, to address the limitation [42]. Essentially, the method considered the posterior distribution of imputed gene expression instead of single point estimate to incorporate the uncertainty. They compared their method with PrediXcan and demonstrated that BAY-TS was able to detect more genes in real data analysis.

Yang *et al.* jointly modeled the imputation step and testing step by utilizing a mixed model that takes individual-level GWAS data and can be performed tissue-by-tissue [20]. Further, they derived a computationally efficient and stable algorithm named PX-EM [43]. In real data analyses of 25 traits in NFBC1966 (dbGAP at http://ncbi.nlm.nih.gov/gap (Study Accession: phs000276.v1.p1) and GERA studies (dbGAP at http://

ncbi.nlm.nih.gov/gap (Study Accession: phs000674.v2.p2), they observed that CoMM identified more genes than PrediXcan, which suggests the necessity and benefit of modeling uncertainties.

Summary-based model

Many TWAS approaches can be adapted to take summary-level GWAS data, in favor of leveraging the massive sample sizes in meta-analysis and making their methods more applicable when individual-level data are not available. After firstly adopting summary-level GWAS data in FUSION, a number of summary-based methods were developed successively. S-PrediXcan [2], S-MultiXcan [13], CoMM-S^2 [15] are the summary-based version of PrediXcan, S-MultiXcan and CoMM, respectively.

Barbeira *et al.* derived the test statistics of PrediXcan with summary-level data, and further showed that the performance of their method, S-PrediXcan, is in high concordance with that of PrediXcan [2]. Along with the work, a general tool named MetaXcan built on S-PrediXcan was proposed. They also showed that FUSION based on summary statistics is equivalent to S-PrediXcan in terms of the Z-score of gene-trait association tests up to a scaling factor, related to the proportion of variance explained by a SNP's allelic dosage and the proportion of variation explained by gene expression. In practice, this factor is very close to 1, leading to similar results between the two methods.

Yang *et al.* has upgraded CoMM to CoMM-S^2 [15]. The basic idea behind CoMM-S^2 is similar to CoMM, but the upgraded framework accommodates individual-level eQTL data and summary-level GWAS data. An efficient algorithm based on variational Bayesian EM was proposed accordingly to fit the model. The performance of CoMM-S^2 is comparable to CoMM, and CoMM-S^2 performs better than CoMM and S-PrediXcan when cellular heritability is low.

SPU test

Xu *et al.* generalized the gene-based test in the TWAS framework [44]. They first showed the test of individual-level data in PrediXcan and FUSION are special cases of a general testing framework, weighted sum test of a generalized linear model [44,45]. Then extended the sum test approach to adaptive sum of powered score (SPU) test, which assigns different power (an integer) to the score vector in a data adaptive manner. Since there is no uniformly powerful test under different situations, the adaptive SPU tests (aSPU) gained more power by calculating the minimal P-value across a set of SPU tests of different weighting schemes. Their method can also be applied to both individual level and summary-level GWAS data.

INTEGRATION OF ANNOTATION DATA

Efforts have been made to integrate genomic annotations to improve gene prediction accuracy in the traditional framework of TWAS [21,22,46–48].

EpiXcan [22] extended TWAS by integrating epigenomic data to increase accuracy in the imputation step. It leverages epigenomic information (DNA methylation histone modification and chromatin accessibility) [46] to learn the prior of SNPs to be regulatory in a Bayesian hierarchical model, and then used the priors to derive penalty factors for SNPs, which is subsequently employed in the weighted Elastic Net approach to predict gene expression. TF-TWAS concerned the effect of transcription-factor (TF) polymorphism on transcription, and integrated this information by including SNPs within the coding regions of TFs in the imputation model, in addition to the cis-SNPs used in PrediXcan. Three models of how the TF polymorphisms affect gene expression were explored. The method was tested in four tissues, and identified 48 genes with improved R^2 comparing to PrediXcan.

Exploration of three-dimensional structure of the human genome enables studies to integrate information such as chromatin interaction data into TWAS. Wu *et al.* first identified the promoter regions and the main body of genes based on a chromatin state model, and then defined the enhancer region based on interactions with promoter regions [47]. Analogous to modeling cis-SNPs in traditional TWAS [1,4], they restricted candidate SNPs in the imputation model to the enhancer-promoter region, which reduced the number of candidate SNPs and may improve prediction accuracy. They conducted gene-based tests and identified novels genes that provide complementary information to results from FUSION.

EXTENSION TO "X-WAS"

The TWAS framework can be naturally extended to "X-WAS", which accommodates other intermediate phenotypes that can be imputed by GWAS data.

For example, the role of DNA methylation, an epigenetic mechanism that is essential for both genomic processes and normal development in human beings [49,50], has been studies in the MWAS framework in schizophrenia [51], bipolar disorder [51], Parkinson's disease [52], and Alzheimer's disease [53]. Methods of methylation imputation and methylation-trait association test are also analogical to TWAS methods. Similarly, Xu *et al.* conducted IWAS, imaging-wide association study (IWAS), which used gray matter volumes of several brain

regions of interest as imaging intermediate phenotypes [54]. The proposed IWAS framework is applicable to both individual-level and summary-level GWAS data, and other imaging intermediate phenotypes.

OTHER RELATED METHODS

Essentially, TWAS approaches integrate GWAS data with eQTL studies to identify disease/trait associated genes. Besides TWAS, there are other methods to integrate information from both GWAS and eQTL studies. Mendelian randomization (MR) based methods and colocalization methods are along this line of work.

MR-based methods

MR-based analysis uses genetic variants as instrumental variables to test for the association between gene expression and complex traits. Oftentimes, MR-base methods have more assumptions than TWAS, which leverage cis-eQTLs as instrument variables to infer gene-trait associations. The method is based on observed gene expressions, which do not involve gene expression prediction like TWAS.

Zhu *et al.* utilizes the most significant associated cis-eQTLs as the instrument variable to detect target genes associated with a complex trait in SMR [55]. Porcu *et al.* proposed a multivariable (multi-gene) multi-instrument MR approach named Transcriptome-Wide Mendelian Randomization (TWMR), which can reduce bias due to pleiotropic effects [56]. Compared to TWAS, MR-based methods (*e.g.*, SMR and TWMR) are more sensitive to co-regulation due to violation of model assumptions. For SMR, genetic variants and the conditioned outcome is assumed to be independent, while InSIDE (Instrument Strength Independent of Direct Effect) was assumed in TWMR.

Barbeira *et al.* systematically evaluated and compared SMR and S-PrediXcan. They showed that SMR P-values tend to be less significant compared with S-PrediXcan, and discussed possible explanations [2]. First, only using a single SNP as the instrument in SMR will render the method less powerful. Second, it is inherent in the SMR framework that the significance is limited by the significance of the eQTL association. Noticeably, the derived statistic in SMR is not well calibrated because the chi-square approximation is only valid in two extreme cases: when the eQTL association is much more significant than the GWAS association or when the GWAS association is much more significant than the eQTL association [2]. Similar limitations apply to TWMR. In addition, TWMR excludes pleiotropic SNPs to avoid estimation bias, the method could suffer from power loss when more SNPs are excluded owing to evidence for mild pleiotropy with increasing GWAS study size [56].

Colocalization

Colocalization analysis identifies genetics variants that are significant in both GWAS and eQTL studies. Unlike TWAS, colocalization does not perform gene expression prediction and gene-trait association tests. Instead, colocalized SNPs are the unit of their concern. However, FUSION is conceptually similar to a test for colocalization of signal between expression and a complex trait [4,26,57]. More importantly, these colocalization methods can serve as a post filtering step to mitigate issues with LD-contamination after TWAS-type gene-level associations (see Discussion).

Nica *et al*. proposed an empirical method, Regulatory Trait Concordance (RTC), to reveal that significant SNPs in GWAS are colocalized with cis- or trans-eQTLs after accounting for LD structure [58]. He *et al.* proposed a Bayesian statistical method, Sherlock, to match the genetic signature of a specific gene with GWAS signals [29]. Sherlock utilizes information in both cis- and trans-SNPs to constitute the genetic signature while accounting for the uncertainty of LD using a block-level Bayes factor.

Plagnol *et al.* proposed a statistical procedure, QTLMatch, to detect whether the colocalization of GWAS and eQTL signals is coincidental by testing the null hypothesis of a sole, causal variant for the GWAS and the eQTL signals [25]. COLOC, a Bayesian method expanded from QTLMatch, considers five scenarios for the colocalization of GWAS and eQTL signals, and subsequently provides the posterior probabilities for the five conditions. From simulation results [4], FUSION and COLOC had similar power under the scenario with a single typed causal variant and COLOC has slightly lower power at small GWAS sizes. However, FUSION could have better performance when the causal variant was untyped or in the presence of allelic heterogeneity, which is likely due to the fact that FUSION explicitly models LD to better capture untyped variants. Multiple-trait-coloc (moloc), extends the COLOC framework to integrate GWAS summary statistics and multiple molecular QTL data to provide evidence how the variants are shared across multiple traits (*e.g.*, complex trait, gene expression and DNA methylation).

eCAVIAR (eQTL and GWAS Causal Variant Identification in Associated Regions) [28] provides the colocalization posterior probability (CLPP) for each variant in a locus for a specific eGene (eGenes are genes that have at least one significant variant (P-value $<$ 1e–5 when corrected for multiple hypothesis) in a tissue to indicate that the variant is causal in both GWAS and eQTL studies. CLPP requires the marginal statistics from GWAS and

eQTL studies while accounting for LD structure of genetic variants in a locus. eCAVIAR allows for multiple variants to be causal in a single locus, which is more accurate than COLOC and RTC in the presence of allelic heterogeneity.

Wen *et al.* developed a unified inference framework, enloc, for enrichment analysis, fine-mapping and colocalized association testing [27]. Their model first estimates the enrichment level of molecular QTLs in the GWAS signals. Then fine-mapping of GWAS signals is conducted to construct the SNP-level priors accounting for the uncertainty of the association status of molecular QTLs. The SNP-level colocalization probability (SCP) and regional colocalization probability (the sum of the SCPs of correlated SNPs within an LD block that harbors a GWAS signal) can be obtained from the previous steps. COLOC and eCAVIAR can be viewed as special cases of enloc, which eliminates the subjective prior specification.

DISCUSSION

In this review, we systematically summarize the development of TWAS, which integrates GWAS and eQTL data under the assumption that gene expression mediates the association between genetic variants and traits. Compared with GWAS, TWAS reduces the testing burden and provides straightforward biological interpretations.

A pipeline for a TWAS analysis should be carefully chosen based on the available data and the trait of interest. Systematic evaluation and comparison of TWAS approaches are needed to guide selection of tools. In particular, Wainberg *et al.* [19] suggested using fine mapping methods in interpreting TWAS results. Moreover, they raised concerns regarding to selection of single tissue and cross-tissue methods. In general, cross-tissue analysis may identify more potential causal genes by leveraging information across tissues, and single-tissue analysis may be more powerful when the mechanism behind the trait of interest and tissues are explicit. Comprehensive comparisons between PrediXcan and UTMOST can be found in [41]. In addition, functional annotation data may also boost statistical power and provide complementary information to prioritize causal genes.

In recent years, TWAS have emerging applications on identifying novel candidate genes of complex human diseases such as cancers [47,59–62], neuropsychiatric diseases [41,63–66] (see Supplementary Table S1). In addition to the macro-phenotypes, micro-phenotypes such as chromatin status has also been considered. Gusev *et al.* defined chromatin status with individual level epigenetics data, and identified genes associated with chromatin status with TWAS framework [48]. After integrating the results from TWAS of schizophrenia and chromatin phenotypes, they identified overlapping genes that bring insights to the regulatory mechanism.

Despite the success of TWAS, there are also arising concerns in its further application. Low prediction accuracy of gene expression will lead to insufficient testing power, and may cause spurious conclusions. Nonetheless, a standard of informative prediction accuracy is not definitive [59]. First, the choice of the reference dataset could be problematic, and the inherent limitations of prediction model arise from the reference transcriptome data. For single tissue TWAS approaches, there is not a principled approach to selecting the tissue to be used. Furthermore, TWAS may suffer low prediction accuracy when data are not available in the right tissue for the trait of interest [5] or sample sizes of disease relevant tissues are relatively small [60,61]. Second, there is lack of assessment whether the prediction models learned in one population can be transferred to another population. Keys *et al.* evaluated the accuracy of trans-ancestral prediction of gene expression [67]. The prediction models were trained in European ancestral panels and applied to predict gene expression in African American individuals. As expected, the R^2 levels were higher in Europeans than that of African-Americans, underscoring the importance of ancestry specific reference eQTL panels. Third, most TWAS methods assume linear models for the prediction of GReX, however, prediction accuracy will be affected when the relationship between GReX and the SNPs is nonlinear, or when trans-regulation plays an important part in modulating gene expression [61]. A straightforward solution is to incorporate trans-eQTLs in the prediction step [68]. However, it is still challenging to decide how to appropriately model trans-eQTLs, as they are more difficult to detect because of a much heavier multiple testing burden and smaller effect sizes than cis-eQTLs [69,70]. Also, trans-eQTLs are more likely to be tissue-specific [68,71,72].

In addition to poor prediction accuracy, interpretation of TWAS associations could also be complicated by confounding factors, especially when SNPs affect the phenotype of interest in non-regulatory mechanisms [73]. Due to LD in the genome, co-regulation in TWAS may lead to prioritization of non-causal genes. Mancuso *et al.* discussed four scenarios of co-regulation, and proposed a fine-mapping method, FOCUS, to address this problem [19,23].

SUPPLEMENTARY MATERIALS

The supplementary materials can be found online with this article at https://10.15302/J-QB-020-0228.

ACKNOWLEDGEMENTS

We thank Zhaolong Yu for suggestions and Michael Farruggia for English language polishing. L. H. acknowledges the following fundings: the

National Natural Science Foundation of China (No. 11601259) and Shanghai Municipal Science and Technology Major Project (No. 2017SHZDZX01). Y. X. and N. S. were supported in part by the China Scholarship Council, and H. Z. was supported in part by NIH grant R01GM122078, NSF grants DMS 1713120 and DMS 1902903.

COMPLIANCE WITH ETHICS GUIDELINES

The authors Yuhan Xie, Nayang Shan, Hongyu Zhao and Lin Hou declare that they have no conflict of interests.

This article is a review article and does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. 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
2. 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
3. Cloney, R. (2016) Integrating gene variation and expression to understand complex traits. Nat. Rev. Genet., 17, 194
4. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., 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
5. Mancuso, N., Shi, H., Goddard, P., Kichaev, G., Gusev, A. and Pasaniuc, B. (2017) Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. Am. J. Hum. Genet., 100, 473–487
6. 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
7. Lappalainen, T., Sammeth, M., Friedländer, M. R., Peter, P. A., 't Hoen, 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
8. 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
9. Boomsma, D. I., de Geus, E. J., Vink, J. M., Stubbe, J. H., Distel, M. A., Hottenga, J. J., Posthuma, D., van Beijsterveldt, T. C., Hudziak, J. J., Bartels, M., *et al.* (2006) Netherlands Twin Register: from twins to twin families. Twin Res. Hum. Genet., 9, 849–857
10. Laakso, M., Kuusisto, J., Stančáková, A., Kuulasmaa, T., Pajukanta, P., Lusis, A. J., Collins, F. S., Mohlke, K. L. and Boehnke, M. (2017) The metabolic syndrome in men study: a resource for studies of metabolic and cardiovascular diseases. J. Lipid Res., 58, 481–493
11. Hoffman, G. E., Bendl, J., Voloudakis, G., Montgomery, K. S., Sloofman, L., Wang, Y. C., Shah, H. R., Hauberg, M. E., Johnson, J. S., Girdhar, K., *et al.* (2019) CommonMind Consortium provides transcriptomic and epigenomic data for Schizophrenia and Bipolar Disorder. Sci. Data, 6, 180
12. 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
13. 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
14. Park, Y., Sarkar, A., Bhutani, K. and Kellis, M. (2017) Multi-tissue polygenic models for transcriptome-wide association studies. bioRxiv, 107623
15. Yang, Y., Shi, X., Jiao, Y., Huang, J., Chen, M., Zhou, X., Sun, L., Lin, X., Yang, C. and Liu, J. (2020) CoMM-S2: a collaborative mixed model using summary statistics in transcriptome-wide association studies. Bioinformatics, 36, 2009–2016
16. 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
17. MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., *et al.* (2017) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res., 45, D896–D901
18. Siva, N. (2008) 1000 Genomes project. Nat. Biotechnol., 26, 256
19. 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
20. 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
21. Tang, Y.-C. and Gottlieb, A. (2018) TF-TWAS: Transcription-factor polymorphism associated with tissue-specific gene expression. bioRxiv, 405936
22. Zhang, W., Voloudakis, G., Rajagopal, V. M., Readhead, B., Dudley, J. T., Schadt, E. E., Björkegren, J. L. M., Kim, Y., Fullard, J. F., Hoffman, G. E., *et al.* (2019) Integrative transcriptome imputation reveals tissue-specific and shared biological mechanisms mediating susceptibility to complex traits. Nat. Commun., 10, 3834
23. 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

24. Giambartolomei, C., Zhenli Liu, J., Zhang, W., Hauberg, M., Shi, H., Boocock, J., Pickrell, J., Jaffe, A. E., Pasaniuc, B., Roussos, P., *et al.* (2018) A Bayesian framework for multiple trait colocalization from summary association statistics. Bioinformatics, 34, 2538–2545
25. Plagnol, V., Smyth, D. J., Todd, J. A. and Clayton, D. G. (2009) Statistical independence of the colocalized association signals for type 1 diabetes and *RPS26* gene expression on chromosome 12q13. Biostatistics, 10, 327–334
26. Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C. and Plagnol, V. (2014) Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet., 10, e1004383
27. Wen, X., Pique-Regi, R. and Luca, F. (2017) Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. PLoS Genet., 13, e1006646
28. Hormozdiari, F., van de Bunt, M., Segrè, A. V., Li, X., Joo, J. W. J., Bilow, M., Sul, J. H., Sankararaman, S., Pasaniuc, B. and Eskin, E. (2016) Colocalization of GWAS and eQTL signals detects target genes. Am. J. Hum. Genet., 99, 1245–1260
29. He, X., Fuller, C. K., Song, Y., Meng, Q., Zhang, B., Yang, X. and Li, H. (2013) Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. Am. J. Hum. Genet., 92, 667–680
30. Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. J. R. Stat. Soc. B, 67, 301–320
31. Zhou, X., Carbonetto, P. and Stephens, M. (2013) Polygenic modeling with bayesian sparse linear mixed models. PLoS Genet., 9, e1003264
32. 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
33. Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., *et al.* (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet., 38, 203–208
34. Hoffman, M. D., Blei, D. M., Wang, C. and Paisley, J. (2013) Stochastic variational inference. J. Mach. Learn. Res., 14, 1303–1347
35. Zeng, P. and Zhou, X. (2017) Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. Nat. Commun., 8, 456
36. Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2017) Variational inference: A review for statisticians. J. Am. Stat. Assoc., 112, 859–877
37. Bennett, D. A., Schneider, J. A., Buchman, A. S., Barnes, L. L., Boyle, P. A. and Wilson, R. S. (2012) Overview and findings from the rush memory and aging project. Curr. Alzheimer Res., 9, 646–663
38. Ng, B., White, C. C., Klein, H. U., Sieberts, S. K., McCabe, C., Patrick, E., Xu, J., Yu, L., Gaiteri, C., Bennett, D. A., *et al.* (2017) An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. Nat. Neurosci., 20, 1418–1426
39. Bennett, D. A., Buchman, A. S., Boyle, P. A., Barnes, L. L., Wilson, R. S. and Schneider, J. A. (2018) Religious orders study and rush memory and aging project. J. Alzheimers Dis., 64, S161–S189
40. Sun, R. and Lin, X. (2017) Set-based tests for genetic association using the generalized berk-jones statistic. ArXiv, 171002469
41. Li, B., Veturi, Y., Bradford, Y., Verma, S. S., Verma, A., Lucas, A. M., Haas, D. W. and Ritchie, M. D. (2019) Influence of tissue context on gene prioritization for predicted transcriptome-wide association studies. Pac. Symp. Biocomput., 24, 296–307
42. Bhutani, K., Sarkar, A., Park, Y., Kellis, M. and Schork, N. J. (2017) Modeling prediction error improves power of transcriptome-wide association studies. bioRxiv, 108316
43. Liu, C., Rubin, D. B. and Wu, Y. N. (1998) Parameter expansion to accelerate em: The px-em algorithm. Biometrika, 85, 755–770
44. Xu, Z., Wu, C., Wei, P. and Pan, W. (2017) A powerful framework for integrating eQTL and GWAS summary data. Genetics, 207, 893–902
45. Pan, W. (2009) Asymptotic tests of association with multiple SNPs in linkage disequilibrium. Genet. Epidemiol., 33, 497–507
46. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., *et al.* (2015) Integrative analysis of 111 reference human epigenomes. Nature, 518, 317–330
47. Wu, L., Shi, W., Long, J., Guo, X., Michailidou, K., Beesley, J., Bolla, M. K., Shu, X. O., Lu, Y., Cai, Q., *et al.* (2018) A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. Nat. Genet., 50, 968–978
48. Gusev, A., Mancuso, N., Won, H., Kousi, M., Finucane, H. K., Reshef, Y., Song, L., Safi, A., McCarroll, S., Neale, B. M., *et al.* (2018) Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. Nat. Genet., 50, 538–548
49. Ardlie, K. G., Deluca, D. S., Segre, A. V., Sullivan, T. J., Young, T. R., Gelfand, E. T., Trowbridge, C. A., Maller, J. B., Tukiainen, T., Lek, M., *et al.* (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science, 348, 648–660
50. DNA methylation. Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title = DNA_methylation&oldid = 984350315. Accessed: April 23, 2020
51. Han, S., Lin, Y., Wang, M., Goes, F. S., Tan, K., Zandi, P., Hyde, T., Weinberger, D. R., Potash, J. B., Kleinman, J. E., *et al.* (2018) Integrating brain methylome with gwas for psychiatric risk gene discovery. bioRxiv, 440206
52. Rawlik, K., Rowlatt, A. and Tenesa, A. (2016) Imputation of DNA methylation levels in the brain implicates a risk factor for Parkinson's disease. Genetics, 204, 771–781
53. Nazarian, A., Yashin, A. I. and Kulminski, A. M. (2018)

Methylation-wide association analysis reveals aim2, dguok, gnai3, and st14 genes as potential contributors to the Alzheimer's disease pathogenesis. bioRxiv, 322503

54. Xu, Z., Wu, C. and Pan, W., and the Alzheimer's Disease Neuroimaging Initiative. (2017) Imaging-wide association study: Integrating imaging endophenotypes in GWAS. Neuroimage, 159, 159–169
55. 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
56. Porcu, E., Rüeger, S., Lepik, K., Santoni, F. A., Reymond, A. and Kutalik, Z., the eQTLGen Consortium, and the BIOS Consortium. (2019) Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. Nat. Commun., 10, 3300
57. Lee, D., Williamson, V. S., Bigdeli, T. B., Riley, B. P., Fanous, A. H., Vladimirov, V. I. and Bacanu, S. A. (2015) JEPEG: a summary statistics based tool for gene-level joint testing of functional variants. Bioinformatics, 31, 1176–1182
58. Nica, A. C., Montgomery, S. B., Dimas, A. S., Stranger, B. E., Beazley, C., Barroso, I. and Dermitzakis, E. T. (2010) Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genet., 6, e1000895
59. Hoffman, J. D., Graff, R. E., Emami, N. C., Tai, C. G., Passarelli, M. N., Hu, D., Huntsman, S., Hadley, D., Leong, L., Majumdar, A., *et al.* (2017) Cis-eQTL-based trans-ethnic meta-analysis reveals novel genes associated with breast cancer risk. PLoS Genet., 13, e1006690
60. Lu, Y., Beeghly-Fadiel, A., Wu, L., Guo, X., Li, B., Schildkraut, J. M., Im, H. K., Chen, Y. A., Permuth, J. B., Reid, B. M., *et al.* (2018) A transcriptome-wide association study among 97,898 women to identify candidate susceptibility genes for epithelial ovarian cancer risk. Cancer Res., 78, 5419–5430
61. Mancuso, N., Gayther, S., Gusev, A., Zheng, W., Penney, K. L., Kote-Jarai, Z., Eeles, R., Freedman, M., Haiman, C. and Pasaniuc, B., and the PRACTICAL consortium. (2018) Large-scale transcriptome-wide association study identifies new prostate cancer risk regions. Nat. Commun., 9, 4079
62. Ioannidis, N. M., Wang, W., Furlotte, N. A., Hinds, D. A., Bustamante, C. D., Jorgenson, E., Asgari, M. M. and Whittemore, A. S., and the 23andMe Research Team. (2018) Gene expression imputation identifies candidate genes and susceptibility loci associated with cutaneous squamous cell carcinoma. Nat. Commun., 9, 4264
63. Huckins, L. M., Dobbyn, A., Ruderfer, D. M., Hoffman, G., Wang, W., Pardiñas, A. F., Rajagopal, V. M., Als, T. D., T Nguyen, H., Girdhar, K., *et al.* (2019) Gene expression imputation across multiple brain regions provides insights into schizophrenia risk. Nat. Genet., 51, 659–674
64. Lamontagne, M., Bérubé, J. C., Obeidat, M., Cho, M. H., Hobbs, B. D., Sakornsakolpat, P., de Jong, K., Boezen, H. M., Nickle, D., Hao, K., *et al.* (2018) Leveraging lung tissue transcriptome to uncover candidate causal genes in COPD genetic associations. Hum. Mol. Genet., 27, 1819–1829
65. Thériault, S., Gaudreault, N., Lamontagne, M., Rosa, M., Boulanger, M. C., Messika-Zeitoun, D., Clavel, M. A., Capoulade, R., Dagenais, F., Pibarot, P., *et al.* (2018) A transcriptome-wide association study identifies PALMD as a susceptibility gene for calcific aortic valve stenosis. Nat. Commun., 9, 988
66. Zhao, B., Shan, Y., Yang, Y., Li, T., Luo, T., Zhu, Z., Li, Y. and Zhu, H. (2019) Transcriptome-wide association analysis of 211 neuroimaging traits identifies new genes for brain structures and yields insights into the gene-level pleiotropy with other complex traits. bioRxiv, 842872
67. Keys, K. L., Mak, A. C. Y., White, M. J., Eckalbar, W. L., Dahl, A. W., Mefford, J., Mikhaylova, A. V., Contreras, M. G., Elhawary, J. R., Eng, C., *et al.* (2019) On the cross-population portability of gene expression prediction models. PLoS Genet, 16, e1008927
68. Wheeler, H. E., Ploch, S., Barbeira, A. N., Bonazzola, R., Andaleon, A., Fotuhi Siahpirani, A., Saha, A., Battle, A., Roy, S. and Im, H. K. (2019) Imputed gene associations identify replicable trans-acting genes enriched in transcription pathways and complex traits. Genet. Epidemiol., 43, gepi.22205
69. Shan, N., Wang, Z. and Hou, L. (2019) Identification of trans-eQTLs using mediation analysis with multiple mediators. BMC Bioinformatics, 20, 126
70. Pierce, B. L., Tong, L., Chen, L. S., Rahaman, R., Argos, M., Jasmine, F., Roy, S., Paul-Brutus, R., Westra, H. J., Franke, L., *et al.* (2014) Mediation analysis demonstrates that trans-eQTLs are often explained by cis-mediation: a genome-wide analysis among 1,800 South Asians. PLoS Genet., 10, e1004818
71. The GTEx Consortium, the Laboratory, Data Analysis & Coordinating Center (LDACC)—Analysis Working Group, the Statistical Methods groups—Analysis Working Group, the Enhancing GTEx (eGTEx) groups, the NIH Common Fund, the NIH/NCI, the NIH/NHGRI, the NIH/NIMH, the NIH/NIDA, the Biospecimen Collection Source Site—NDRI *et al.* (2017) Genetic effects on gene expression across human tissues. Nature, 550, 204–213
72. Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M. J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Kasela, S., *et al.* (2018) Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. bioRxiv, 447367
73. Liao, C., Laporte, A. D., Spiegelman, D., Akçimen, F., Joober, R., Dion, P. A. and Rouleau, G. A. (2019) Transcriptome-wide association study of attention deficit hyperactivity disorder identifies associated genes and phenotypes. Nat. Commun., 10, 4450