Power analysis of transcriptome-wide association study

Bowei Ding[1,*], Chen Cao[2,*], Qing Li[2], Jingjing Wu[1,#], Quan Long[1,2,3,4,#]

[1]Department of Mathematics & Statistics, University of Calgary, Calgary, Canada.

[2]Department of Biochemistry & Molecular Biology, Alberta Children's Hospital Research Institute, University of Calgary, Calgary, Canada.

[3]Department of Medical Genetics, University of Calgary, Calgary, Canada.

[4]Hotchkiss Brain Institute, O'Brien Institute for Public Health, University of Calgary, Calgary, Canada.

* = Joint first authors: B.D. and C.C.

= Joint corresponding authors: J.W. (jinwu@ucalgary.ca) and Q.L. (quan.long@ucalgary.ca)

ABSTRACT

Standard Genome-wide association study (GWAS) discovers genetic variants explaining phenotypic variance by directly associate them. With the availability of other omics data such as gene expression, the field is stepping into an exciting era of multi-scale omics integration. An emerging technique is transcriptome-wide association study (TWAS) that conducts association mapping by utilizing gene expression data from a separate reference dataset based on which a model predicting expression by genotype is trained. Despite its success in practice, two fundamental questions have been unaddressed yet. First, in practice, the accuracy of predicting expression by genotype is generally low, which is bounded by the expression heritability. So, the question is whether such a low accuracy may impact the power of TWAS, and what level of

accuracy is sufficient. Second, since predicting expression is a critical step in TWAS, one may ask what if we have actual expression assessed by a real experiment, and whether that will improve or deteriorate power. Answering these questions will bring thorough understanding of TWAS and practical guidelines in association mapping. To address the above questions, we conducted power analysis for GWAS, TWAS, and expression medicated GWAS (emGWAS). Specifically, we derived non-centrality parameters (NCPs), enabling closed-form derivation of statistical power to facilitate a thorough power analysis without relying on particular implementations. We assessed the power of the three protocols with respect to two representative scenarios: causality (genotype contributes to phenotype through expression) and pleiotropy (genotype contributes directly to both phenotype and expression). For both scenarios, we tested various properties including expression heritability. Our analysis led to two main outcomes: (1) TWAS utilizing predicted expression enjoys higher power than emGWAS that has actual expressions in the pleiotropy scenario, revealing a deep insight into TWAS models as well as a practical guideline of applying TWAS even in cases when expressions are available in a GWAS dataset. (2) TWAS is suboptimal compared to GWAS when expression heritability is too low. The superiority ordering of TWAS and GWAS disclosed a turn-point in each of the causality and pleiotropy scenarios. Analysis of published discoveries shows the selection of protocols might be improved based on the identified turn-points.

Keywords: Power analysis, GWAS, TWAS, Non-centrality parameter, Expression heritability

INTRODUCTION

High-throughput sequencing instruments have provided unprecedented opportunities to assess 'omics' data for humans and model organisms. For instance, current experimental techniques can quickly profile transcriptome (the RNA expression of all genes) (WANG *et al.* 2009; TRAPNELL *et al.* 2010; OZSOLAK AND MILOS 2011; MCGETTIGAN 2013), proteome (the level of all proteins) (BELL *et al.* 2009; PIBLE AND ARMENGAUD 2015; SELEVSEK *et al.* 2015) and other omics data (ZHANG *et al.* 2012; TEPERINO *et al.* 2013; COATS AND RUMPHO 2014). From the perspective of genome-wide association study (GWAS) which links genetic variants to disease phenotypes (OZAKI *et al.* 2002; KLEIN *et al.* 2005; MILLS AND RAHAL 2019), these 'omics' could bridge the gap between genotype and phenotype, offering leverages to further our understanding of the genetic basis of complex traits. How to integrate these omics data is a challenging research field that has attracted attention from many researchers (FUKUSHIMA *et al.* 2009; HASIN *et al.* 2017; YAN *et al.* 2018; EDDY *et al.* 2020).

To this end, transcriptome-wide association study (TWAS) provides an efficient protocol to quantitatively aggregate multiple genetic variants into a single test by utilizing transcriptome data. Pioneered by Gamazon *et al* (GAMAZON *et al.* 2015), the typical pipeline of TWAS is composed of two steps: First, a model is trained to predict gene expressions by local genetic variants surrounding the focal genes. More specifically, based on a reference dataset which contains both expression and genotype, one can train a predictive model that predicts gene expressions using local genotypes. Second, when analyzing the main dataset for genotype-phenotype association mapping, which contains genotype and phenotype (but not expression), one can use the pretrained model to predict expressions with each subject's genotypes; and then associate each predicted expression to the phenotype. Initiated by Gusev *et al* (GUSEV *et al.* 2016), methods using summary statistics in GWAS datasets (i.e., meta-analysis) to conduct

TWAS were also developed (BARBEIRA *et al.* 2018). Since then, TWAS has achieved significant success in identifying the genetic basis of complex traits (GONG *et al.* 2018; GUSEV *et al.* 2018; MANCUSO *et al.* 2018; THERIAULT *et al.* 2018; ATKINS *et al.* 2019; RATNAPRIYA *et al.* 2019; ZHANG *et al.* 2019) and became a popular method in practice. Inspired by TWAS, researchers have developed multiple similar protocols using other endophenotypes to conduct association mappings. For example, IWAS utilizes images (XU *et al.* 2017) and PWAS leverages proteins (BRANDES *et al.* 2020).

Despite the recent success of TWAS, there are puzzling issues to be answered to pave the way for broader applications of TWAS without concerns. **First**, a fundamental process in TWAS is the prediction of expressions. However, as many methodological papers show, the average R^2 between predicted and actual expressions is very low (e.g., around 0.02 ~ 0.05), partly due to the limited expression heritability (GAMAZON *et al.* 2015), an upper bound of the predictive accuracy. Naturally, one may ask: if the expression heritability is too low, will TWAS be still preferable compared to standard GWAS? Indeed, in real data analysis, the discovered genes with significant TWAS *p*-values do enjoy a higher R^2 than average (GAMAZON *et al.* 2015; GUSEV *et al.* 2016; MANCUSO *et al.* 2017; MOGIL *et al.* 2018; WAINBERG *et al.* 2019). So, an immediate question will be, what level of expression heritability, possibly in conjunction with other properties of data, are required for TWAS to be powerful. **Second**, as pointed out by the first TWAS paper (GAMAZON *et al.* 2015), the key insight of TWAS is to aggregate the sensible genetic variants in order to estimate the "genetically regulated gene expression", or GReX (GAMAZON *et al.* 2015), for a downstream GWAS. With this in mind, one may ask how powerful TWAS is when compared with GWAS under a hypothetical protocol in which one has actual expression data (in addition to genotype and phenotype). In this paper, we refer to such a protocol as "expression mediated GWAS", or emGWAS. This is not a trivial question. From a positive perspective, the actual expression may capture GReX better than a predicted one that

102 suffers from prediction accuracy; from a negative point of view, the actual expression may
103 contain non-genetic components contributed by experimental artifacts or environments,
104 therefore impairing the power. Although emGWAS is impractical in many situations due to the
105 difficulties of accessing tissues (e.g., in the studies of brain diseases), it is still feasible for some
106 diseases where tissues are routinely available (e.g., blood or cancerous tissues). More
107 importantly, studying the power issue of emGWAS will allow a crystal understanding of what
108 TWAS is exactly and how we should apply it optimally in practice.
109
110 In this work, to address the above questions that will lead to both theoretical understanding and
111 practical guidelines of TWAS, we conducted power analysis of TWAS, GWAS, and emGWAS.
112 There is an existing study that investigated power comparisons between GWAS, TWAS, and
113 another protocol that integrates eQTL in GWAS (VETURI AND RITCHIE 2018). Unlike that study
114 which is purely simulation-based, we used a more traditional closed-form analysis. In particular,
115 our power analysis was based on deriving the non-centrality parameters (NCPs) for the relevant
116 statistical tests to calculate the closed-form powers (**Methods**). To our best knowledge, there is
117 no derived closed-form of the NCP in current literature for linear mixed model (LMM), as the
118 main-stream model for GWAS (as well as some of the regressions in emGWAS and TWAS). As
119 such, we developed a novel method that converts an LMM to a linear regression by
120 decorrelating the covariance structure of the response variable in the LMM (**Methods**), which
121 made the closed-form NCP derivation feasible.
122
123 Our closed-form derivation is advantageous over pure simulation-based method as the latter
124 directly calculates the power for a particular configuration of available data without repetitive
125 stochastic resampling from alternative hypothesis. As a result, our derivation saves
126 computational resources, yields more accurate estimations, and can be easily adapted to power
127 analysis of other similar protocols such as IWAS (XU *et al.* 2017) and PWAS (OKADA *et al.* 2016;

128 BRANDES *et al.* 2020). Moreover, closed-form derivations estimate power without conducting the
129 actual analysis, therefore the outcome of the analysis is not restricted to specific
130 implementations. For example, there are many implementations of GWAS and TWAS models,
131 and they may yield different results due to variations in filtering input data and optimizing
132 parameters. Our work provides a theoretical conclusion of the expected power assuming no
133 other artifacts caused by specific data or tools, although it does not capture the power loss due
134 to practical issues.
135
136 Our power analysis disclosed guidelines on the appropriateness of GWAS, TWAS, and
137 emGWAS with respect to various properties including the underlying genetic architecture, the
138 expression heritability, and the number of causal genes. In the rest of the paper, we will first
139 describe our methods including the novel derivation of NCPs for the relevant models in GWAS,
140 TWAS and emGWAS. After that, in-depth presentations and discussions of the results will be
141 provided. Finally, we will conclude the paper by recapping the take home messages and
142 discussing the potential limitation and future work.
143
144 **MATERIALS & METHODS**
145
146 *Mathematical definitions of GWAS, TWAS, and emGWAS protocols*
147
148 There are many existing protocols conducting GWAS, TWAS, and potentially emGWAS
149 (AULCHENKO *et al.* 2007; PURCELL *et al.* 2007; KANG *et al.* 2010; GOGARTEN *et al.* 2012; ZHOU
150 AND STEPHENS 2012; GAMAZON *et al.* 2015; GUSEV *et al.* 2016). In this work, it is assumed that
151 (1) the phenotypic variation is contributed by multiple genes, and (2) for each causal gene, there
152 are multiple genetic variants contributing to gene expressions and also the phenotype. This
153 setting is motivated by the fact that most complex traits are contributed by multiple loci, and a

fundamental assumption of TWAS is that there are multiple causal local genetic variants in the gene under consideration. Specifically, the models are defined below:

GWAS. For GWAS, we adopted the standard LMM similar to EMMAX (KANG *et al.* 2010):

$$Y = \beta_{j0}\mathbf{1} + X_j\beta_{j1} + u + \varepsilon, \qquad j = 1,2,\dots,n_x, \tag{1}$$

where n_x is the total number of genetic variants, Y is an $n \times 1$ vector of phenotypes (n represents the number of individuals), **1** is an $n \times 1$ vector of all 1's, for the jth genetic variant X_j is an $n \times 1$ genotype vector with $X_{ij} \in \{0,1,2\}$ representing the number of copies of minor allele for the ith individual and β_{j0} and β_{j1} are the intercept and effect size respectively of the genetic variant, u is an $n \times 1$ vector of random effects following multivariate normal distribution, i.e. $u \sim N(0, \sigma_g^2 K_x)$, ε is an $n \times 1$ vector of errors with $\varepsilon \sim N(0, \sigma_e^2 I)$, σ_g^2 and σ_e^2 are variance components, *I* is an $n \times n$ identity matrix, K_x is a known $n \times n$ real symmetric genomic relationship matrix (GRM). Following Patterson *et al* (PATTERSON *et al.* 2006), K_x is calculated by

$$K_x = \frac{1}{n_x}\tilde{X}\tilde{X}^T, \tag{2}$$

where n_x is the total number of genetic variants, $\tilde{X}$ is an $n \times n_x$ matrix after standardization. For example, for each element in the jth column ($j = 1,2,\dots,n_x$), $\tilde{X}_{ij}$ can be calculated by

$$\tilde{X}_{ij} = \frac{X_{ij} - \bar{X}_{.j}}{S_{Xj}}, \tag{3}$$

where $\bar{X}_{.j} = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$ and $S_{Xj}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_{ij} - \bar{X}_{.j}\right)^2$ are the sample mean and sample variance of the jth variant.

emGWAS. For emGWAS, we conducted two regressions sequentially. First the phenotype was regressed on the actual (not predicted) expression. In the second step, the expression was regressed on each individual local genetic variant, similar to a cis-eQTL analysis. We chose to use LMM for the association between phenotype and expression as there are multiple genes from the whole genome contributing to the phenotype and the effects of non-focal genes in the model may be captured by the random term. The GRM is calculated by DNA instead of expressions because we assume that phenotype is largely contributed by genetics. For the association between expression and local genetic variants (i.e., cis-eQTL), we chose to use linear regression (LM), the frequently adopted method for conducting cis-eQTL analysis. More specifically, the first model is

$$Y = \beta_{l0}\mathbf{1} + \beta_{l1}Z_l + u + \varepsilon, \qquad l = 1,2,\dots,n_z, \tag{4}$$

where Y, $\mathbf{1}$, u and ε have exactly the same interpretation and assumption as in (1), n_z is the total number of genes, for the lth gene Z_l is an $n \times 1$ gene expression vector and β_{l0} and β_{l1} are the intercept and effect size respectively of the gene.

For the second part, we fit a simple linear regression model for gene expression on its local genetic variants:

$$Z_l = \beta_{lk0}\mathbf{1} + \beta_{lk1}X_{lk} + \varepsilon_{el}, \qquad l = 1,2,\dots,n_z, k = 1,2,\dots,n_{el}, \tag{5}$$

where for the lth gene the new notations X_{lk} is an $n \times 1$ vector of the kth local genetic variant, $\varepsilon_{el} \sim N(0, \sigma_{el}^2 I)$ is a $n \times 1$ vector of errors with σ_{el}^2 the variance component, X_{lk} is a $n \times 1$ vector of the kth local genetic variant, n_{el} is the total number of local genetic variants in the lth gene, β_{lk0} and β_{lk1} are the intercept and effect size respectively of the variant, and n_{el} is the total number of local genetic variants in the gene.

TWAS. For TWAS, the analytic protocol is similar to that of emGWAS, except for that the gene expressions are predicted using a pretrained ElasticNet model. More specifically,

$$Y = \beta_{Pl0}\mathbf{1} + \beta_{Pl1}\hat{Z}_l + u + \varepsilon, \qquad l = 1,2,\ldots,n_z, \tag{6}$$

where for the lth gene the new notations $\hat{Z}_l$ is an $n \times 1$ vector of the predicted gene expressions, and β_{Pl0} and β_{Pl1} are the intercept and effect size respectively of the predicted gene expression.

There are many approaches to estimate gene expression, including least absolute shrinkage and selection operator (LASSO), elastic net, etc. In Gamazon *et al.* [18], it has been shown that the elastic net performs well and is more robust against minor changes in input variants. So, we applied the elastic net to train a predictive model with use of the “glmnet” package in R. The objective function in “glmnet” is

$$L_{enet}(\beta) = \frac{1}{2n}\|Z - X\beta\|^2 + \lambda\left(\frac{1-\alpha}{2}\|\beta\|^2 + \alpha\|\beta\|_1\right), \tag{7}$$

where λ and α are tuning parameters. The penalty term is a convex (linear) combination of the LASSO and ridge penalties. When $\alpha = 1$, (7) is the objective function of LASSO, while $\alpha = 0$ makes it the objective function of the ridge. The optimal tuning parameters were chosen by minimizing the cross-validated squared-error. Readers are referred to **Appendix A** for more details.

Closed-form derivation of NCP and power calculation

Non-centrality parameter (NCP) is the parameter related to non-central distributions measuring the departure from central distribution under an alternative hypothesis. NCP enables the calculation of the probability of rejecting the null hypothesis when the alternative hypothesis is correct (under the non-central distribution). As such, deriving NCP is a nature way to calculate the power of a statistical test in a closed form. To our best knowledge, there is no literature reporting closed-form NCP derivation for LMM, a dominant model used in GWAS. We have developed a novel method to calculate NCP, which is described below.

First, for a standard simple linear regression, the NCP of a *t*-test of the coefficient of the predictor can be derived in a similar way as for a one-sample *t* test statistic. Suppose we have a simple random sample $X_1, \ldots, X_n \sim N(\mu, \sigma)$, then the one-sample *t* test statistic for testing the null hypothesis $H_0: \mu = \mu_0$ is

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu_0)/\sigma}{\sqrt{\frac{(n-1)S^2/\sigma^2}{n-1}}} \sim t_{n-1}, \tag{8}$$

where $\bar{X}$ and S are the sample mean and sample standard deviation (the unbiased version) respectively, and under H_0, $\sqrt{n}(\bar{X} - \mu_0)/\sigma \sim N(0,1)$, $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$ and thus $T \sim t_{n-1}$.

Under the alternative hypothesis $H_a: \mu = \mu_a$, the test statistic $T = \frac{\sqrt{n}[(\bar{X}-\mu_a)+(\mu_a-\mu_0)]/\sigma}{\sqrt{\frac{(n-1)S^2/\sigma^2}{n-1}}}$ follows a non-central t distribution with NCP given by

$$v = \frac{\mu_a - \mu_0}{\sigma/\sqrt{n}}. \tag{9}$$

Now the main obstacle is how to derive closed-form NCP under the setting of the very different LMM. We got around of this difficulty by converting an LMM to a linear regression without intercept by decorrelating the response variable and the predictors, a technique that has been used for mixed models (LONG *et al.* 2013; RAKITSCH *et al.* 2013). The procedure is given as follows. First, we solve the null model $Y_c = u + \varepsilon$ that contains no genetic variants. (This is an established innovation to reduce computational burden of factorizing the GRM many times when analyzing large number of variants (KANG *et al.* 2010; LONG *et al.* 2013)) Using the Newton-Raphson method detailed in **Appendix B**, we can estimate σ_g^2. Denoting the eigen decomposition of the GRM as $K_x = U_x \Lambda_x U_x^{-1}$, we can construct the de-correlation matrix as

$$D_x = \left(\sigma_g^2 \Lambda_x + \sigma_e^2 I\right)^{-\frac{1}{2}} U_x^T. \tag{10}$$

By left multiplying X and Y both by D_x and denoting $X^* = D_x X = (X_1^*, X_2^*, \dots, X_n^*)^T$ and $Y^* = D_x Y = (Y_1^*, Y_2^*, \dots, Y_n^*)^T$, the covariance structure in Y^* has been then removed. Then conducting linear regression of Y^* on X^* is equivalent to the above LMM model. The proof of the validity of this decorrelation structure is presented in **Appendix C**.

Then, based on the closed form NCP for linear regression, we derive the estimated NCP of LMM (1) which is given by

$$\hat{\upsilon}_{Gj} = \frac{\sum_{i=1}^{n} \hat{X}_{ij}^{*} \hat{Y}_{i}^{*} \sum_{i=1}^{n} \hat{D}_{xi\cdot}^{2} - \sum_{i=1}^{n} \hat{Y}_{i}^{*} \hat{D}_{xi\cdot} \sum_{i=1}^{n} \hat{X}_{ij}^{*} \hat{D}_{xi\cdot}}{\sqrt{\sum_{i=1}^{n} \left(\hat{X}_{ij}^{*}\right)^{2} \left(\sum_{i=1}^{n} \hat{D}_{xi\cdot}\right)^{2} - \left(\sum_{i=1}^{n} \hat{D}_{xi\cdot} \hat{X}_{ij}^{*}\right)^{2} \sum_{i=1}^{n} \hat{D}_{xi\cdot}^{2}}}, \tag{11}$$

where $\hat{X}_{j}^{*} = \hat{D}_{x} X_{j} = (\hat{X}_{1j}^{*}, \hat{X}_{2j}^{*}, \dots, \hat{X}_{nj}^{*})^{T}$, $\hat{Y}^{*} = \hat{D}_{x} Y = (\hat{Y}_{1}^{*}, \hat{Y}_{2}^{*}, \dots, \hat{Y}_{n}^{*})^{T}$, and $\hat{D}_{x\mathrm{i}\cdot} = \sum_{j=1}^{n} \hat{D}_{xij}$. The proof of this expression of NCP for LMM is in **Appendix D.**

The above result for NCP is instrumental for deriving the powers of all the three models, i.e., GWAS, emGWAS and TWAS, which are detailed below.

For GWAS, for each individual test we use significance level $\alpha_{x} = \frac{0.05}{n_{x}}$ which is after the Bonferroni correction (SHAFFER 1995) that makes adjustment for multiple testing, where n_{x} is the total number of SNPs. Throughout this paper, we use $f(t; \upsilon)$ to denote the probability density function (PDF) of the non-central *t* distribution with degrees of freedom *n-2* and NCP υ. Then the statistical power of the jth SNP can be estimated by $P_{Gj} = \int_{F_0^{-1}(1-\alpha_x)}^{+\infty} f\left(t; \hat{\upsilon}_{Gj}\right) dt$ after plugging in the estimated NCP $\hat{\upsilon}_{Gj}$, where $F_0(t)$ is the cumulative distribution function (CDF) of the central t distribution; $F_0^{-1}(1-\alpha_x)$ gives the critical value from the central *t* distribution with degrees of freedom *n-2*. The computation of power can be directly implemented in R via the function “pt”, which requires the critical value, NCP and degrees of freedom.

For emGWAS, which consists of two regression models (4) and (5), we assume these two regression models are independent of each one. For model $Y = \beta_{l0}\mathbf{1} + Z_{l}\beta_{l1} + u + \varepsilon$ given in

(4), we left multiply the estimated $\widehat{D}_x$ to both sides of the equation, then the estimated NCP for the lth gene expression is given by

$$\hat{\upsilon}_{eZl} = \frac{\sum_{i=1}^{n} \hat{Z}_{il}^{*}\, \hat{Y}_i^{*} \sum_{i=1}^{n} \widehat{D}_{xi\cdot}^{2} - \sum_{i=1}^{n} \hat{Y}_i^{*}\, \widehat{D}_{xi\cdot} \sum_{i=1}^{n} \hat{Z}_{il}^{*}\, \widehat{D}_{xi\cdot}}{\sqrt{\sum_{i=1}^{n} \left(\hat{Z}_{il}^{*}\right)^2 \left(\sum_{i=1}^{n} \widehat{D}_{xi\cdot}^{2}\right)^2 - \left(\sum_{i=1}^{n} \widehat{D}_{xi\cdot}\, \hat{Z}_{il}^{*}\right)^2 \sum_{i=1}^{n} \widehat{D}_{xi\cdot}^{2}}}, \tag{12}$$

where $\hat{Z}_l^{*} = \widehat{D}_x Z_l = \left(\hat{Z}_{1l}^{*}, \hat{Z}_{2l}^{*}, \dots, \hat{Z}_{nl}^{*}\right)^T$. For model (4), we use the significance level $\alpha_z = \frac{0.05}{n_z}$ for each individual test, where n_z is the total number of genes. Then the statistical power of detecting the lth gene expression is estimated by $P_{eZl} = \int_{F_0^{-1}(1-\alpha_z)}^{+\infty} f\,(t; \hat{\upsilon}_{eZl}) dt$.

The second step in emGWAS is simply based on the estimated NCP for the standard linear regression (5), given by

$$\hat{\upsilon}_{eXlk} = \frac{\sum_{i=1}^{n} (X_{ilk} - \bar{X}_{\cdot lk})\, Z_{il}}{\sqrt{\sum_{i=1}^{n} (X_{ilk} - \bar{X}_{\cdot lk})^2\, \hat{\sigma}_{el}}}, \tag{13}$$

where

$$\hat{\sigma}_{el} = \frac{1}{n-2} \sum_{i=1}^{n} \left(Z_{il} - \bar{Z}_{\cdot l} + \hat{\beta}_{lk}(X_{ilk} - \bar{X}_{\cdot lk})\right)^2. \tag{14}$$

The power of detecting X_{lk} in the second step is estimated in a similar way. Specifically, for model (5) we use the significance level $\alpha_{el} = \frac{0.05}{n_{el}}$, where n_{el} is the total number of local genetic variants in the lth gene. The power is then estimated by $P_{eXlk} = \int_{F_0^{-1}(1-\alpha_{el})}^{+\infty} f\,(t; \hat{\upsilon}_{eXlk}) dt$. Since

we assume models (4) and (5) are independent, the power of detecting the lth gene and the kth variants in the lth gene simultaneously is estimated by $P_{eZl}P_{eXlk}$.

For TWAS, the NCP is estimated similarly to that in the first step of emGWAS, i.e.

$$\hat{\upsilon}_{Tl} = \frac{\sum_{i=1}^{n} \hat{\hat{Z}}_{il}^{*} \hat{Y}_i^{*} \sum_{i=1}^{n} \widehat{D}_{xi\cdot}^{2} - \sum_{i=1}^{n} \hat{Y}_i^{*} \widehat{D}_{xi\cdot} \sum_{i=1}^{n} \hat{\hat{Z}}_{il}^{*} \widehat{D}_{xi\cdot}}{\sqrt{\sum_{i=1}^{n} \left(\hat{\hat{Z}}_{il}^{*}\right)^2 \left(\sum_{i=1}^{n} \widehat{D}_{xi\cdot}^{2}\right)^2 - \left(\sum_{i=1}^{n} \widehat{D}_{xi\cdot} \hat{\hat{Z}}_{il}^{*}\right)^2 \sum_{i=1}^{n} \widehat{D}_{xi\cdot}^{2}}}, \quad (15)$$

where the only difference between (12) and (15) is that $\hat{Z}_{il}^{*} = \widehat{D}_x Z_{il}$ in (15) is replaced by $\hat{\hat{Z}}_{il}^{*} = \widehat{D}_x \hat{Z}_{il}$ in (15). The significance level used for each individual test is $\alpha_z = \frac{0.05}{n_z}$ and the power is estimated by $P_{Tl} = \int_{F_0^{-1}(1-\alpha_z)}^{+\infty} f\,(t; \hat{\upsilon}_{Tl})dt$.

Simulation of phenotype and expression

Statistical powers of the three protocols under investigation are dependent on the actual effect size. As such, we conducted simulations to generate data with various levels of effect size. Here the effect size depends on a complicated combination of many factors. We chose to focus on the following three aspects. (1) We considered two scenarios: causality and pleiotropy (**Figure 1**). In the causality scenario, genotype contributes to phenotype via the mediation of expression (**Figure 1a**); whereas in pleiotropy scenario, genotype contributes to both expression and phenotype directly (**Figure 1b**). We did not consider the scenario in which phenotype is causal to expression. (2) We considered the strength of three different variant components: trait heritability (the phenotypic variance component explained by genotype, denoted by $h^2_{\mathrm{x=>y}}$),

expression heritability (the expression variance component explained by genotype, denoted by $h^2_{\mathrm{x=>z}}$), and the phenotypic variance component explained by expression, abbreviated as *PVX* and denoted by $h^2_{\mathrm{z=>y}}$. (3) We considered the number of genes contributing to phenotype and the number of local genetic variants contributing to expression.

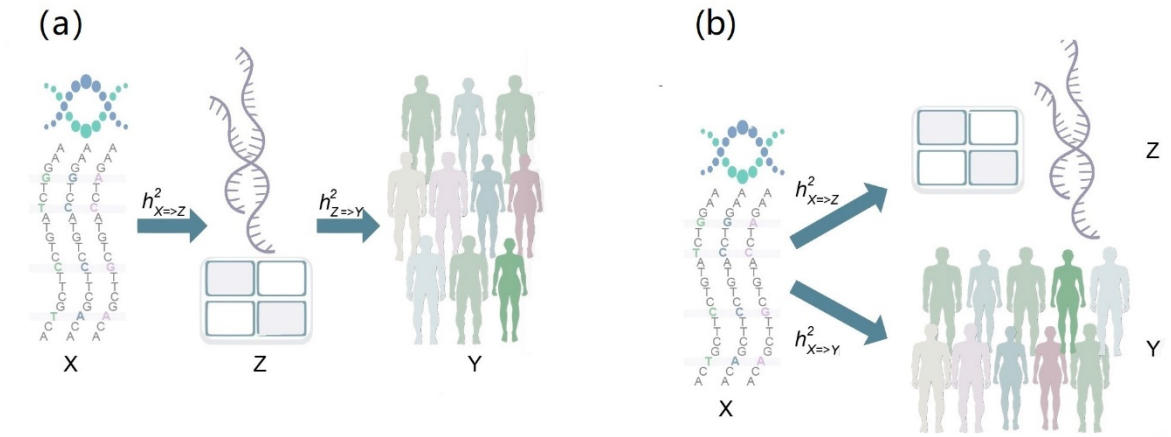

Figure 1: Causality (a) and Pleiotropy (b) scenarios for genotype (X), expression (Z) and phenotype (Y).

In all the simulations, we use the real genotype from the 1000 Genomes Project data (*N* = 2504). Although there are multiple datasets with both expressions and genotypes available, we chose to simulate expressions instead of using the existing real expressions and genotypes. This is because that it is difficult to match the relationship between expression and genotype to be exactly the same as prespecified properties, such as expression heritability, number of contributing genetic variants, etc. So, simulating expressions kept our power analysis straightforward and comprehensive.

In the causality scenario, the phenotypes were simulated based on the following procedure. First, a number of genes (n_{z-sig}= 4, 9, or 13) were selected as causal genes. Then, in each

359 gene (indexed by $l = 1,2,\ldots,n_{z-sig}$), a number of genetic variants ($n_{z(l)-sig}$ = 4 ~9) that are
360 common and independent ($MAF > 0.05\ and\ R^2 < 0.01$) were selected as causal variants. A
361 linear combination of the local variants in the lth gene is generated to produce the expression
362 values ($Z_{(l)}$), and then a linear combination of these gene expressions $\boldsymbol{Z}$ is simulated as the
363 genomic contribution to the phenotype. Note that during the course, we prespecified the
364 expression heritability ($h^2_{x=>z}$) and the PVX ($h^2_{\mathrm{z=>y}}$). We ensured that the above simulations
365 using linear combinations did yield the desired heritability and variance component (**Appendix**
366 **E**).
367
368 In the pleiotropy scenario, the simulations generally followed the same procedure as the
369 causality scenario, except for a fundamental difference in that the phenotype Y was directly
370 generated from a combination of genotypes instead of through expression (**Appendix F**). Note
371 that, although expressions Z and phenotype Y were generated separately, they are still based
372 on the same genetic variants, therefore are statistically correlated (although not based on
373 genuine biological causality). Therefore, if the trait heritability and the expression heritability is
374 reasonably large, TWAS can still identify the causal genes based on this statistical correlation.
375
376 In the simulations, the expression heritability was tested using a spectrum of values: $h^2_{x=>z}$ in
377 (2.5%; 3%; 4%; 6%; 8%; 10%; 30%) for both scenarios. The trait heritability $h^2_{x=>y}$ in the
378 pleiotropy scenario and the PVX $h^2_{z=>y}$ in the causality scenario is in (0.5%; 1%; 2.5%; 5%;
379 10%). Notably, our selection of expression heritability is at the high-end of the observed average
380 in real data (GAMAZON *et al.* 2015), and our trait heritability is lower than the expected heritability
381 in normal GWAS. In our study, we have actually tested broader values, however, as we will
382 present in **Results**, the turn-points for TWAS being advantageous over GWAS fall in-between of
383 above values, and the other values that are outside of the range show the same pattern. That is

why we chose to present the outcome of the above listed values to demonstrate the main message of this work.

Finally, as there are many variants and genes involved in the studies, the overall power for each of the three protocols are defined as

$$P_{GWAS} = 1 - \prod_{j=1}^{n_{x-sig}} \left(1 - P_{G(j)}\right), \tag{16}$$

$$P_{emGWAS} = 1 - \prod_{l=1}^{n_{z-sig}} \left(1 - P_{eZ(l)} P_{eX(l)}\right), \; where \; P_{eX(l)} = 1 - \prod_{k=1}^{n_{z(l)-sig}} \left(1 - P_{eX(l)(k)}\right), \tag{17}$$

$$P_{TWAS} = 1 - \prod_{l=1}^{n_{z-sig}} \left(1 - P_{T(l)}\right), \tag{18}$$

where n_{x-sig}, n_{z-sig} and $n_{z(l)-sig}$ denote the numbers of significant SNPs, significant genes, and significant SNPs in the lth significant gene, respectively, $G(j)$ denotes the jth sigficiant SNP identified by GWAS, $Z(l)$ and $X(l)(k)$ denote the lth significant gene and the kth significant SNP in the lth significant gene respectively identified by emGWAS, and $T(l)$ denotes the lth sigficiant gene identified by TWAS. That is, the power of GWAS is defined as the probability of detecting at least one causal variant (in any causal gene); the power of emGWAS is the probability of detecting at least one gene and one local SNP of that gene simultaneously; and the power of TWAS is the probability of at least one predicted gene expression is identified as significant.

RESULTS

First, as a quality control of the expression prediction, we compared the expression heritability and the mean R^2 between predicted and actual expressions (**Table 1**). Consistent with our intuition, the R^2 is closer to the actual heritability when the heritability is larger.

Table 1: Comparisons of R^2 of imputed gene expression under different levels of expression heritability and number of genetic variants.

	Mean of R^2	Sample Standard Deviation of R^2
$h_1^2 = 0.025$	0.007847616	0.007415877
$h_1^2 = 0.03$	0.01259302	0.008410582
$h_1^2 = 0.04$	0.02319834	0.009481371
$h_1^2 = 0.06$	0.04415579	0.01083593
$h_1^2 = 0.08$	0.06465895	0.01175991
$h_1^2 = 0.1$	0.08518152	0.01264175
$h_1^2 = 0.3$	0.2886779	0.01514781

Causality scenario

First, we analyzed the conditions when expression heritability is high ($h^2_{x=>z}$ = 0.1 or 0.3) but the PVX is low (**Figure 2**). Overall, it is evident that emGWAS outperforms both GWAS and TWAS

with a large margin. Also, we observed that TWAS generally outperforms GWAS. Note that in this figure, the PVX is low, which is not favorable for TWAS; however, TWAS is still more powerful because the high expression heritability is a more important factor than PVX.

Additionally, we observed several trends that are consistent with intuition. The powers of GWAS and TWAS increase as expression heritability increases, however, this trend is not very pronounced for emGWAS. When the number of causal genes becomes larger, the powers of GWAS and emGWAS become lower; however, the number of causal genes doesn't impact TWAS too much. This is because that TWAS uses GReX ($\hat{Z}$) to aggregate genetic effects, and therefore does not suffer from the burden of multiple-testing correction.

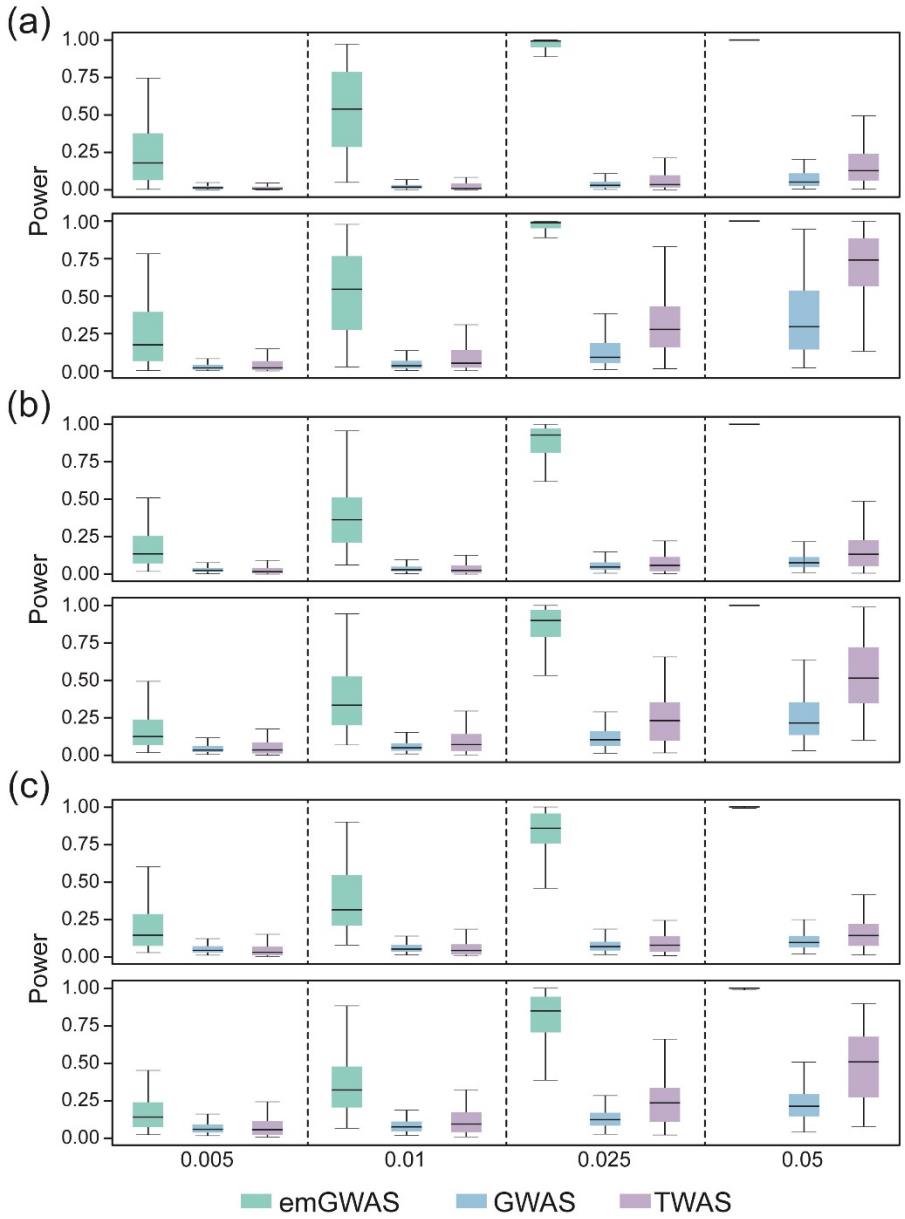

Figure 2: Causality scenario when expression heritability is high and PVX is low.

The PVX is 0.005, 0.01, 0.025, and 0.05 in the four columns as indicated by the X-axis labels. The numbers of genes contributed to phenotype for (a), (b) and (c) are 4, 9 and 13 respectively. The expression heritability for the top and bottom panels in (a), (b) and (c) are 0.1 and 0.3 respectively. The number of causal variants per gene is randomly sampled from the interval [4,9].

Second, we looked at the cases when the PVX is high, but the expression heritability is relatively low ($h^2_{x=>z}$= 0.025, 0.03, 0.04 or 0.08). Evidently, the emGWAS is still the most powerful with its powers all being 1.0. The comparison between TWAS and GWAS is a bit tricky: TWAS is suboptimal compared with GWAS when the expression heritability is as low as 0.025 or 0.03 (**Figure 3a,b**); however the superiority order is reversed when the expression heritability is moderate, such as 0.04 (**Figure 3c**); and TWAS is obviously better than GWAS when the expression heritability is as high as 0.08 (**Figure 3d**). This is an important discovery because we quantified the turn-point of the relative power between GWAS and TWAS. In other words, when the expression heritability is lower than 0.04, even if PVX is high (a factor that is favorable for TWAS), one should not use TWAS; instead, the traditional GWAS is more powerful and should be utilized.

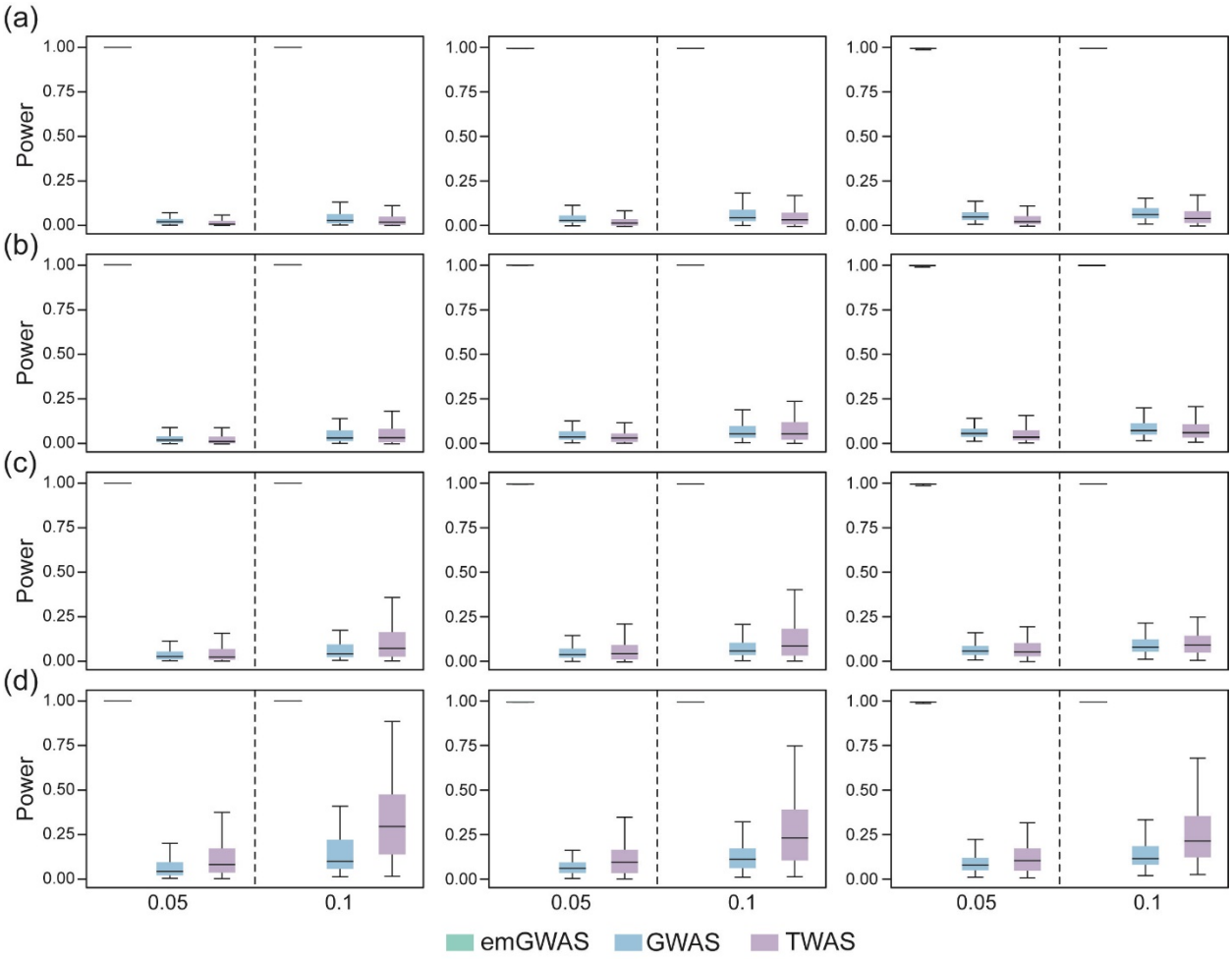

Figure 3: Causality scenario when expression heritability is low and PVX is high.

The PVX is 0.05 and 0.1 in the two columns as indicated by the X-axis labels. The numbers of genes contributed to phenotype for the left, middle and right panels are 4, 9 and 13 respectively. The expression heritability levels in (a), (b), (c) and (d) are 0.025, 0.03, 0.04 and 0.08 respectively. The number of causal variants per gene is randomly sampled from the interval [4,9].

Pleiotropy scenario

Again, we first look at the conditions when expression heritability is high and trait heritability is low (**Figure 4**). In contrast to the cases in the causality scenario, emGWAS has very low power

467 compared to TWAS and GWAS. The underlying explanation of this observation is that, when the
468 genetic variants do not affect phenotype through expressions, the non-genetic effects in actual
469 expressions cause noises. In contrast, in TWAS, if the elastic net training accurately captures
470 the genetic contribution to the expressions, the predictive expressions used in TWAS will be
471 advantageous over actual expressions. Unavoidably there are errors in this elastic net training,
472 as revealed by **Table 1**. However, our results showed that the loss of power due to non-genetic
473 effects is overwhelmingly larger than the problem caused by errors in the elastic net training.
474
475 Similar to the cases in the casualty scenario, TWAS in general outperforms GWAS except for
476 the cases when the trait heritability is extremely low and the number of contributing genes is
477 large, which is rare in practice. So the conclusion is that, again, TWAS is more powerful over
478 GWAS when expression heritability is high (in the pleiotropy scenario).
479

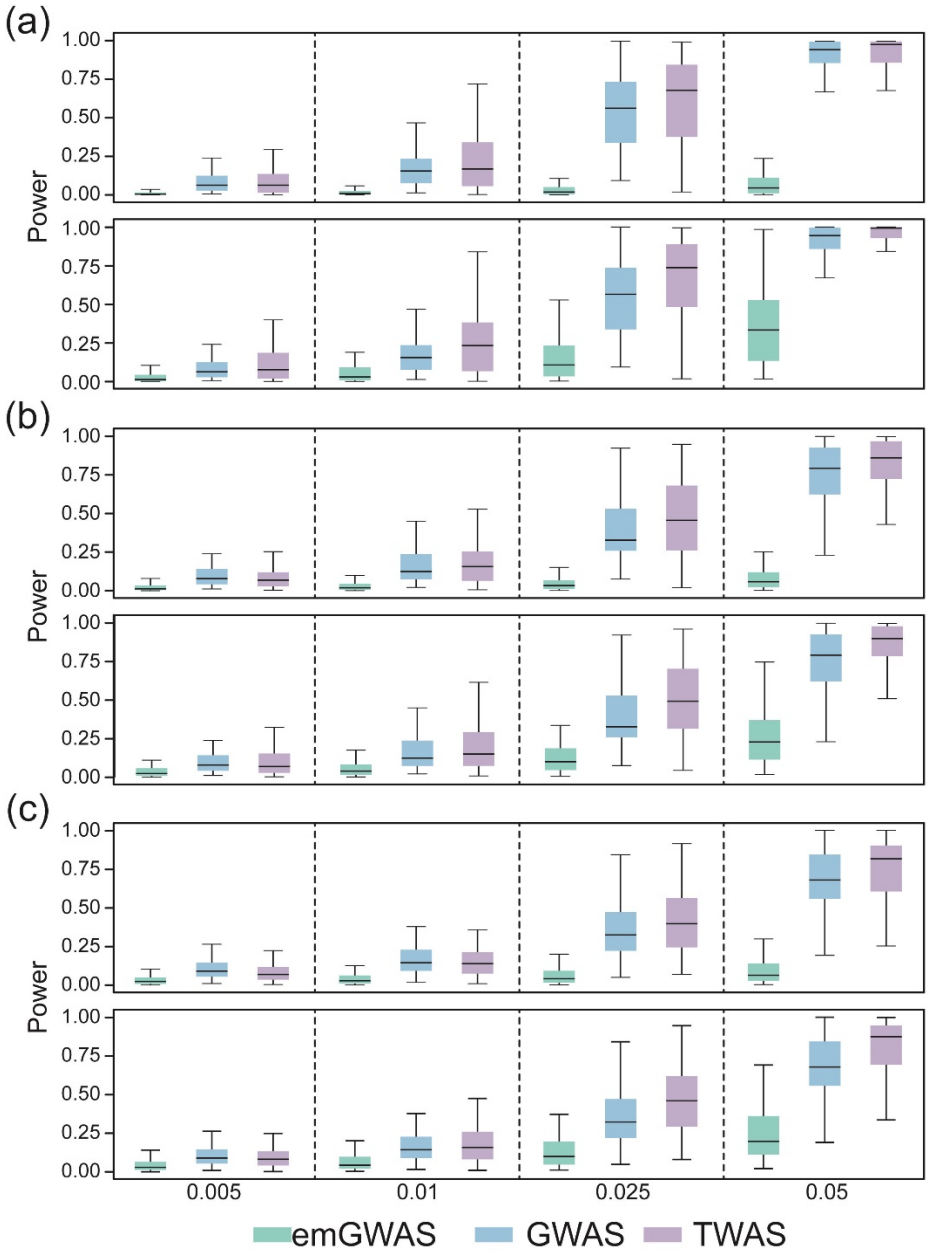

Figure 4: Pleiotropy scenario when expression heritability is high and trait heritability is low. The trait heritability is 0.005, 0.01, 0.025, and 0.05 in the four columns as indicated by the X-axis labels. The numbers of genes contributing to phenotype for (a), (b) and (c) are 4, 9 and 13 respectively. The expression heritability for the top and the bottom panels in (a), (b) and (c) are 0.1 and 0.3 respectively. The number of causal variants per gene is randomly sampled from the interval [4,9].

Finally, we looked at the cases when expression heritability is low but the trait heritability is high. Again, it is obvious that the emGWAS has the lowest power among the three models. Similar to the causality scenario, the comparison between TWAS and GWAS also shows a turn-point: the power of TWAS is lower than GWAS when the expression heritability is 0.025 or 0.04 (**Figure 5a, b**), TWAS shows similar performance to GWAS when the expression heritability is 0.06 (**Figure 5c**), whereas TWAS outperforms GWAS when the expression heritability is 0.08 (**Figure 5d**).

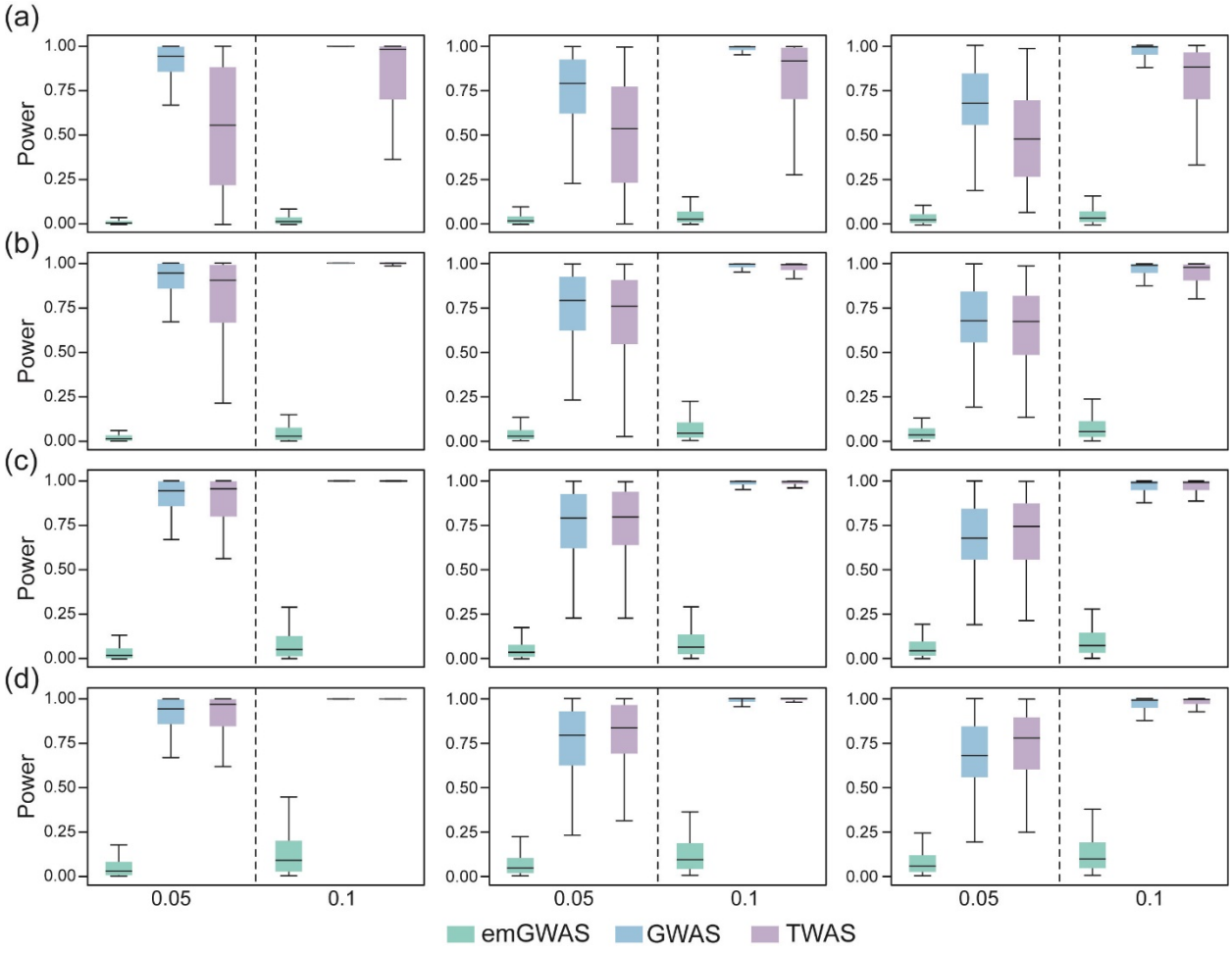

Figure 5: Pleiotropy scenario when expression heritability is low and trait heritability is high.

498 The PVX is 0.05 and 0.1 in the two columns as indicated by the X-axis labels. The numbers of genes
499 contributing to phenotype for the left, middle and right panels are 4, 9 and 13 respectively. The
500 expression heritability levels in (a), (b), (c) and (d) are 0.025, 0.04, 0.06 and 0.08 respectively. The
501 number of causal variants per gene is randomly sampled from the interval [4,9].
502
503
504 To recap, we observed the following two phenomena as take-home messages (**Figure 6**):
505 emGWAS is more powerful than TWAS and GWAS in the casualty scenario, while it is less
506 powerful in the pleiotropy scenario regardless of the accuracy level of the predicted expressions
507 (quantified in **Table 1**). This demonstrates the advantage of using predicted expression in
508 TWAS over using the actual expression in emGWAS. When the non-genetic components in
509 expression do not contribute to phenotype (i.e., in the pleiotropy scenario), the predicted
510 expressions turn out to be better as they capture the genetic contributions better than actual
511 expressions that are influenced by non-genetic contributions. The comparisons between GWAS
512 and TWAS revealed two turn-points 0.04 and 0.06 in the causality and pleiotropy scenarios
513 respectively. If the expression heritability is lower than 0.04 (in causality scenario) or 0.06 (in
514 pleiotropy scenario), one should resort to traditional GWAS, instead of using TWAS.
515

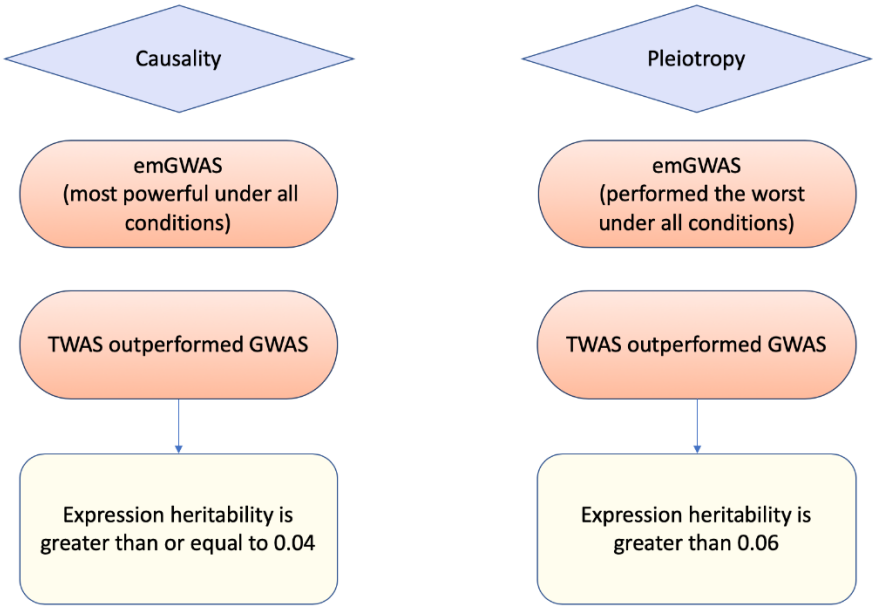

Figure 6: Overall comparison between emGWAS, GWAS and TWAS.

These turn-points have an immediate impact on practical association mappings. As there are not many publications disclosing the estimated expression heritability, we analyzed published R^2 (between predicted and actual expressions) as an approximation of the expression heritability. Although in practice different tools may use different predictive models, **Table 1** roughly indicates the difference between expression heritability and R^2 in our procedure, i.e. R^2 of 0.023 and 0.044 for the corresponding expression heritability 0.04 and 0.06 respectively. **Table 2** shows a list of published significant results in the PrediXcan paper (GAMAZON *et al.* 2015); and **Table 3** shows the average R^2 of genes included in recently published TWAS projects. One could see that the R^2 of a proportion of discovered genes are much lower than 0.044 or 0.023 (**Table 2**), and most of the average R^2 over different projects are also lower than 0.044 0r 0.023 (**Table 3**). Indicated by our power analysis, conducting TWAS on genes with low expression heritability is questionable, and for such genes an analysis using the standard GWAS protocol may yield higher power. Although here we are not able to figure out whether

532 these genes follow causality or pleiotropy scenario by only searching for the discovered genes,
533 in practice, one may use other advanced statistical models (SCHADT *et al.* 2005) to choose an
534 appropriate threshold to distinguish between pleiotropy and causality.
535
536 Therefore, we suggest a modified TWAS protocol as follows. First, one needs to estimate
537 expression heritability in the reference panel and filter out genes with expression heritability
538 lower than 0.04. Second, after conducting the TWAS, one analyzes the causality to decide
539 which threshold (0.04 or 0.06) should be used based on the identified underlying scenario
540 (causality or pleiotropy). If the significant gene's expression heritability is around the threshold,
541 then one may also carry out GWAS directly as a supplement.
542
543
544 **Table 2: Top R^2 and P-values of significant genes published by PrediXcan (GAMAZON *et al.* 2015).**
545 The R^2 are in bold and underlined if the values are less than 0.023 (corresponding to the expression
546 heritability of 0.04); R^2 are in bold if the values are between 0.023 and 0.44 (corresponding to the
547 expression heritability of 0.06).

Gene	PrediXcan P value	Cross-validated prediction R^2
DCLRE1B	2.46×10^{-11}	**0.0388**
PTPN22	1.44×10^{-8}	0.0795
PTPRE	7.71×10^{-7}	**0.0355**
ATG16L1	1.94×10^{-10}	0.0638
IL23R	1.74×10^{-7}	**0.0378**
APEH	2.77×10^{-7}	0.1164
ZNF300	6.29×10^{-7}	**0.0387**
NKD1	8.91×10^{-7}	0.0693
BSN	2.89×10^{-6}	0.2336

GPX1	3.87×10^{-6}	**<u>0.0211</u>**
SLC22A5	5.75×10^{-6}	0.6356
KCNN4	2.62×10^{-6}	0.4655
DCLRE1B	4.34×10^{-15}	**0.0388**
ZNF165	2.92×10^{-13}	**0.0374**
ERBB3	1.01×10^{-11}	0.2206
EGFL8	2.52×10^{-10}	*0.0558*
C6orf136	2.52×10^{-10}	**<u>0.0137</u>**
HCG27	2.52×10^{-10}	0.3721
GTF2H4	2.52×10^{-10}	0.0982
DDR1	2.52×10^{-10}	0.1427
AGER	2.52×10^{-10}	*0.0502*
POU5F1	2.52×10^{-10}	0.2874
ATP6V1G2	2.52×10^{-10}	0.2543
TUBB	2.52×10^{-10}	**0.0295**
AIF1	2.52×10^{-10}	**0.039**
CYP21A2	2.52×10^{-10}	0.229
LSM2	2.52×10^{-10}	**0.0317**
VARS2	2.52×10^{-10}	0.3628
APOM	2.52×10^{-10}	0.0699
DDAH2	2.52×10^{-10}	0.1943
NCR3	2.52×10^{-10}	0.2548
ZSCAN16	7.37×10^{-10}	**0.0291**
ZKSCAN4	7.73×10^{-10}	0.0991
PTPN22	5.41×10^{-9}	0.0795

RPS26	6.00×10^{-9}	0.0719
GDF11	9.11×10^{-9}	**0.0341**
SUOX	4.49×10^{-8}	0.1339
BTN3A2	3.30×10^{-7}	0.7662
PRSS16	1.34×10^{-6}	0.1639
FAM109A	1.94×10^{-6}	0.0665
SH2B3	3.05×10^{-6}	**0.0368**

548

549 **Table 3: Average R^2 in published TWAS projects.**

Title of the publication	Description of prediction accuracy
Large-scale transcriptome-wide association study identifies new prostate cancer risk regions (MANCUSO *et al.* 2018)	The mean R^2 = 0.07 for measured and predicted gene expression for TCGA normal prostate samples using models fitted in GTEx normal prostate.
A framework for transcriptome-wide association studies in breast cancer in diverse study populations (BHATTACHARYA *et al.* 2020)	The median CV R^2 for the 153 genes is 0.011 in both African American and white women.
Evaluation of PrediXcan for prioritizing GWAS associations and predicting gene Expression (LI *et al.* 2018)	The average of prediction accuracy (R^2) is 0.023 for the DGN model and 0.02 for the GTEx model, with both using whole blood model.
A gene-based association method for mapping traits using reference transcriptome data (GAMAZON *et al.* 2015)	The average prediction R^2 value is 0.0197 for GEUVADIS LCLs. For GTEx tissues, the prediction R^2 values are 0.0367 (adipose), 0.0358 (tibial artery), 0.0356 (left-ventricular heart), 0.0359 (lung), 0.0269 (muscle), 0.0422 (tibial nerve), 0.0374 (sun-exposed skin), 0.0398 (thyroid) and 0.0458 (whole blood).

550

551

552

553 **CONCLUSION & DISCUSSION**

554

555 In this work, based on the decorrelation procedure frequently adopted in solving LMM, we
556 derived the NCP for three association mapping protocols: GWAS, TWAS and emGWAS. This
557 innovative derivation enabled us to derive closed-form power estimation for the three protocols.
558 Our power analysis revealed two critical insights that may lead to substantial impact in practice.
559 **First**, in the pleiotropy scenario, TWAS that is based on predicted expressions is
560 overwhelmingly more powerful than emGWAS that uses actual expressions, regardless of the
561 performance of the prediction *per se* (**Table 1**). This outcome suggests that, in practice, even if
562 expressions are assessed by real experiments, TWAS still has its advantage and may be
563 carried out. Theoretically, this conclusion is not too surprising for statisticians but may be a bit
564 counterintuitive for experimentalists. Nevertheless, this result strengthens some people's
565 understanding of what TWAS is exactly: the key insight into TWAS is that it utilizes expression
566 as an objective function to select a linear combination of weighted genetic variants, rather than
567 predicting expressions to a good extent. This is similar to the idea of denoising in the field of
568 machine learning (TIAN *et al.* 2019). **Second**, the condition that TWAS is more powerful than
569 GWAS is dependent on expression heritability. When the expression heritability is lower than
570 0.06 (in the pleiotropy scenario) or 0.04 (in the casualty scenario), GWAS is more powerful than
571 TWAS, although GWAS does not use any information in gene expression that TWAS does. This
572 finding indicates that many TWAS conducted in practice might be suboptimal when expression
573 heritability is low (**Table 2 & 3**). As such, our power analysis provided quantified guidelines in
574 practice to direct researchers to an appropriate choice of association mapping protocol.

575

576 A recent publication (VETURI AND RITCHIE 2018) compared the statistical powers of GWAS and
577 TWAS. Methodologically, to calculate the powers, they utilized pure simulations whereas we
578 derived the closed form using NCP which does not depend on specific implementations of the
579 protocols and therefore may reveal more theoretical properties that are universally true. The
580 focus of their work is also different from ours: they put more attention on multiple competing
581 techniques used to train the genotype-expression predictive model and the impact of actual
582 number of causal genetic variants, whereas we focused more on revealing the superiority
583 ordering of GWAS, TWAS and emGWAS and applying it to guide the practice of TWAS. They
584 did not include emGWAS, whereas we did not include eGWAS (a method using eQTL to assist
585 association mapping) that they considered. Our and their conclusions are also not identical.
586 Although we both highlighted the importance of expression heritability, they concluded that
587 expression heritability affects power only under the causality scenario, while it does not under
588 the pleiotropy scenario. In contrast, we concluded that expression heritability is influential under
589 both scenarios.
590
591 Finally, our closed-form derivation can be easily adapted to other middle omics (or
592 endophenotype) directed methods such as IWAS (XU *et al.* 2017) and PWAS (OKADA *et al.*
593 2016; BRANDES *et al.* 2020). In fact, the Z in our formulas could represent images or proteins or
594 any other omics, and thus no further modifications of NCPs are necessary when adapting our
595 work to other omics.
596
597 There are three limitations in the present study. Although our closed-form derivation brought
598 benefits of its easy adoption and its ability of revealing universal principles independent of
599 specific implementations, it has its own limitations. For instance, our closed-form derivation by
600 its nature is not able to capture power loss due to limitations of implementations or the bias of a

specific dataset. Also, closed-form derivation is more sensitive to model assumptions than empirical methods based on simulations. Another issue is that our NCP has approximation involved. In the decorrelation procedure, one needs to estimate the variance component σ_g^2 in order to form the decorrelation matrix D_x. The estimated σ_g^2 introduces extra variability and therefore may decrease the power of the three protocols. We have omitted this extra variability because the estimation of σ_g^2 is generally mature and accurate in practice when sample size is at the level of thousands. Finally, in our comparison we only used linear models for GWAS and TWAS, and, in a totally different direction, kernel-based nonparametric and semiparametric methods were not considered. As a future work, we may explore the more robust analysis based on kernel methods for both GWAS (WU *et al.* 2010; WU *et al.* 2011) and TWAS (CAO *et al.* 2020).

DECLARATIONS

Abbreviations & Notation table

GWAS: Genome-wide association study;

TWAS: Transcriptome-wide association study;

MAF: Minor allele frequency;

X: Genotype

Z: Expression

Y: Phenotype

v: NCP

β: coefficient for a fixed effect term

u: random effect term in a mixed model

h^2: heritability or variance component coefficient

σ^2: variance component to be estimated in a mixed model

Availability of data and materials.

The 1000 Genomes Project genotype data is downloaded from the project website: https://www.internationalgenome.org

All the code to conduct the analysis is in our GitHub: https://github.com/boweiding/power-analysis-of-TWAS

Acknowledgements

J.W. is supported by an NSERC Discovery Grant (RGPIN-2018-04328). Q.L. is supported by an NSERC Discovery Grant (RGPIN-2017-04860), a Canada Foundation for Innovation JELF grant (36605), a New Frontiers in Research Fund (NFRFE-2018-00748), a Clinical Research Fund and a Startup grant supported by Alberta Children's Hospital Research Institute (ACHRI). C.C. is supported by ACHRI postdoctoral scholarship.

Authors' contributions

Conceived the study: Long Q, Wu J; Derived formula: Ding B, Wu J; Conducted analysis: Ding B, Cao C, Li Q; Wrote the manuscript: Long Q, Wu J, Cao C, Ding B.

APPENDICES

Appendix A: Solving the ElasticNet model for predicting expression using genotype

In our study, we optimize tuning parameters through a 10-fold cross-validation. First we set up a sequence of candidate α values with $\alpha \in (0.1, 0.15, 0.2, 0.25, \cdots, 0.9)$, under each α, we conduct the 10-fold cross-validation to select the optimal λ value which minimizes the objective function. For 10-fold cross-validation, the sample is randomly divided into 10 folds of equal size. Then the model is fit 10 times under a specific λ with each time omitting one of the folds without repetition for estimating parameters. For each time, the objective function will be computed in the omitted fold (test set). Then the mean objective function is the mean of the values among ten different test sets. Therefore, the optimal λ is the one under which the objective function is minimized among the whole sequence of λ under a specific α. After obtaining an optimal λ under each α, we compare the objective function of each combination of λ and α and choose the minimized error among all combinations. Thus, the tuning parameters λ and α with minimized objective function are the optimal ones.

Appendix B: Solving the variance component σ_g^2

In the decorrelation matrix D_x given by (10), although the kinship matrix K_x is known and further Λ_x and U_x can be decomposed from K_x, the variance components σ_g^2 and σ_e^2 are unknown parameters and should be estimated in order to obtain an estimated decorrelation matrix. The estimation of variance components σ_g^2 and σ_e^2 is generally mature and accurate in practice when sample size is at the level of thousands and the effect of each SNP or expression on the phenotype is small [35,42].

To estimate σ_g^2 and σ_e^2, following the efficient mixed-model association eXpedited (EMMAX) [35], we use a variance component model

$$Y_c = u + \varepsilon \tag{19}$$

to obtain the maximum likelihood estimator (MLE) of σ_g^2 and σ_e^2, where u and ε are given in (1.2), $Y_c = Y - \bar{Y}\mathbf{1}$ is an $n \times 1$ vector, Y is the $n \times 1$ phenotype vector, and $\bar{Y}$ is the sample mean $\bar{Y} = \sum_{i=1}^{n} Y_i$. Here Y_c is essentially the centralized Y with sample mean zero in order to have model (19) being appropriate [35]. In order to facilitate our following derivation and calculation, we define

$$\delta = \frac{\sigma_e^2}{\sigma_g^2}. \tag{20}$$

Now it holds approximately that $Y_c \sim N\left(0, \sigma_g^2(K_x + \delta I)\right)$. The next claim presents the MLEs of σ_g^2 and σ_e^2.

Claim 1. *Under models (19) and (20), the MLEs of σ_g^2 and σ_e^2 are given by*

$$\hat{\sigma}_g^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{V_{xi}^2}{\lambda_{xi} + \hat{\delta}}, \tag{21}$$

$$\hat{\sigma}_e^2 = \frac{\hat{\delta}}{n}\sum_{i=1}^{n} \frac{V_{xi}^2}{\lambda_{xi} + \hat{\delta}}, \tag{22}$$

where $V_x = U_x^T Y_c = (V_{x1}, \cdots, V_{xn})^T$ and $\hat{\delta}$ is the solution to the estimating equation

$$\sum_{i=1}^{n}\left[\frac{nV_{xi}^2}{\left(\sum_{i=1}^{n}\frac{V_{xi}^2}{\lambda_{xi}+\delta}\right)\left(\lambda_{xi}+\hat{\delta}\right)^2}-\frac{1}{\lambda_{xi}+\hat{\delta}}\right]=0. \tag{23}$$

Proof. Let $\Sigma=\sigma_g^2(K_x+\delta I)$, then by the eigen-decomposition $K_x=U_x\Lambda_x U_x^T$ and the facts that $U_xU_x^T=U_x^TU_x=I$ we have

$$\begin{aligned}U_x^T\Sigma U_x&=\sigma_g^2U_x^TK_xU_x+\sigma_g^2\delta U_x^TU_x\\&=\sigma_g^2(\Lambda_x+\delta I)\\&=\begin{pmatrix}\sigma_g^2(\lambda_{x1}+\delta) & 0 & 0 & \cdots & 0\\ 0 & \sigma_g^2(\lambda_{x2}+\delta) & 0 & \cdots\cdots & 0\\ \cdots & \cdots & \cdots & \cdots\cdots & \cdots\\ 0 & 0 & 0 & \cdots\cdots & \sigma_g^2(\lambda_{xn}+\delta)\end{pmatrix},\end{aligned}$$

and then

$$|\Sigma|=|U_x^T\Sigma U_x|=\prod_{i=1}^{n}\sigma_g^2\,(\lambda_{xi}+\delta).$$

With $V_x=U_x^TY_c$, we rewrite $Y_c^T\Sigma^{-1}Y_c$ as

$$\begin{aligned}Y_c^T\Sigma^{-1}Y_c&=(U_x^TY_c)^TU_x^T\Sigma^{-1}U_xU_x^TY_c\\&=V_x^T(U_x^T\Sigma U_x)^{-1}V_x\\&=\sum_{i=1}^{n}V_{xi}^2\left[\sigma_g^2(\lambda_{xi}+\delta)\right]^{-1}.\end{aligned}$$

Now the log-likelihood function of σ_g^2 and δ is given by

$$
\begin{aligned}
l\left(\sigma_g^2, \delta\right) &= \ln[f(Y_{c1}, Y_{c2}, \cdots, Y_{cn})] \\
&= \ln\left[(2\pi)^{-\frac{n}{2}}|\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} Y_c^T \Sigma^{-1} Y_c\right)\right] \\
&= -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma| - \frac{1}{2} Y_c^T \Sigma^{-1} Y_c \\
&= -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln\left[\prod_{i=1}^{n} \sigma_g^2\,(\lambda_{xi} + \delta)\right] - \frac{1}{2}\sum_{i=1}^{n} V_{xi}^2 \left[\sigma_g^2(\lambda_{xi} + \delta)\right]^{-1} \\
&= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln \sigma_g^2 - \frac{1}{2}\sum_{i=1}^{n} \ln(\lambda_{xi} + \delta) - \frac{1}{2}\left(\sigma_g^2\right)^{-1} \sum_{i=1}^{n} V_{xi}^2\,(\lambda_{xi} + \delta)^{-1}.
\end{aligned}
$$

Then the MLE estimating equations for σ_g^2 and δ are

$$
\frac{\partial}{\partial\left(\sigma_g^2\right)} l\left(\sigma_g^2, \delta\right) = -\frac{n}{2\sigma_g^2} + \frac{1}{2\sigma_g^4}\sum_{i=1}^{n} \frac{V_{xi}^2}{\lambda_{xi} + \delta} = 0, \tag{24}
$$

$$
\frac{\partial}{\partial\delta} l\left(\sigma_g^2, \delta\right) = -\frac{1}{2}\sum_{i=1}^{n} \frac{1}{\lambda_{xi} + \delta} + \frac{1}{2}\sum_{i=1}^{n} \frac{V_{xi}^2}{\sigma_g^2(\lambda_{xi} + \delta)^2} = \frac{1}{2}\sum_{i=1}^{n}\left[\frac{V_{xi}^2}{\sigma_g^2(\lambda_{xi} + \delta)^2} - \frac{1}{\lambda_{xi} + \delta}\right] = 0. \tag{25}
$$

From (24) we obtain

$$
\sigma_g^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{V_{xi}^2}{\lambda_{xi} + \delta},
$$

and plugging this into (25) we obtain

$$
\sum_{i=1}^{n}\left[\frac{nV_{xi}^2}{\left(\sum_{i=1}^{n} \frac{V_{xi}^2}{\lambda_{xi} + \delta}\right)(\lambda_{xi} + \delta)^2} - \frac{1}{(\lambda_{xi} + \delta)}\right] = 0.
$$

Finally, by (20), hence the result.

It has been shown [35] that the MLEs of σ_g^2 and σ_e^2 given in (2.4) and (2.5) respectively have good accuracy. In order to solve for δ in (2.6), one may use Newton-Raphon's method.

Specifically, let $g(\delta) = \sum_{i=1}^{n}\left[\frac{nV_{xi}^2}{\left(\sum_{i=1}^{n}\frac{V_{xi}^2}{\lambda_{xi}+\delta}\right)(\lambda_{xi}+\delta)^2} - \frac{1}{\lambda_{xi}+\delta}\right]$, then with the current δ_n, we update it by $\delta_{n+1} = \delta_n - \frac{g(\delta_n)}{g'(\delta_n)}$ until $|\delta_{n+1} - \delta_n| \leq 10^{-6}$, say. Then the converged δ_{n+1} will be the MLE of δ. Finally, plugging the MLEs $\hat{\sigma}_g^2$ and $\hat{\sigma}_e^2$ back into (10) gives the estimated decorrelation matrix

$$\widehat{D}_x = \left(\hat{\sigma}_g^2\Lambda_x + \hat{\sigma}_e^2 I\right)^{-\frac{1}{2}}U_x^T. \tag{26}$$

Appendix C: Proof of the soundness of the decorrelation procedure

Since K_x is assumed of full rank, by eigen-decomposition we denote $K_x = U_x\Lambda_x U_x^T$, where $\Lambda_x = diag(\lambda_{x1}, \dots, \lambda_{xn}) \in \mathbb{R}^{n\times n}$ is a diagonal matrix with diagonal elements the eigenvalues λ_{xi}'s of K_x in decreasing order, $U_x \in \mathbb{R}^{n\times n}$ is the eigenvector matrix with each column the eigenvector associated with the corresponding eigenvalues such that $U_x U_x^T = U_x^T U_x = I$. We assume $(\sigma_g^2\Lambda_x + \sigma_e^2 I)$ is of full rank. Then for all the three analytical models, (1) for GWAS, (4) for emGWAS and (6) for TWAS, we have the following claim on their decorrelation matrix D_x.

Claim 2. *If*

$$D_x = \left(\sigma_g^2\Lambda_x + \sigma_e^2 I\right)^{-\frac{1}{2}}U_x^T, \tag{10}$$

then $D_x(u+\varepsilon) \sim N(0, I)$.

Proof. For the LMMs given in (1), (4), and (6), the covariance matrix of the n dimensional phenotype vector Y is

$$\begin{aligned} Var(Y) &= \sigma_g^2 K_x + \sigma_e^2 I \\ &= \sigma_g^2 U_x \Lambda_x U_x^T + \sigma_e^2 U_x U_x^T \\ &= U_x\left(\sigma_g^2 \Lambda_x + \sigma_e^2 I\right) U_x^T. \end{aligned}$$

With $D_x = \left(\sigma_g^2 \Lambda_x + \sigma_e^2 I\right)^{-\frac{1}{2}} U_x^T$, we have

$$\begin{aligned} Var(D_x Y) &= D_x Var(Y) D_x^T \\ &= D_x U\left(\sigma_g^2 \Lambda_x + \sigma_e^2 I\right) U_x^T D_x^T \\ &= \left(\sigma_g^2 \Lambda_x + \sigma_e^2 I\right)^{-\frac{1}{2}} U_x^T U_x \left(\sigma_g^2 \Lambda_x + \sigma_e^2 I\right) U_x^T \left(\left(\sigma_g^2 \Lambda_x + \sigma_e^2 I\right)^{-\frac{1}{2}} U_x^T\right)^T \\ &= \left(\sigma_g^2 \Lambda_x + \sigma_e^2 I\right)^{-\frac{1}{2}} U_x^T U_x \left(\sigma_g^2 \Lambda_x + \sigma_e^2 I\right) U_x^T U_x \left(\sigma_g^2 \Lambda_x + \sigma_e^2 I\right)^{-\frac{1}{2}} \\ &= \left(\sigma_g^2 \Lambda_x + \sigma_e^2 I\right)^{-\frac{1}{2}} I \left(\sigma_g^2 \Lambda_x + \sigma_e^2 I\right) I \left(\sigma_g^2 \Lambda_x + \sigma_e^2 I\right)^{-\frac{1}{2}} \\ &= \left(\sigma_g^2 \Lambda_x + \sigma_e^2 I\right)^0 \\ &= I. \end{aligned}$$

Hence the result.

Appendix D: Proof of the NCP in an LMM

For LMM (1), in order to identify the significant genetic variants, if we left multiply each term in (1) by the decorrelation matrix D_x given by (10), then for the jth genetic variant vector, model (1) is reduced to

$$Y^* = \beta_{j0}X_0 + \beta_{j1}X_j^* + \varepsilon^*, \tag{27}$$

Where $Y^* = D_x Y = (Y_1^*, Y_2^*, \dots, Y_n^*)^T$ is the transformed phenotype vector, $X_0 = D_x \mathbf{1} = (D_{x1\cdot}, D_{x2\cdot}, \dots, D_{xn\cdot})^T$ with $D_{xi\cdot} = \sum_{j=1}^n D_{xij}$, $X_j^* = D_x X_j = (X_{1j}^*, X_{2j}^*, \dots, X_{nj}^*)^T$ is the transformed j^{th} genetic variant vector, and $\varepsilon^* = D_x(u + \varepsilon)$ is the random vector such that $\varepsilon^* \sim N(0, I)$.

Claim 3. *For GWAS (1), the MLEs of β_{j0} and β_{j1} are*

$$\beta_{j1_mle} = \frac{\sum_{i=1}^n X_{ij}^* Y_i^* \sum_{i=1}^n \mathrm{D}_{xi\cdot}^2 - \sum_{i=1}^n Y_i^* D_{xi\cdot} \sum_{i=1}^n X_{ij}^* D_{xi\cdot}}{\sum_{i=1}^n \left(X_{ij}^*\right)^2 \sum_{i=1}^n \mathrm{D}_{xi\cdot}^2 - \left(\sum_{i=1}^n X_{ij}^* \mathrm{D}_{xi\cdot}\right)^2}, \tag{28}$$

$$\beta_{j0_mle} = \frac{\sum_{i=1}^n Y_i^* \mathrm{D}_{xi\cdot}}{\sum_{i=1}^n \mathrm{D}_{xi\cdot}^2} - \frac{\beta_{j1_mle} \sum_{i=1}^n X_{ij}^* \mathrm{D}_{xi\cdot}}{\sum_{i=1}^n \mathrm{D}_{xi\cdot}^2}, \tag{29}$$

And the variance of β_{j1_mle} is

$$Var(\beta_{j1_mle}) = \frac{\sum_{i=1}^n \mathrm{D}_{xi\cdot}^2}{\sum_{i=1}^n \left(X_{ij}^*\right)^2 \sum_{i=1}^n \widehat{D}_{xi\cdot}^2 - \left(\sum_{i=1}^n X_{ij}^* D_{xi\cdot}\right)^2}. \tag{30}$$

Proof. For LMM (1), or equivalently the transformed model (27), the log-likelihood function is

$$l(\beta_{j0};\beta_{j1}) = \ln\left(\prod_{i=1}^{n} f_{Y^*}\,(Y_1^*, Y_2^*, \cdots, Y_n^*)\right)$$
$$= -n\ln\sqrt{2\pi} - \sum_{i=1}^{n}\frac{\left(Y_i^* - \beta_{j0}\mathrm{D}_{xi\cdot} - \beta_{j1}X_{ij}^*\right)^2}{2}.$$

Then the MLE estimating equations for β_{j0} and β_{j1} are:

$$\frac{\partial}{\partial\beta_{j0}} l\left(\beta_{j0},\beta_{j1}\right) = \sum_{i=1}^{n}\mathrm{D}_{xi\cdot}\left(Y_i^* - \beta_{j0}\mathrm{D}_{xi\cdot} - \beta_{j1}X_{ij}^*\right) = \sum_{i=1}^{n} Y_i^*\mathrm{D}_{xi\cdot} - \beta_{j0}\sum_{i=1}^{n}\mathrm{D}_{xi\cdot}^2 - \beta_{j1}\sum_{i=1}^{n} X_{ij}^*\mathrm{D}_{xi\cdot} = 0, \quad (31)$$

$$\frac{\partial}{\partial\beta_{j1}} l\left(\beta_{j0},\beta_{j1}\right) = \sum_{i=1}^{n} X_{ij}^*\left(Y_i^* - \mathrm{D}_{xi\cdot}\beta_{j0} - \beta_{j1}X_{ij}^*\right) = \sum_{i=1}^{n} X_{ij}^*\,Y_i^* - \beta_{j0}\sum_{i=1}^{n} X_{ij}^*\,\mathrm{D}_{xi\cdot} - \beta_{j1}\sum_{i=1}^{n}\left(X_{ij}^*\right)^2 = 0. \quad (32)$$

From (31) we obtain (29). Plugging (29) into (32) gives:

$$0 = \sum_{i=1}^{n} X_{ij}^*\,Y_i^* - \frac{\sum_{i=1}^{n} Y_i^*\mathrm{D}_{xi\cdot}\sum_{i=1}^{n} X_{ij}^*\,\mathrm{D}_{xi\cdot}}{\sum_{i=1}^{n}\mathrm{D}_{xi\cdot}^2} + \frac{\beta_{j1}(\sum_{i=1}^{n} X_{ij}^*\mathrm{D}_{xi\cdot})^2}{\sum_{i=1}^{n}\mathrm{D}_{xi\cdot}^2} - \beta_{j1}\sum_{i=1}^{n}\left(X_{ij}^*\right)^2$$

$$= \frac{\sum_{i=1}^{n} X_{ij}^*\,Y_i^*\sum_{i=1}^{n}\mathrm{D}_{xi\cdot}^2 - \sum_{i=1}^{n} Y_i^*\,\mathrm{D}_{xi\cdot}\sum_{i=1}^{n} X_{ij}^*\,\mathrm{D}_{xi\cdot}}{\sum_{i=1}^{n}\mathrm{D}_{xi\cdot}^2} - \beta_{j1}\frac{\sum_{i=1}^{n}\left(X_{ij}^*\right)^2\sum_{i=1}^{n}\mathrm{D}_{xi\cdot}^2 - \left(\sum_{i=1}^{n}\mathrm{D}_{xi\cdot}\,X_{ij}^*\right)^2}{\sum_{i=1}^{n}\mathrm{D}_{xi\cdot}^2},$$

and thus

$$\beta_{j1_{\mathrm{mle}}} = \frac{\sum_{i=1}^{n} X_{ij}^*\,Y_i^*\sum_{i=1}^{n}\mathrm{D}_{xi\cdot}^2 - \sum_{i=1}^{n} Y_i^*\,D_{xi\cdot}\sum_{i=1}^{n} X_{ij}^*\,D_{xi\cdot}}{\sum_{i=1}^{n}\left(X_{ij}^*\right)^2\sum_{i=1}^{n}\mathrm{D}_{xi\cdot}^2 - \left(\sum_{i=1}^{n} X_{ij}^*\mathrm{D}_{xi\cdot}\right)^2}$$

$$= \frac{\sum_{i=1}^{n} \mathrm{D}_{xi\cdot}^2 \sum_{i=1}^{n} \left(X_{ij}^* - D_{xi\cdot} \frac{\sum_{i=1}^{n} X_{ij}^* \mathrm{D}_{xi\cdot}}{\sum_{i=1}^{n} \mathrm{D}_{xi\cdot}^2} \right) Y_i^*}{\sum_{i=1}^{n} \left(X_{ij}^* \right)^2 \sum_{i=1}^{n} \mathrm{D}_{xi\cdot}^2 - \left(\sum_{i=1}^{n} \mathrm{D}_{xi\cdot} X_{ij}^* \right)^2},$$

i.e. (28) holds. Since by Claim 2 we have $Var(D_x Y) = Var(Y^*) = I$, direct calculation gives the variance of β_{j1_mle} as

$$\begin{aligned}
Var(\beta_{j1_mle}) &= \left(\frac{\sum_{i=1}^{n} D_{xi\cdot}^2}{\sum_{i=1}^{n} \left(X_{ij}^* \right)^2 \sum_{i=1}^{n} D_{xi\cdot}^2 - \left(\sum_{i=1}^{n} X_{ij}^* D_{xi\cdot} \right)^2} \right)^2 Var\left(\sum_{i=1}^{n} \left(X_{ij}^* - D_{xi\cdot} \frac{\sum_{i=1}^{n} X_{ij}^* D_{xi\cdot}}{\sum_{i=1}^{n} \mathrm{D}_{xi\cdot}^2} \right) Y_i^* \right) \\
&= \left(\frac{\sum_{i=1}^{n} \mathrm{D}_{xi\cdot}^2}{\sum_{i=1}^{n} \left(X_{ij}^* \right)^2 \sum_{i=1}^{n} \mathrm{D}_{xi\cdot}^2 - \left(\sum_{i=1}^{n} X_{ij}^* D_{xi\cdot} \right)^2} \right)^2 \sum_{i=1}^{n} \left(X_{ij}^* - \mathrm{D}_{xi\cdot} \frac{\sum_{i=1}^{n} X_{ij}^* D_{xi\cdot}}{\sum_{i=1}^{n} \mathrm{D}_{xi\cdot}^2} \right)^2 Var(Y_i^*) \\
&= \frac{\sum_{i=1}^{n} \left(X_{ij}^* \sum_{i=1}^{n} \mathrm{D}_{xi\cdot}^2 - \mathrm{D}_{xi\cdot} \sum_{i=1}^{n} X_{ij}^* \mathrm{D}_{xi\cdot} \right)^2}{\left(\sum_{i=1}^{n} \left(X_{ij}^* \right)^2 \sum_{i=1}^{n} \mathrm{D}_{xi\cdot}^2 - \left(\sum_{i=1}^{n} X_{ij}^* D_{xi\cdot} \right)^2 \right)^2} \\
&= \frac{\sum_{i=1}^{n} \left[\left(X_{ij}^* \right)^2 \left(\sum_{i=1}^{n} \mathrm{D}_{xi\cdot}^2 \right)^2 - 2 X_{ij}^* \mathrm{D}_{xi\cdot} \sum_{i=1}^{n} \mathrm{D}_{xi\cdot}^2 \sum_{i=1}^{n} X_{ij}^* \mathrm{D}_{xi\cdot} + \mathrm{D}_{xi\cdot}^2 \left(\sum_{i=1}^{n} X_{ij}^* \mathrm{D}_{xi\cdot} \right)^2 \right]}{\left(\sum_{i=1}^{n} \left(X_{ij}^* \right)^2 \sum_{i=1}^{n} \mathrm{D}_{xi\cdot}^2 - \left(\sum_{i=1}^{n} X_{ij}^* D_{xi\cdot} \right)^2 \right)^2} \\
&= \frac{\sum_{i=1}^{n} \left(X_{ij}^* \right)^2 \left(\sum_{i=1}^{n} \mathrm{D}_{xi\cdot}^2 \right)^2 - 2 \sum_{i=1}^{n} \widehat{D}_{xi\cdot}^2 \left(\sum_{i=1}^{n} X_{ij}^* \mathrm{D}_{xi\cdot} \right)^2 + \sum_{i=1}^{n} \mathrm{D}_{xi\cdot}^2 \left(\sum_{i=1}^{n} X_{ij}^* D_{xi\cdot} \right)^2}{\left(\sum_{i=1}^{n} \left(X_{ij}^* \right)^2 \sum_{i=1}^{n} \widehat{D}_{xi\cdot}^2 - \left(\sum_{i=1}^{n} X_{ij}^* D_{xi\cdot} \right)^2 \right)} \\
&= \frac{\sum_{i=1}^{n} \left(X_{ij}^* \right)^2 \left(\sum_{i=1}^{n} \mathrm{D}_{xi\cdot}^2 \right)^2 - \sum_{i=1}^{n} \mathrm{D}_{xi\cdot}^2 \left(\sum_{i=1}^{n} X_{ij}^* D_{xi\cdot} \right)^2}{\left(\sum_{i=1}^{n} \left(X_{ij}^* \right)^2 \sum_{i=1}^{n} \mathrm{D}_{xi\cdot}^2 - \left(\sum_{i=1}^{n} X_{ij}^* D_{xi\cdot} \right)^2 \right)^2} \\
&= \frac{\sum_{i=1}^{n} \mathrm{D}_{xi\cdot}^2}{\sum_{i=1}^{n} \left(X_{ij}^* \right)^2 \sum_{i=1}^{n} \widehat{D}_{xi\cdot}^2 - \left(\sum_{i=1}^{n} X_{ij}^* D_{xi\cdot} \right)^2}.
\end{aligned}$$

Since the true variance components σ_g^2 and σ_e^2 are unknown and thus the decorrelation matrix D_x is unknown either, we use the estimated $\widehat{D}_x$ in (26) instead. Plugging $\widehat{D}_x$ into (28) and (30), we obtain an estimated MLE and an estimated standard deviation of the MLE given respectively by

$$\hat{\beta}_{j1_mle} = \frac{\sum_{i=1}^{n} \hat{X}_{ij}^{*}\, \hat{Y}_{i}^{*} \sum_{i=1}^{n} \widehat{\mathrm{D}}_{xi\cdot}^{2} - \sum_{i=1}^{n} \hat{Y}_{i}^{*}\, \widehat{D}_{xi\cdot} \sum_{i=1}^{n} \hat{X}_{ij}^{*}\, \widehat{D}_{xi\cdot}}{\sum_{i=1}^{n} \left(\hat{X}_{ij}^{*}\right)^{2} \sum_{i=1}^{n} \widehat{\mathrm{D}}_{xi\cdot}^{2} - \left(\sum_{i=1}^{n} \hat{X}_{ij}^{*} \widehat{\mathrm{D}}_{xi\cdot}\right)^{2}},$$

$$\widehat{SD}\left(\beta_{j1_mle}\right) = \sqrt{\frac{\sum_{i=1}^{n} \widehat{D}_{xi\cdot}^{2}}{\sum_{i=1}^{n} \left(X_{ij}^{*}\right)^{2} \sum_{i=1}^{n} \widehat{D}_{xi\cdot}^{2} - \left(\sum_{i=1}^{n} \widehat{D}_{xi\cdot}\, X_{ij}^{*}\right)^{2}}},$$

where $\hat{X}_{j}^{*} = \widehat{D}_{x} X_{j} = (\hat{X}_{1j}^{*}, \hat{X}_{2j}^{*}, \ldots, \hat{X}_{nj}^{*})^{T}$, $\hat{Y}^{*} = \widehat{D}_{x} Y = (\hat{Y}_{1}^{*}, \hat{Y}_{2}^{*}, \ldots, \hat{Y}_{n}^{*})^{T}$, and $\widehat{D}_{x\mathrm{i}\cdot} = \sum_{j=1}^{n} \widehat{D}_{xij}$.

Recall the NCP pf the t test statistic for the slope β in a simple linear regression model is given by $\upsilon = \frac{\beta}{SD(\hat{\beta})}$. Then the NCP υ_{Gj} based on MLE under the true alternative β_{j1} is given by

$$\upsilon_{Gj} = \frac{\beta_{j1}}{SD(\beta_{j1_mle})} = \frac{\beta_{j1}\sqrt{\sum_{i=1}^{n} \left(X_{ij}^{*}\right)^{2} \sum_{i=1}^{n} \mathrm{D}_{xi\cdot}^{2} - \left(\sum_{i=1}^{n} X_{ij}^{*} D_{xi\cdot}\right)^{2}}}{\sqrt{\sum_{i=1}^{n} \mathrm{D}_{xi\cdot}^{2}}}. \quad (33)$$

Since β_{j1} and D_{x} are unknown, when replacing them with the estimated MLE $\hat{\beta}_{j1_mle}$ and $\widehat{D}_{x}$ in (26) respectively into (33), an estimated NCP $\hat{\upsilon}_{Gj}$ for GWAS is given by (11).

Appendix E: Simulating expression and phenotype that have prespecified variance components under the causality scenario

For a certain l^{th} selected gene $\left(l = 1,2,\cdots, n_{z-sig}\right)$, the gene expression $Z_{(l)}$ was simulated by its corresponding local $n_{z(l)-sig}$ randomly picked genetic variants.

$$Z_{(l)} = \left(\frac{1}{Var\left(\sum_{k=1}^{n_{z(l)-sig}} X_{1(l)(k)} b_{(l)(k)}\right) + Var(\varepsilon_{(l)})} \right)^{\frac{1}{2}} \left(X_{(l)} b_{(l)} + \varepsilon_{(l)}\right), \tag{34}$$

where the $n \times 1$ vector $Z_{(l)}$ is the lth significant gene expression of n individuals, $X_{(l)}$ is an $n \times n_{z(l)-sig}$ matrix of randomly picked $n_{z(l)-sig}$ qualified variants in the lth significant gene, $X_{i(l)(k)}$ denotes the kth causal variant in the lth causal gene for the ith individual with $i = 1, \dots, n$, $b_{(l)} = (b_{(l)(1)}, b_{(l)(2)}, \dots, b_{(l)(n_{z(l)-sig})})^T$ is a $n_{z(l)-sig} \times 1$ vector of effect sizes of the causal variants, $\varepsilon_{(l)}$ is an $n \times 1$ vector of errors such that $\varepsilon_{(l)} \sim N(0, \mathrm{Var}(\varepsilon_{(l)})$. In our simulation $b_{(l)}$ was generated with each element a random value between 0 and 1. The gene expression was simulated based on the contribution from $X_{(l)} b_{(l)}$ and residuals from the expression heritability $h^2_{\mathrm{x=>z}}$. The expression heritability quantifies the contribution of genetic variants to the gene expression by

$$h^2_{\mathrm{x=>z}} = \frac{Var\left(\sum_{k=1}^{n_{z(l)-sig}} X_{1(l)(k)} b_{(l)(k)}\right)}{Var\left(\sum_{k=1}^{n_{z(l)-sig}} X_{1(l)(k)} b_{(l)(k)}\right) + Var(\varepsilon_{(l)})}, \tag{35}$$

where $Var\left(\sum_{k=1}^{n_{z(l)-sig}} X_{1(l)(k)} b_{(l)(k)}\right)$ was estimated by the sample variance given by

$$\widehat{Var}\left(\sum_{k=1}^{n_{z(l)-sig}} X_{1(l)(k)} b_{(l)(k)}\right) = \frac{1}{n-1} \sum_{i=1}^{n} \left(\sum_{k=1}^{n_{z(l)-sig}} X_{i(l)(k)} b_{(l)(k)} - \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{n_{z(l)-sig}} X_{i(l)(k)} b_{(l)(k)} \right)^2. \tag{36}$$

Plugging the estimated variance (36) into (35), and estimate of $Var(\varepsilon_{(l)})$ was given by

$$\widehat{Var}(\varepsilon_{(l)}) = \left(\frac{1}{h^2_{x=>z}} - 1 \right) \widehat{Var}\left(\sum_{k=1}^{n_{z(l)-sig}} X_{1(l)(k)} b_{(l)(k)}\right). \tag{37}$$

Then the errors $\varepsilon_{(l)}$ are generated randomly from the normal distribution $N\sim\left(0,\widehat{Var}(\varepsilon_{(l)})I\right)$. After replacing $Var\left(\sum_{k=1}^{n_{z(l)-sig}} X_{1(l)(k)}b_{(l)(k)}\right)$ and $Var(\varepsilon_{(l)})$ in model (34) with their estimates given in (36) and (37) respectively, we can then generate $Z_{(l)}$.

Following the steps above, the n_{z-sig} causal genes' expression $Z_{(l)}$'s are simulated. Since in causality scenario gene expression explains phenotypes, the latter were generated from the n_{z-sig} causal genes. More specifically,

$$Y_{ca} = \left(\frac{1}{Var\left(\sum_{l=1}^{n_{z-sig}} Z_{1(l)}b_{c(l)}\right) + Var(\varepsilon_c)}\right)^{\frac{1}{2}} (Zb_c + \varepsilon_c), \tag{38}$$

where the $n \times 1$ vector Y_{ca} is the phenotypes under causality scenario, $Z = (Z_{(1)}, Z_{(2)}, \dots, Z_{(n_{z-sig})})$ is an $n \times n_{z-sig}$ matrix of significant genes' expressions, $b_c = (b_{c(1)}, b_{c(2)}, \dots, b_{c(n_{z-sig})})^T$ is an $n_{z-sig} \times 1$ vector of effect sizes of the gene expressions, and the $n \times 1$ vector of errors $\varepsilon_c \sim N(0, Var(\varepsilon_c))$. Similarly, b_c was generated with each element a random value between 0 and 1. We estimated $Var\left(\sum_{l=1}^{n_{z-sig}} Z_{(l)}b_{c(l)}\right)$ by sample variance

$$\widehat{Var}\left(\sum_{l=1}^{n_{z-sig}} Z_{1(l)}b_{c(l)}\right) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\sum_{l=1}^{n_{z-sig}} Z_{i(l)}b_{c(l)} - \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{n_{z-sig}} Z_{i(l)}b_{c(l)}\right)^2. \tag{39}$$

Following the relationship between PVX and the variance component, i.e.

$$h^2_{\mathrm{z=>y}} = \frac{Var\left(\sum_{l=1}^{n_{z-sig}} Z_{1(l)}b_{c(l)}\right)}{Var\left(\sum_{l=1}^{n_{z-sig}} Z_{1(l)}b_{c(l)}\right) + Var(\varepsilon_c)} \tag{40}$$

$Var(\varepsilon_c)$ can be thus estimated by

$$\widehat{Var}(\varepsilon_c) = \left(\frac{1}{h^2_{z=>y}} - 1\right)\widehat{Var}\left(\sum_{l=1}^{n_{z-sig}} Z_{1(l)} b_{c(l)}\right). \quad (41)$$

Then the errors ε_{ca} were generated randomly from the normal distribution $N \sim \left(0, \widehat{Var}(\varepsilon_c)\right)$. Finally we replaced $Var\left(\sum_{l=1}^{n_{z-sig}} Z_{1(l)} b_{c(l)}\right)$ and $Var(\varepsilon_c)$ in model (38) with their estimates given in (39) and (41) respectively, and then Y_{ca} in (38) can be simulated.

Appendix F: Simulating expression and phenotype that have prespecified variance components under the pleiotropy scenario

Similar to the process in **Appendix E**, the phenotypes were simulated from a combination of n_{x-sig} significant genetic variants directly as follow

$$Y_p = \left(\frac{1}{Var\left(\sum_{k=1}^{n_{x-sig}} X_{1(k)} b_{p(k)}\right) + Var(\varepsilon_p)}\right)^{1/2} \left(\mathrm{X} b_p + \varepsilon_p\right), \quad (42)$$

where the $n \times 1$ vector Y_p is the phenotypes under causality scenario, $X = (X_{(1)}, X_{(2)}, \ldots, X_{(n_{x-sig})})$ is an $n \times n_{x-sig}$ matrix of significant genetic variants, b_p is a $n_{x-sig} \times 1$ vector of effect sizes, and ε_p is an $n \times 1$ vector of residuals. Similarly, we first generated the slopes b_p randomly with each element between 0 and 1, then we obtained $\mathrm{X} b_p$ and further the sample variance

$$\widehat{Var}\left(\sum_{k=1}^{n_{x-sig}} X_{1(k)} b_{p(k)}\right) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\sum_{k=1}^{n_{x-sig}} X_{i(k)} b_{p(k)} - \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{n_{x-sig}} X_{i(k)} b_{p(k)}\right)^2. \quad (43)$$

Based on the prespecified trait heritability $h^2_{\mathrm{x=>y}}$ and its relationship with variance components given by

$$h^2_{\mathrm{x=>y}} = \frac{Var\left(\sum_{k=1}^{n_{x-sig}} X_{1(k)} b_{p(k)}\right)}{Var\left(\sum_{k=1}^{n_{x-sig}} X_{1(k)} b_{p(k)}\right) + Var(\varepsilon_p)}, \quad (44)$$

$Var(\varepsilon_p)$ can then be estimated by

$$\widehat{Var}(\varepsilon_p) = \left(\frac{1}{h^2_{x=>\mathrm{y}}} - 1\right)\widehat{Var}\left(\sum_{k=1}^{n_{x-sig}} X_{1(k)} b_{p(k)}\right). \quad (45)$$

Then the errors ε_p were generated randomly from the normal distribution $N \sim \left(0, \widehat{Var}(\varepsilon_p)\right)$. Finally we replaced $Var\left(\sum_{k=1}^{n_{x-sig}} X_{1(k)} b_{p(k)}\right)$ and $Var(\varepsilon_p)$ in model (42) with their estimates given in (43) and (45) respectively, and then Y_p in (42) can be simulated.

References

Atkins, I., B. Kinnersley, Q. T. Ostrom, K. Labreche, D. Il'yasova *et al.*, 2019 Transcriptome-Wide Association Study Identifies New Candidate Susceptibility Genes for Glioma. Cancer Res 79**:** 2065-2071.

Aulchenko, Y. S., S. Ripke, A. Isaacs and C. M. van Duijn, 2007 GenABEL: an R library for genome-wide association analysis. Bioinformatics 23**:** 1294-1296.

Barbeira, A. N., S. P. Dickinson, R. Bonazzola, J. Zheng, H. E. Wheeler *et al.*, 2018 Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat Commun 9**:** 1825.

Bell, A. W., E. W. Deutsch, C. E. Au, R. E. Kearney, R. Beavis *et al.*, 2009 A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. Nat Methods 6**:** 423-430.

Bhattacharya, A., M. Garcia-Closas, A. F. Olshan, C. M. Perou, M. A. Troester *et al.*, 2020 A framework for transcriptome-wide association studies in breast cancer in diverse study populations. Genome Biol 21**:** 42.

Brandes, N., N. Linial and M. Linial, 2020 PWAS: Proteome-Wide Association Study, pp. 237-239. Springer International Publishing, Cham.

Cao, C., D. Kwok, S. Edie, Q. Li, B. Ding *et al.*, 2020 kTWAS: integrating kernel-machine with transcriptome-wide association studies improves statistical power and reveals novel genes. bioRxiv.

Coats, V. C., and M. E. Rumpho, 2014 The rhizosphere microbiota of plant invaders: an overview of recent advances in the microbiomics of invasive plants. Front Microbiol 5**:** 368.

Eddy, S., L. H. Mariani and M. Kretzler, 2020 Integrated multi-omics approaches to improve classification of chronic kidney disease. Nat Rev Nephrol.

Fukushima, A., M. Kusano, H. Redestig, M. Arita and K. Saito, 2009 Integrated omics approaches in plant systems biology. Curr Opin Chem Biol 13**:** 532-538.

Gamazon, E. R., H. E. Wheeler, K. P. Shah, S. V. Mozaffari, K. Aquino-Michaels *et al.*, 2015 A gene-based association method for mapping traits using reference transcriptome data. Nat Genet 47**:** 1091-1098.

Gogarten, S. M., T. Bhangale, M. P. Conomos, C. A. Laurie, C. P. McHugh *et al.*, 2012 GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. Bioinformatics 28**:** 3329-3331.

Gong, L., D. Zhang, Y. Lei, Y. Qian, X. Tan *et al.*, 2018 Transcriptome-wide association study identifies multiple genes and pathways associated with pancreatic cancer. Cancer Med 7**:** 5727-5732.

Gusev, A., A. Ko, H. Shi, G. Bhatia, W. Chung *et al.*, 2016 Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet 48**:** 245-252.

Gusev, A., N. Mancuso, H. Won, M. Kousi, H. K. Finucane *et al.*, 2018 Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. Nat Genet 50**:** 538-548.

Hasin, Y., M. Seldin and A. Lusis, 2017 Multi-omics approaches to disease. Genome Biol 18**:** 83.

Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S. Y. Kong *et al.*, 2010 Variance component model to account for sample structure in genome-wide association studies. Nat Genet 42**:** 348-354.

Klein, R. J., C. Zeiss, E. Y. Chew, J. Y. Tsai, R. S. Sackler *et al.*, 2005 Complement factor H polymorphism in age-related macular degeneration. Science 308**:** 385-389.

969 Li, B., S. S. Verma, Y. C. Veturi, A. Verma, Y. Bradford *et al.*, 2018 Evaluation of PrediXcan for
970 prioritizing GWAS associations and predicting gene expression. Pac Symp Biocomput
971 23**:** 448-459.
972 Long, Q., Q. Zhang, B. J. Vilhjalmsson, P. Forai, U. Seren *et al.*, 2013 JAWAMix5: an out-of-
973 core HDF5-based java implementation of whole-genome association studies using
974 mixed models. Bioinformatics 29**:** 1220-1222.
975 Mancuso, N., S. Gayther, A. Gusev, W. Zheng, K. L. Penney *et al.*, 2018 Large-scale
976 transcriptome-wide association study identifies new prostate cancer risk regions. Nat
977 Commun 9**:** 4079.
978 Mancuso, N., H. Shi, P. Goddard, G. Kichaev, A. Gusev *et al.*, 2017 Integrating Gene
979 Expression with Summary Association Statistics to Identify Genes Associated with 30
980 Complex Traits. Am J Hum Genet 100**:** 473-487.
981 McGettigan, P. A., 2013 Transcriptomics in the RNA-seq era. Curr Opin Chem Biol 17**:** 4-11.
982 Mills, M. C., and C. Rahal, 2019 A scientometric review of genome-wide association studies.
983 Commun Biol 2**:** 9.
984 Mogil, L. S., A. Andaleon, A. Badalamenti, S. P. Dickinson, X. Guo *et al.*, 2018 Genetic
985 architecture of gene expression traits across diverse populations. PLoS Genet 14**:**
986 e1007586.
987 Okada, H., H. A. Ebhardt, S. C. Vonesch, R. Aebersold and E. Hafen, 2016 Proteome-wide
988 association studies identify biochemical modules associated with a wing-size phenotype
989 in Drosophila melanogaster. Nat Commun 7**:** 12649.
990 Ozaki, K., Y. Ohnishi, A. Iida, A. Sekine, R. Yamada *et al.*, 2002 Functional SNPs in the
991 lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction.
992 Nat Genet 32**:** 650-654.
993 Ozsolak, F., and P. M. Milos, 2011 RNA sequencing: advances, challenges and opportunities.
994 Nat Rev Genet 12**:** 87-98.
995 Patterson, N., A. L. Price and D. Reich, 2006 Population structure and eigenanalysis. PLoS
996 Genet 2**:** e190.
997 Pible, O., and J. Armengaud, 2015 Improving the quality of genome, protein sequence, and
998 taxonomy databases: a prerequisite for microbiome meta-omics 2.0. Proteomics 15**:**
999 3418-3423.
1000 Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira *et al.*, 2007 PLINK: a tool set
1001 for whole-genome association and population-based linkage analyses. Am J Hum Genet
1002 81**:** 559-575.
1003 Rakitsch, B., C. Lippert, O. Stegle and K. Borgwardt, 2013 A Lasso multi-marker mixed model
1004 for association mapping with population structure correction. Bioinformatics 29**:** 206-214.
1005 Ratnapriya, R., O. A. Sosina, M. R. Starostik, M. Kwicklis, R. J. Kapphahn *et al.*, 2019 Retinal
1006 transcriptome and eQTL analyses identify genes associated with age-related macular
1007 degeneration. Nat Genet 51**:** 606-610.
1008 Schadt, E. E., J. Lamb, X. Yang, J. Zhu, S. Edwards *et al.*, 2005 An integrative genomics
1009 approach to infer causal associations between gene expression and disease. Nat Genet
1010 37**:** 710-717.
1011 Selevsek, N., C. Y. Chang, L. C. Gillet, P. Navarro, O. M. Bernhardt *et al.*, 2015 Reproducible
1012 and consistent quantification of the Saccharomyces cerevisiae proteome by SWATH-
1013 mass spectrometry. Mol Cell Proteomics 14**:** 739-749.
1014 Shaffer, J. P., 1995 Multiple hypothesis testing. Annual review of psychology 46**:** 561-584.
1015 Teperino, R., A. Lempradl and J. A. Pospisilik, 2013 Bridging epigenomics and complex
1016 disease: the basics. Cell Mol Life Sci 70**:** 1609-1621.
1017 Theriault, S., N. Gaudreault, M. Lamontagne, M. Rosa, M. C. Boulanger *et al.*, 2018 A
1018 transcriptome-wide association study identifies PALMD as a susceptibility gene for
1019 calcific aortic valve stenosis. Nat Commun 9**:** 988.

1020 Tian, C., L. Fei, W. Zheng, Y. Xu, W. Zuo *et al.*, 2019 Deep learning on image denoising: An
1021 overview. arXiv preprint arXiv:1912.13171.
1022 Trapnell, C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan *et al.*, 2010 Transcript assembly
1023 and quantification by RNA-Seq reveals unannotated transcripts and isoform switching
1024 during cell differentiation. Nat Biotechnol 28**:** 511-515.
1025 Veturi, Y., and M. D. Ritchie, 2018 How powerful are summary-based methods for identifying
1026 expression-trait associations under different genetic architectures? Pac Symp Biocomput
1027 23**:** 228-239.
1028 Wainberg, M., N. Sinnott-Armstrong, N. Mancuso, A. N. Barbeira, D. A. Knowles *et al.*, 2019
1029 Opportunities and challenges for transcriptome-wide association studies. Nat Genet 51**:**
1030 592-599.
1031 Wang, Z., M. Gerstein and M. Snyder, 2009 RNA-Seq: a revolutionary tool for transcriptomics.
1032 Nat Rev Genet 10**:** 57-63.
1033 Wu, M. C., P. Kraft, M. P. Epstein, D. M. Taylor, S. J. Chanock *et al.*, 2010 Powerful SNP-set
1034 analysis for case-control genome-wide association studies. Am J Hum Genet 86**:** 929-
1035 942.
1036 Wu, M. C., S. Lee, T. Cai, Y. Li, M. Boehnke *et al.*, 2011 Rare-variant association testing for
1037 sequencing data with the sequence kernel association test. Am J Hum Genet 89**:** 82-93.
1038 Xu, Z., C. Wu, W. Pan and I. Alzheimer's Disease Neuroimaging, 2017 Imaging-wide
1039 association study: Integrating imaging endophenotypes in GWAS. Neuroimage 159**:**
1040 159-169.
1041 Yan, J., S. L. Risacher, L. Shen and A. J. Saykin, 2018 Network approaches to systems biology
1042 analysis of complex disease: integrative methods for multi-omics data. Brief Bioinform
1043 19**:** 1370-1381.
1044 Zhang, A., H. Sun, P. Wang, Y. Han and X. Wang, 2012 Modern analytical techniques in
1045 metabolomics analysis. Analyst 137**:** 293-300.
1046 Zhang, W., G. Voloudakis, V. M. Rajagopal, B. Readhead, J. T. Dudley *et al.*, 2019 Integrative
1047 transcriptome imputation reveals tissue-specific and shared biological mechanisms
1048 mediating susceptibility to complex traits. Nat Commun 10**:** 3834.
1049 Zhou, X., and M. Stephens, 2012 Genome-wide efficient mixed-model analysis for association
1050 studies. Nat Genet 44**:** 821-824.
1051