IGREX for quantifying the impact of genetically regulated expression on phenotypes

Mingxuan Cai[1], Lin S. Chen[2], Jin Liu[3*]& Can Yang[1*]

[1]Department of Mathematics, The Hong Kong University of Science and Technology

[1]Department of Public Health Sciences, The University of Chicago

[3]Center for Quantitative Medicine, Duke-NUS Medical School

Abstract

By leveraging existing GWAS and eQTL resources, transcriptome-wide association studies (TWAS) have achieved many successes in identifying trait-associations of genetically-regulated expression (GREX) levels. TWAS analysis relies on the shared GREX variation across GWAS and the reference eQTL data, which depends on the cellular conditions of the eQTL data. Considering the increasing availability of eQTL data from different conditions and the often unknown trait-relevant cell/tissue-types, we propose a method and tool, IGREX, for precisely quantifying the proportion of phenotypic variation attributed to the GREX component. IGREX takes as input a reference eQTL panel and individual-level or summary-level GWAS data. Using eQTL data of 48 tissue types from the GTEx project as a reference panel, we evaluated the tissue-specific IGREX impact on a wide spectrum of phenotypes. We observed strong GREX effects on immune-related protein biomarkers. By incorporating trans-eQTLs and analyzing genetically-regulated alternative splicing events, we evaluated new potential directions for TWAS analysis.

*Correspondence should be addressed to Jin Liu (jin.liu@duke-nus.edu.sg) and Can Yang (macyang@ust.hk)

Introduction

Genome-wide association studies (GWAS) have successfully identified tens of thousands of unique associations between single-nucleotide polymorphisms (SNPs) and a wide range of complex traits/diseases (`http://www.ebi.ac.uk/gwas/`). More than 90% of identified risk variants are located in non-coding regions [1], making it challenging to understand their functional mechanisms. Increasing evidence [2, 3, 4, 5, 6, 7, 8, 9] has suggested that many of those risk variants may affect traits/diseases via the modulation of their cis gene expression levels. For example, a study of 18 complex traits revealed an enrichment for expression quantitative trait loci (eQTLs) in 11% of 729 tissue-trait pairs [10]. There is great interest in precisely characterizing the specific role of genetically regulated gene expression (GREX) in human traits and diseases.

It is well known that the effects of genetic variation on gene expressions depend on cellular contexts [11]. The rapidly increasing availability of eQTL data from different tissue types, cell types, populations and other conditions provides an unprecedented opportunity to study and evaluate GREX effects in a variety of conditions. For example, the V7 release of the Genotype-Tissue Expression (GTEx) project (`https://gtexportal.org/home/`) has collected gene expression samples from 53 non-diseased tissues across 714 individuals [11]. Multiple blood eQTL resources comprising thousands of individuals are made publicly available [12, 13]; and other ongoing projects such as Genetics of DNA Methylation Consortium (GoDMC) and eQTLGen consortium are collecting expression data with sample sizes larger than $10,000$ [14, 15]. Those data serve as rich eQTL resources for a comprehensive evaluation of GREX effects.

The vast amount of publicly available eQTL and GWAS data resources enables an integrative framework, transcriptome-wide association studies (TWAS), for mapping gene-level trait associations and evaluating GREX effects on human traits and diseases. Using a reference eQTL panel (e.g., GTEx), gene-specific expression prediction models can be built based on cis-acting genetic factors. Then the gene expression levels of a GWAS cohort can be predicted based on individual genetic profiles, and the genetically-regulated and predicted expression levels are further associated with the phenotype of interest in the GWAS study to map gene-level trait-associations [16, 17, 18, 19, 20, 21, 22, 23, 24]. Existing methods have been proposed [8, 25], including PrediXcan [16], TWAS [17], FOCUS [19], S-PrediXcan [21], UTMOST [26]

and CoMM [22]. Through applications to a wide variety of phenotypes, these methods have successfully identified specific gene-trait associations, whereas a comprehensive and precise evaluation of the impact of GREX variation on various traits and the trait-relevant cellular context is still needed [27].

TWAS-types of integrative analysis rely on a key assumption: there exists a steady-state GREX variation shared across reference eQTL data and GWAS data, and the steady-state GREX variation can further induce phenotypic variation. The multi-tissue eQTL data from the GTEx project is commonly used as the reference eQTL panel [16, 21, 26]. The GTEx project has collected data from post-mortem donors and has provided a source of largely non-diseased tissues for general purposes. The GTEx reference may or may not have considerable shared GREX variation with GWAS data of specific phenotypes in specific populations. Given the often unknown disease/trait-relevant tissue types and the increasing availability of eQTL data resources from different conditions, there is a need for new methods and tools that can be used to assess the proportion of the shared GREX variation in the phenotypic variation from a global perspective, and guide the selection of eQTL reference data and tissue-types for specific phenotypes and populations.

The heritability measure has been widely used to quantify the impact of genetic variation on phenotypic variation, and has served as a preliminary yet insightful assessment of the potential of genetic studies on various phenotypes [28, 29]. Analogous to the heritability measure, the estimation of proportion of GREX on phenotypic variation can also be used to evaluate the impact of the genetic regulatory effects on phenotypes mediated by expression levels, and inform trait-relevant tissue types or conditions in specific populations. To the best of our knowledge, there are two methods that have been proposed for this purpose [23, 24]. The RhoGE method [23] estimates the proportion of phenotypic variation explained by GREX based on linkage-disequilibrium (LD) score regression (LDSC) [30]. Since it ignores the uncertainty in predicting gene expression levels, the proportion of variance explained by GREX could be substantially under-estimated by RhoGE. The other method, known as the gene expression co-score regression (GECS) [24], requires the analyzed SNPs not being in LD to ensure unbiasedness, which greatly limits its applicability in real data analysis.

In this work, we propose a unified framework, named IGREX, for quantifying the impact of genetically regulated expression, while accounting for uncertainty in predicted gene expression

levels in the presence of moderate to weak eQTL effects. IGREX requires only summary-level GWAS data as input, greatly enhancing the applicability of the method. We evaluated the performance of IGREX with comprehensive simulation studies, highlighting the importance of accounting for expression estimation uncertainty. Using 48 tissue types from the GTEx project as the reference panel, we applied IGREX to both individual-level and summary-level GWAS data sets, and evaluated the tissue-specific IGREX impact on a wide spectrum of cellular and organismal phenotypes. Our results provide new biological insights into the role of gene expression in the genetic architecture of complex traits. We also demonstrate the reproducibility of results. By incorporating trans-eQTLs and analyzing genetically-regulated alternative splicing events, we evaluated new potential directions for TWAS analysis.

Results

Method overview. IGREX is a two-stage method for quantifying the proportion of phenotypic variation that can be attributed to GREX variation. The method can be applied to both individual-level (IGREX-i) and summary-level (IGREX-s) GWAS data. It first evaluates the posterior distribution of GREX effects based on an eQTL reference panel and then estimates the proportion of variance explained by GREX using the 'predicted' gene expression in the GWAS data. Here, we briefly introduce the statistical model of IGREX-i and present additional technical details in the Methods Section.

Consider a reference eQTL data set $\mathcal{D}_r$ and an individual-level GWAS data set $\mathcal{D}_i$. The eQTL data $\mathcal{D}_r = \{\mathbf{Y}, \mathbf{X}_r\}$ is comprised of an $n_r \times G$ gene expression matrix, $\mathbf{Y}$, and an $n_r \times M$ genotype matrix, $\mathbf{X}_r$, where G is the number of genes, M is the number of SNPs and n_r is the sample size. The GWAS data $\mathcal{D}_i = \{\mathbf{t}, \mathbf{X}\}$ contains a phenotype vector $\mathbf{t} \in \mathbb{R}^n$ and a genotype matrix $\mathbf{X} \in \mathbb{R}^{n \times M}$, where n is the sample size of the GWAS data. Let $\mathbf{y}_g$ and $\mathbf{X}_{r,g}$ be the vector of expression levels of the g-th gene and the genotype matrix corresponding to its local (cis) SNPs from the reference panel, respectively. We first relate $\mathbf{y}_g$ to $\mathbf{X}_{r,g}$ with a linear model:

$$\mathbf{y}_g = \mathbf{X}_{r,g}\boldsymbol{\beta}_g + \mathbf{e}_{r,g}, \quad g = 1, \ldots, G, \tag{1}$$

where $\boldsymbol{\beta}_g \sim \mathcal{N}(\mathbf{0}, \sigma^2_{\beta_g}\mathbf{I}_{M_g})$ is the vector of genetic effects of M_g cis SNPs on the expression levels of the g-th gene, and $\mathbf{e}_{r,g} \sim \mathcal{N}(0, \sigma^2_{r,g}\mathbf{I}_{n_r})$ is the error term. Since we are interested in the steady-state component of gene expression levels regulated by genetic variants, $\boldsymbol{\beta}_g$ is

assumed to be the same for individuals in both datasets, $\mathcal{D}_r$ and $\mathcal{D}_i$. Consequently, the GREX component of individuals in the GWAS data can be evaluated by $\mathbf{X}_g\boldsymbol{\beta}_g$. Meanwhile, we assume that the genetic effects on the phenotype of interest $\mathbf{t}$ can be decomposed into two parts, i.e. the effects mediated via GREX and the effects through alternative pathways not mediated by gene expression levels:

$$\mathbf{t} = \sum_{g=1}^{G} \alpha_g \mathbf{X}_g \boldsymbol{\beta}_g + \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{2}$$

where $\alpha_g \sim \mathcal{N}(0, \sigma_\alpha^2)$ is the effect size of $\mathbf{X}_g\boldsymbol{\beta}_g$ on $\mathbf{t}$, $\boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{0}, \sigma_\gamma^2\mathbf{I}_M)$ is the vector of alternative genetic effects, and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma_\epsilon^2\mathbf{I}_n)$ is the error term. In this model, $\sum_{g=1}^{G} \alpha_g \mathbf{X}_g\boldsymbol{\beta}_g$ and $\mathbf{X}\boldsymbol{\gamma}$ correspond to the overall impact of the GREX component and the alternative genetic effects on $\mathbf{t}$, respectively. Thus, the impact of GREX on the phenotype can be quantified by the proportion of phenotypic variance explained by the GREX component: $\text{PVE}_{\text{GREX}} = \frac{\text{Var}(\sum_{g=1}^{G} \alpha_g \mathbf{x}_g^T \boldsymbol{\beta}_g)}{\text{Var}(t)}$.

To estimate this quantity, we propose a two-stage procedure: In the first stage, we estimate $\sigma_{\beta_g}^2$ and $\sigma_{r,g}^2$ using an efficient algorithm and evaluate the posterior distribution $\boldsymbol{\beta}_g|\mathbf{y}_g, \mathbf{X}_{r,g} \sim \mathcal{N}(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ for all genes. In the second stage, by treating the posterior obtained in the first stage as the prior distribution of $\boldsymbol{\beta}_g$ in model (2), we can obtain estimated values of σ_α^2, σ_γ^2 and σ_ϵ^2 using either method of moments (MoM) or restricted maximum likelihood (REML). Following this procedure, the resulting estimate of PVE_{GREX} is obtained (with details in the Methods Section) by

$$\widehat{\text{PVE}}_{\text{GREX}} = \frac{\text{tr}(\sum_{g=1}^{G} \hat{\sigma}_\alpha^2 \mathbf{X}_g(\boldsymbol{\mu}_g\boldsymbol{\mu}_g^T + \boldsymbol{\Sigma}_g)\mathbf{X}_g^T)}{\text{tr}(\sum_{g=1}^{G} \hat{\sigma}_\alpha^2 \mathbf{X}_g(\boldsymbol{\mu}_g\boldsymbol{\mu}_g^T + \boldsymbol{\Sigma}_g)\mathbf{X}_g^T + \hat{\sigma}_\gamma^2\mathbf{X}\mathbf{X}^T + \hat{\sigma}_\epsilon^2\mathbf{I}_n)}.$$

In the above estimation, the substitution of posterior $\boldsymbol{\beta}_g|\mathbf{y}_g, \mathbf{X}_{r,g}$ accounts for the posterior variance $\boldsymbol{\Sigma}_g$ and naturally results in the adjustment of estimation uncertainty associated with $\boldsymbol{\beta}_g$. This is important because in the GWAS data, the gene expression levels are not directly measured, but rather are predicted or imputed based on genetic variants. It is known that the prediction accuracy and uncertainty vary substantially among genes. For most of the genes in the genome, the genetically regulated expression variation accounts for only a small to moderate proportion of total expression variation. Thus, the prediction may not be accurate and could be subject to high uncertainty. In contrast, our model accounts for the estimation uncertainty by $\boldsymbol{\Sigma}_g$ and can yield unbiased estimation for $\widehat{\text{PVE}}_{\text{GREX}}$. In addition, the standard error of $\widehat{\text{PVE}}_{\text{GREX}}$ can be obtained based on the delta method (see Supplementary Note). The IGREX framework can also be used to test $H_0 : \text{PVE}_{\text{GREX}} = 0$ for the phenotype of interest in

specific populations given an eQTL reference with a specific tissue type or cellular context.

In real applications, individual-level GWAS data may not be accessible. We have further developed IGREX-s which requires only summary-level GWAS data as input (See Methods). Based on MoM, IGREX-s can approximate IGREX-i while requiring only SNP-level z-scores from GWAS and a reference genotype matrix $\tilde{\mathbf{X}} \in \mathbb{R}^{m\times M}$ of a similar LD pattern to $\mathbf{X}$, where m is the number of samples in the reference panel. Using simulations, we showed that with a few hundreds of samples in the eQTL reference data, the estimation of IGREX-s with summary statistics well approximates IGREX-i using individual level data. In practice, $\tilde{\mathbf{X}}$ can be $\mathbf{X}_r$ or a subset of $\mathbf{X}$. The estimate of PVE_{GREX} given by IGREX-s is

$$\widehat{\text{PVE}}_{\text{GREX}} = \frac{\hat{\sigma}_\alpha^2}{\hat{\sigma}_t^2}\text{tr}(\sum_{g=1}^{G}(\boldsymbol{\mu}_g\boldsymbol{\mu}_g^T + \boldsymbol{\Sigma}_g)\hat{\mathbf{R}}_g),$$

where $\hat{\mathbf{R}}_g = \tilde{\mathbf{X}}_g^T\tilde{\mathbf{X}}_g/(m-1)$ is the estimated LD matrix associated with the g-th gene and $\tilde{\mathbf{X}}_g$ is the corresponding columns of $\tilde{\mathbf{X}}$. IGREX also allows for the adjustment of covariates including sex, age and genotype principal components (See details in Supplementary Note).

Simulation studies. We conducted extensive simulation studies to evaluate the performance of IGREX. For all the simulated data, we fixed $n = 4,000$, $G = 200$, $M = 20,000$ (i.e., 100 cis SNPs for each gene). The total phenotypic heritability was set as $h_t^2 = \frac{\text{Var}(\sum_{g=1}^{G}\alpha_g\mathbf{x}_g^T\boldsymbol{\beta}_g + \mathbf{x}^T\boldsymbol{\beta}_g)}{\text{Var}(t)} = 0.5$, where $\text{PVE}_{\text{GREX}} = 0.2$ and the proportion explained by the alternative genetic effects, $\text{PVE}_{\text{Alternative}} = \frac{\text{Var}(\mathbf{x}_g^T\boldsymbol{\gamma})}{\text{Var}(t)} = 0.3$ (results for other scenarios are shown in Supplementary Fig. 1-3). To simulate the genotype data, we first sampled the minor allele frequencies (MAF) from uniform distribution $\mathcal{U}(0.05, 0.5)$ and data matrices from normal distribution $\mathcal{N}(\mathbf{0}, \Sigma(\rho))$, where $\Sigma_{jj'} = \rho^{|j-j'|}$ characterizes the LD patterns between SNPs. Then, the genotype matrices $\mathbf{X}_r$ and $\mathbf{X}$ were obtained by categorizing the entries of generated data matrices into $0, 1, 2$ according to MAF. Given the genotype matrices, $\boldsymbol{\beta}_g$ and α_g, the gene expression $\mathbf{y}_g$ and phenotype $\mathbf{t}$ were simulated following models (1) and (2). We will discuss the details for generating $\boldsymbol{\beta}_g$ and α_g later. To assess IGREX-s, we calculated the z-score of each SNP and randomly subsetted $m = 500$ samples from $\mathbf{X}$ for estimating LD matrix $\hat{\mathbf{R}}_g$ (results for other settings of m are shown in Supplementary Fig. 4).

We first evaluated the estimation performance of IGREX for different settings of eQTL reference data. Specifically, we varied n_r at $\{800, 1000, 2000\}$, $\text{PVE}_y = \frac{\text{Var}(\mathbf{x}^T\boldsymbol{\beta}_g)}{\text{Var}(\mathbf{y}_g)}$ at $\{0.1, 0.2, 0.3\}$, where PVE_y quantifies the gene expression heritability explained by its local SNPs. To mimic the

scenario in which the expression estimation uncertainty was incorrectly ignored, we obtained the posterior mean of $\boldsymbol{\beta}_g$ in the first stage, and replaced the true effect size $\boldsymbol{\beta}_g$ by its posterior mean $\boldsymbol{\mu}_g$ while specifying the posterior variance to be $\boldsymbol{\Sigma}_g = \mathbf{0}$ in the second stage, and then conducted REML and MoM as before. We denoted these methods as REML_0 and MoM_0. The simulation results summarized in Fig.1a showed that both PVE_{GREX} and $\text{PVE}_{\text{Alternative}}$ were accurately estimated using REML-based IGREX-i in all settings. The MoM-based IGREX-i slightly underestimated PVE_{GREX} when both sample size n_r and PVE_y were very small, but steadily achieved similar performance as REML-based estimation when either n_r or PVE_y increased. In all settings, IGREX-s well approximated MoM, producing nearly identical estimation results. In contrast, both REML_0 and MoM_0 did not account for estimation uncertainty in the expression prediction, and they showed poor estimation performance even when sample size was large and PVE_y value was high.

Next we conducted simulations to evaluate the situation that the IGREX model was misspecified. Here we considered the situation where genetic effects $\boldsymbol{\beta}_g$ and $\boldsymbol{\alpha}$ were sparse while we assumed dense effect sizes in the IGREX model. This was designed to mimic the real data situation that the architecture of eQTL signals is often sparse [31]. Let π_α and π_β be the sparsity of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_g$, i.e., $\pi_\alpha = (\#\text{ Nonzero entries in } \boldsymbol{\alpha})/G$ and $\pi_\beta = (\#\text{ Nonzero entries in } \boldsymbol{\beta}_g)/M_g$, respectively. To evaluate the influence of different sparsity patterns on our method, we varied π_α and π_β at $\{0.2, 0.5, 0.8\}$. The nonzero entries in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_g$ were simulated form a normal distribution. As shown in Fig. 1b-c, all three methods of IGREX produced accurate estimates in the presence of sparse genetic effects, implying the robustness of IGREX to model misspecification. Moreover, the estimation performance was not influenced by the degree of sparsity. Next, we investigated the influence of LD patterns by letting ρ vary at $\{0.1, 0.3, 0.5, 0.8\}$. From Fig.1d, we observed that IGREX produced accurate estimation in the presence of LD. In contrast, REML_0 and MoM_0 consistently underestimated PVE_{GREX} as a result of ignoring estimation uncertainty.

We also compared IGREX with an existing method in the literature, RhoGE [23]. RhoGE is an LDSC-based approach for estimating PVE_{GREX}. However, this method does not adjust for estimation uncertainty. The results are shown in Fig. 1e. As expected, IGREX yielded unbiased estimation while RhoGE substantially underestimated PVE_{GREX} in most settings. It achieved similar accuracy as IGREX only when the genetically regulated expression accounted

for most of the expression variation, $\text{PVE}_y \geq 0.9$. In other words, RhoGE only works well when the genetically-predicted expression levels are very close to the true underlying expression levels for most of the genes, which may not be realistic for real data analysis.

Real data applications with individual-level GWAS data. With eQTL data of 48 human tissues from the GTEx project as reference, we applied IGREX to two individual-level GWAS datasets, the Northern Finland Birth Cohorts program 1966 (NFBC) [32] and the Wellcome Trust Case Control Consortium (WTCCC) [33]. The details of the datasets and the data pre-processing procedures are described in the Methods Section.

In analyzing the NFBC data, we focused on six quantitative traits with statistically significant heritability, based on $5,123$ individuals and $309,245$ genotyped SNPs. Those six traits are Glucose ($h_t^2 = 14.2\% \pm 5.3\%$), high-density lipoprotein cholesterol (HDL, $h_t^2 = 32.9\% \pm 5.6\%$), low-density lipoprotein cholesterol (LDL, $h_t^2 = 29.0\% \pm 5.5\%$), triglycerides (TG, $h_t^2 = 13.6\% \pm 5.3\%$), total cholesterol (TC, $h_t^2 = 20.1\% \pm 5.4\%$) and systolic blood pressure (SysBP, $h_t^2 = 17.1\% \pm 5.4\%$). Fig. 2a-b shows the tissue-specific $\widehat{\text{PVE}}_{\text{GREX}}$ estimates of the six traits. The REML and MoM methods yielded similar estimates in most of the tissues.

IGREX can also be used to inform trait-relevant tissue types. By testing $H_0 : \text{PVE}_{\text{GREX}} = 0$ in each tissue type, we observed significant GREX components in liver for both LDL and TC. As shown in Fig. 2a, $\widehat{\text{PVE}}_{\text{GREX}}$ for LDL in liver is as high as 14.3% (with standard error 2.6%), capturing 52.6% of total heritability defined as $\text{PVE}_{\text{GREX}}/h_t^2$; and TC also has $\widehat{\text{PVE}}_{\text{GREX}} = 13.7\%$ (with standard error 2.5%) in liver, which captures 79.4% of total heritability (see Supplementary Fig. 6). It is known that LDL synthesized in liver is an important lipoprotein particle for transporting cholesterol in the blood [34, 35]. Our findings suggest that genetic variants affect LDL through regulating their corresponding gene targets and liver is the most relevant tissue involved in gene regulation. Next, we analyzed the impact of ignoring the estimation uncertainty (with the complete results given in the Supplementary Fig. 5). As shown in Fig. 2c-d, the $\widehat{\text{PVE}}_{\text{GREX}}$ declined substantially as a result of ignoring expression estimation uncertainty. In Fig. 2e, we compared the estimates based on individual level data using IGREX-i versus those based on IGREX-s with summary statistics. For all six of the traits, the IGREX-s estimates well approximated the estimates using the individual level data, which is consistent with our simulation results.

Next we investigated the role of GREX in complex human traits and diseases, using the

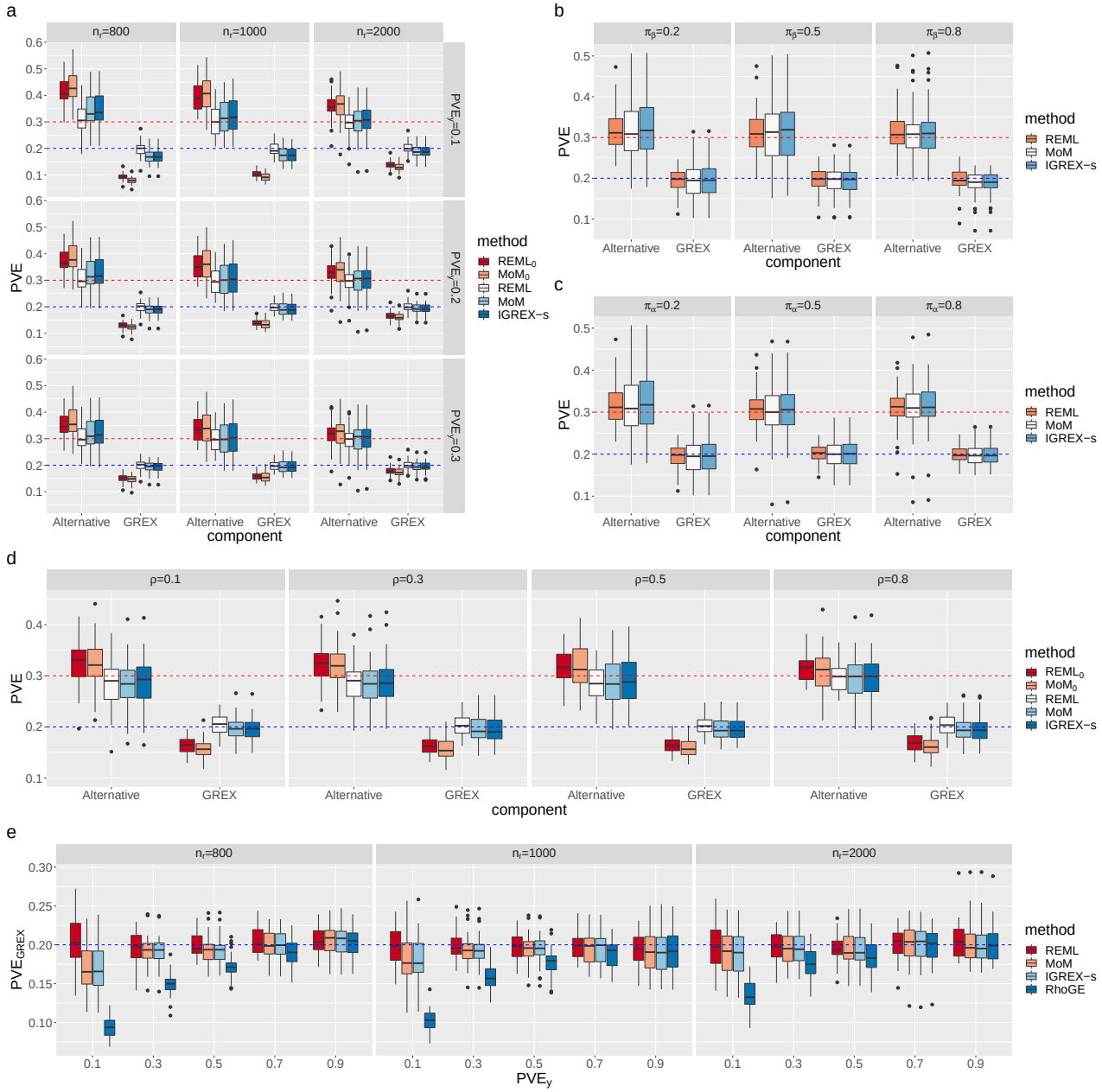

Figure 1: Simulation studies to compare estimation accuracies of IGREX with other methods. REML and MoM in the legend are abbreviations of the IGREX-i estimation methods. The blue and red dashed lines represent the true values of PVE_{GREX} and $\text{PVE}_{\text{Alternative}}$, respectively. We averaged the results over 30 replications and generated box plots for evaluating the estimation performance of: **a** the three models of IGREX ,REML_0 and MoM_0 when n_r was varied at $\{800, 1000, 2000\}$ and PVE_y was varied at$\{0.1, 0.2, 0.3\}$; (**b**) the three models of IGREX when $\pi_\alpha = 0.2$ and π_β was varied at $\{0.2, 0.5, 0.8\}$; (**c**) the three models of IGREX when $\pi_\beta = 0.2$ and π_α were varied at $\{0.2, 0.5, 0.8\}$; (**d**) the three models of IGREX, REML_0 and MoM_0 when ρ was varied at $\{0.1, 0.3, 0.5, 0.8\}$; (**e**) the three models of IGREX and RhoGE when n_r was varied at $\{800, 1000, 2000\}$.

WTCCC dataset [33]. We applied IGREX to estimate the tissue-specific PVE_{GREX} of seven diseases including bipolar disorder (BD), coronary artery disease (CAD), Crohn's disease (CD),

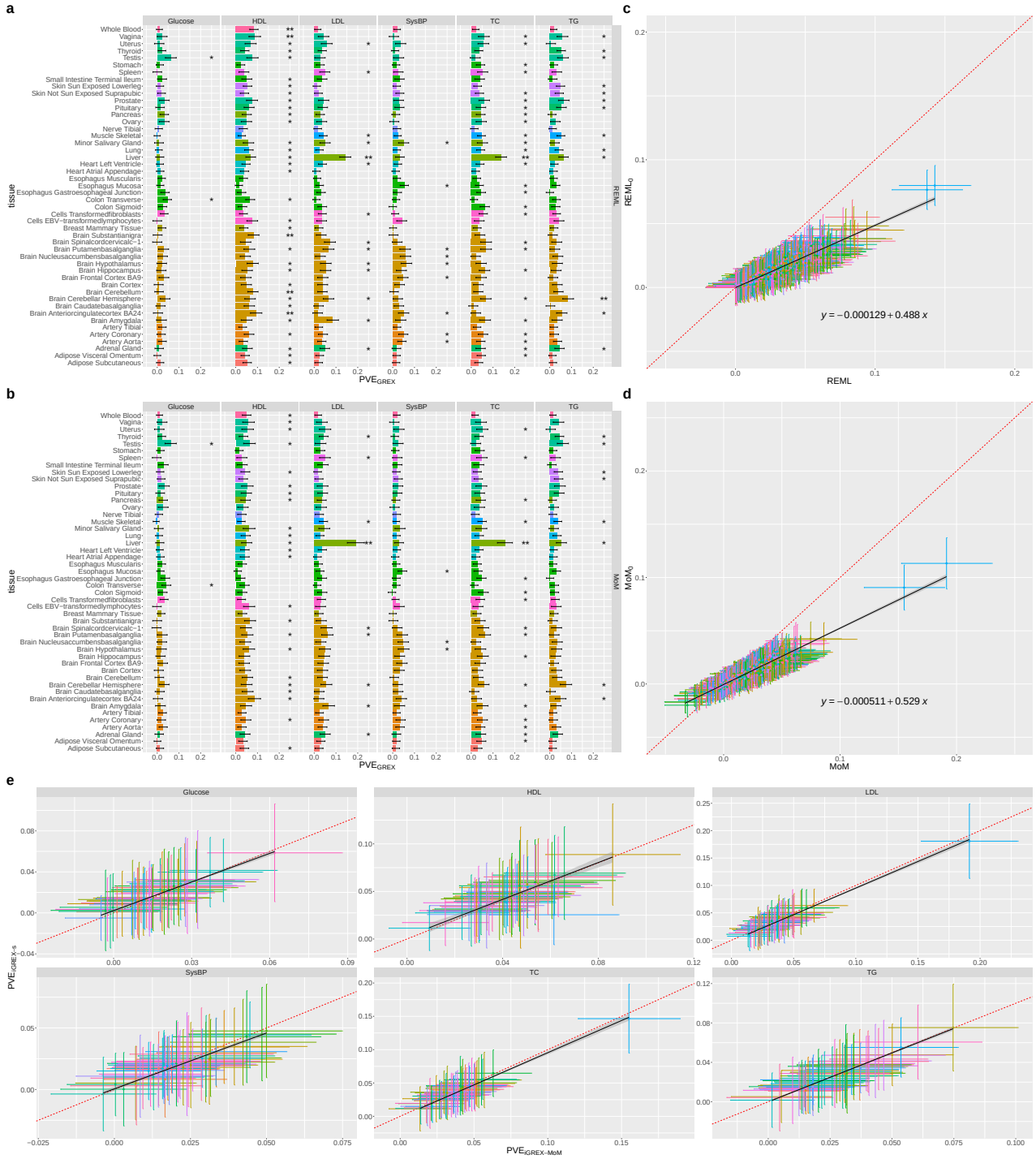

Figure 2: Tissue-specific $\widehat{\text{PVE}}_{\text{GREX}}$ of the six traits from NFBC data set. (**a**-**b**) $\widehat{\text{PVE}}_{\text{GREX}}$ obtained by REML and MoM. Tissues are colored according to their categories. The number of asterisks represents the significance level: p-value< 0.05 is annotated by $*$; p-value$< 0.05/48$ is annotated by $**$. (**c**-**d**) All pairs of estimates generated by REML and MoM against their counterparts without accounting for uncertainty. A regression line is fitted and the estimated coefficients are given in the plot. (**e**) Each panel is a plot of $\widehat{\text{PVE}}_{\text{GREX}}$ generated by IGREX-s against those generated by MoM for all 48 tissues in one of the six traits.

hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D) and type 2 diabetes (T2D). The estimates of $\text{PVE}_{\text{GREX}}/h_t^2$ obtained by REML are shown in Supplementary Fig. 8. The top GREX components measured by $\text{PVE}_{\text{GREX}}/h_t^2$ are 12.8% for BD in amygdala, 21.2% for CAD in spinal cord, 18.4% for CD in amygdala, 16.7% for HT in spleen and 17.9% for T2D in anterior cingulate cortex. The average estimates of $\text{PVE}_{\text{GREX}}/h_t^2$ across 48 tissues for RA and T1D are as high as 34.1% and 71.2%, respectively. Both RA and T1D are autoimmune diseases, with well-established strong associations in the major histocompatibility complex (MHC) region [33, 36]. After removing the MHC region, we observed a substantial reduction in the $\text{PVE}_{\text{GREX}}/h_t^2$ estimates: the mean $\widehat{\text{PVE}}_{\text{GREX}}$ dropped from 34.1% to 7.6% for RA and from 71.2% to 11.7% for T1D, as shown in Fig. 3a. Additionally, the tissue-specific comparisons presented in Fig. 3b showed an extensive reduction of PVE_{GREX} in all tissue types for T1D and RA, while such changes were not observed for other traits. This finding suggests the heavy involvement of GREX variation in the immune functions related to the MHC region for both RA and T1D. Here we illustrate that IGREX can be used to inform disease/trait-relevant tissue types or cellular contexts.

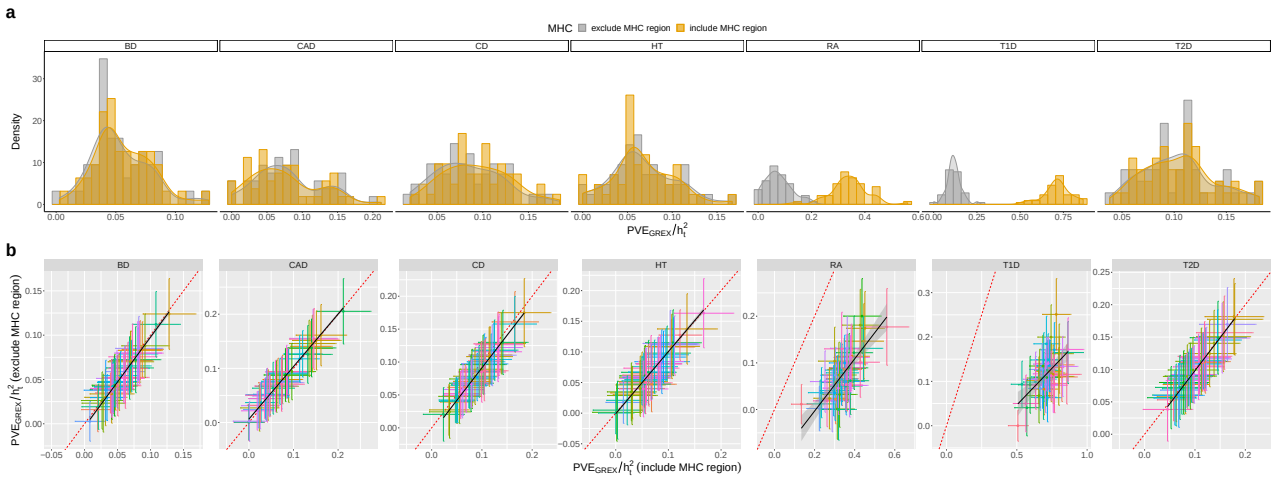

Figure 3: Percentage of heritability explained by GREX ($\text{PVE}_{\text{GREX}}/h_t^2$) of the seven traits from WTCCC data. (**a**) The distributions of estimated $\text{PVE}_{\text{GREX}}/h_t^2$ across 48 GTEx tissues. (**b**) Tissue-specific comparisons of $\text{PVE}_{\text{GREX}}/h_t^2$ estimated by whole genome with those estimated by excluding the MHC region.

Analysis of a wide spectrum of phenotypes using IGREX-s with summary-level GWAS data The vast amount of publicly available summary-level GWAS data and their easy accessibility allow us to conduct a comprehensive evaluation of the impact of GREX on a wide spectrum of phenotypes using IGREX-s, from molecular traits such as proteins and metabolites

to various complex phenotypes including schizophrenia, height, and body mass index (BMI). In the following analysis, we used the genotypes of the 635 GTEx samples as the LD reference $\tilde{\mathbf{X}}$ in the IGREX-s estimation.

First, we estimated PVE_{GREX} in 249 proteins with significantly nonzero heritabilities using summary statistics from a plasma protein quantitative trait loci (pQTL) study [38], as summarized in Fig. 4a. In Supplementary Fig. 10, the heritabilities estimated by IGREX ($\hat{h}_t^2 = \widehat{\text{PVE}}_{\text{GREX}} + \widehat{\text{PVE}}_{\text{Alternative}}$) are shown to be highly consistent with those estimates obtained using MoM [37]. From this perspective, heritability can be attributed to two components: the GREX component and its alternative effects. Then, we grouped 48 tissue types into 16 groups by their functions and tested the significance of tissue-specific GREX effects on the 249 proteins. We observed a significant GREX contribution in many tissue-protein pairs (Fig. 4b and Supplementary Fig. 11-13). In particular, 9 out of the 249 proteins had significant GREX components in at least one tissue type at 0.05 level after Bonferroni correction. As shown in Fig. 4d-e, some proteins, including CD96, DEFB119, MICB and PDE4D, exhibit cross-tissue GREX impacts; meanwhile other proteins, namely CFB, CXCL11, EVI2B, IDUA and LRPAP1, have tissue-specific GREX effect patterns. We found these tissue-specific patterns to be consistent with protein functions. For example, the CFB protein, which is implicated in the growth of preactivated B-lymphocytes, is found to be most associated with GREX in EBV-transformed lymphocytes ($\widehat{\text{PVE}}_{\text{GREX}} = 22.7\%$). As another example, the CXCL11 protein has the highest $\widehat{\text{PVE}}_{\text{GREX}} = 20.0\%$ in pancreas, and the *CXCL11* gene is often over-expressed in pancreas tissue [39]. We also noted that 6 out of the 9 proteins were immune-related, echoing our previous implications of the important role of GREX in immune processes. In addition to the proteins, metabolic traits are also important intermediate traits for complex biological processes. We applied IGREX-s to a summary level data set of circulating metabolites [40], and studied the impact of GREX on metabolic traits. The results are discussed in the Supplementary Materials.

Then we applied IGREX-s to the summary data of complex human traits. Here we analyzed three traits: schizophrenia (SCZ), height, and BMI. We considered four datasets of schizophrenia with increasing and overlapping samples: SCZ subset [41], SCZ1 [42], SCZ1+Sweden (SCZ1Swe)[43] and SCZ2 [44]. We found that the estimated $\text{PVE}_{\text{GREX}}/h_t^2$ in all four SCZ datasets have higher values in the brain tissues than in other tissue types (Fig. 5b). As expected, the statistical power increases with sample size of GWAS (Fig. 5a). Additionally, we

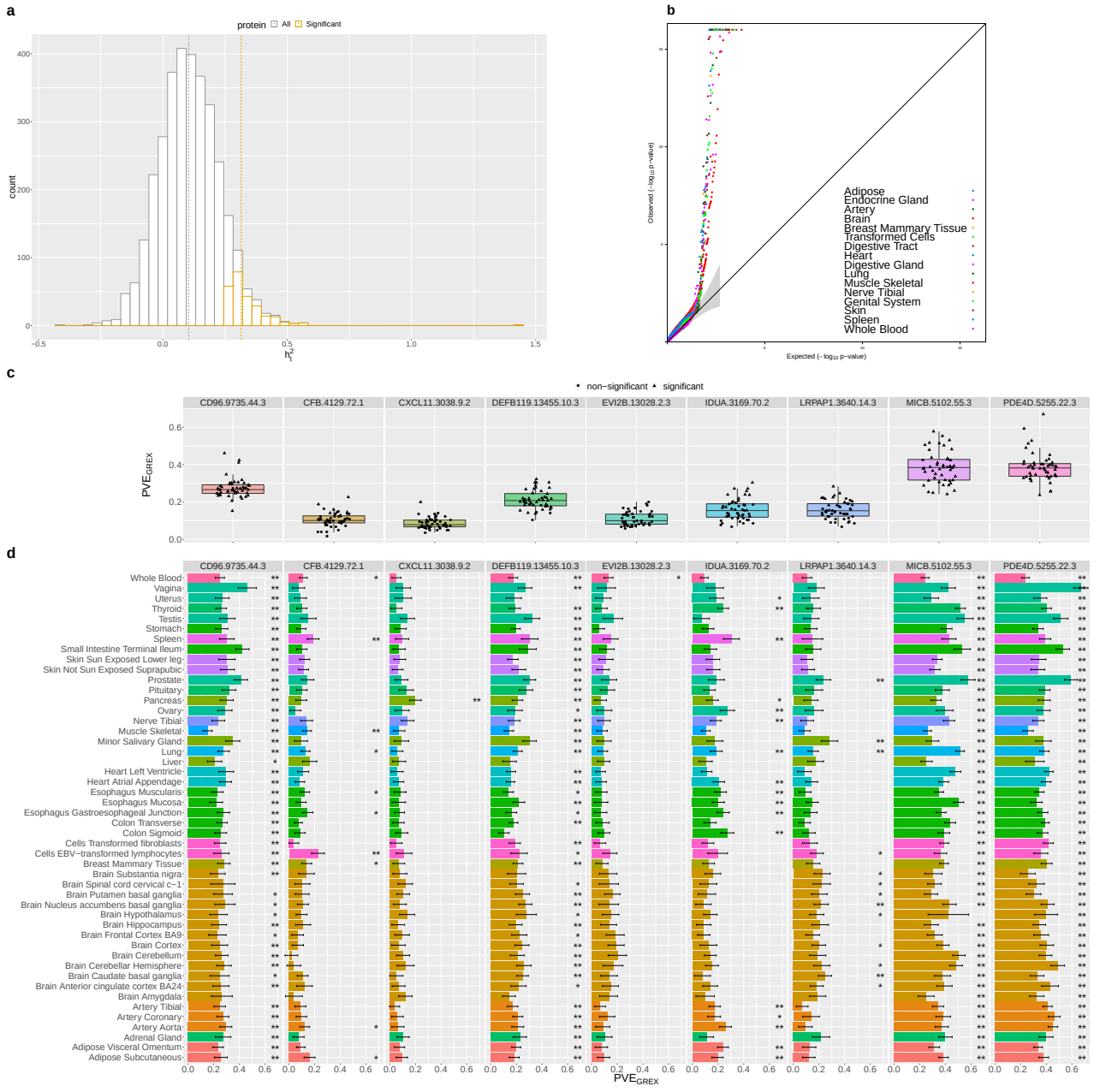

Figure 4: Analysis of plasma pQTL summary statistics. (**a**) The distribution of estimated heritabilities of $3,283$ proteins estimated using [37]. The whole study is colored in grey, while the 249 proteins with significant heritabilities are colored in yellow. Dashed lines represent the means of corresponding distributions. (**b**) QQ-plot of PVE_{GREX} p-values of tissue-protein pairs. GTEx tissues are categorized into 16 types and colored accordingly. (**c**) The Manhattan plot of the protein encoding genes in aorta, cerebellum, liver and whole blood. Each point represents a tissue-protein pair. (**d**) $\widehat{\text{PVE}}_{\text{GREX}}$ in the 9 proteins whose $\widehat{\text{PVE}}_{\text{GREX}}$ are significant in at leat one tissue at 0.05 level using Bonferrni correction. (**e**) $\widehat{\text{PVE}}_{\text{GREX}}$ obtained by IGREX-s. Tissues are colored according to their categories. The number of asterisks represents the significance level: p-value$< 0.05/48$ is annotated by $*$; p-value$< 0.05/(48 * 9)$ is annotated by $**$.

also analyzed the human height and BMI phenotypes using pairs of independent GWAS data for replication purposes. The obtained estimates, $\widehat{\text{PVE}}_{\text{GREX}}$, from pairs of independent GWAS

data are highly consistent. Although the analysis results are reproducible in several different data sets, we noted the estimated percentages of heritability explained by GREX for all three complex traits are less than 10% (8.7% for schizophrenia, 8.7% for height and 3.7% for BMI in the most expressed tissue types. See Fig. 5c and Supplementary Fig. 15).

The relatively low GREX contribution to complex traits other than lipid or molecular traits can be attributed to multiple reasons. First, it is known that trans-acting genetic effects can explain a substantial proportion of expression variation [8, 12]. However, trans-eQTL effects are often tissue-specific and can be harder to detect and replicate across studies [45]. In TWAS-types of analysis, generally the prediction of gene expression is based on only cis genetic variants of each gene. As such, the PVE_{GREX} values reported here, also based on only cis genetic variants, may be underestimated. In the next section, we will further explore the contribution of trans-eQTLs. Second, the genetic effects on gene expression may not be steady across the reference GTEx data with largely non-diseased tissues for general purposes and the GWAS data with diseased individuals from specific populations [46]. From this perspective, before analyzing specific complex traits and diseases via TWAS, it would be helpful to first estimate the impact of GREX and select the most informative available eQTL reference data.

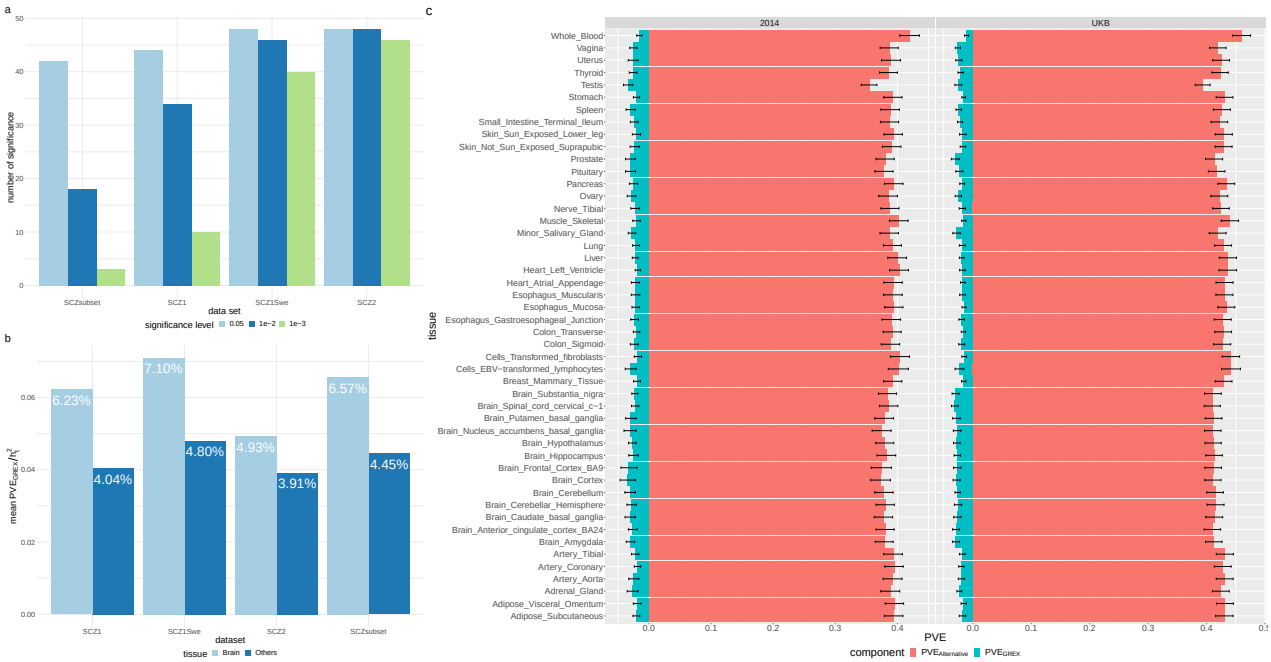

Figure 5: Analyses of complex traits: schizophrenia and height. (**a**) Number of significant GREX components revealed under different significance levels for the four schizophrenia datasets. (**b**) Mean estimated percentages of heritability for schizophrenia explained by GREX in brain tissues and in other tissues. (**c**) $\widehat{\text{PVE}}_{\text{GREX}}$ and $\widehat{\text{PVE}}_{\text{Alternative}}$ of height estimated using height2014 and UKB datasets, respectively.

Additional insights on GREX considering trans-eQTLs and genetically-regulated alternative splicing events. The cis-acting genetic effects on local gene expression levels are often shared across tissue types and are often replicable across studies [47]. It is also reported that a substantial proportion (up to 70%) of gene expression heritability can be attributed to trans-acting genetic effects which act predominantly in a tissue-specific manner and have a lower rate of replication across studies [48, 49]. More recently, the eQTLGen consortium [15] has conducted a blood-eQTL meta-analysis and has reported 6,298 (31%) trans-eQTL genes for 10,317 trait-associated SNPs using 31,684 blood samples from 37 datasets. The results suggest that trans-eQTLs are prevalent in the genome, while it is still underpowered to detect them for tissues other than whole blood given the often tissue-specific nature of trans-genetic effects and the limited sample sizes for most tissue types.

Although it is still unrealistic to account for all trans-eQTLs in the estimation of PVE_{GREX} due to the limitation of sample sizes, it is possible to explore the potential by incorporating the blood-based trans-eQTLs reported by the eQTLGen consortium and re-estimating $\text{PVE}_{\text{GREX}}/h_t^2$. We first analyzed 13 datasets comprised of 12 phenotypes that have significant $\text{PVE}_{\text{GREX}}/h_t^2$ estimates in the whole blood, including 7 proteins, 1 lipid trait and 4 complex diseases (with 2 SCZ datasets). We observed an increasing trend of $\text{PVE}_{\text{GREX}}/h_t^2$ in the blood for all 13 datasets (Fig. 6a), by accounting for only $\sim 1,700$ unique trans-eQTLs that are not cis-eQTLs. As a comparison, we applied the same procedure to 13 GTEx brain tissues of the two largest SCZ datasets, and did not observe an increase in $\text{PVE}_{\text{GREX}}/h_t^2$ (Fig. 6b). This is not surprising because the trans-eQTLs incorporated above were detected and reported based on whole blood samples and may not be trans-eQTLs in the brain tissues. Our results suggest that the estimation of GREX impacts on traits can be further boosted by incorporating robust trans-eQTLs from the same tissue types.

In addition to the gene expression level, we also evaluated the effects of alternative splicing on complex trait heritability. We applied IGREX to quantify the impact of genetically regulated alternative splicing on multiple phenotypes. Alternative splicing is an important gene regulatory process that results in multiple transcripts from a single multi-exon gene. It is commonly observed in humans and plays an essential role in cellular differentiation [50, 51]. Differential variations in splicing may also result in phenotypic variation and contribute to the development of complex diseases including cancer [52, 53, 54]. In a recent work, by extending the TWAS

framework to analyze splicing events and associating 40 complex traits with genetically-predicted splicing quantification, novel putative disease-associated genes were detected [55]. Here, using multi-tissue splicing quantification data from GTEX as reference, we applied IGREX to study the impact of genetically-regulated splicing events on four trait-tissue pairs that were found to have a high $\text{PVE}_{\text{GREX}}/h_t^2$. We estimated the proportion of phenotypic variation explained by genetically-regulated splicing to be 12.5%, 13.5%, 1.0% and 1.1% for LDL in liver, TC in liver, SCZ in amygdala and SCZ in cerebellar hemisphere, respectively. Unlike eQTLs that are often found to be near transcription starting sites, most of the sQTLs were found to be enriched within gene bodies, in particular within the introns they regulate, and have little to no effects on cis gene expression levels [55, 56]. In other words, sQTLs are often independent of eQTLs. Therefore, integrating genetically-regulated splicing quantification may partially explain the phenotypic variation attributed to alternative genetic factors, $\text{PVE}_{\text{Alternative}}$. We argue that with the proper multi-omics reference data, similar analyses can be conducted to quantify the impact of genetically-regulated methylation, protein, and other multi-omics variation on phenotype [51].

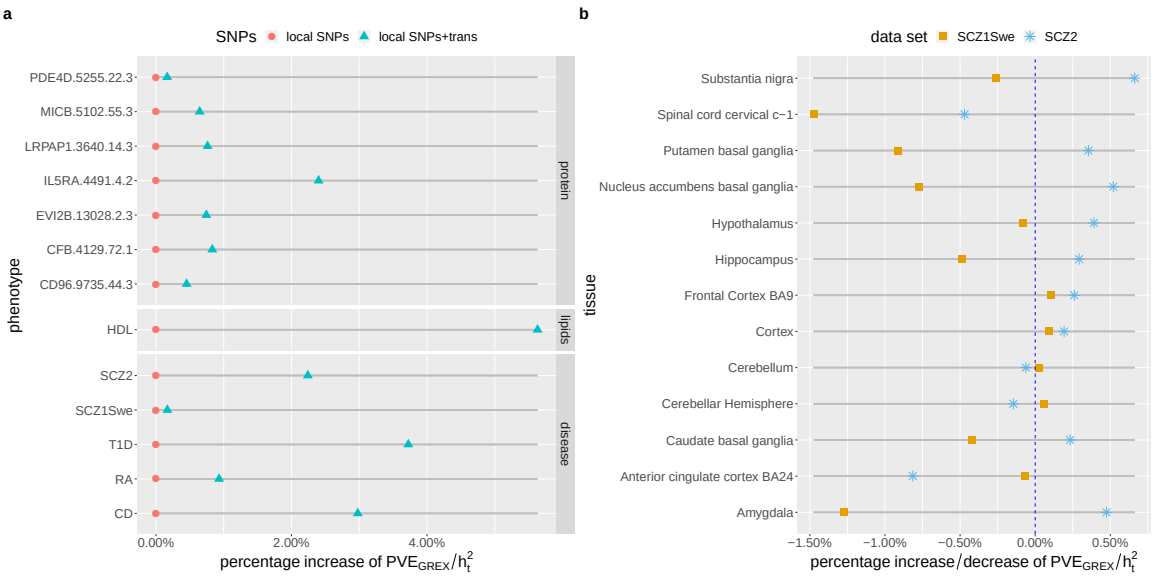

Figure 6: Comparison of $\text{PVE}_{\text{GREX}}/h_t^2$'s estimated with only local SNPs and those estimated with additional trans-eQTLs. (**a**) Estimated $\text{PVE}_{\text{GREX}}/h_t^2$ of 13 datasets in blood. All these datasets have significant $\text{PVE}_{\text{GREX}}/h_t^2$ in blood at 0.05 nominal level using only local SNPs. (**b**) Estimated $\text{PVE}_{\text{GREX}}/h_t^2$ of two largest SCZ data sets in 13 GTEx brain tissues. All these tissues have significant $\text{PVE}_{\text{GREX}}/h_t^2$ at 0.05 nominal level in both datasets using only local SNPs.

Discussion

In this work, we proposed a method, IGREX, for integrating GWAS and eQTL reference data to quantify the GREX impact on phenotype. IGREX can be applied to both individual-level and summary-level GWAS data, and was shown to achieve estimation accuracy even when the eQTL effects are weak. IGREX can be used in many ways: it can inform the role of GREX variation in various phenotypes and/or the role of GREX in known pathways; it can guide the selection of eQTL reference data and suggest trait-relevant tissues/cell-types/contexts; and it is generally applicable to the integration of GWAS with other omics data types to examine the role of genetically-regulated multi-omics traits.

IGREX is closely related to several existing methods and here we briefly discuss the connections and distinctions. By also integrating an eQTL reference and GWAS data, methods including TWAS [17], PrediXcan [16], and the more general MetaXcan [21] aim to identify specific trait-associated genes. In contrast, IGREX estimates the impact of genetically regulated expression from a global perspective by quantifying the phenotypic variation that can be attributed to the GREX component. Since both the TWAS-type of analyses and IGREX rely on the shared GREX variation across eQTL and GWAS data, we argue that with the increasing availability of eQTL resources in different populations, conditions and contexts, the proper selection of eQTL reference panels via IGREX will greatly promote the chances of successes in the subsequent TWAS-type of analyses.

There are also existing methods, such as RhoGE, designed for identifying and estimating correlations between gene expression and complex traits. RhoGe provides an LDSC-based approach for estimating PVE_{GREX}. Unlike IGREX, this method does not adjust for estimation uncertainty. Consequently, it significantly underestimates the PVE_{GREX} when the eQTL effects on expression levels are weak or moderate. In fact, RhoGE estimated the PVE_{GREX} for the majority of $1,350$ tissue-trait pairs to be almost negligible, with the first quantile, the median, and the third quantile being 0.00125%, 0.162% and 0.616%, respectively [23]. In contrast, as demonstrated via simulation studies, IGREX can accurately estimate PVE_{GREX} in various scenarios by accounting for the estimation uncertainty.

Based on estimating PVE_{GREX} for a wide-array of tissue-trait pairs, we observed a stronger impact of GREX on molecular intermediate traits and lipid traits in trait-relevant tissue types. We also observed a relatively low PVE_{GREX} for complex traits in general. The big picture

suggests the attenuated impact on downstream phenotypes (e.g, height and SCZ), which is consistent with the result from a pioneer study [57]. However, we noted that the PVE_{GREX} estimates could be improved. A substantial amount of expression heritability is explained by trans-acting genetic factors while current TWAS and IGREX analyses are mainly using only cis-eQTLs. We explored the potential of incorporating trans-eQTLs in TWAS analysis by re-estimating PVE_{GREX} for selected traits in blood tissues with significant trans-eQTLs independently derived from the blood-based eQTLGen Consortium. We observed consistent increases in PVE_{GREX} for blood-related traits. In contrast, such an increase was not observed in the PVE_{GREX} estimates for other tissue types, again illustrating the importance of considering trait-relevant tissue types/conditions in the TWAS-type of analyses. Additionally, we extended the IGREX analysis to quantify the impact of genetically-regulated alternative splicing events on selected traits. Our results suggested the potential for extending TWAS-type of analysis to integrate reference multi-omics QTL data with GWAS in mapping novel disease/trait-associated genes with mechanisms via other omics traits (such as splicing, methylation, protein, etc.).

A key assumption in applying IGREX or TWAS methods with a general-purpose eQTL data as reference is the existence of steady-sate component in GREX, i.e., the genetic effects on gene expression $\boldsymbol{\beta}_g$ are shared across the eQTL reference and GWAS data. However, there are many situations in which this assumption is violated. For example, it has been observed that CAD-risk SNPs have a larger overlap with cis-eQTLs isolated from disease-relevant tissues than those from GTEx tissues [46], implying the existence of a dynamic component. In the presence of this dynamic component, the accuracy of $\widehat{\text{PVE}}_{\text{GREX}}$ based on GTEx is reduced. In those cases, we suggest exploring other trait-relevant or condition-specific eQTL reference panels using IGREX for a better understanding of the role of GREX and before conducting TWAS analysis.

Methods

The IGREX-i for individual-level GWAS data. First, let $\mathcal{D}_r = \{\mathbf{Y}, \mathbf{X}_r\}$ denote the reference data set from an eQTL study, where $\mathbf{Y} \in \mathbb{R}^{n_r \times G}$ is the gene expression matrix, $\mathbf{X}_r \in \mathbb{R}^{n_r \times M}$ is the genotype matrix, n_r is the sample size of the eQTL study, G is the number of genes and M is the number of single-nucleotide polymorphisms (SNPs). Suppose we have individual-level GWAS data $\mathcal{D}_i = \{\mathbf{t}, \mathbf{X}\}$ comprised of phenotype vector $\mathbf{t} \in \mathbb{R}^n$ and

genotype matrix $\mathbf{X} \in \mathbb{R}^{n \times M}$, where n is the GWAS sample size. For $g = 1, ..., G$, we let the g-th gene expression vector $\mathbf{y}_g \in \mathbb{R}^{n_r}$ denote the corresponding column of $\mathbf{Y}$, local genotype matrices $\mathbf{X}_{r,g} \in \mathbb{R}^{n_r \times M_g}$ and $\mathbf{X}_g \in \mathbb{R}^{n \times M_g}$ denote the corresponding M_g columns in $\mathbf{X}_r$ and $\mathbf{X}$, respectively, where M_g is the number of local SNPs for g-th gene. To make the notation uncluttered, we further assume that $\mathbf{X}_{r,g}$ and $\mathbf{X}_g$ have been standardized and both $\mathbf{y}_g$ and $\mathbf{t}$ have been properly adjusted for covariates. The complete model that accounts for covariates is described in the Supplementary Materials. Now, we consider linear model (1) that associates the gene expression vector $\mathbf{y}_g$ to $\mathbf{X}_{r,g}$:

$$\mathbf{y}_g = \mathbf{X}_{r,g}\boldsymbol{\beta}_g + \mathbf{e}_{r,g},$$

where $\boldsymbol{\beta}_g$ is an $M_g \times 1$ vector of genetic effects on the gene expression levels, $\mathbf{e}_{r,g} \sim \mathcal{N}(0, \sigma^2_{r,g}\mathbf{I}_{n_r})$ is a vector of independent noise and $\mathbf{I}$ is the identity matrix with the subscript being its size. Assuming that there is a steady-state component in gene expression regulated by genetic variants, individuals in $\mathcal{D}_r$ and $\mathcal{D}_i$ share the same $\boldsymbol{\beta}_g$. Hence, the GREX in $\mathcal{D}_i$ can be evaluated by $\mathbf{X}_g\boldsymbol{\beta}_g$. Then, we assume that the phenotype $\mathbf{t}$ can be decomposed into two parts, i.e., the genetic effects via GREX and the genetic effects through alternative pathways, as in model (2):

$$\mathbf{t} = \sum_{g=1}^{G} \alpha_g \mathbf{X}_g\boldsymbol{\beta}_g + \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon},$$

where α_g is the effect of $\mathbf{X}_g\boldsymbol{\beta}_g$ on $\mathbf{t}$, $\boldsymbol{\gamma}$ is an $n \times 1$ vector of alternative genetic effects and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2_\epsilon\mathbf{I}_n)$ is a vector of independent errors. The term $\sum_{g=1}^{G} \alpha_g\mathbf{X}_g\boldsymbol{\beta}_g$ can be viewed as the overall impact of GREX on the phenotype and $\mathbf{X}\boldsymbol{\gamma}$ represents the alternative impact of genotypes on the phenotype. Given a genotype vector $\mathbf{x} \in \mathbb{R}^M$ and a phenotype $t \in \mathbb{R}$, the impact of GREX can be quantified by the proportion of variance explained by the GREX component:

$$\text{PVE}_{\text{GREX}} = \frac{\text{Var}(\sum_{g=1}^{G} \alpha_g \mathbf{x}_g^T\boldsymbol{\beta}_g)}{\text{Var}(t)}, \tag{3}$$

where $\mathbf{x}_g$ is the subvector of genotype $\mathbf{x}$ corresponding to the g-th gene.

To estimate PVE_{GREX}, we introduce the following probabilistic structure for the effects in model (1) and (2):

$$\boldsymbol{\beta}_g \sim \mathcal{N}(\mathbf{0}, \sigma^2_{\beta_g}\mathbf{I}_{M_g}),\ \alpha_g \sim \mathcal{N}(\mathbf{0}, \sigma^2_\alpha),\ \boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{0}, \sigma^2_\gamma\mathbf{I}_M), \tag{4}$$

which is motivated by a recent theoretical justification [58] for heritability estimation on a mis-specified linear mixed model (LMM). This prior specification in (4) provides a great com-

putational advantage as well as a stable performance for IGREX under model mis-specification, as demonstrated in the simulation studies.

The proposed method for individual-level GWAS data, IGREX-i, provides a two-stage framework for estimating PVE_{GREX}. In the first stage, we estimate the parameters $\sigma^2_{\beta_g}$ and $\sigma^2_{r,g}$ in model (1) via a fast expectation-maximization (EM)-type algorithm, the parameter-expanded EM (PX-EM) algorithm [59]. Based on the estimates, denoted as $\hat{\sigma}^2_{\beta_g}$ and $\hat{\sigma}^2_{r,g}$, the posterior distribution of $\boldsymbol{\beta}_g$ is given by

$$\boldsymbol{\beta}_g|\mathbf{y}_g, \mathbf{X}_{r,g} \sim \mathcal{N}(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \text{ where } \boldsymbol{\Sigma}_g = \left(\frac{1}{\hat{\sigma}^2_{r,g}}\mathbf{X}^T_{r,g}\mathbf{X}_{r,g} + \frac{1}{\hat{\sigma}^2_{\beta_g}}\mathbf{I}_{M_g}\right)^{-1}, \ \boldsymbol{\mu}_g = \boldsymbol{\Sigma}_g \frac{1}{\hat{\sigma}^2_{r,g}}\mathbf{X}^T_{r,g}\mathbf{y}_g. \quad (5)$$

In the second stage, we treat the posterior distribution obtained in (5) as the prior distribution of $\boldsymbol{\beta}_g$ in model (2). This substitution naturally accounts for the uncertainty in estimating $\boldsymbol{\beta}_g$ which has been captured by $\boldsymbol{\Sigma}_g$. To evaluate the covariance of $\mathbf{t}$, we first note that $\mathbb{E}(\mathbf{t}|\boldsymbol{\alpha}) = \sum_{g=1}^G \alpha_g \mathbf{X}_g\boldsymbol{\mu}_g$ and $\text{Cov}(\mathbf{t}|\boldsymbol{\alpha}) = \sum_{g=1}^G \alpha_g^2 \mathbf{X}_g\boldsymbol{\Sigma}_g\mathbf{X}_g^T + \sigma^2_\gamma\mathbf{X}\mathbf{X}^T + \sigma^2_\epsilon\mathbf{I}_n$; then, using the law of total expectation and total variance, we obtain $\mathbb{E}(\mathbf{t}) = \mathbb{E}(\mathbb{E}(\mathbf{t}|\boldsymbol{\alpha})) = \mathbf{0}$ and

$$\text{Cov}(\mathbf{t}) = \text{Cov}(\mathbb{E}(\mathbf{t}|\boldsymbol{\alpha})) + \mathbb{E}(\text{Cov}(\mathbf{t}|\boldsymbol{\alpha})) = \sum_{g=1}^G \sigma^2_\alpha \mathbf{X}_g(\boldsymbol{\mu}_g\boldsymbol{\mu}_g^T + \boldsymbol{\Sigma}_g)\mathbf{X}_g^T + \sigma^2_\gamma\mathbf{X}\mathbf{X}^T + \sigma^2_\epsilon\mathbf{I}_n, \quad (6)$$

respectively. By observing the form of (6), it is clear that the i-th diagonal element of $\sum_{g=1}^G \sigma^2_\alpha \mathbf{X}_g(\boldsymbol{\mu}_g\boldsymbol{\mu}_g^T + \boldsymbol{\Sigma}_g)\mathbf{X}_g^T$ and $\sigma^2_\gamma\mathbf{X}\mathbf{X}^T$ represents the variance explained by GREX and alternative genetic effects, respectively. Therefore, the PVE_{GREX} defined in (3) can be estimated by

$$\widehat{\text{PVE}}_{\text{GREX}} = \frac{\text{tr}(\sum_{g=1}^G \hat{\sigma}^2_\alpha \mathbf{X}_g(\boldsymbol{\mu}_g\boldsymbol{\mu}_g^T + \boldsymbol{\Sigma}_g)\mathbf{X}_g^T)}{\text{tr}(\sum_{g=1}^G \hat{\sigma}^2_\alpha \mathbf{X}_g(\boldsymbol{\mu}_g\boldsymbol{\mu}_g^T + \boldsymbol{\Sigma}_g)\mathbf{X}_g^T + \hat{\sigma}^2_\gamma\mathbf{X}\mathbf{X}^T + \hat{\sigma}^2_\epsilon\mathbf{I}_n)}, \quad (7)$$

where $\hat{\sigma}^2_\alpha$, $\hat{\sigma}^2_\gamma$ and $\hat{\sigma}^2_\epsilon$ are the estimated values of σ^2_α, σ^2_γ and σ^2_ϵ, respectively.

IGREX-i provides two methods for estimating the parameters and $\widehat{\text{PVE}}_{\text{GREX}}$ in the second stage. Let $\boldsymbol{\psi} = \left[\sigma^2_\alpha, \sigma^2_\gamma, \sigma^2_\epsilon\right]^T$ be the vector of parameters to be estimated, $\mathbf{K}_\alpha = \sum_{g=1}^G \mathbf{X}_g(\boldsymbol{\mu}_g\boldsymbol{\mu}_g^T + \boldsymbol{\Sigma}_g)\mathbf{X}_g^T$ and $\mathbf{K}_\gamma = \mathbf{X}\mathbf{X}^T$. The first method is based on MoM, which minimizes the distance between the second moment of $\mathbf{t}$ at the population level and that at the sample level $f(\boldsymbol{\psi}) = ||\mathbf{t}\mathbf{t}^T - (\sigma^2_\alpha\mathbf{K}_\alpha + \sigma^2_\gamma\mathbf{K}_\gamma + \sigma^2_\epsilon\mathbf{I}_n)||^2$. By setting $\frac{\partial f(\boldsymbol{\psi})}{\partial \sigma^2_\alpha} = \frac{\partial f(\boldsymbol{\psi})}{\partial \sigma^2_\gamma} = \frac{\partial f(\boldsymbol{\psi})}{\partial \sigma^2_\epsilon} = 0$, we obtain the estimating equation

$$\mathbf{S}\boldsymbol{\psi} = \mathbf{q}, \quad (8)$$

$$\text{with } \mathbf{S} = \begin{bmatrix} \text{tr}(\mathbf{K}_\alpha^2) & \text{tr}(\mathbf{K}_\alpha\mathbf{K}_\gamma) & \text{tr}(\mathbf{K}_\alpha) \\ \text{tr}(\mathbf{K}_\alpha\mathbf{K}_\gamma) & \text{tr}(\mathbf{K}_\gamma^2) & \text{tr}(\mathbf{K}_\gamma) \\ \text{tr}(\mathbf{K}_\alpha) & \text{tr}(\mathbf{K}_\gamma) & n \end{bmatrix}, \ \boldsymbol{\psi} = \begin{bmatrix} \sigma^2_\alpha \\ \sigma^2_\gamma \\ \sigma^2_\epsilon \end{bmatrix}, \ \mathbf{q} = \begin{bmatrix} \mathbf{t}^T\mathbf{K}_\alpha\mathbf{t} \\ \mathbf{t}^T\mathbf{K}_\gamma\mathbf{t} \\ \mathbf{t}^T\mathbf{t} \end{bmatrix}.$$

The solution of Equation (8) is given by $\hat{\boldsymbol{\psi}} = \mathbf{S}^{-1}\mathbf{q}$. And the variance-covariance matrix of $\hat{\boldsymbol{\psi}}$ is given by $\text{Cov}(\hat{\boldsymbol{\psi}}) = \mathbf{S}^{-1}\text{Cov}(\mathbf{q})\mathbf{S}^{-1}$ using the sandwich estimator. Then, the standard error of $\widehat{\text{PVE}}_{\text{GREX}}$ can be obtained by the delta method (see Supplementary Materials). The second method applies the restricted maximum likelihood (REML) by further assuming the normal distribution of $\mathbf{t}$: $\mathbf{t} \sim \mathcal{N}(\mathbf{0}, \sigma_\alpha^2\mathbf{K}_\alpha + \sigma_\gamma^2\mathbf{K}_\gamma + \sigma_\epsilon^2\mathbf{I}_n)$. The variance components are estimated by the Minorization-Maximization (MM) algorithm [60].

The IGREX-s for summary-level GWAS data. The special formulation of method of moments allows IGREX to be extended (IGREX-s) to handle summary-level GWAS data (i.e. z-scores) when the individual-level data $\mathcal{D}_i$ is not accessible. Suppose we only have the z-scores from summary-level GWAS data $\{z_j\}_{j=1}^M$ generated from $\mathcal{D}_i$. The definition of the z-score is $z_j = \frac{(\mathbf{x}_j^T\mathbf{x}_j)^{-1}\mathbf{x}_j^T\mathbf{t}}{\sqrt{\hat{\sigma}_j^2(\mathbf{x}_j^T\mathbf{x}_j)^{-1}}}$, where $\mathbf{x}_j$ is the j-th column of $\mathbf{X}$ and $\hat{\sigma}_j^2$ is the estimate of residual variance by regressing $\mathbf{x}_j$ on $\mathbf{t}$. By assuming that z-scores are calculated from a standardized genotype matrix $\mathbf{X}$, we have $\mathbf{x}_j^T\mathbf{x}_j = n$. Besides, the polygenicity assumption implies that $\hat{\sigma}_j^2 \approx \hat{\sigma}_t^2$, where $\hat{\sigma}_t^2$ is the estimate of $\text{Var}(t)$. Hence, we have

$$z_j \approx \frac{\mathbf{x}_j^T\mathbf{t}}{\sqrt{n\hat{\sigma}_t^2}}, \tag{9}$$

and PVE_{GREX} defined in (3) can be estimated by

$$\widehat{\text{PVE}}_{\text{GREX}} = \frac{\frac{1}{n}\text{tr}(\sum_{g=1}^G \hat{\sigma}_\alpha^2\mathbf{X}_g(\boldsymbol{\mu}_g\boldsymbol{\mu}_g^T + \boldsymbol{\Sigma}_g)\mathbf{X}_g^T)}{\hat{\sigma}_t^2} \approx \frac{\hat{\sigma}_\alpha^2}{\hat{\sigma}_t^2}\text{tr}(\sum_{g=1}^G(\boldsymbol{\mu}_g\boldsymbol{\mu}_g^T + \boldsymbol{\Sigma}_g)\hat{\mathbf{R}}_g), \tag{10}$$

where $\hat{\mathbf{R}}_g = \tilde{\mathbf{X}}_g^T\tilde{\mathbf{X}}_g/(m-1)$ is the estimated LD matrix associated with the g-th gene and $\tilde{\mathbf{X}}_g$ is the corresponding columns of a reference genotype matrix $\tilde{\mathbf{X}}$. In practice, $\tilde{\mathbf{X}} \in \mathbb{R}^{m\times M}$ can be the genotype matrix either from the GTEx Project or the 1000 Genomes Project. Now, we consider MoM in the estimating equation (8) to obtain $\frac{\hat{\sigma}_\alpha^2}{\hat{\sigma}_t^2}$. By eliminating σ_ϵ^2 and dividing both sides by n^2, we have

$$\begin{bmatrix} \frac{\text{tr}(\mathbf{K}_\alpha^2) - \frac{\text{tr}^2(\mathbf{K}_\alpha)}{n}}{n^2} & \frac{\text{tr}(\mathbf{K}_\alpha\mathbf{K}_\gamma) - \frac{\text{tr}(\mathbf{K}_\alpha)\text{tr}(\mathbf{K}_\gamma)}{n}}{n^2} \\ \frac{\text{tr}(\mathbf{K}_\alpha\mathbf{K}_\gamma) - \frac{\text{tr}(\mathbf{K}_\alpha)\text{tr}(\mathbf{K}_\gamma)}{n}}{n^2} & \frac{\text{tr}(\mathbf{K}_\gamma^2) - \frac{\text{tr}^2(\mathbf{K}_\gamma)}{n}}{n^2} \end{bmatrix} \begin{bmatrix} \sigma_\alpha^2 \\ \sigma_\gamma^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{n^2}\mathbf{t}^T\mathbf{K}_\alpha\mathbf{t} - \frac{\text{tr}(\mathbf{K}_\alpha)}{n^3}\mathbf{t}^T\mathbf{t} \\ \frac{1}{n^2}\mathbf{t}^T\mathbf{K}_\gamma\mathbf{t} - \frac{\text{tr}(\mathbf{K}_\gamma)}{n^3}\mathbf{t}^T\mathbf{t} \end{bmatrix}. \tag{11}$$

The terms on the left hand side do not involve $\mathbf{t}$ and thus can be approximated using $\tilde{\mathbf{X}}$ [37]. For example, $\frac{\text{tr}(\mathbf{K}_\alpha^2) - \frac{\text{tr}^2(\mathbf{K}_\alpha)}{n}}{n^2}$ can be well approximated by $\frac{\text{tr}(\tilde{\mathbf{K}}_\alpha^2) - \frac{\text{tr}^2(\tilde{\mathbf{K}}_\alpha)}{m}}{m^2}$, where $\tilde{\mathbf{K}}_\alpha = \sum_{g=1}^G \tilde{\mathbf{X}}_g(\boldsymbol{\mu}_g\boldsymbol{\mu}_g^T + \boldsymbol{\Sigma}_g)\tilde{\mathbf{X}}_g^T$. Other terms on the left hand side can be approximated in the same way. For the right hand side, each term can be approximated using $\hat{\mathbf{R}}_g$ and z-scores from

approximation (9): $\mathbf{t}^T\mathbf{K}_\alpha\mathbf{t} \approx n\hat{\sigma}_t^2 \sum_g \mathbf{z}_g^T(\boldsymbol{\mu}_g\boldsymbol{\mu}_g^T+\boldsymbol{\Sigma}_g)\mathbf{z}_g$, where $\mathbf{z}_g \in \mathbb{R}^{M_g}$ is the vector of z-scores corresponding to the g-th gene; $\frac{\mathrm{tr}(\mathbf{K}_\alpha)}{n}\mathbf{t}^T\mathbf{t} \approx n\hat{\sigma}_t^2\mathrm{tr}(\sum_g(\boldsymbol{\mu}_g\boldsymbol{\mu}_g^T+\boldsymbol{\Sigma}_g)\hat{\mathbf{R}}_g)$; $\mathbf{t}^T\mathbf{K}_\gamma\mathbf{t} \approx n\hat{\sigma}_t^2\sum_{j=1}^M z_j^2$; and $\frac{\mathrm{tr}(\mathbf{K}_\gamma)}{n}\mathbf{t}^T\mathbf{t} \approx n\hat{\sigma}_t^2$. With these approximations, Equation (11) becomes

$$\begin{bmatrix} \frac{\mathrm{tr}(\tilde{\mathbf{K}}_\alpha^2)-\frac{\mathrm{tr}^2(\tilde{\mathbf{K}}_\alpha)}{m}}{m^2} & \frac{\mathrm{tr}(\tilde{\mathbf{K}}_\alpha\tilde{\mathbf{K}}_\gamma)-\frac{\mathrm{tr}(\tilde{\mathbf{K}}_\alpha)\mathrm{tr}(\tilde{\mathbf{K}}_\gamma)}{m}}{m^2} \\ \frac{\mathrm{tr}(\tilde{\mathbf{K}}_\alpha\tilde{\mathbf{K}}_\gamma)-\frac{\mathrm{tr}(\tilde{\mathbf{K}}_\alpha)\mathrm{tr}(\tilde{\mathbf{K}}_\gamma)}{m}}{m^2} & \frac{\mathrm{tr}(\tilde{\mathbf{K}}_\gamma^2)-\frac{\mathrm{tr}^2(\tilde{\mathbf{K}}_\gamma)}{m}}{m^2} \end{bmatrix} \begin{bmatrix} \frac{\hat{\sigma}_\alpha^2}{\hat{\sigma}_t^2} \\ \frac{\hat{\sigma}_\gamma^2}{\hat{\sigma}_t^2} \end{bmatrix} = \begin{bmatrix} \frac{\sum_g \mathbf{z}_g^T(\boldsymbol{\mu}_g\boldsymbol{\mu}_g^T+\boldsymbol{\Sigma}_g)\mathbf{z}_g-\mathrm{tr}(\sum_g(\boldsymbol{\mu}_g\boldsymbol{\mu}_g^T+\boldsymbol{\Sigma}_g)\hat{\mathbf{R}}_g)}{n} \\ \sum_{j=1}^M \frac{z_j^2-1}{n} \end{bmatrix}.$$

Then, $\frac{\hat{\sigma}_\alpha^2}{\hat{\sigma}_t^2}$ can be obtained by solving this equation. Plugging this estimate into Equation (10) gives the $\widehat{\mathrm{PVE}}_{\mathrm{GREX}}$. The standard errors of $\widehat{\mathrm{PVE}}_{\mathrm{GREX}}$ can be estimated by the block jackknife method [61] (Supplementary Materials).

IGREX can incorporate fixed effects to adjust for possible confounding factors, such as population structure. Details are provided in the Supplementary Note.

GTEx eQTL dataset. We used the gene expression data from the V7 release of GTEx Consortium as our reference dataset. We analyzed the 48 tissues with number of genotyped samples $\geq$ 70, which are collected from 620 donors with total sample size 10,294. The sample size of each tissue ranges from 80 to 491 (details provided in Supplementary Table 4). We set the mappability cutoff at 0.9 to filter gene expression measures with lower quality, leaving $16,333 - 27,378$ genes to be included in our analysis. Based on the third phase of the International HapMap project phase 3 (HapMap3), 1,189,556 SNPs were included from the GTEx genotyped data for analysis. For each gene, we included only the SNPs within 500kb of the transcription start and end of each protein coding genes. In real data analysis, we used the covariates provided by the GTEx consortium, including genotype principal components (PCs), Probabilistic Estimation of Expression Residuals (PEER) factors, genotyping platform and sex (as described in `https://gtexportal.org/home/documentationPage`).

Additionally, the GTEx genotype data was used as an LD reference panel when applying IGREX-s to GWAS summary statistics. In this application, we used top 5 PCs as covariates.

Individual level GWAS datasets. The NFBC dataset is comprised of 5,402 individuals with ten continuous phenotypes related to cardiovascular disease including Glucose, body mass index (BMI), C-reactive protein (CRP), insulin, high-density lipoprotein cholesterol (HDL), low-density lipoprotein cholesterol (LDL), triglycerides (TG), total cholesterol (TC), diastolic blood pressure (DiaBP) and systolic blood pressure (SysBP). There are 364,590 genotyped SNPs in this dataset. We first excluded the individuals whose reported sex differed from their

sex determined from the X chromosome. We then excluded the SNPs with minor allele frequency less than 1%, with missing values in more than 1% of the individuals or with Hardy-Weinberg equilibrium (HWE) p-value below 0.0001. This quality control process yields $5,123$ individuals with $319,147$ SNPs in NFBC dataset for our analysis. We evaluated the genetic relatedness matrix (GRM) using the processed genotype data and selected the top 20 PCs as covariates in the study.

The WTCCC dataset contains seven disease phenotypes including bipolar disorder (BD), coronary artery disease (CAD), Crohn's disease (CD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D) and type 2 diabetes (T2D). It includes $\sim 2,000$ cases per phenotype and $3,004$ controls with $490,032$ genotyped SNPs. We first removed the individuals with genotyping rate less than 5%. Then we excluded the SNPs satisfying at least one of the following: minor allele frequency less than 5%; genotypes missing in more than 1% samples; HWE p-value is below 0.001. We also removed the individuals with estimated genetic correlation larger than 2.5%. After quality control, around $4,700$ individuals with $300,000$ SNPs were retained for our analysis (See Supplementary Table 1). Based on the obtained data, we calculated the GRM and extracted top 20 PCs as covariates to be included in our analysis.

GWAS summary statistics. We analyzed ten summary level GWAS datasets: human plasma pQTL data [38], circulating metabolite data [40], four schizophrenia datasets [41, 42, 43, 44], two independent height datasets [62] and European ancestry of BMI datasets with sample age ≤ 50 separated by men and women [63]. The SNPs with missing information (i.e. chromosome, minor allele, allele frequency) were first removed. Following the practice of LDSC [30], we checked the χ^2 statistic of each SNP and excluded those with extreme values ($\chi^2 > 80$) to prevent the outliers that may unduly affect the results. The detailed information is provided in Supplementary Table 2. After pre-processing, the remaining SNPs were further matched with reference data, and this step is automatically conducted in our IGREX software.

The eQTLGen summary data. We used the trans-eQTLs in blood provided by the eQTLGen Consortium [15]. The trans-eQTL analysis were restricted to known complex trait-associated SNPs. The significant trans-eQTLs were identified by controlling the FDR at 0.05. There were $5,4786$ gene-SNP pairs composed of $6,298$ genes and $3,853$ SNPs. The remaining pairs after matching with both reference and GWAS datasets are summarized in Supplementary Table 5.

Data availability. The GTEx gene expression data was downloaded from GTEx Consortium website `https://gtexportal.org/home/datasets`. The GTEx genotype data can be accessed from dbGAP with accession number phs000424.v7.p2. The HapMap3 genotype data is available at `ftp://ftp.ncbi.nlm.nih.gov/hapmap/`. The NFBC study was downloaded from dbGAP using accession number phs000276.v1.p1. The WTCCC data was obtained from its consortium website `https://www.wtccc.org.uk/info/access_to_data_samples.html`. The GWAS summary statistics can be accessed using the links provided in Supplementary Table 2. The eQTLGen data can be downloaded from `http://www.eqtlgen.org`.

Software. The R software package IGREX is publicly available on GitHub repository: `https://github.com/mxcai/iGREX`.

Acknowledgements. We thank Mr. Kevin J. Gleason for proof-reading the work. This work was supported in part by the National Science Funding of China [61501389]; the Hong Kong Research Grant Council [12316116, 12301417 and 16307818]; The Hong Kong University of Science and Technology [startup grant R9405 and IGN17SC02]; Duke-NUS Medical School WBS [R-913-200-098-263]; Ministry of Education, Singapore. AcRF Tier 2 [MOE2016-T2-2-029, MOE2018-T2-1-046 and MOE2018-T2-2-006]. LSC was independently supported by the National Institutes of Health (NIH) grant R01GM108711. The computational work for this article was (fully or partially) performed on resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg).

References

[1] Matthew T. Maurano, Richard Humbert, Eric Rynes, Robert E. Thurman, Eric Haugen, Hao Wang, Alex P. Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, Anthony Shafer, Fidencio Neri, Kristen Lee, Tanya Kutyavin, Sandra Stehling-Sun, Audra K. Johnson, Theresa K. Canfield, Erika Giste, Morgan Diegel, Daniel Bates, R. Scott Hansen, Shane Neph, Peter J. Sabo, Shelly Heimfeld, Antony Raubitschek, Steven Ziegler, Chris Cotsapas, Nona Sotoodehnia, Ian Glass, Shamil R. Sunyaev, Rajinder Kaul, and John A. Stamatoyannopoulos. Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099):1190–1195, 2012.

[2] William Cookson, Liming Liang, Gonçalo Abecasis, Miriam Moffatt, and Mark Lathrop. Mapping complex disease traits with global gene expression. *Nature Reviews Genetics*, 10(3):184, 2009.

[3] Mark M Pomerantz, Nasim Ahmadiyeh, LI Jia, Paula Herman, Michael P Verzi, Harshavardhan Doddapaneni, Christine A Beckwith, Jennifer A Chan, Adam Hills, Matt Davis, et al. The 8q24 cancer risk variant rs6983267 shows long-range interaction with myc in colorectal cancer. *Nature genetics*, 41(8):882, 2009.

[4] Kiran Musunuru, Alanna Strong, Maria Frank-Kamenetsky, Noemi E Lee, Tim Ahfeldt, Katherine V Sachs, Xiaoyu Li, Hui Li, Nicolas Kuperwasser, Vera M Ruda, et al. From noncoding variant to phenotype via sort1 at the 1p13 cholesterol locus. *Nature*, 466(7307):714, 2010.

[5] Olivier Harismendy, Dimple Notani, Xiaoyuan Song, Nazli G Rahim, Bogdan Tanasa, Nathaniel Heintzman, Bing Ren, Xiang-Dong Fu, Eric J Topol, Michael G Rosenfeld, et al. 9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response. *Nature*, 470(7333):264, 2011.

[6] Dan L Nicolae, Eric Gamazon, Wei Zhang, Shiwei Duan, M Eileen Dolan, and Nancy J Cox. Trait-associated SNPs are more likely to be eqtls: annotation to enhance discovery from gwas. *PLoS genetics*, 6(4):e1000888, 2010.

[7] Lucia A. Hindorff, Praveen Sethupathy, Heather A. Junkins, Erin M. Ramos, Jayashri P.

Mehta, Francis S. Collins, and Teri A. Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367, 2009.

[8] Frank W Albert and Leonid Kruglyak. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16(4):197, 2015.

[9] Kevin J. Gleason, Fan Yang, Brandon L. Pierce, Xin He, and Lin S. Chen. Primo: integration of multiple GWAS and omics QTL summary statistics for elucidation of molecular mechanisms of trait-associated snps and detection of pleiotropy in complex traits. *bioRxiv*, 579581, 2019.

[10] Eric R Gamazon, Ayellet V Segrè, Martijn van de Bunt, Xiaoquan Wen, Hualin S Xi, Farhad Hormozdiari, Halit Ongen, Anuar Konkashbaev, Eske M Derks, François Aguet, et al. Using an atlas of gene regulation across 44 human tissues to inform complex disease-and trait-associated variation. *Nature genetics*, 50(7):956, 2018.

[11] GTEx Consortium et al. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204, 2017.

[12] Luke R Lloyd-Jones, Alexander Holloway, Allan McRae, Jian Yang, Kerrin Small, Jing Zhao, Biao Zeng, Andrew Bakshi, Andres Metspalu, Manolis Dermitzakis, et al. The genetic architecture of gene expression in peripheral blood. *The American Journal of Human Genetics*, 100(2):228–237, 2017.

[13] Harm-Jan Westra, Marjolein J Peters, Tõnu Esko, Hanieh Yaghootkar, Claudia Schurmann, Johannes Kettunen, Mark W Christiansen, Benjamin P Fairfax, Katharina Schramm, Joseph E Powell, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature genetics*, 45(10):1238, 2013.

[14] Ting Qi, Yang Wu, Jian Zeng, Futao Zhang, Angli Xue, Longda Jiang, Zhihong Zhu, Kathryn Kemper, Loic Yengo, Zhili Zheng, et al. Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. *Nature communications*, 9(1):2282, 2018.

[15] Urmo Võsa, Annique Claringbould, Harm-Jan Westra, Marc Jan Bonder, Patrick Deelen, Biao Zeng, Holger Kirsten, Ashis Saha, Roman Kreuzhuber, Silva Kasela, et al. Unraveling

608 the polygenic architecture of complex traits using blood eQTL meta-analysis. *bioRxiv*,
609 page 447367, 2018.

610 [16] Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-
611 Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, Dan L Nicolae, Nancy J Cox,
612 et al. A gene-based association method for mapping traits using reference transcriptome
613 data. *Nature genetics*, 47(9):1091, 2015.

614 [17] Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda WJH
615 Penninx, Rick Jansen, Eco JC De Geus, Dorret I Boomsma, Fred A Wright, et al.
616 Integrative approaches for large-scale transcriptome-wide association studies. *Nature*
617 *genetics*, 48(3):245, 2016.

618 [18] Zhihong Zhu, Futao Zhang, Han Hu, Andrew Bakshi, Matthew R Robinson, Joseph E
619 Powell, Grant W Montgomery, Michael E Goddard, Naomi R Wray, Peter M Visscher,
620 et al. Integration of summary data from GWAS and eQTL studies predicts complex trait
621 gene targets. *Nature genetics*, 48(5):481, 2016.

622 [19] Nicholas Mancuso, Malika K Freund, Ruth Johnson, Huwenbo Shi, Gleb Kichaev, Alexan-
623 der Gusev, and Bogdan Pasaniuc. Probabilistic fine-mapping of transcriptome-wide
624 association studies. *Nature Genetics*, 51(4):675, 2019.

625 [20] Kunal Bhutani, Abhishek Sarkar, Yongjin Park, Manolis Kellis, and Nicholas J Schork.
626 Modeling prediction error improves power of transcriptome-wide association studies.
627 *bioRxiv*, page 108316, 2017.

628 [21] Alvaro N Barbeira, Scott P Dickinson, Rodrigo Bonazzola, Jiamao Zheng, Heather E
629 Wheeler, Jason M Torres, Eric S Torstenson, Kaanan P Shah, Tzintzuni Garcia, Todd L
630 Edwards, et al. Exploring the phenotypic consequences of tissue specific gene expression
631 variation inferred from GWAS summary statistics. *Nature communications*, 9(1):1825,
632 2018.

633 [22] Can Yang, Xiang Wan, Xinyi Lin, Mengjie Chen, Xiang Zhou, and Jin Liu. CoMM:
634 a collaborative mixed model to dissecting genetic contributions to complex traits by
635 leveraging regulatory information. *Bioinformatics*, 35(1644-1652):865, 2018.

[23] Nicholas Mancuso, Huwenbo Shi, Pagé Goddard, Gleb Kichaev, Alexander Gusev, and Bogdan Pasaniuc. Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. *The American Journal of Human Genetics*, 100(3):473–487, 2017.

[24] Luke J O'Connor, Alexander Gusev, Xuanyao Liu, Po-Ru Loh, Hilary K Finucane, and Alkes L Price. Estimating the proportion of disease heritability mediated by gene expression levels. *BioRxiv*, page 118018, 2017.

[25] Daniel J Schaid, Wenan Chen, and Nicholas B Larson. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, 19(8):491, 2018.

[26] Yiming Hu, Mo Li, Qiongshi Lu, Haoyi Weng, Jiawei Wang, Seyedeh M. Zekavat, Zhaolong Yu, Boyang Li, Jianlei Gu, Sydney Muchnik, Yu Shi, Brian W. Kunkle, Shubhabrata Mukherjee, Pradeep Natarajan, Adam Naj, Amanda Kuzma, Yi Zhao, Paul K. Crane, Hui Lu, Hongyu Zhao, and Alzheimer's Disease Genetics Consortium. A statistical framework for cross-tissue transcriptome-wide association analysis. *Nature Genetics*, 51(3):568–576, 2019.

[27] Michael Wainberg, Nasa Sinnott-Armstrong, Nicholas Mancuso, Alvaro N. Barbeira, David A. Knowles, David Golan, Raili Ermel, Arno Ruusalepp, Thomas Quertermous, Ke Hao, Johan L. M. Björkegren, Hae Kyung Im, Bogdan Pasaniuc, Manuel A. Rivas, and Anshul Kundaje. Opportunities and challenges for transcriptome-wide association studies. *Nature Genetics*, 51(4):592–599, 2019.

[28] Kanix Wang, Hallie Gaitsch, Hoifung Poon, Nancy J Cox, and Andrey Rzhetsky. Classification of common human diseases derived from shared genetic and environmental determinants. *Nature genetics*, 49(9):1319, 2017.

[29] Chirag M Lakhani, Braden T Tierney, Arjun K Manrai, Jian Yang, Peter M Visscher, and Chirag J Patel. Repurposing large health insurance claims data to estimate genetic and environmental contributions in 560 phenotypes. *Nature genetics*, 51(2):327, 2019.

[30] Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, Benjamin M Neale, Schizophrenia Working Group

of the Psychiatric Genomics Consortium, et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*, 47(3):291, 2015.

[31] Heather E Wheeler, Kaanan P Shah, Jonathon Brenner, Tzintzuni Garcia, Keston Aquino-Michaels, Nancy J Cox, Dan L Nicolae, Hae Kyung Im, GTEx Consortium, et al. Survey of the heritability and sparse architecture of gene expression traits across human tissues. *PLoS genetics*, 12(11):e1006423, 2016.

[32] Chiara Sabatti, Anna-Liisa Hartikainen, Anneli Pouta, Samuli Ripatti, Jae Brodsky, Chris G Jones, Noah A Zaitlen, Teppo Varilo, Marika Kaakinen, Ulla Sovio, et al. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature genetics*, 41(1):35, 2009.

[33] Wellcome Trust Case Control Consortium et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661, 2007.

[34] John M Dietschy, Stephen D Turley, and David K Spady. Role of liver in the maintenance of cholesterol and low density lipoprotein homeostasis in different animal species, including humans. *Journal of lipid research*, 34(10):1637–1659, 1993.

[35] Petri T. Kovanen, Michael Scott Brown, and Joseph L Goldstein. Increased binding of low density lipoprotein to liver membranes from rats treated with 17 alpha-ethinyl estradiol. *The Journal of biological chemistry*, 254 22:11367–73, 1979.

[36] Tao Feng and Xiaofeng Zhu. Genome-wide searching of rare genetic variants in WTCCC data. *Human genetics*, 128(3):269–280, 2010.

[37] Xiang Zhou. A unified framework for variance component estimation with summary statistics in genome-wide association studies. *The annals of applied statistics*, 11(4):2027, 2017.

[38] Benjamin B Sun, Joseph C Maranville, James E Peters, David Stacey, James R Staley, James Blackshaw, Stephen Burgess, Tao Jiang, Ellie Paige, Praveen Surendran, et al. Genomic atlas of the human plasma proteome. *Nature*, 558(7708):73, 2018.

[39] Katherine E. Cole, Christine A. Strick, Timothy J. Paradis, Kevin T. Ogborne, Marcel Loetscher, Ronald P. Gladue, Wen Lin, James G. Boyd, Bernhard Moser, Douglas E. Wood, Barbara G. Sahagan, and Kuldeep Neote. Interferon–inducible T cell alpha chemoattractant (I-TAC): A novel non-ELR CXC chemokine with potent activity on activated T cells through selective high affinity binding to CXCR3. *Journal of Experimental Medicine*, 187(12):2009–2021, 1998.

[40] Johannes Kettunen, Ayşe Demirkan, Peter Würtz, Harmen HM Draisma, Toomas Haller, Rajesh Rawal, Anika Vaarhorst, Antti J Kangas, Leo-Pekka Lyytikäinen, Matti Pirinen, et al. Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nature communications*, 7:11122, 2016.

[41] Cross Disorder Group of the Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet*, 381(9875):1360–1360, 2013.

[42] S Ripke, AR Sanders, KS Kendler, DF Levinson, P Sklar, PA Holmans, DY Lin, J Duan, RA Ophoff, OA Andreassen, et al. Schizophrenia psychiatric genome-wide association study (gwas) consortium genome-wide association study identifies five new schizophrenia loci. *Nature Genetics*, 43:969–976, 2011.

[43] Stephan Ripke, Colm O'Dushlaine, Kimberly Chambert, Jennifer L Moran, Anna K Kähler, Susanne Akterin, Sarah E Bergen, Ann L Collins, James J Crowley, Menachem Fromer, et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature genetics*, 45(10):1150, 2013.

[44] Stephan Ripke, Benjamin M Neale, Aiden Corvin, James TR Walters, Kai-How Farh, Peter A Holmans, Phil Lee, Brendan Bulik-Sullivan, David A Collier, Hailiang Huang, et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421, 2014.

[45] Chen Yao, Roby Joehanes, Andrew D Johnson, Tianxiao Huan, Chunyu Liu, Jane E Freedman, Peter J Munson, David E Hill, Marc Vidal, and Daniel Levy. Dynamic role of trans regulation of gene expression in relation to complex traits. *The American Journal of Human Genetics*, 100(4):571–580, 2017.

[46] Oscar Franzén, Raili Ermel, Ariella Cohain, Nicholas K Akers, Antonio Di Narzo, Husain A Talukdar, Hassan Foroughi-Asl, Claudia Giambartolomei, John F Fullard, Katyayani Sukhavasi, et al. Cardiometabolic risk loci share downstream cis-and trans-gene regulation across tissues and diseases. *Science*, 353(6301):827–830, 2016.

[47] Center LDACCAnalysis Working Group Coordinating, Common Fund NIH, GTEx Consortium, Statistical Methods groupsAnalysis Working Group, et al. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204, 2017.

[48] Elin Grundberg, Kerrin S Small, Åsa K Hedman, Alexandra C Nica, Alfonso Buil, Sarah Keildson, Jordana T Bell, Tsun-Po Yang, Eshwar Meduri, Amy Barrett, et al. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nature genetics*, 44(10):1084, 2012.

[49] Fred A Wright, Patrick F Sullivan, Andrew I Brooks, Fei Zou, Wei Sun, Kai Xia, Vered Madar, Rick Jansen, Wonil Chung, Yi-Hui Zhou, et al. Heritability and genomics of gene expression in peripheral blood. *Nature genetics*, 46(5):430, 2014.

[50] Walter Gilbert. Why genes in pieces? *Nature*, 271(5645):501, 1978.

[51] Arianne J Matlin, Francis Clark, and Christopher WJ Smith. Understanding alternative splicing: towards a cellular code. *Nature reviews Molecular cell biology*, 6(5):386, 2005.

[52] Yang I Li, Bryce van de Geijn, Anil Raj, David A Knowles, Allegra A Petti, David Golan, Yoav Gilad, and Jonathan K Pritchard. Rna splicing is a primary link between genetic variation and disease. *Science*, 352(6285):600–604, 2016.

[53] Atsushi Takata, Naomichi Matsumoto, and Tadafumi Kato. Genome-wide identification of splicing qtls in the human brain and their enrichment among schizophrenia-associated loci. *Nature communications*, 8:14519, 2017.

[54] Rolf I Skotheim and Matthias Nees. Alternative splicing in cancer: noise, functional, or systematic? *The international journal of biochemistry & cell biology*, 39(7-8):1432–1449, 2007.

[55] Yang I Li, David A Knowles, Jack Humphrey, Alvaro N Barbeira, Scott P Dickinson,

748 Hae Kyung Im, and Jonathan K Pritchard. Annotation-free quantification of rna splicing
749 using leafcutter. *Nature genetics*, 50(1):151, 2018.

750 [56] Maria Gutierrez-Arcelus, Halit Ongen, Tuuli Lappalainen, Stephen B Montgomery, Alfonso
751 Buil, Alisa Yurovsky, Julien Bryois, Ismael Padioleau, Luciana Romano, Alexandra
752 Planchon, et al. Tissue-specific effects of genetic and epigenetic variation on gene regulation
753 and splicing. *PLoS genetics*, 11(1):e1004958, 2015.

754 [57] Alexis Battle, Zia Khan, Sidney H Wang, Amy Mitrano, Michael J Ford, Jonathan K
755 Pritchard, and Yoav Gilad. Impact of regulatory variation from RNA to protein. *Science*,
756 347(6222):664–667, 2014.

757 [58] Jiming Jiang, Cong Li, Debashis Paul, Can Yang, Hongyu Zhao, et al. On high-dimensional
758 misspecified mixed model analysis in genome-wide association study. *The Annals of*
759 *Statistics*, 44(5):2127–2160, 2016.

760 [59] Chuanhai Liu, Donald B Rubin, and Ying Nian Wu. Parameter expansion to accelerate
761 EM: the PX-EM algorithm. *Biometrika*, 85(4):755–770, 1998.

762 [60] Hua Zhou, Liuyi Hu, Jin Zhou, and Kenneth Lange. MM algorithms for variance compo-
763 nents models. *Journal of Computational and Graphical Statistics*, pages 1–12, 2019.

764 [61] M. H. Quenouille. Notes on bias in estimation. *Biometrika*, 43(3/4):353–360, 1956.

765 [62] Andrew R Wood, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson,
766 Audrey Y Chu, Karol Estrada, Jian'an Luan, Zoltán Kutalik, et al. Defining the role
767 of common variation in the genomic and biological architecture of adult human height.
768 *Nature genetics*, 46(11):1173, 2014.

769 [63] Thomas W Winkler, Anne E Justice, Mariaelisa Graff, Llilda Barata, Mary F Feitosa,
770 Su Chu, Jacek Czajkowski, Tõnu Esko, Tove Fall, Tuomas O Kilpeläinen, et al. The
771 influence of age and sex on genetic associations with adult body size and shape: a large-scale
772 genome-wide interaction study. *PLoS genetics*, 11(10):e1005378, 2015.