Quantifying the impact of genetically regulated expression on complex traits and diseases

Mingxuan Cai[1], Lin 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

About 90% of risk variants identified from genome-wide association studies (GWAS) are located in non-coding regions, highlighting the regulatory role of genetic variants. We propose a unified statistical framework, IGREX, for quantifying the impact of genetically regulated expression (GREX). This is achieved by estimating proportion of phenotypic variations that can be explained by the GREX component. IGREX only requires summary-level GWAS data and a gene expression reference panel as input. In real data analysis, using 48 tissues from the GTEx project as the reference panel, we applied IGREX to a wide spectrum of phenotypes in GWAS, and observed a significant proportion of phenotypic variations could be attributed to the GREX component. In particular, the results given by IGREX revealed tissue-across and tissue-specific patterns of the GREX effects. We also observed strong association between GREX effect and immune-related proteins, further supporting the relevance between GREX and the immune processes.

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

Over the last decade, genome-wide association studies(GWASs) have successfully identified about 90,000 significant associations (p-value $< 5 \times 10^{-8}$) between single-nucleotide polymorphisms (SNPs) and a wide range of complex traits/diseases (http://www.ebi.ac.uk/gwas/). Nevertheless, more than 90% of identified risk variants are located in non-coding regions [1], leading to difficulties in understanding the biological basis of GWAS findings. Increasing evidence [2, 3, 4, 5, 6, 7, 8] suggests that the path from genotypes to phenotypes involves gene regulatory mechanisms. For example, a study of 18 complex traits revealed significant enrichment for expression quantitative trait loci (eQTLs) in 11% of 729 tissue-trait pairs [9], implying the pervasive involvement of regulation effects in a wide spectrum of human traits. These observations lead to a scientific hypothesis that a vast proportion of genetic variants affect phenotypes by regulating the gene expression levels. To test this hypothesis, there is a need to comprehensively characterize the role of genetically regulated gene expression (GREX) in human genetics.

Fortunately, the advent of cellular-level data generated by genomic consortia provides an unprecedented chance to study the behavior of GREX effects. For example, the current 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 generated by Illumina Sequencing platforms [10], allowing for tissue-specific analysis. Multiple blood eQTL resources comprising thousands of individuals are made available for open access [11, 12]; other ongoing projects such as Genetics of DNA Methylation Consortium (GoDMC) and eQTLGen consortium are collecting expression data with sample size larger than $10,000$ [13], serving as promising resources for comprehensive analysis.

The availability of these data sets along with GWAS data enables an integrative framework for studying the GREX effects: the gene expressions of the GWAS cohort can be first 'imputed' based on statistical models fitted using a reference panel (e.g. GTEx) and then related to phenotypes [14, 15, 16, 17, 18, 19, 20, 21, 22]. This framework enjoys several benefits. First, it does not require the availability of gene expression information for GWAS data, which makes it applicable to a wide spectrum of phenotypes. Second, the prediction process naturally filters out the environmental noise and confounding variations that are ubiquitous in gene expression measurement, allowing the analysis to be focused on GREX effects. Third, the reverse influence on gene expression caused by phenotypic variation is eliminated. However, the DNA variations (i.e., SNPs) and gene expression available from the reference panel (e.g., GTEx) are often collected from non-diseased individuals for general use. Therefore, the integrative analysis of general-purposed expression data with GWAS data of a specific phenotype depends on an assumption: there exists a steady-state component in gene expression regulated by genetic variants, and the variation of this steady-state component can further induce phenotypice variations. Based on this assumption, multiple statistical models have been proposed to test the association between a given phenotype and the 'imputed' gene expression [8, 23]. Examples include PrediXcan [14], TWAS [15], FOCUS [17], MetaXcan [19] and CoMM [20].

While all the above methods can localize gene-trait associations based on the predicted gene expression, how much of the variance of a phenotype can be attributed to GREX remains unkonwn. As heritability of a phenotype that is defined as the proportion of phenotypic variance explained by DNA variations is often used to quantify the overall genetic effects, it is of great interest to characterize the impact of gene regulation on phenotypic variation from a global perspective. For example, how much of the phenotypic variations at the cellular level (e.g., glucose) and the organismal level (e.g., height) can be attributed to GREX? Are there any cross-tissue patterns or tissue-specific characteristics of GREX in different levels of phenotypes? To the best of our knowledge, there are two literatures that have attempted to address part of these problems [21, 22]. The first method (RhoGE) [21] estimates the proportion of phenotypic variation explained by GREX based on the idea of linkage-disequilibrium (LD) score regression (LDSC) [24]. Since it ignores the uncertainty in predicting gene expression, the proportion of variance explained by GREX could be substantially under-estimated. Another method, known as gene expression co-score regression (GECS) [22], have very stringent requirements that the analyzed SNPs are not in LD to ensure unbiasedness, which greatly limits its application in real data analysis.

In this article, we propose a unified framework, named IGREX, for quantifying the impact of genetically regulated expression, while accounting for uncertainty in predicted gene expression under weak signal. IGREX only requires summary-level GWAS data as its input, greatly enhancing the applicability of the model to a wide range of phenotypes. We investigated the performance of IGREX with comprehensive simulation, which highlights the importance of accounting for uncertainty. Then, using 48 tissues from the GTEx project as the reference panel, we applied IGREX to both individual-level and summary-level GWAS data sets comprised of various cellular and organismal phenotypes. Our results provide new biological insights regarding the function of gene expression in the genetic architecture of complex traits. We also demonstrate the reproducibility using independent datasets.

Results

Method overview. IGREX is a two-stage method that first evaluates the posterior distribution of GREX effects from a gene expression reference panel and then estimates the proportion of variance explained by GREX using the 'predicted' gene expression of GWAS data. It can be applied to both individual-level (IGREX-i) and summary-level (IGREX-s) GWAS data. Here, we briefly introduce the statistical formulation of IGREX-i and leave the technical details in the Methods Section.

Suppose we have the reference eQTL data set $\mathcal{D}_r$ and individual-level GWAS data set $\mathcal{D}_i$: $\mathcal{D}_r = \{\mathbf{Y}, \mathbf{X}_r\}$ is comprised of $n_r \times G$ gene expression matrix $\mathbf{Y}$ and $n_r \times M$ 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; $\mathcal{D}_i = \{\mathbf{t}, \mathbf{X}\}$ 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 GWAS sample size. We first link each gene expression to its local SNPs by the following linear model:

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

where the subscript $_g$ represents the g-th gene, $\boldsymbol{\beta}_g \sim \mathcal{N}(\mathbf{0}, \sigma^2_{\beta_g}\mathbf{I}_{M_g})$ is the genetic effects of M_g local SNP, $\mathbf{e}_{r,g} \sim \mathcal{N}(0, \sigma^2_{r,g}\mathbf{I}_{n_r})$ is the independent noise, and the local SNPs are defined as SNPs around the target gene (e.g. ±1 Mb around the transcription start site). Because we are interested in the steady-state component of gene expression regulated by genetic variants, $\boldsymbol{\beta}_g$ is assumed to be the same for individuals in both $\mathcal{D}_r$ and $\mathcal{D}_i$. Consequently, the GREX of individuals in GWAS data can be evaluated by $\mathbf{X}_g\boldsymbol{\beta}_g$. Next, we assume that the genetic effects on $\mathbf{t}$ can be decomposed into two parts, i.e. the genetic effect through GREX and the genetic effect through alternative ways:

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

where $\alpha_g \sim \mathcal{N}(\mathbf{0}, \sigma^2_\alpha)$ is the effect size of $\mathbf{X}_g\boldsymbol{\beta}_g$ on $\mathbf{t}$, $\boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{0}, \sigma^2_\gamma\mathbf{I}_M)$ is the alternative genetic effects vector of length M and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2_\epsilon\mathbf{I}_n)$ is the independent noise. 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 component on $\mathbf{t}$, respectively. Thus, given a genotype vector $\mathbf{x}$ and a phenotype t, the impact of GREX can be quantified by the proportion of variance explained by the GREX component: $\text{PVE}_{\text{GREX}} = \frac{\text{Var}(\sum_{g=1}^{G} \alpha_g \mathbf{x}_g^T \boldsymbol{\beta}_g)}{\text{Var}(t)}$. To estimate this quantity, the inference procedure of IGREX is decomposed into two stages. At the first stage, we estimate $\sigma_{\beta_g}^2$ and $\sigma_{r,g}^2$ using a fast 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. At the second stage, by treating the posterior obtained in the stage one 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 given in the Methods Section) by

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

where the parameters with hat represent their corresponding estimates. As we can observe from the above estimation, the substitution of posterior $\boldsymbol{\beta}_g | \mathbf{y}_g, \mathbf{X}_{r,g}$ naturally results in the adjustment of uncertainty associated with $\boldsymbol{\beta}_g$, which is quantified by the posterior variance $\boldsymbol{\Sigma}_g$. Besides the point estimate, the standard error of $\widehat{\text{PVE}}_{\text{GREX}}$ can be obtained by the delta method (see Supplementary Note).

In real applications, individual-level GWAS data may not be accessible. Hence, we have further developed IGREX-s for handling summary-level GWAS data (See Methods). Based on the MoM, IGREX-s can well approximate IGREX-i while requiring only the z-scores of SNPs and a reference genotype matrix $\tilde{\mathbf{X}} \in \mathbb{R}^{m \times M}$ of a similar LD pattern with $\mathbf{X}$, where the sample size m can be as small as a few hundreds. In practice, $\tilde{\mathbf{X}}$ can be a random subsample of individuals in $\mathbf{X}$ or a reference panel of the same ethnic origin. 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 $\hat{\mathbf{R}}_g = \tilde{\mathbf{X}}_g^T \tilde{\mathbf{X}}_g / m$ is the estimated LD matrix associated with the g-th gene and $\tilde{\mathbf{X}}_g$ is the corresponding columns of $\tilde{\mathbf{X}}$. Our method IGREX also allows to incorporate sex, age and principal components as covariates to minimize the influence of confounding factors (See details in Supplementary Note).

Simulation. We conducted comprehensive 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 SNPs in 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 $\text{PVE}_{\text{Alternative}} = \frac{\text{Var}(\mathbf{x}_g^T \boldsymbol{\gamma})}{\text{Var}(t)} = 0.3$ (results for other scenarios are shown in Supplementary Figs. 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, the gene expression $\mathbf{y}_g$ and phenotype $\mathbf{t}$ were simulated following the generative models (1) and (2). To assess IGREX-s, we calculated the z-score of each SNP and randomly subsetted $m = 500$ rows 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 situation that 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 specified posterior variance $\boldsymbol{\Sigma}_g = \mathbf{0}$ at the second stage, and then conducted REML and MoM as before. We denote these methods as REML_0 and MoM_0. The simulation results summarized in Fig. 1a show that both PVE_{GREX} and $\text{PVE}_{\text{Alternative}}$ are accurately estimated using REML-based IGREX-i under all circumstances. The MoM-based IGREX-i slightly underestimates PVE_{GREX} when both sample size n_r and signal strength PVE_y are very small, but steadily converges to the same performance of REML as either n_r or PVE_y increases. For all settings, IGREX-s well approximates MoM, producing almost identical estimations. In contrast, as both REML_0 and MoM_0 do not account for uncertainty arising in the first stage, they have poor estimation performance even with very large sample size and very strong signal in our simulation study.

Next, we conducted simulations to evaluate the situation that the IGREX model was mis-specified. First, we considered the situation where genetic effects $\boldsymbol{\beta}_g$ and $\boldsymbol{\alpha}$ were sparse. To evaluate the influence of different sparsity patterns on our method, we first fixed the proportion of non-zero effects $\pi_\alpha = (\text{NO. of nonzero entries in } \boldsymbol{\alpha})/G$ at 0.2 and varied $\pi_\beta = (\text{NO. of nonzero entries in } \boldsymbol{\beta}_g)/M_g$ at $\{0.2, 0.5, 0.8\}$, then we fixed $\pi_\beta = 0.2$ and varied π_α at $\{0.2, 0.5, 0.8\}$. As shown in Figs. 1b-c, all three methods of IGREX produce accurate estimates in the presence of sparse genetic effects, imlying the robustness of IGREX to model misspecification. Besides, the estimation performances are not influenced by the degree of sparsity. Second, we investigated the influence of LD pattern by setting ρ varied at $\{0.1, 0.3, 0.5, 0.8\}$. From Fig. 1d., we can observe that IGREX produces accurate estimations despite the magnitude of LD. On the other hand, REML_0 and MoM_0 consistently underestimate PVE_{GREX} as a result of ignoring estimation uncertainty.

In addition, we made comparisons between IGREX and the method proposed in RhoGE [21], which provides an LDSC-based approach for estimating PVE_{GREX}. However, this model does not adjust for estimation uncertainty. The results are shown in Fig. 1e. As we can expect, the pattern of IGREX is consistent with that in Fig. 1a. On the other hand, RhoGE substantially underestimates PVE_{GREX} for most cases despite the reference sample size. It only achieves the same accuracy as IGREX when the signal strength $\text{PVE}_y \geq 0.9$, which is not realistic for eQTL data.

Real data application on individual-level GWAS data. We applied our approaches to two individual-level GWAS datasets, the Northern Finland Birth Cohorts program 1966 (NFBC) [25] and the Wellcome Trust Case Control Consortium (WTCCC) [26], with eQTL data from 48 human tissues in GTEx project. The details of the datasets and the data preprocessing procedures are described in the Methods Section.

After sample quality control of the NFBC dataset, we have ten quantitative traits from $5,123$ individuals with $309,245$ SNPs. We first estimated the heritabilities of the ten traits and then exluded four traits of very small heritabilities including body mass index (BMI), C-reactive protein (CRP), insulin and diastolic blood pressure (DiaBP) and restricted our analysis within the remaining six traits with high heritabilities: high-density lipoprotein cholesterol (HDL), low-density lipoprotein cholesterol (LDL), triglycerides (TG), total cholesterol (TC) and systolic blood pressure (SysBP). Figs. 2a-b show the $\widehat{\text{PVE}}_{\text{GREX}}$ of the six traits on 48 GTEx tissues obtained using REML and MoM, respectively. We can observe that the two methods produced quite similar estimates in most of the tissues. Although the REML estimates are slightly higher than the MoM estimates in some cases, the discrepancy is not significant. Of the outcomes shown in the figures, LDL and TC deserve special attention: both of them have a large proportion of variations can be explained by the GREX component in liver. According to the REML approach (Fig. 2a), the $\widehat{\text{PVE}}_{\text{GREX}}$ for LDL in liver is as high as 14.3% (with

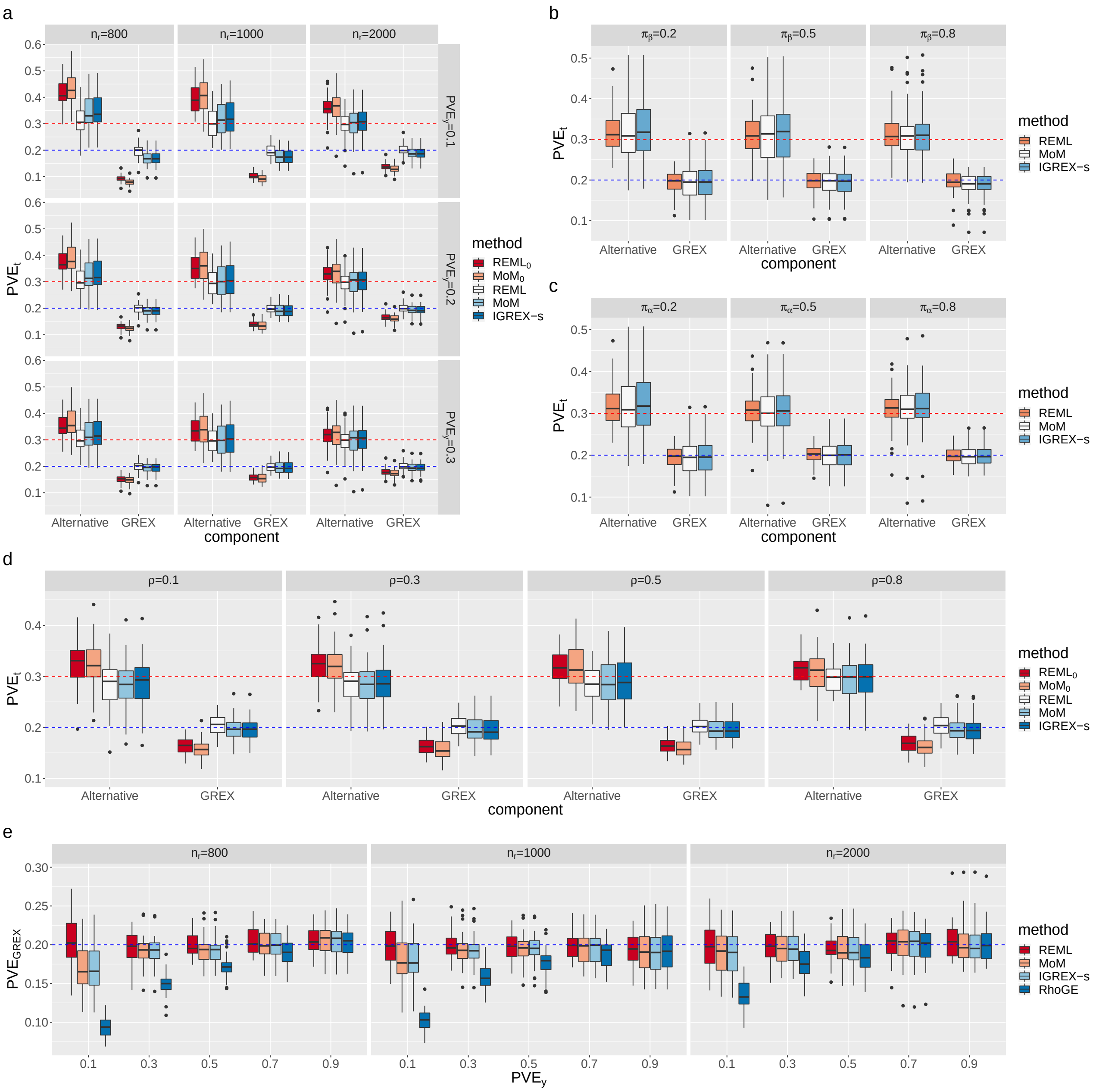

Figure 1: Simulation studies to compare estimation accuracies of IGREX with other methods. REML and MoM in the legend are abbreviations of methods on which IGREX-i is based. The blue and red dashed lines represent the true values of $\mathrm{PVE}_{\mathrm{GREX}}$ and $\mathrm{PVE}_{\mathrm{Alternative}}$, respectively. We conducted 30 replications and generated box plots for analyzing 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 is varied at $\{0.2, 0.5, 0.8\}$; (**d**) the three models of IGREX, REML_0 and MoM_0 when ρ is varied at $\{0.1, 0.3, 0.5, 0.8\}$; (**e**) the three models of IGREX and RhoGE when n_r is varied at $\{800, 1000, 2000\}$.

standard error 2.6%), capturing 52.6% of total heritability defined as $\mathrm{PVE}_{\mathrm{GREX}}/h^2$; TC also has high $\widehat{\mathrm{PVE}}_{\mathrm{GREX}} = 13.7\%$ (with standard error 2.5%), which captures 79.4% of total heritability (see Supplementary Fig. 6). These resuts are verified by the MoM (Fig. 2b). In fact, LDL synthesized in liver is an important lipoprotein particle for transporting cholesterol in the blood. Our finding suggests that the genetic architecture of LDL synthesis in liver extensively involves the gene regulation mechanism, which provides a new insight of this biological process. Additionally, we analyzed the impact of ignoring the uncertainty (with the complete results given in the Supplementary Fig. 5). By observing the slopes of fitted regression lines in Figs. 2c-d, it is clear that half of $\widehat{\text{PVE}}_{\text{GREX}}$ is lost because of ignoring the uncertainty. To evaluate the performance of IGREX-s, we also generated z-scores from NFBC data and applied IGREX-s based on the summary statistics. The resulting estimates are then compared to MoM estimates in Fig. 2e. For all six traits, IGREX-s estimates well approximate the MoM estimates using the individual level data, which is consistent with our simulation result.

Now we investigate the role of GREX in complex human traits and diseases, using the WTCCC dataset [26]. We applied IGREX to estimate the PVE_{GREX} of seven diseases 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). For diseases, we analyzed percentage of heritability explained by GREX ($\text{PVE}_{\text{GREX}}/h^2$) to avoid the influence of ascertainment bias. The estimated $\text{PVE}_{\text{GREX}}/h^2$ obtained by REML are shown in Supplementary Fig. 8. The results show that all the diseases have moderate to high estimated $\text{PVE}_{\text{GREX}}/h^2$ in some subsets of the tissues. The top $\text{PVE}_{\text{GREX}}/h^2$'s 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. Two diseaes that deserve special attention are RA and T1D, whose average $\text{PVE}_{\text{GREX}}/h^2$ estimates are as high as 34.1% and 71.2%, respectively. It is well known that RA and T1D are both autoimmune diseases whose strong associations with major histocompatibility complex (MHC) region have been well established in previous studies [26, 27]. To have a better unserstanding of our observations, we compared the estimated $\text{PVE}_{\text{GREX}}/h^2$ with those obtained by removing the MHC region (results are given in the Supplementary Fig. 9). The distributions of $\text{PVE}_{\text{GREX}}/h^2$ estimates are shown in Fig. 3a. We observed a substantial downward shift of the distribution after removing the MHC region in RA and T1D: 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. In addition, the tissue-specific comparisons shown in Fig. 3b reveal an extensive reduction of PVE_{GREX} in all tissues for T1D and RA, while such change does

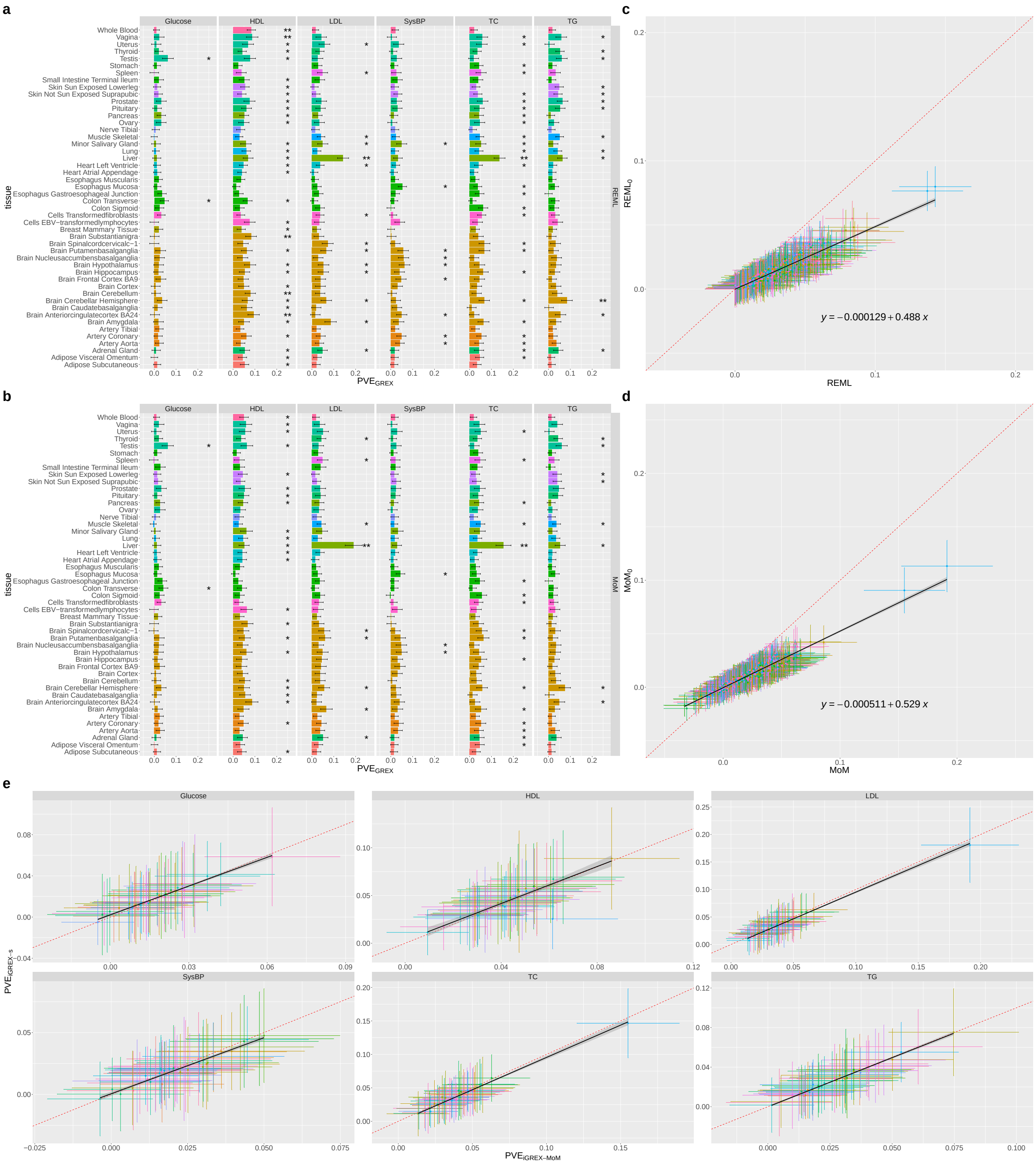

Figure 2: Tissue-specific $\widehat{\mathrm{PVE}}_{\mathrm{GREX}}$ of the six traits from NFBC data set. (**a**-**b**) $\widehat{\mathrm{PVE}}_{\mathrm{GREX}}$ obtained by REML and MoM. Tissues are colored according to their categories. The number of asterisks represents the significance level: p-value< 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 $\widehat{\mathrm{PVE}}_{\mathrm{GREX}}$ generated by IGREX-s against those generated by MoM for all 48 tissues in one of the six traits.

not appear in other traits. This finding implies that the steady-state gene regulation process pervasively participates in the immune functionality of the MHC region for RA and T1D. We note that this discovery reveals a potential rationale behind the etiologies of the MHC-related autoimmune diseases such as RA and T1D.

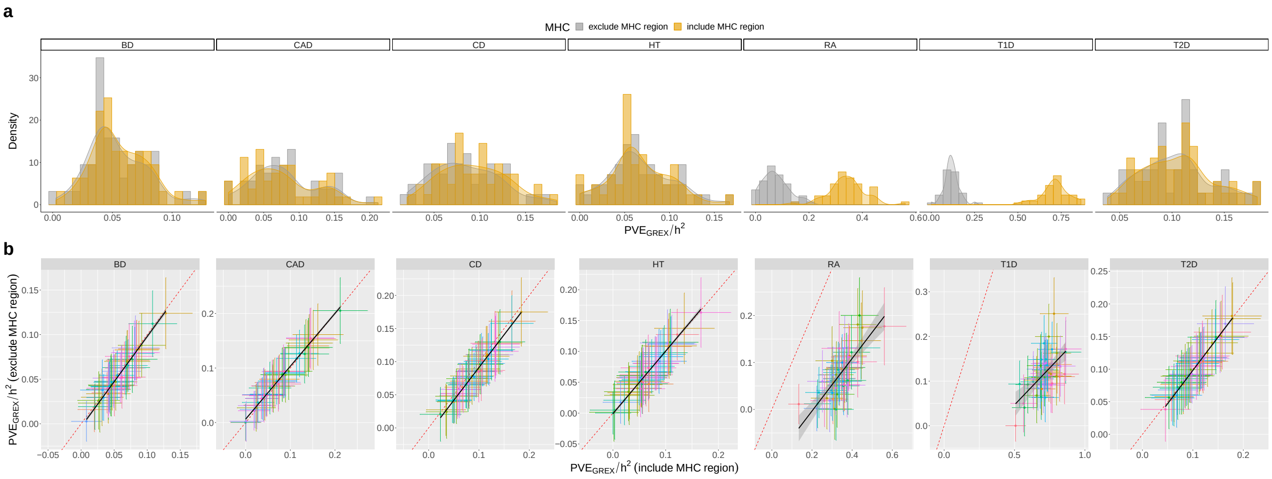

Figure 3: Percentage of heritability explained by GREX ($\text{PVE}_{\text{GREX}}/h^2$) of the seven traits from WTCCC data. (**a**) The distributions of estimated $\text{PVE}_{\text{GREX}}/h^2$ across 48 GTEx tissues. (**b**) Tissue-specific comparisons of $\text{PVE}_{\text{GREX}}/h^2$ estimated by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

Analysis results using the summary-level GWAS data. Since the summary statistics are much easier to access than the individual-level GWAS data, we are allowed to analyze a wider spectrum of phenotypes using IGREX-s. To study the pattern of GREX impact in multiple levels of human traits, we applied our method to proteins, metabolites as well as high-level complex phenotypes such as schizophrenia, height and waist-to-hip ratio adjusted BMI (WHRadjBMI). In the following analyses, we used the genotypes of 379 individuals of European ancestry from the 1,000 genome project as the reference panel.

Firstly, we quantified PVE_{GREX} in the protein level using the summary statistics from a plasma protein quantitative trait loci (pQTL) study [28]. Fig .4a shows the heritability distributions of all $3,283$ proteins in the dataset estimated using MQS [29]. Protens with insignificant heritabilities were excluded and 249 remained for inclusion in our analysis (See Supplementray Table 3). The outcomes show that the heritabilities estimated by IGREX ($\hat{h}_t^2 = \widehat{\text{PVE}}_{\text{GREX}} + \widehat{\text{PVE}}_{\text{Alternative}}$) are strongly consistent to those estimated by MQS (See Supplementary Fig. 10). The p-values for testing the significance of GREX effects on these proteins are shown by the QQ-plot in Fig. 4b, where the tissues were categorized into 16 groups (tissue-specific QQ-plots are given in Supplementary Fig. 11, Manhattan plot and heatmap of all tissue-protein pairs are given in Supplementary Figs. 12-13). As we can observe, the GREX components have significant contribution in many tissue-protein pairs. In particular, 9 out of 249 proteins have significant GREX components in at least one tissue at 0.05 level using Bonferroni correction. As illustrated in Fig. 4d-e, the contribution of GREX components shows heterogeneous across-tissue patterns in the nine proteins: CD96, DEFB119, MICB and PDE4D have high $\widehat{\mathrm{PVE}}_{\mathrm{GREX}}$ regardless of the tissue type; on the other hand, significant GREX impacts for CFB, CXCL11, EVI2B, IDUA and LRPAP1 exist only in some subsets of tissues. We found that these tissue-specific patterns are consistent with the protein functions. For example, the CFB protein, which is implicated in the growth of preactivated B-lymphocytes, is found most associated with GREX in EBV-transformed lymphocytes ($\widehat{\mathrm{PVE}}_{\mathrm{GREX}} = 18.7\%$); besides, the CXCL11 with its highest $\widehat{\mathrm{PVE}}_{\mathrm{GREX}} = 16.6\%$ in pancreas is known to have a high expression level in pancreas. We also noted that 6 out of the 9 proteins were immune-related, suggesting that the genetics of immune process could be more related to gene regulation effects.

Besides the proteins, metabolic phenotypes also serve as an important intermediate for high level biological processes. To understand the role of gene regulation in the genetics of such traits, we applied IGREX-s to a summary level data set of circulating metabolites [30], which was comprised of meta-analysis of 123 metabolites. We focused our analysis on the 21 metabolites that were highly heritable (estimated $h^2 > 10\%$) including glycine, various features of HDL, LDL, very low-density lipoprotein (VLDL) and intermediate-density lipoprotein (IDL) and other polyunsaturated fatty acids (otPUFA). The distributions of $\mathrm{PVE}_{\mathrm{GREX}}/h^2$ estimates in different tissues are given in Fig. 5a. The median values of percentage estimates are higher than 10% in 6 out of the 48 tissues and only higher than 15% in liver and spinal cord (cervical c-1). According to the estimated values shown in the heat map of Fig. 5b, we can see that the features associated with IDL, LDL and VLDL have estimated $\mathrm{PVE}_{\mathrm{GREX}}/h^2$ around 20% in liver and 16% in spinal cord, suggesting that they are more related to the GREX effects in these two tissues. On the contrary, there is no signal of GREX components detected under the nominal level 0.05 in any GTEx tissue for HDL associated features and glycine.

We also applied IGREX-s to the summary data of complex human traits. Here we analyzed schizophrenia (SCZ), height and WHRadjBMI. We considered four datasets of schizophrenia with increasing sample sizes: SCZ subset [31], SCZ1 [32], SCZ1+Sweden (SCZ1Swe)[33] and

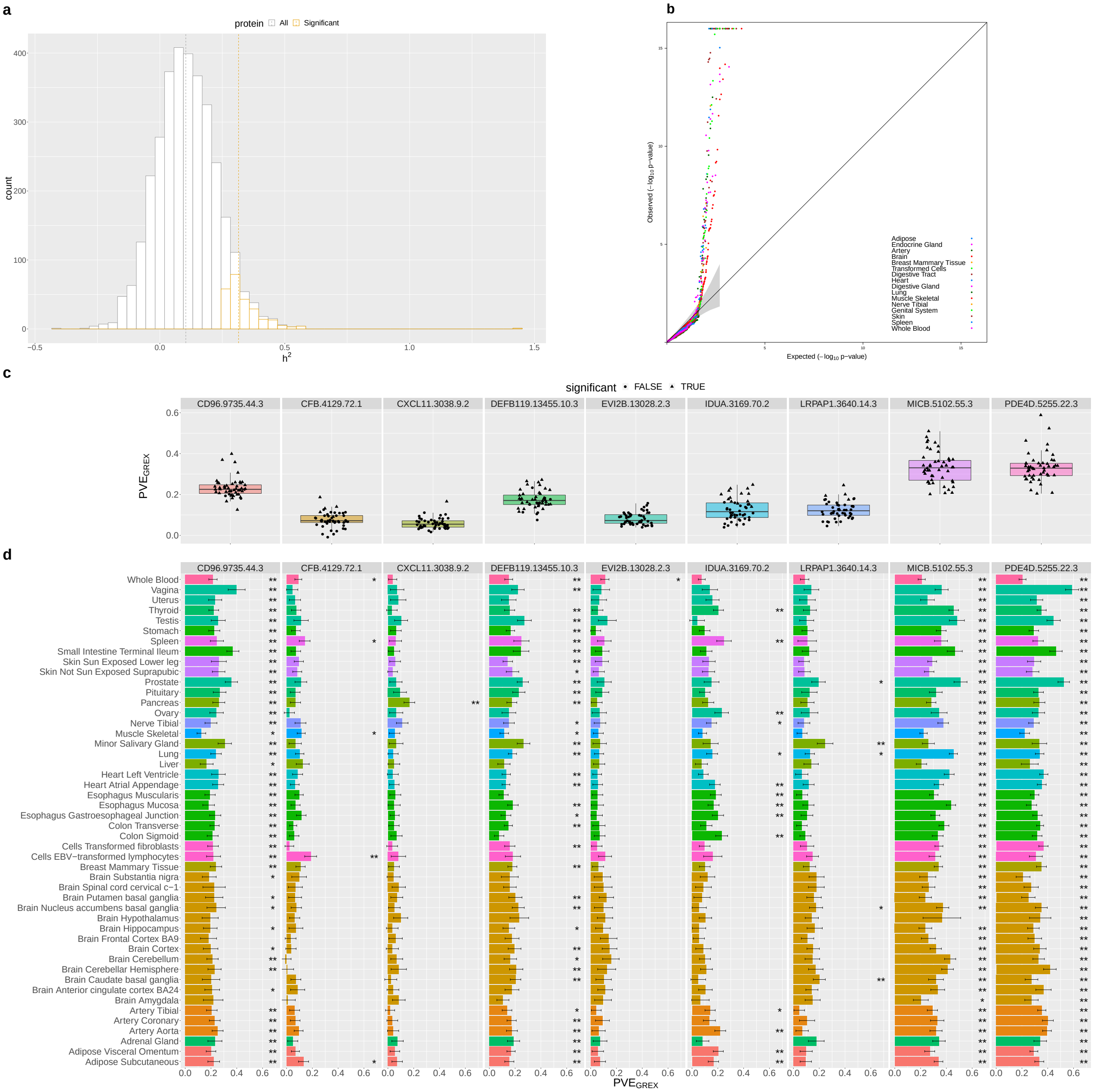

Figure 4: Analysis of plasma pQTL summary statistics. (**a**) The distribution of 3,283 proteins estimated using MQS. The whole study is colored in grey, while the 249 proteins with significant heritabilities are colored in yellow. Dashed lines represent the means of corresponding distributions. (**b**) 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 $**$.

SCZ2 [34]. We found that the estimated $\text{PVE}_{\text{GREX}}/h^2$ in all four SCZ datasets have higher values in brain than in other tissues (Fig. 6b), implying stronger GREX effects for SCZ in

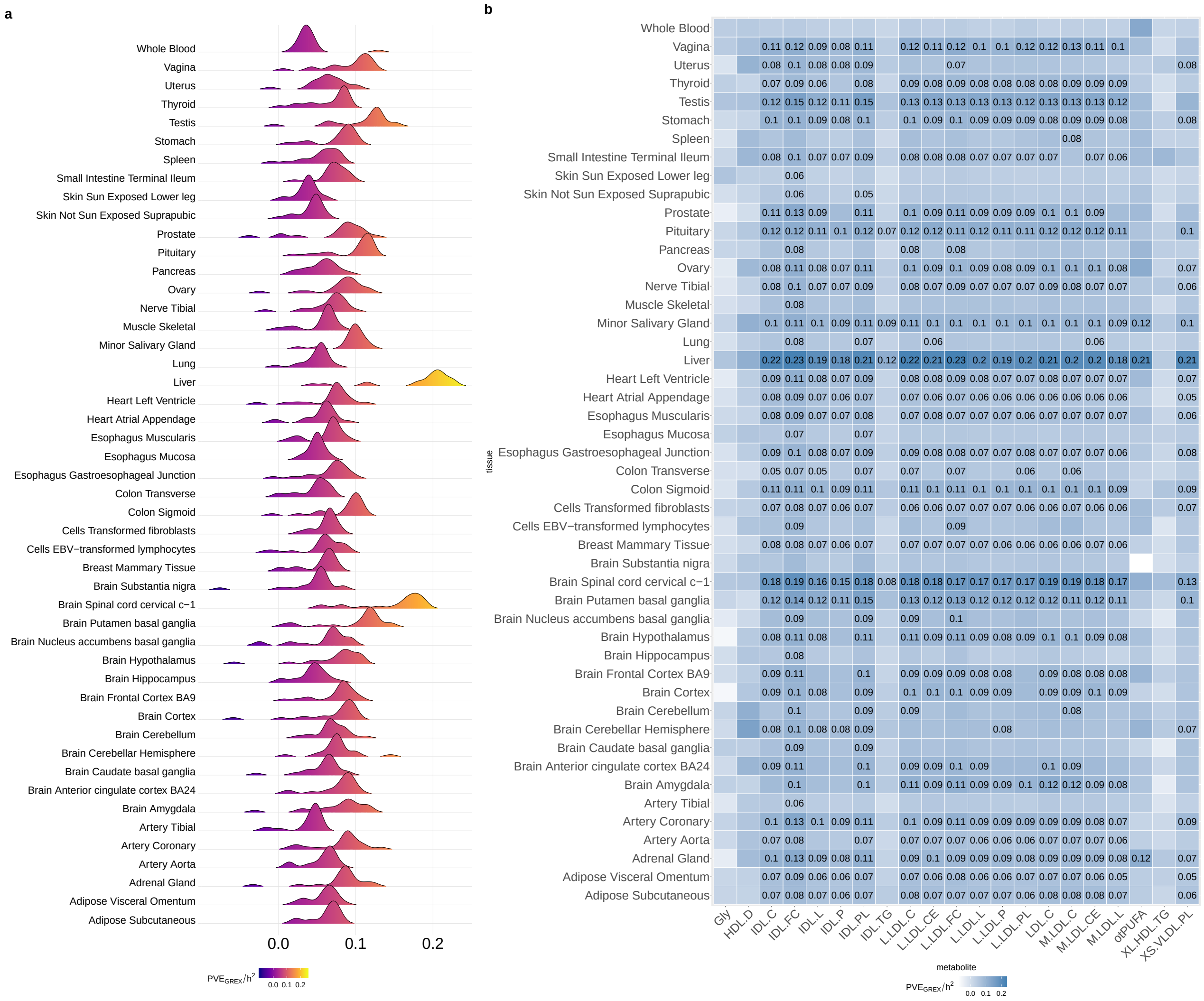

Figure 5: $\widehat{\mathrm{PVE}}_{\mathrm{GREX}}/h^2$ for 21 circulating metabolites. (**a**) The distributions of estimated $\mathrm{PVE}_{\mathrm{GREX}}/h^2$ in different tissues. (**b**) The heat map of estimated $\mathrm{PVE}_{\mathrm{GREX}}/h^2$. Entries that are significant at nominal level (0.05) are labeled with their estimate values.

brain. Besides, increasing number of tissues are found to have a significant impact of the GREX component as the sample size increases, as shown in Fig. 6a. This trend is also observed by comparing the significance levels of $\mathrm{PVE}_{\mathrm{GREX}}/h^2$ estimates in the four datasets (Supplementary Fig. 14), where the estimation accuracy increases with the sample size. For the human height and WHRadjBMI, we considered pairs of independent datasets for replication purpose: height datasets included GWAS anthropoetric 2014 (height2014) [35] and UK Biobank (UKB) summary statistics provided by Neale Lab (`http://www.nealelab.is/uk-biobank/`), WHRadjBMI datasets include summary statistics obtained by analyzing men and women, seperately [36]. By comparing the panels of Fig. 6c, we can observe that IGREX produced similar results in the two independent datasets. While the outcomes are reproducible, we noted the estimated percentages of heritability explained by GREX for all three complex traits are less than 10% (6.7% for schizophrenia, 7.1% for height and 3.7% for WHRadjBMI in the most expressed tissue. See Fig. 6c and Supplementary Fig. 15), lower than those of other phenotypes. There are two possible reasons of this observation. First, IGREX only takes account of the local genetic effects on gene expression due to the limited sample size of eQTL studies. However, the gene regulation mechanisms of some complex traits may involve distant SNPs, resulting in underestimated $\widehat{\mathrm{PVE}}_{\mathrm{GREX}}$. With a large eQTL sample size, this problem can be addressed by accounting for the regulation effects across the whole genome. Second, the genetic effects on gene expression may not be steady-state but rely on the biological status of tissues and individuals. As GTEx data serve as a general-puposed reference, the dynamics of genetically regulated gene expression may not be captured [37]. For example, the schizophrenia patients may have different gene expression patterns and mechanisms from healthy individuals in disease related tissues. Similarly, the genetic effects on gene expression associated with height may vary between adult tissues and teenage tissues. In this scenario, condition-specific gene expression data are demanded to provide more reliable estimates of $\mathrm{PVE}_{\mathrm{GREX}}$.

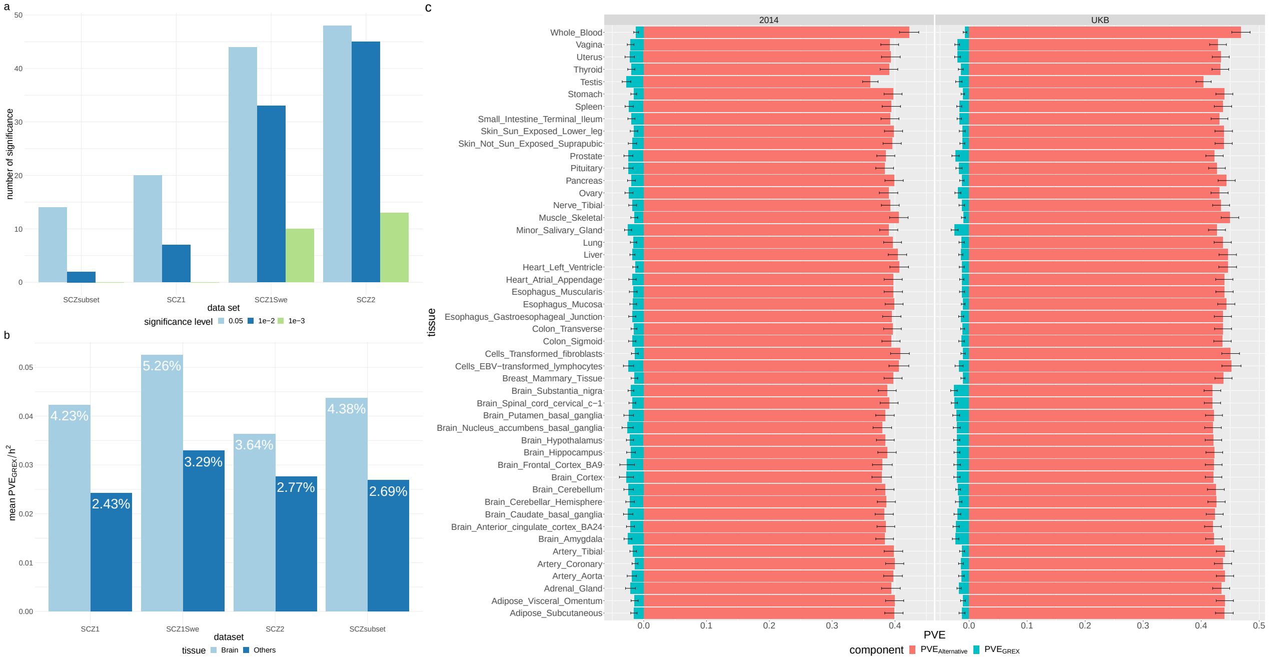

Figure 6: Analyses of complex traits: schizophrenia and height. (**a**) Number of significant GREX components revealed under different significance level for the four schizophrenia datasets. (**b**) Mean estimated percentages of heritability for schizophrenia explained by GREX in brain tissues and in other tissues. (**c**) $\widehat{\mathrm{PVE}}_{\mathrm{GREX}}$ and $\widehat{\mathrm{PVE}}_{\mathrm{Alternative}}$ of height estimated using height2014 and UKB datasets, respectively.

Discussion

Despite the great success of GWAS in the past 10 years, the biological basis of a large proportion of discovered genetic variants locating in the non-coding regions remains unknown. As cumulated evidence suggests the involvement of gene regulation mechanism for these genetic variants, there is a pressing need to characterize the role of gene regulation in the genetics of various phenotypes. By leveraging the general-purposed eQTL data (e.g. GTEx) with GWAS, our proposed method, IGREX, quantifies the impact of genetically regulated expression and provides new insights for the genetic architectures of extensive phenotypes.

IGREX is closely related to several existing methods such as TWAS [15], PrediXcan [14] and RhoGE [21]. Here we briefly discuss the relationship between IGREX and these methods. TWAS and PrediXcan can be considered within a more general MetaXcan framework that integrates eQTL information with GWAS results and identifies trait-associated genes. While both IGREX and MetaXcan 'impute' the gene expression based on eQTL reference, IGREX is distinct from MetaXcan in two perspectives:

- First, MetaXcan aims at identifying genes whose expressions are associated with phenotypes. In contrast, IGREX explores the impact of genetically regulated expression from a global perspective by quantifying the phenotypic variation that can be attributed to the GREX component.

- Second, while MetaXcan increases the power of gene-based association mapping by incorporating the eQTL information, the identified signals may not be totally attributed to GREX effects. In fact, when the signal from SNP to gene expression is weak, the posterior distribution of $\boldsymbol{\beta}_g$ will not change a lot from its prior (i.e., $\boldsymbol{\mu}_g \approx \mathbf{0}$ and $\boldsymbol{\Sigma}_g \approx \sigma^2_{\beta_g}\mathbf{I}_{M_g}$). Consequently, $\mathbf{X}_g(\boldsymbol{\mu}_g\boldsymbol{\mu}_g^T + \boldsymbol{\Sigma}_g)\mathbf{X}_g^T$ and $\mathbf{X}_g\mathbf{X}_g^T$ are numarically very close, resulting in a tagging effect between the two relatedness matrices. If the alternative genetic component is not adjusted for, the GREX effects can absorve the signals from the alternative genetic effects. This hampers MetaXcan from distinguishing the GREX effects and alternative genetic effects (See Supplementary Fig. 16). On the other hand, IGREX filters out the alternative genetic component by accounting for the alternative impact $\mathbf{X}\boldsymbol{\gamma}$ and captures the GREX signal only. This feature allows IGREX to produce results that are more biologically interpretable.

RhoGE is designed for identifying and estimating correlation between gene expression and trait. It also provides an LDSC-based approach for estimating $\mathrm{PVE}_{\mathrm{GREX}}$. Unlike IGREX, this method does not adjust for estimation uncertainty. Consequently, it significantly underestimates the $\mathrm{PVE}_{\mathrm{GREX}}$ when the signal is weak. In fact, RhoGE estimated the $\mathrm{PVE}_{\mathrm{GREX}}$ for the majority of $1,350$ tissue-trait pairs to be almost negligible (the first quantile, the median, and the third quantile are 0.00125%, 0.162% and 0.616%, respectively. See Table S9 of [21]). On the contrary, by accounting for the estimation uncertainty, IGREX can accurately estimate $\mathrm{PVE}_{\mathrm{GREX}}$ under weak signal. Through simulation studies, we have demonstrated that IGREX has better performance than RhoGE under various signal strengths.

A key assumption in applying IGREX to general-purposed eQTL data is the existence of steady-sate component in GREX, i.e., the genetic effects on gene expression $\boldsymbol{\beta}_g$ should be the same in eQTL reference and GWAS data. However, there are situations where this assumption is violated. For example, it has been observed that more gene regulatory effects of CAD-risk SNPs are identified in the disease tissues than in the healthy GTEx tissues [37]. In the presence of this dynamic component, the $\widehat{\mathrm{PVE}}_{\mathrm{GREX}}$ based on GTEx tissues may not be accurate enough, and substituting the gene expression reference by those derived from trait associated tissues is expected to produce better estimates.

In conclusion, we have presented a statistical approach, IGREX, that integrates GWAS data and eQTL reference to quantify the GREX impact in multiple levels of phenotypes. Not only does IGREX have better estimation accuracy than related methods, it also provides biological insights into the role of gene regulatory mechanisms in the genetics of various traits. Besides, IGREX enjoys a high practicality because it can be applied to both individual-level and summary-level GWAS data. We have successfully applied our method to both cellular level and organismal level traits and revealed cross-tissue and tissue-specific patterns of GREX in these traits. We have also applied IGREX to independent datasets of same traits, demonstrating the results given by our approach can be replicated.

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 some eQTL studies, 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 eQTL study, G is the number of genes and M is the number of single-neucleotide polymorphisms (SNPs). Then, suppose we have individual-level GWAS data set $\mathcal{D}_i = \{\mathbf{t}, \mathbf{X}\}$ 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 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 colums 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 confounding factors. The complete model that accounts for confounders is described in the supplementary. Now, we consider linear model (1) that links 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, $\mathbf{e}_{r,g} \sim \mathcal{N}(0, \sigma^2_{r,g}\mathbf{I}_{n_r})$ 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 genetically regulated expression (GREX) in $\mathcal{D}_i$ can be evaluated by $\mathbf{X}_g\boldsymbol{\beta}_g$. Then we assume that the pehontype $\mathbf{t}$ can be decomposed into two parts, i.e., the genetic effects through GREX and the genetic effects through alternative ways, 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 over-all impact of GREX on the phenotype and $\mathbf{X}\boldsymbol{\gamma}$ represents the alternative impact. 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:

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

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

To estimate $\mathrm{PVE}_{\mathrm{GREX}}$, 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 [38] for heritability estimation on mis-specified linear mixed model (LMM). This prior specification in (4) provides a great computational advantage as well as a stable performance for IGREX under model mis-specification, as demonstrated in the simulation study.

The proposed method for individual-level GWAS data, IGREX-i, provides a two-stage framework for estimating PVE_{GREX}. At the first stage, we estimate the parameters $\sigma^2_{\beta_g}$ and $\sigma^2_{r,g}$ in model (1) by a fast expectation-maximization (EM)-type algorithm, the parameter-expanded EM (PX-EM) algorithm [39]. 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

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

At 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 associated with $\boldsymbol{\beta}_g$ 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 $\text{Cov}(\mathbf{t}|\boldsymbol{\alpha}) = \sum_{g=1}^{G} \alpha_g^2 \mathbf{X}_g \boldsymbol{\Sigma}_g \mathbf{X}_g^T + \sigma^2_\gamma \mathbf{X}_g \mathbf{X}_g^T + \sigma^2_\epsilon \mathbf{I}_n$; 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

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

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 $\sigma^2_\gamma \mathbf{X}_g \mathbf{X}_g^T$ represents the variance explained by GREX and alternative genetic effects, respectively. Therefore, the PVE_{GREX} defined in (3) can be estimated by

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

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 approaches for estimating the parameters and $\widehat{\text{PVE}}_{\text{GREX}}$ at 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 $\mathbf{K}_\gamma = \mathbf{X}_g\mathbf{X}_g^T$. The first method is based on the method of moments (MoM), which minizes the distance between the second moment of $\mathbf{t}$ at the population level and that at the sample level $f(\boldsymbol{\psi}) = ||\mathbf{t}\mathbf{t}^T - (\sigma^2_\alpha \mathbf{K}_\alpha + \sigma^2_\gamma \mathbf{K}_\gamma + \sigma^2_\epsilon \mathbf{I}_n)||^2$. Let $\frac{\partial f(\boldsymbol{\psi})}{\partial \sigma^2_\alpha} = \frac{\partial f(\boldsymbol{\psi})}{\partial \sigma^2_\gamma} = \frac{\partial f(\boldsymbol{\psi})}{\partial \sigma^2_\epsilon} = 0$, we obtain the estimating equation

$$\mathbf{S}\boldsymbol{\psi} = \mathbf{q}, \tag{8}$$

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

The solution of Equation (8) is given by $\hat{\boldsymbol{\psi}} = \mathbf{S}^{-1}\mathbf{q}$. And $\mathrm{Cov}(\hat{\boldsymbol{\psi}}) = \mathbf{S}^{-1}\mathrm{Cov}(\mathbf{q})\mathbf{S}^{-1}$ by sandwich estimator. Then, the standard error of $\widehat{\mathrm{PVE}}_{\mathrm{GREX}}$ can be obtained by delta method (Supplementary). 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 Minorization-Maximization (MM) algorithm [40].

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 available. 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 $\mathrm{Var}(t)$. Hence, we have

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

and $\mathrm{PVE}_{\mathrm{GREX}}$ defined in (3) can be estimated by

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

where $\hat{\mathbf{R}}_g = \tilde{\mathbf{X}}_g^T\tilde{\mathbf{X}}_g/m$ is the estimated LD matrix associated with the g-th gene and $\tilde{\mathbf{X}}_g$ is the corresponding columns of some genotype matrix $\tilde{\mathbf{X}}$. In practice, $\tilde{\mathbf{X}} \in \mathbb{R}^{m\times M}$ can be the genotype matrix either from reference panel (e.g. eQTL studies such as GTEx) or the 1000 genome project. Now, we consider the method of moments 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

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

The terms on the left hand side does not involve $\mathbf{t}$ and thus can be approximated using $\tilde{\mathbf{X}}$ [29]. For example, $\frac{\mathrm{tr}(\mathbf{K}_\alpha^2) - \frac{\mathrm{tr}^2(\mathbf{K}_\alpha)}{n}}{n^2}$ can be well approximated by $\frac{\mathrm{tr}(\tilde{\mathbf{K}}_\alpha^2) - \frac{\mathrm{tr}^2(\tilde{\mathbf{K}}_\alpha)}{m}}{m^2}$, 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 block jackknife (Supplementary).

IGREX can incorporate fixed effectsto adjust 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. This data is comprised of 48 tissues 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 expressions, leaving 16, 333 $\sim$ 27, 378 genes for inclusion in our analysis. The genotype data were obtained from the third phase of the International HapMap project phase 3 (HapMap3) with 1, 189, 556 genotyped SNPs. 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 covarites provided by the GTEx consortium, including top 3 principal components (PC), Probabilistic Estimation of Expression Residuals (PEER) factors, genotyping platform and sex (as described in `https://gtexportal.org/home/documentationPage`).

Individual level GWAS datasets. The NFBC dataset is comprised of 5, 402 individuals with ten continuous phenotypes related to cardiovascular diseases including 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. The individuals with contradictory in reported sex and sex determined from the X chromosome were first excluded. 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) with , coronary artery disease (CAD), Crohn's disease (CD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D) and type 2 diabetes (T2D). It includes around $2,000$ cases 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 is less than 5%; genotypes are missed 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 remained for further analysis (See Supplementary Table 1). Based on the obtained data, we calculated the GRM and extracted top 20 PCs as covariates included in our study.

GWAS summary statistics. We analyzed ten summary level GWAS datasets: human plasma pQTL data [28], circulating metabolite data [30], four schizophrenia datasets [31, 32, 33, 34], two independent height datasets [35] and European ancestry of WHRadjBMI datasets separated by men and women [36]. The SNPs with missing information (i.e. chromosome, minor allele, allele frequency) were first removed. Following the practice of LDSC [24], we checked the χ^2 statistic of each SNP and excluded those with extreme values ($\chi^2 > 80$) to prevent dominant effect. The detailed information is provided in Supplementary Table 2. After preprocess, the remaining SNPs were further matched with reference data during analysis, which is automatically processed using our IGREX software.

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

Data availability. The GTEx gene expression data was downloaded from GTEx Consortium website `https://gtexportal.org/home/datasets`. 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 caccessed using the links provided in Supplementary Table 2.

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] 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.

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

[11] 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.

[12] 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.

[13] 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, 2018.

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

[15] Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda WJH Penninx, Rick Jansen, Eco JC De Geus, Dorret I Boomsma, Fred A Wright, et al.

Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics*, 48(3):245, 2016.

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

[17] Nicholas Mancuso, Gleb Kichaev, Huwenbo Shi, Malika Freund, Claudia Giambartolomei, Alexander Gusev, and Bogdan Pasaniuc. Probabilistic fine-mapping of transcriptome-wide association studies. *bioRxiv*, page 236869, 2017.

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

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

[20] Can Yang, Xiang Wan, Xinyi Lin, Mengjie Chen, Xiang Zhou, and Jin Liu. Comm: a collaborative mixed model to dissecting genetic contributions to complex traits by leveraging regulatory information. *Bioinformatics*, page bty865, 2018.

[21] 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.

[22] 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.

[23] Daniel J Schaid, Wenan Chen, and Nicholas B Larson. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature reviews. Genetics*, 2018.

[24] 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.

[25] 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.

[26] 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.

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

[28] 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.

[29] 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.

[30] 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.

[31] JW Smoller. Cross disorder group of the psychiatric genomics consortium. identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis (vol 381, pg 1371, 2013). *Lancet*, 381(9875):1360–1360, 2013.

[32] 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.

[33] 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.

[34] 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.

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

[36] Dmitry Shungin, Thomas W Winkler, Damien C Croteau-Chonka, Teresa Ferreira, Adam E Locke, Reedik Mägi, Rona J Strawbridge, Tune H Pers, Krista Fischer, Anne E Justice, et al. New genetic loci link adipose and insulin biology to body fat distribution. *Nature*, 518(7538):187, 2015.

[37] 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.

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

[39] Chuanhai Liu, Donald B Rubin, and Ying Nian Wu. Parameter expansion to accelerate em: the px-em algorithm. *Biometrika*, 85(4):755–770, 1998.

[40] Hua Zhou, Liuyi Hu, Jin Zhou, and Kenneth Lange. Mm algorithms for variance components models. *arXiv preprint arXiv:1509.07426*, 2015.