www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Cis-SNPs Set Testing and PrediXcan Analysis for Gene Expression Data using Linear Mixed Models

Received: 12 June 2017
Accepted: 19 October 2017
Published online: 10 November 2017

Ping Zeng[1,2], Ting Wang[1] & Shuiping Huang[1]

Understanding the functional mechanism of SNPs identified in GWAS on complex diseases is currently a challenging task. The studies of expression quantitative trait loci (eQTL) have shown that regulatory variants play a crucial role in the function of associated SNPs. Detecting significant genes (called eGenes) in eQTL studies and analyzing the effect sizes of cis-SNPs can offer important implications on the genetic architecture of associated SNPs and interpretations of the molecular basis of diseases. We applied linear mixed models (LMM) to the gene expression level and constructed likelihood ratio tests (LRT) to test for eGene in the Geuvadis data. We identified about 11% genes as eGenes in the Geuvadis data and found some eGenes were enriched in approximately independent linkage disequilibrium (LD) blocks (e.g. MHC). We further performed PrediXcan analysis for seven diseases in the WTCCC data with weights estimated using LMM and identified 64, 5, 21 and 1 significant genes ($p < 0.05$ after Bonferroni correction) associated with T1D, CD, RA and T2D. We found most of the significant genes of T1D and RA were also located within the MHC region. Our results provide strong evidence that gene expression plays an intermediate role for the associated variants in GWAS.

Since the first study of age-related macular degeneration (AMD) was published in 2005[1], the past few years have witnessed a remarkably fast development of genome-wide association studies (GWAS)[2]. A large number of genetic susceptibility loci (mostly single nucleotide polymorphisms, [SNPs]) have been identified for many complex diseases[3–6], including human cancers[7–12], psychiatric disorders[13–16], autoimmune-related diseases[17–23], and many others. However, for most complex diseases, the identified variants only account for a minority of heritable variation, resulting in the so-called missing heritability problem[24]. Additionally, the majority of identified SNPs in GWAS are located within the non-coding regions (e.g. approximately 88% lie in intergenic or intronic regions[4]) and their causal genetic function remains largely unknown. Understanding the functional effects of the non-coding genetic variants is currently one of the main challenges. Recent advances of sequencing technologies have allowed researchers to quickly and cheaply type every genetic variant across the genome. A lot of large scale expression quantitative trait locus (eQTLs) studies[19,25–27] have been implemented and revealed that many variants identified in GWAS are also regulatory SNPs, which have an important influence on the molecular-level phenotypes (e.g. gene expression)[25,28–31]. This suggests that eQTLs mediate the effects of risk variants in GWAS and hold the fundamental important role to understand the genetic mechanism of disease susceptibility and phenotypic variation[27,32].

In GWAS literature, linear mixed models (LMM) are one of the most popular approaches, and widely used for multilocus association analysis[33–41], adjustment for individual relatedness and population stratification[42–44], genome-wide SNP heritability estimation or heritability partition[45,46] and genetic prediction[47,48]. LMM is also applied to eQTL studies, including fine mapping[49–51], predication of gene expression[52–54] and heritability estimation using cis-SNPs[55,56]. Motivated by the wide flexibility and applicability of LMM and the biologically functional importance of cis-SNPs mentioned above, in the present study based on LMM we develop an efficient likelihood ratio test (LRT) to examine whether a set of cis-SNPs are jointly related to the expression level of the gene that they are located within. We further perform PrediXcan analysis[57] for seven diseases from Wellcome Trust Case Control Consortium (WTCCC)[17] by making full use of the estimated effects of cis-SNPs yielded via LMM. We carry out numerical studies to evaluate the power of LRT and adopt the Geuvadis gene expression data[26] to illustrate our analysis framework.

[1]Xuzhou Medical University, Department of Epidemiology and Biostatistics, Xuzhou, 221004, China. [2]University of Michigan, Department of Biostatistics, Ann Arbor, MI, 48104, USA. Correspondence and requests for materials should be addressed to P.Z. (email: zpstat@xzhmu.edu.cn) or S.H. (email: hsp@xzhmu.edu.cn)

Methods

Overview of linear mixed models. Let **e** be an n-vector of continuous phenotype (e.g. gene expression level) measured on n independent samples and assume **e** is centered so that we ignore the intercept in the model. Let **X** be an n by q matrix for q covariates, **Z** is an n by p matrix of genotypes for p variants (e.g. cis-SNPs within a predefined gene or other well-defined genetic region). We formulate the relationship between **e**, **X** and **Z** via the following linear mixed model[45,58]

$$\begin{aligned} \mathbf{e} &= \mathbf{X}\boldsymbol{b} + \mathbf{Z}\boldsymbol{\beta} + \varepsilon,\ \varepsilon \sim N(0, \sigma^2\mathbf{I}_n), \\ \beta_k &\sim N(0, \tau^2), \end{aligned} \tag{1}$$

where $\boldsymbol{b}$ and $\boldsymbol{\beta}$ are the effects of covariates and cis-SNPs and are assumed to be fixed and random, respectively; ε is the n-vector of independent and identically distributed residual with variance σ^2 and $\mathbf{I}_n$ is an n by n identify matrix. In equation (1) the phenotype **e** has marginal mean $\mathbf{X}\boldsymbol{b}$ and variance $\boldsymbol{\Sigma} = \tau^2\mathbf{ZZ}' + \sigma^2\mathbf{I}_n = \sigma^2\mathbf{V}_\lambda$ with $\mathbf{V}_\lambda = \lambda\mathbf{ZZ}' + \mathbf{I}_n$ and $\lambda = \tau^2/\sigma^2$. Note that λ is the signal-noise ratio in equation (1) and is an important quantity related to the SNP-based heritability (denoted as h^2) by $h^2 = \lambda/(1+\lambda)$. Efficient estimation algorithms and software (e.g. GCTA[59]) have been designed for large scale applications of LMM to genome-wide genetic data.

Two applications of LMM. As mentioned above, LMM in equation (1) has widely important applications in genetics, and is the foundation of variance-component based association test, population structure control, phenotypic prediction and heritability estimation. In this paper we are particularly interested in two applications of LMM under the context of gene expression data.

Cis-SNPs set association test. The first application of LMM is the cis-SNPs set association test. That is, our objective is to detect whether the p cis-SNPs located within a given gene are simultaneously related to the expression level **e** of that gene. Here we only focus on cis-SNPs due to the fact that in terms of previous work most eQTLs are near the regulated gene and only a few eQTLs are trans-acting[49,60,61] and the effects of trans-SNPs are usually too weak to be detected with a reasonably high power[62]. By treating the effects of cis-SNPs $\boldsymbol{\beta}$ as random, the cis-SNPs set association test is equivalent to examining H_0: $\lambda = 0$ in equation (1). However, it is a nonstandard hypothesis test in the sense that the parameter of interest λ is on the boundary of the parameter space. Under this situation, the commonly-used asymptotic null chi-square distribution does not necessarily hold[63–67].

We use likelihood ratio test (LRT) to test H_0: $\lambda = 0$ and define the LRT statistic as

$$T = 2\ \sup_{\lambda\geq 0}\ [L(\lambda) - L(\lambda = 0)], \tag{2}$$

where $L(\lambda)$ is the profile log-likelihood function[68,69] of equation (1). While the score-based test[34,70] can be also employed for testing for H_0: $\lambda = 0$, we prefer LRT as it has been shown previously that: (i) LRT is more powerful than the score test[40,71]; (ii) in addition to p value for significance test, LRT provides additional useful estimates of unknown parameters (e.g. the estimates of λ and the effects $\boldsymbol{\beta}$ of cis-SNPs) for downstream data analyses; while the score test cannot offer such estimates as it only fits the null model (i.e. the simple linear model). The unknown parameter λ is obtained by restricted maximum likelihood estimation (REML)[68,72] and the exact null distribution of the LRT statistic T in equation (2) is obtained via a simulation-based manner (Algorithm 1 in Supporting Information) using the spectral representation[40,73–75].

In previous work LRT was applied to examine the variance component for multilocus genetic association studies[40]. Although efficient algorithms have been developed[71], LRT still has a high computational cost because it needs to fit both the null model (i.e. a simple linear model) and the alternative model (i.e. a linear mixed model, fitted using REML via Newton-Raphson iterations). Additionally, the null distribution of the LRT statistic is obtained using a simulation-based algorithm[40,74] (Algorithm 1 in Supporting Information). Thus, These limit LRT more widespread application to large scale association studies. For genes with relatively large p values (e.g. greater than 0.05), the simulation-based algorithm is fast and needs only a few simulations to yield stable p value estimates. However, it is computationally expensive for genes that have very small p values (e.g. less than 10^{-6}). For example, assume there are a total of 20,000 genes, then at least 10^7 simulations are required to obtain stable p values at the significance level of $\alpha = 2.5 \times 10^{-6}$ corrected by the Bonferroni method for multiple hypothesis testing, making LRT infeasible for large scale gene-based association studies. Furthermore, for more extremely small p values (e.g. less than 10^{-10}), the resulting p value estimates are typically zero due to limited simulations in the simulation-based algorithm, which is less informative for subsequent data analyses. To reduce the computation burden of the simulation-based algorithm and generate more informative p values for these most significant genes, we approximate the exact distribution with an appropriate mixture as previously considered in[76–78]. Specifically, assume the approximate distribution has a mixture form of

$$T \sim \varphi\chi_0^2 + (1-\varphi)\kappa\chi_1^2, \tag{3}$$

where χ_0^2 is a point mass at zero and χ_1^2 is a chi-square distribution with one degree of freedom, φ is the proportion parameter and κ is the scale parameter. The unknown parameters φ and κ can be estimated by the method of moment, the quantile regression or the method of local probability[78]. The corresponding p value of T is yielded from the estimated approximate distribution of equation (3) (Algorithm 2 in Supporting Information).

PrediXcan analysis based on BLUE. Once λ in equation (1) is estimated by REML, say $\widehat{\lambda}$, we obtain the best linear unbiased estimator (BLUE) for the random effects of the cis-SNPs

$$\widehat{\boldsymbol{\beta}} = \widehat{\lambda}\mathbf{Z}^{T}(\widehat{\lambda}\mathbf{Z}\mathbf{Z}^{T} + \mathbf{I}_{n})^{-1}\left(\mathbf{e} - \mathbf{X}_{(}\mathbf{X}^{T}\widehat{\mathbf{V}}_{\widehat{\lambda}}^{-1}\mathbf{X}_{)}^{-1}\mathbf{X}^{T}\widehat{\mathbf{V}}_{\widehat{\lambda}}^{-1}\mathbf{e}\right). \tag{4}$$

The BLUE $\widehat{\boldsymbol{\beta}}$ can be employed for genetic prediction, called the best linear unbiased prediction (BLUP)[79], in both GWAS[47,48] and gene expression data[54]. Here we apply $\widehat{\boldsymbol{\beta}}$ as weights in the recently developed PrediXcan analysis[57] for gene-based test in transcriptome-wide association studies[80]. Specifically, let **G** be the same set of cis-SNPs as **Z** for a given gene and **y** be the phenotype in the GWAS. The basic idea of PrediXcan analysis is first to impute the unobserved gene expression level using weights (e.g. $\widehat{\boldsymbol{\beta}}$ in equation (4)) estimated from a reference transcriptome data[57], say, $\hat{\mathbf{e}}_{\text{gwas}} = \mathbf{G}\widehat{\boldsymbol{\beta}}$. Note that, here we explicitly adopt the subscript "gwas" to emphasize that we are predicting the unmeasured gene expression in the given GWAS data with **G** and $\widehat{\boldsymbol{\beta}}$ rather than predicting gene expression with **Z** and $\widehat{\boldsymbol{\beta}}$. Then, we test for the relationship between **y** and $\hat{\mathbf{e}}_{\text{gwas}}$ via a linear model or logistic model depending on **y** is a continuous or binary (e.g. case-control) phenotype. Of note, to generate weights in the PrediXcan analysis, the elastic net (ENET)[57] and Bayesian sparse linear mixed model (BSLMM)[80] were also used previously. We will compare the performance of various weights (generated from LMM, ENET and BSLMM) in our real data applications.

Numerical Studies. We first evaluated the performance of the approximate LRT (aLRT, based on Algorithm 2) and compared with the exact LRT (eLRT, based on Algorithm 1) on SNPs set testing. To make our numerical studies as real as possible, we selected a region of continuous genotypes **Z** from the Geuvadis data[26] (see below). The selected genotypes **Z** included 100 cis-SNPs with minor allele frequency (MAF) larger than 0.05 and the sample size n was 465. For the type I error control, we randomly selected 10, 25, 50, 75 or 100 markers included into equation (1), and simulated gene expression levels from a standard normal distribution and set the cis-SNPs effect $\boldsymbol{\beta}$ to zero. For the statistical power evaluation, we generated $\boldsymbol{\beta}$ from a normal distribution with mean zero and varying variances (i.e. $\tau^2 = 0.03^2$, 0.08^2 or 0.10^2; these values were adopted to ensure a reasonable power); again we randomly selected 10, 25, 50, 75 or 100 markers included into equation (1), and simulated gene expression levels from a normal distribution with mean $\mathbf{Z}\boldsymbol{\beta}$ and variance 1. We set M to 10^6 in Algorithm 1 and L to 10^4, 5×10^3, 10^3 or 500 in Algorithm 2. Here M and L are respectively the number of simulations used in Algorithm 1 and Algorithm 2 in Supporting Information. The number of replicates was 10^6 and 10^4 for the type I error control and statistical power evaluation, respectively. Following previous work[34], the significance level α was set to 10^{-4}, and the type I error and power were estimated as the proportion of p values less than α.

Cis-regulatory variants set detection in Geuvadis data. We applied LRT (both aLRT and eLRT) to the Geuvadis data[26] to perform cis-SNPs set detection. The gene whose expression level is related to at least one cis-SNP is referred to as eGene[81]. Detecting eGene is one of the most important tasks in eQTL studies. Briefly, our aim is to examine whether a set of cis-SNPs that locate within a 10 kb genomic region centered at the transcription start site (TSS) of that gene are related to its gene expression level. These markers are referred to as cis-regulatory variants or cis-expression quantitative trait loci (cis-eQTL) and have important implications for understanding gene regulation and interpreting the genetic basis of variation for complex diseases and traits[4,25,82–85]. In the Geuvadis project[26] a total of 465 individuals were sequenced on lymphoblastoid cell lines (LCL) from five different populations: CEU, FIN, GBR, TSI and YRI. The genotypes were measured in the 1000 Genomes project. The PEER normalization[26,86–88] was first used to remove technical variations and then each gene expression measurement was quantile normalized to a standard normal distribution. According to GENCODE[89] release 12, following[49] we focused on 15,771 protein coding genes that were expressed on at least half individuals and had at least 10 cis-SNPs, resulting in an average of 75 cis-SNPs (MAF $>$ 0.05) per gene.

PrediXcan analysis for WTCCC data based on LMM and Geuvadis data. We performed PrediXcan analysis for the Wellcome Trust Case Control Consortium (WTCCC) data[17]. The WTCCC data consists of 2,938 shared controls and about 14,000 cases from seven common diseases: 1,963 individuals with type 1 diabetes (T1D), 1,748 individuals with Crohn's disease (CD), 1,860 individuals with rheumatoid arthritis (RA), 1,868 individuals with bipolar disorder (BD), 1,924 individuals with type 2 diabetes (T2D), 1,926 individuals with coronary artery disease (CAD), and 1,952 individuals with hypertension (HT). We first imputed missing genotypes of WTCCC using BIMBAM[90], and further imputed SNPs using the Europe population of 1000 Genomes as the reference panel[91] with SHAPEIT[92–95] and IMPUTE2[95]. Finally, we yielded about 2,000,000 SNPs shared across all individuals after stringent quality control (i.e. Hardy-Weinberg equilibrium p value $< 10^{-4}$ and MAF $<$ 0.05). For PrediXcan analysis[57] we focused on the same 15,771 genes as in the Geuvadis data, and for each pair of genes in the WTCCC and Geuvadis data we matched their cis-SNPs. We predicted the expression level of each gene in the WTCCC data with the weights as the BLUE of cis-SNPs of the corresponding gene in the Geuvadis data, and performed logistic regression for each gene in turn as the WTCCC is a case control study.

Data availability. Our study did not generate any data and made use of data generated by Wellcome Trust Case Control Consortium. The datasets of WTCCC can be available by application to the Consortium Data Access Committee at https://www.wtccc.org.uk/. The Geuvadis gene expression data can be publicly available at http://www.Geuvadis.org/. The R function implementing LRT (for both aLRT and eLRT) is freely available at https://github.com/biostatpzeng/LRT.

Results

Evaluation of type I error and power for numerical studies. Note that evaluating the performance of the approximate LRT (aLRT) is equivalent to evaluating the approximate distribution generated using the

No	eLRT	aLRT			
		10^4	5×10^3	10^3	500
$\tau^2 = 0.03^2$					
10	0.0000	0.0006	0.0006	0.0000	0.0006
25	0.0034	0.0047	0.0050	0.0044	0.0050
50	0.0082	0.0098	0.0095	0.0095	0.0092
75	0.0177	0.0210	0.0210	0.0189	0.0210
$\tau^2 = 0.08^2$					
100	0.0395	0.0395	0.0395	0.0395	0.0368
10	0.0614	0.0669	0.0675	0.0663	0.0675
25	0.2510	0.2598	0.2595	0.2625	0.2611
50	0.5497	0.5641	0.5625	0.5668	0.5638
75	0.7765	0.7880	0.7880	0.7837	0.7783
$\tau^2 = 0.10^2$					
100	0.8801	0.8887	0.8881	0.8834	0.8868
10	0.1733	0.1872	0.1850	0.1856	0.1944
25	0.4873	0.4964	0.4987	0.4984	0.5013
50	0.8058	0.8181	0.8194	0.8197	0.8184
75	0.9289	0.9336	0.9349	0.9336	0.9355
100	0.9698	0.9712	0.9712	0.9707	0.9717

Table 1. Estimated power for eLRT and aLRT in the numerical studies. Note: The No column denotes the number of cis-SNPs included in the gene; we set M to 10^6 in Algorithm 1 (generate the exact null distribution for eLRT) and L to 10^4, 5×10^3, 1×10^3 and 500 in Algorithm 2 (generate the approximate null distribution for aLRT). The significant level was set to 10^{-4}. aLRT: the approximate likelihood ratio test; eLRT: the exact likelihood ratio test.

mixture method in Algorithm 2. Figure S1 shows the approximate mixture distribution is very consistent to the exact one which is generated using the simulation-based method in Algorithm 1 and has been previously proved to control for the type I error efficiently[40,74]. While we also note that sometimes (e.g. $L = 500$) the approximate distribution tends to be slightly liberal. Table 1 shows aLRT maintains a similar statistical power as eLRT under a range of scenarios. It is seen when the association signal is strong (e.g. $\tau^2 = 0.10^2$), aLRT with $L = 500$ generally leads to a slightly higher power than eLRT, corresponding to the finding that the approximate distribution tends to be slightly liberal when L is small (e.g. $L = 500$) in Fig. S1. Nevertheless, the inflation of power due to the approximation is acceptable; for example, the greatest difference between the power of aLRT and eLRT is less than 0.017 (Table S1).

We further compared the computation time for eLRT and aLRT. A total of 10^3 genes were tested and each gene included 50 cis-SNPs. The sample size was set to 10^3. We again set M to 10^6 in Algorithm 1 and L to 10^4, 5×10^3, 10^3 or 500 in Algorithm 2. The computation was implemented on a personal computer with 3.09 GHz and 3.16 Gb memory and the computation time was averaged over 50 repeats. It shows that eLRT needs about 4.5 hours under this setting, while aLRT needs less than 800 seconds (i.e. about 767, 690, 624 and 616 seconds for $L = 10^4$, 5×10^3, 10^3 or 500, respectively), about 20 times faster than the exact counterpart (i.e. eLRT).

Detection of eGene in the Geuvadis data. Figure 1a displays the p values of aLRT and eLRT. It shows aLRT and eLRT generate comparable results as shown in the numerical studies; the correlation p values (−log10 scale) of aLRT and eLRT is 0.991 (standard error [se] is 8.1×10^{-4}). We used the Bonferroni method to control for the family wise error rate at 0.05 significance level. After Bonferroni correction, aLRT and eLRT respectively identify 1,665 (10.56%) and 1,707 (10.82%) eGenes. The number of shared eGenes between aLRT and eLRT is 1,653. As a comparison, we also performed the score test, discovering 1,189 eGenes (7.54%), much less than these of aLRT and eLRT. We list the eGenes identified by aLRT but not by eLRT in Table S2, where it shows the p values from eLRT are unstable because of limited simulations (i.e. $M = 10^6$) in eLRT in Algorithm 1, whereas the p values from aLRT are relatively stable. As mentioned before, it is computationally expensive to obtain believable p values for genes with extremely small p values for eLRT using Algorithm 1; in contrast, aLRT avoids this limitation and offers useful p values, demonstrating the benefit of the approximation strategy. Thereby, the following results are mainly based on aLRT.

To check the distribution pattern of these eGenes, we plot the p values of all genes against the estimated heritability, the number of cis-SNPs included in each gene and the length of the gene in Fig. 1b–d. As expected, it is more likely to be an eGene for a gene with larger heritability (Fig. 1b); the correlation between the p values (−log10 scale) and estimated heritability values is 0.856 (se $= 3.2 \times 10^{-3}$). Nevertheless, we do also see that some genes with large heritability fail to be identified as eGenes (e.g. the blue region in Fig. 1b), which may be the direct consequence of the small sample size (i.e. $n = 465$) for the Geuvadis data. We do not see any pattern between the p values (−log10 scale) with the number of cis-SNPs included in each gene (Fig. 1c), and with the length of the gene (Fig. 1d). These observations suggest that a more heritable gene has a higher likelihood to be an eGene, but not all cis-SNPs in a gene have influences on the expression level, and further imply that the genetic architecture of gene

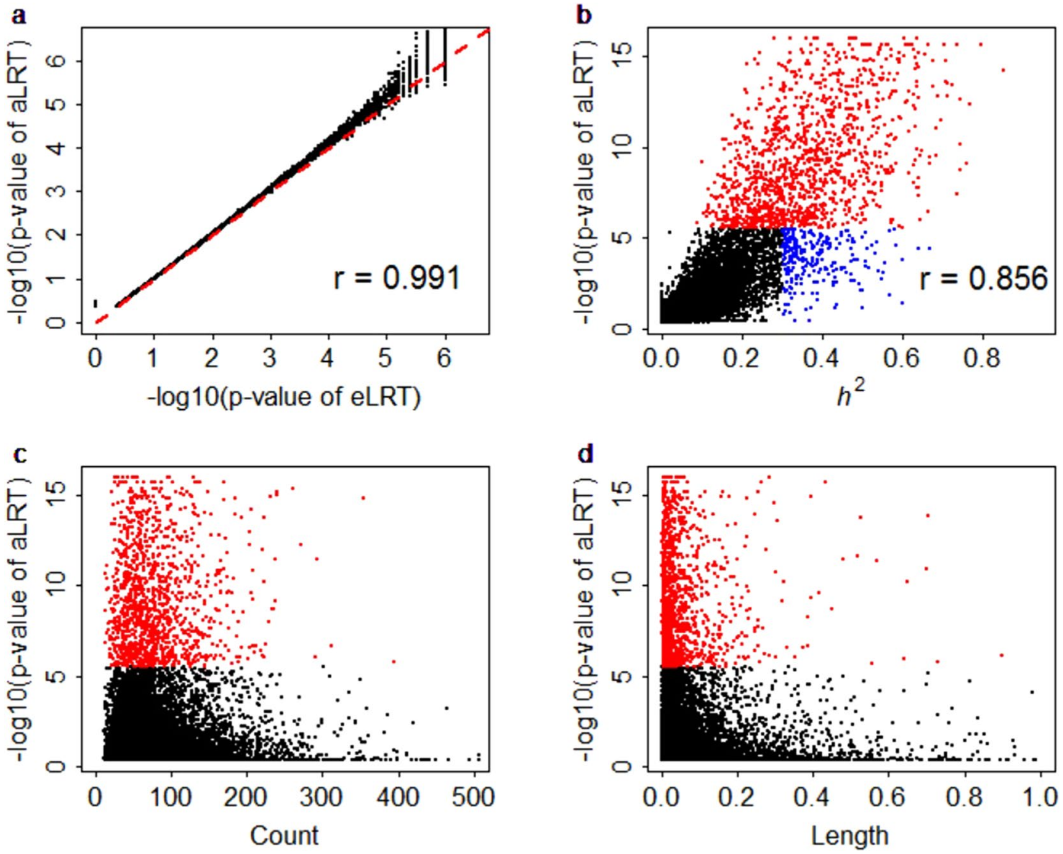

Figure 1. The p values of eLRT and aLRT for all the analyzed genes in the Geuvadis data. (**a**) The scatter plot of p values (with −log10 scale) between eLRT and aLRT across all the genes. (**b**) The scatter plot of p value (with −log10 scale) of aLRT with the estimated heritability of each gene. (**c**) The scatter plot of p values (with −log10 scale) of aLRT with the number of cis-SNPs in each gene. (**d**) The scatter plot of p value (with −log10 scale) of aLRT with the length of cis-SNPs in each gene. eLRT: the exact likelihood ratio test, aLRT: the approximate likelihood ratio test.

expression levels may be less polygenic than that assumed by LMM[54,55,96]. We show the distribution of p values of aLRT for all genes in Fig. 2a, the proportion of eGene for each chromosome in Fig. 2b and the proportion of eGene against the proportion of genes distributed in each chromosome in Fig. 2c. It is seen that chromosomes 1, 2, 6, 11, 17 and 19 include more eGenes, and the proportion of eGene is positively proportional to the proportion of genes of chromosome (the correlation is 0.922 and se = 0.062).

We further examine the enrichment of eGene for approximately independent linkage disequilibrium (LD) blocks across chromosomes. For the Geuvadis data we obtain 1,435 independent LD blocks[97]. We calculate the enrichment fold for each LD block following a similar way as in[98]. In particular, the enrichment fold is computed as the ratio of the proportion of eGene and the proportion of length for the given LD block. We observe enrichments of eGene in some special genetic regions (Fig. 3a) and list these LD blocks with enrichment-fold larger than 20 in Table S3. Here we use the major histocompatibility complex (MHC) region (Chr 6: 26–34 Mb) as an illustrative example. There are 134 eGenes in chromosome 6, among which 36 are located within the MHC region (denoted in blue in Fig. 3c). The total length of chromosome 6 is about 171 Mb, and the length of the MHC region is 8 Mb. Then the enrichment fold is 5.74, which is the ratio of the proportion of eGene in the MHC region (i.e. 0.27 = 36/134) and the proportion of the length of MHC (i.e. 0.05 = 8/171). It is significantly higher (p value is 4.32×10^{-3} using an approximate z test) than the average enrichment fold (the median is 1.35) of other LD blocks in chromosome 6. It has long been recognized that the MHC region has importantly biological function on many human diseases and traits[99]. For example, in terms of the NHGRI-EBI GWAS Catalog (http://www.ebi.ac.uk/gwas/, until 05/25/2017), we find that a total of 1,044 (2.72% among all 38,369 variants) identified markers are located within in the MHC region and are associated with as many as about 320 (16.9% among all 1,890 phenotypes) diseases and traits (e.g. type I diabetes, Crohn's disease, rheumatoid arthritis and infectious diseases)[17,100,101]. However, like most of other identified SNPs, the genetic function of these identified SNPs in the MHC region is also not well understood to date[102]. Therefore, the enrichment of eGene in the MHC region (Fig. 3c) offers a useful understanding for the functional mechanism for these identified SNPs in GWAS.

PrediXcan analysis results for WTCCC. We now turn to the PrediXcan analysis of the seven diseases (i.e. T1D, CD, RA, HT, CAD, BD and T2D) in the WTCCC data. Following[57] we focus on genes with estimated heritability larger than 0.01, finally resulting in 9,418 genes. Briefly, the BLUE of the cis-SNPs were used to predict

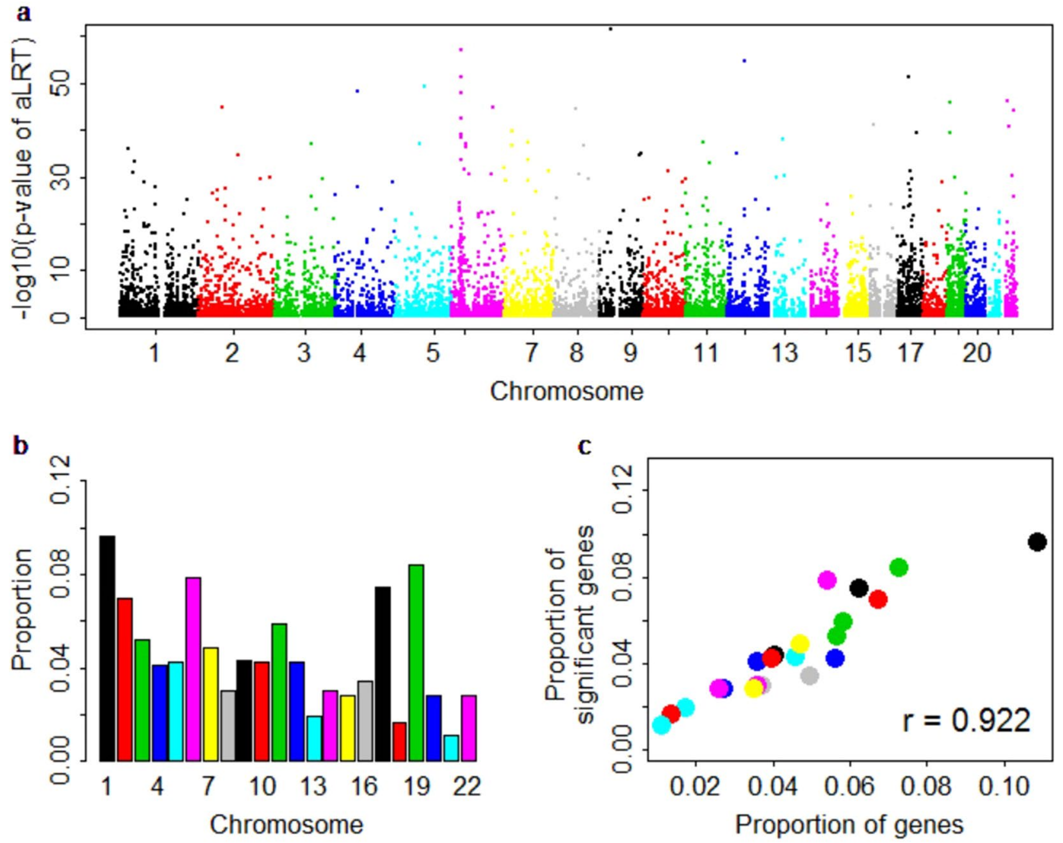

Figure 2. The distribution of p values of aLRT for all genes. (**a**) The Manhattan plot shows p values (with −log10 scale) and gene positions across chromosomes, in which the y-axis is −log10 (p values) for each gene, the x-axis is the gene position and the various colors represent different chromosomes. (**b**) The barplot shows the proportion of significant genes for each chromosome. (**c**) The scatter plot of the proportion of significant genes against the proportion of genes distributed in each chromosome. aLRT: the approximate likelihood ratio test.

the gene expression level using the genotypes of WTCCC; then the predicted gene expression was tested for association with the case-control phenotypes of WTCCC using logistic regression. Manhattan plots summarizing genome-wide association results for the seven diseases are shown in Fig. S2. After Bonferroni correction at 0.05 significance level, 64, 5, 21 and 1 genes are identified that are related to T1D, CD, RA and T2D, respectively. Among these, we observe 57 (89.1%) for T1D and 19 (90.5%) for RA are located within the MHC region (Fig. 4a,b), and all the 76 (57 + 19) genes include risk SNPs that were discovered in previous GWAS (Table S4). Using weights of BSLMM in PrediXcan analysis, 64, 5, 17, 1 and 1 genes are identified that are related to T1D, CD, RA, CAD and T2D, comparable to those identified with LMM; while using weights of ENET in PrediXcan analysis, only 9 and 1 genes are identified that are associated with T1D and CD, much less than those yielded from LMM or BSLMM. Note that the original PrediXcan analysis[57] based on ENET identified much more significant genes, mainly due to a larger reference data used there[25,57] — 922 samples were sequenced RNA from whole blood[25]. The venn diagram (Fig. 5) shows the identified genes of T1D, CD and RA are shared among the three methods, especially between LMM and BSLMM. Presumably, the different genes identified with various weights are attributed to the distinct genetic architecture of the gene expression and the diseases as well as the assumptions underlying various models. In summary, together with the enrichments of eGene in the MHC region observed in Fig. 3c in the Geuvadis data, the observations that the significant genes identified by PrediXcan analysis for T1D and RA are also enriched in the same region offer strong supports that gene expression level plays an intermediate role for the risk variants identified in GWAS and the two diseases (i.e. T1D and RA).

Discussion

In this paper we have applied the popular LMM to the gene expression data. We mainly focus on eGene detection and PrediXcan analysis based on the BLUE of the effects of cis-SNPs. Based on LMM we have employed LRT to discover the eGene in gene expression data, and developed an approximate LRT (aLRT) to speed up the computation. Both numerical studies and real data applications have shown that aLRT works equally well compared with the exact LRT (eLRT) and demonstrated that aLRT can offer more useful estimates for extremely small p values. Importantly, we have shown that aLRT achieves substantial gains in computation while maintaining the effective type I error control and the statistical power. As shown, aLRT is orders of magnitude faster than eLRT depending on the choice of L. For example, if $M = 10^7$ in Algorithm 1 and $L = 10^3$ in Algorithm 2, theoretically, aLRT can

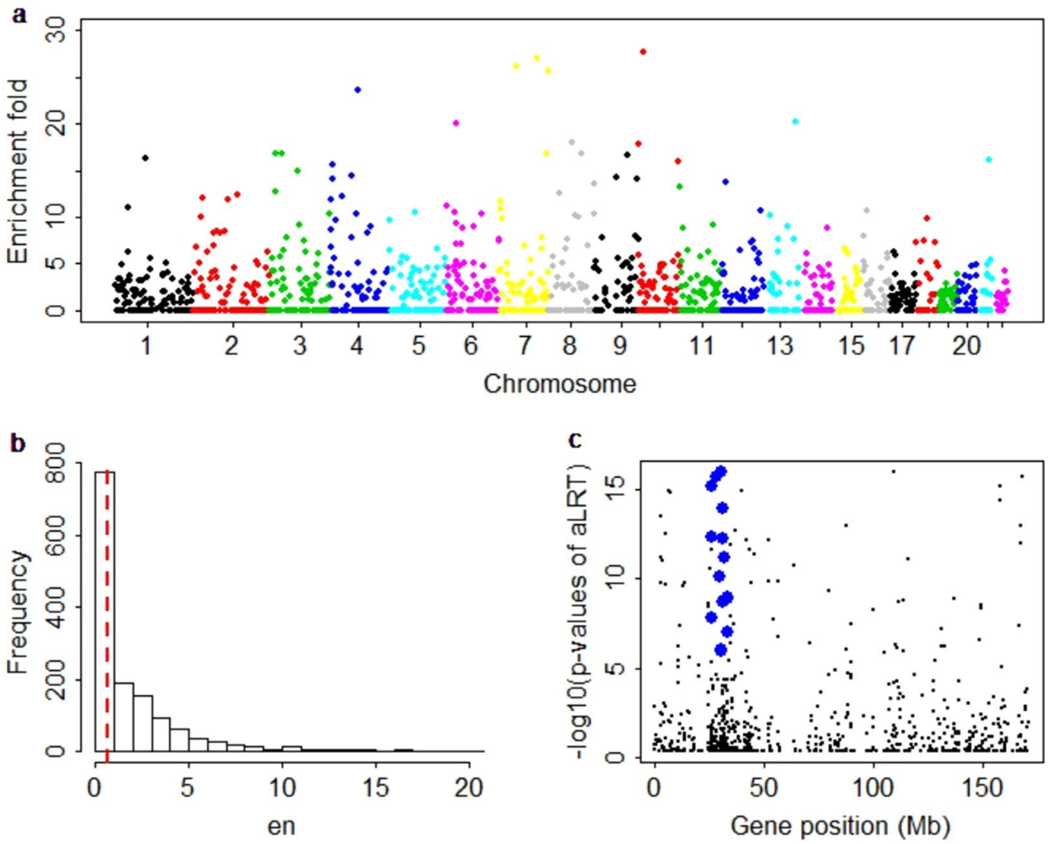

Figure 3. Distribution of enrichment fold for 1,400 approximately independent LD blocks for Geuvadis data. (**a**) A Manhattan-type plot shows enrichment fold for each independent LD block across chromosomes, in which the y-axis is enrichment fold for each LD block, the x-axis is the position of that LD block and the various colors represent different chromosomes. (**b**) The histogram plot shows the distribution of enrichment fold, the median (0.65) of enrichment fold is denoted with a red line. (**c**) The pattern of p values of aLRT (with −log10 scale) for the MHC region (Chr 6: 26–34 Mb). MHC: major histocompatibility complex, T1D: type 1 diabetes, RA: rheumatoid arthritis, LD linkage disequilibrium.

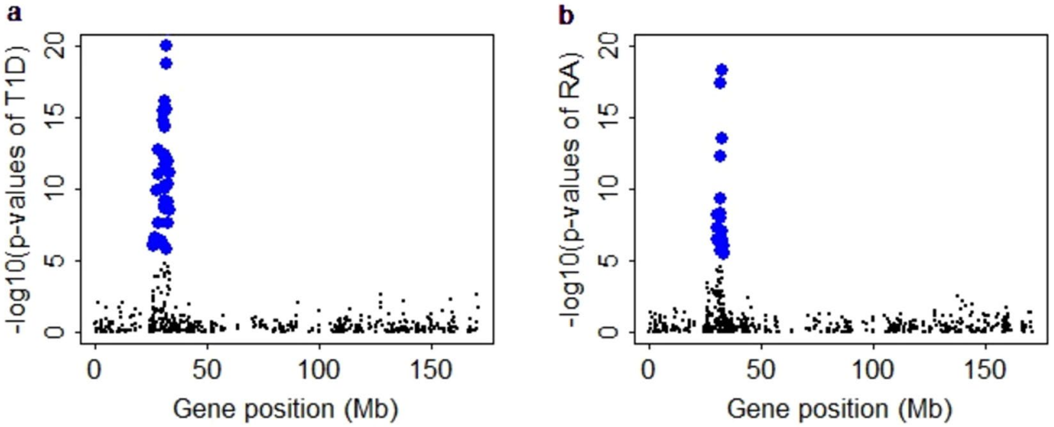

Figure 4. The pattern of p values (with −log10 scale) of PrediXcan analysis of (**a**) T1D and (**b**) RA for chromosome 6. T1D: type 1 diabetes, RA: rheumatoid arthritis.

improve the computation approximately 10^4 times relative to eLRT if ignoring the estimation of the approximate mixture null distribution. For the balance between accuracy and computational cost, in practice we recommend using $L = 10^4$ since empirically this choice has a higher accuracy compared with smaller values of L while not resulting in the increase of the computation burden significantly.

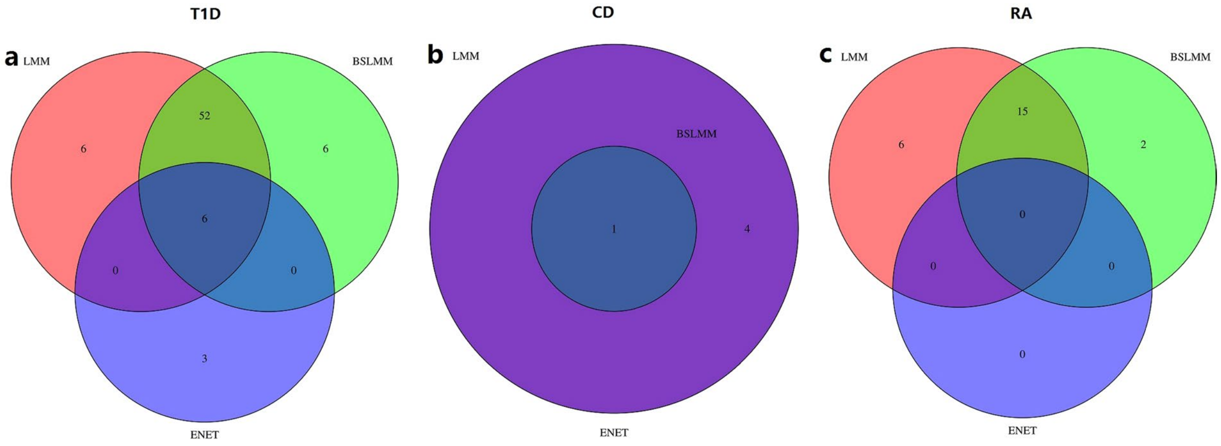

Figure 5. The venn diagram for identified genes of T1D, CD and RA using LMM, ENET and BSLMM. T1D: type 1 diabetes, CD: Crohn's disease, RA: rheumatoid arthritis, LMM: linear mixed model, ENET: elastic net, BSLMM: Bayesian sparse linear mixed model.

In the Geuvadis data we have shown that eGenes enrich in some special genetic regions (e.g. the MHC region), consistent with the previous finding from a perspective of the prediction of gene expression level[54]. However, we note that the power of eGene detection is still underpowered (e.g. less than 11% in the Geuvadis data with LRT, and less than 8% with the score test) because of the small simple size (i.e. 465 in the Geuvadis data). Incorporating functional annotations of cis-SNPs into the test is a potential way to improve the power[49,81] and is an active area in eQTL studies. The enrichment of eGenes in some specific genetic regions can offer important implications for SNPs that are identified in GWAS since it is now believed that the function of SNPs on phenotypes works by at least partially regularizing gene expression levels in a cis- or trans-acting manner[27,30,49,61,81].

Our analysis on two (i.e. T1D and RA) of seven diseases in the WTCCC data has shown that the PrediXcan analysis is an efficient way bridging SNPs, gene expressions and diseases. Especially, the PrediXcan analysis shows the same region (i.e. MHC) of enrichment of significant genes in PrediXcan analysis as that for the eGenes in the gene expression data. This is not likely by chance since there is a lot of evidence that the MHC region has important impacts on T1D and RA[17,100,102]. Nevertheless, we caution that the results of PrediXcan analysis for a given disease may be tissue-specific (e.g. the gene expression of the Geuvadis data used in the present paper was measured from lymphoblastoid cell lines) as it has been shown that the gene expression level is tissue-specific even for biologically developmentally close tissues[27,103–105]. Investigating the performance of PrediXcan analysis on diseases using tissue-specific gene expression level is an interesting problem in the further.

Finally, we recognize that different weights computed using various methods (e.g. LMM, ENET and BSLMM) can be used in PrediXcan analysis[57,80]. Although it has shown the genetic architecture of gene expression is less polygenic compared to most human complex diseases[55,96] and the sparse LMM has a better performance to capture the variation of gene expression[54], the optimal weights in PrediXcan analysis is not fully clear and may be case-specific. The property of PrediXcan analysis is also not fully studied and its power relies on many factors, such as the used reference transcriptome data (e.g. the Geuvadis data in the present paper), the genetic architecture of gene expression and the diseases. The weights estimated from LMM may not be the best choice, but in the real applications, we indeed found that the PrediXcan analysis based on LMM behaves comparably relatively to other competing methods. Performing a comprehensive comparison of PrediXcan analysis based on larger reference transcriptome data with various weights on large-scale GWAS phenotypes is our ongoing work.

References

1. Klein, R. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–389, https://doi.org/10.1126/science.1109557 (2005).
2. Bush, W. S. & Moore, J. H. Chapter 11: Genome-Wide Association Studies. *PLoS Comput. Biol.* **8**, e1002822, https://doi.org/10.1371/journal.pcbi.1002822 (2012).
3. Visscher, P., Brown, M., McCarthy, M. & Yang, J. Five Years of GWAS Discovery. *Am. J. Hum. Genet.* **90**, 7–24, https://doi.org/10.1016/j.ajhg.2011.11.029 (2012).
4. Hindorff, L. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **106**, 9362–9367, https://doi.org/10.1073/pnas.0903103106 (2009).
5. Zeng, P. *et al.* Statistical analysis for genome-wide association study. *Journal of Biomedical Research* **29**, 285–297, https://doi.org/10.7555/jbr.29.20140007 (2015).
6. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic Mapping in Human Disease. *Science* **322**, 881–888, https://doi.org/10.1126/science.1156409 (2008).
7. Dong, J. *et al.* Association analyses identify multiple new lung cancer susceptibility loci and their interactions with smoking in the Chinese population. *Nat. Genet.* **44**, 895–899, https://doi.org/10.1038/ng.2351 (2012).
8. Henderson, B. E., Lee, N. H., Seewaldt, V. & Shen, H. The influence of race and ethnicity on the biology of cancer. *Nat. Rev. Cancer* **12**, 648–653, https://doi.org/10.1038/nrc3341 (2012).
9. Gudmundsson, J. *et al.* A genome-wide association study yields five novel thyroid cancer risk loci. *Nat. Commun.* **8**, 14517, https://doi.org/10.1038/ncomms14517 (2017).
10. Al Olama, A. A. *et al.* A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. *Nat. Genet.* **46**, 1103–1109 (2014).

11. Schumacher, F. R. *et al*. Genome-wide association study of colorectal cancer identifies six new susceptibility loci. *Nat. Commun.* **6**, 7138, https://doi.org/10.1038/ncomms8138 (2015).
12. Al-Tassan, N. A. *et al.* A new GWAS and meta-analysis with 1000Genomes imputation identifies novel risk variants for colorectal cancer. *Scientific Reports* **5**, 10442, https://doi.org/10.1038/srep10442 (2015).
13. Cross-Disorder Group of the Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *The Lancet* **381**, 1371–1379, https://doi.org/10.1016/S0140-6736(12)62129-1 (2013).
14. Lane, J. M. *et al*. Genome-wide association analyses of sleep disturbance traits identify new loci and highlight shared genetics with neuropsychiatric and metabolic traits. *Nat. Genet.* **49**, 274–281, https://doi.org/10.1038/ng.823 (2017).
15. Lo, M.-T. *et al*. Genome-wide analyses for personality traits identify six genomic loci and show correlations with psychiatric disorders. *Nat. Genet.* **49**, 152–156, https://doi.org/10.1038/ng.823 (2017).
16. Cuyvers, E. & Sleegers, K. Genetic variations underlying Alzheimer's disease: evidence from genome-wide association studies and beyond. *The Lancet Neurology* **15**, 857–868, https://doi.org/10.1016/S1474-4422(16)00127-7 (2016).
17. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678, https://doi.org/10.1038/nature05911 (2007).
18. Sawcer, S. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476**, 214–219, https://doi.org/10.1038/nature10251 (2011).
19. Farh, K. K.-H. *et al*. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343, https://doi.org/10.1038/nature13835 (2015).
20. Lenz, T. L. *et al*. Widespread non-additive and interaction effects within HLA loci modulate the risk of autoimmune diseases. *Nat. Genet.* **47**, 1085–1090 (2015).
21. Li, Y. R. *et al*. Meta-analysis of shared genetic architecture across ten pediatric autoimmune diseases. *Nat. Med.* **21**, 1018–1027, https://doi.org/10.1038/nm.3933 (2015).
22. Jin, Y. *et al*. Genome-wide association studies of autoimmune vitiligo identify 23 new risk loci and highlight key pathways and regulatory variants. *Nat. Genet.* **48**, 1418–1424, https://doi.org/10.1038/ng.3680 (2016).
23. de Lange, K. M. *et al*. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**, 256–261, https://doi.org/10.1038/ng.3760 (2017).
24. Manolio, T. *et al*. Finding the missing heritability of complex diseases. *Nature* **461**, 747–753, https://doi.org/10.1038/nature08494 (2009).
25. Battle, A. *et al*. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* **24**, 14–24, https://doi.org/10.1101/gr.155192.113 (2014).
26. Lappalainen, T. *et al*. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511, https://doi.org/10.1038/nature12531 (2013).
27. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648–660, https://doi.org/10.1126/science.1262110 (2015).
28. Stranger, B. E. *et al*. Population genomics of human gene expression. *Nat. Genet.* **39**, 1217–1224, https://doi.org/10.1038/ng2142 (2007).
29. Pickrell, J. K. *et al*. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772, https://doi.org/10.1038/nature08872 (2010).
30. Nicolae, D. L. *et al*. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* **6**, https://doi.org/10.1371/journal.pgen.1000888 (2010).
31. Vockley, C. M., Barrera, A. & Reddy, T. E. Decoding the role of regulatory element polymorphisms in complex disease. *Curr. Opin. Genet. Dev.* **43**, 38–45, https://doi.org/10.1016/j.gde.2016.10.007 (2017).
32. Montgomery, S. B. & Dermitzakis, E. T. From expression QTLs to personalized transcriptomics. *Nature reviews Genetics* **12**, 277–282, https://doi.org/10.1038/nrg2969 (2011).
33. Tzeng, J. Y., Zhang, D., Chang, S.-M., Thomas, D. C. & Davidian, M. Gene-Trait Similarity Regression for Multimarker-Based Association Analysis. *Biometrics* **65**, 822–832, https://doi.org/10.1111/j.1541-0420.2008.01176.x (2009).
34. Wu, M. C. *et al*. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *Am. J. Hum. Genet.* **89**, 82–93, https://doi.org/10.1016/j.ajhg.2011.05.029 (2011).
35. Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13**, 762–775, https://doi.org/10.1093/biostatistics/kxs014 (2012).
36. Sun, J., Zheng, Y. & Hsu, L. A Unified Mixed-Effects Model for Rare-Variant Association in Sequencing Studies. *Genet. Epidemiol.* **37**, 334–344, https://doi.org/10.1002/gepi.21717 (2013).
37. Wang, X., Epstein, M. P. & Tzeng, J. Analysis of Gene-Gene Interactions Using Gene-Trait Similarity Regression. *Hum. Hered.* **78**, 17–26, https://doi.org/10.1159/000360161 (2014).
38. Hasegawa, T. *et al*. AP-SKAT: highly-efficient genome-wide rare variant association test. *BMC Genomics* **17**, 1–8, https://doi.org/10.1186/s12864-016-3094-3 (2016).
39. Nicolae, D. L. Association Tests for Rare Variants. *Annu. Rev. Genomics Hum. Genet.* **17**, 117–130, https://doi.org/10.1146/annurev-genom-083115-022609 (2016).
40. Zeng, P. *et al*. Likelihood Ratio Tests in Rare Variant Detection for Continuous Phenotypes. *Ann. Hum. Genet.* **78**, 320–332, https://doi.org/10.1111/ahg.12071 (2014).
41. Zeng, P. & Wang, T. Bootstrap Restricted Likelihood Ratio Test for the Detection of Rare Variants. *Curr. Genomics* **16**, 194–202, https://doi.org/10.1371/journal.pone.0093355 (2015).
42. Joo, J. W. J., Hormozdiari, F., Han, B. & Eskin, E. Multiple testing correction in linear mixed models. *Genome Biol.* **17**, 62 (2016).
43. Kang, H. M. *et al*. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354, https://doi.org/10.1038/ng.548 (2010).
44. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* **46**, 100–106, https://doi.org/10.1038/ng.2876 (2014).
45. Yang, J. *et al*. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569, https://doi.org/10.1038/ng.608 (2010).
46. Yang, J. *et al*. Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* **43**, 519–525, https://doi.org/10.1038/ng.823 (2011).
47. Makowsky, R. *et al*. Beyond Missing Heritability: Prediction of Complex Traits. *PLoS Genet.* **7**, e1002051, https://doi.org/10.1371/journal.pgen.1002051 (2011).
48. Wray, N. R. *et al*. Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14**, 507–515 (2013).
49. Wen, X., Luca, F. & Pique-Regi, R. Cross-Population Joint Analysis of eQTLs: Fine Mapping and Functional Annotation. *PLoS Genet.* **11**, e1005176, https://doi.org/10.1371/journal.pgen.1005176 (2015).
50. Wen, X., Lee, Y., Luca, F. & Pique-Regi, R. Efficient Integrative Multi-SNP Association Analysis via Deterministic Approximation of Posteriors. *Am. J. Hum. Genet.* **98**, 1114–1129, https://doi.org/10.1016/j.ajhg.2016.03.029 (2016).
51. Spain, S. L. & Barrett, J. C. Strategies for fine-mapping complex traits. *Hum. Mol. Genet.* **24**, R111–R119 (2015).

52. Manor, O. & Segal, E. GenoExp: a web tool for predicting gene expression levels from single nucleotide polymorphisms. *Bioinformatics* **31**, 1848–1850, https://doi.org/10.1093/bioinformatics/btv050 (2015).
53. Manor, O. & Segal, E. Robust prediction of expression differences among human individuals using only genotype information. *PLoS Genet.* **9**, e1003396 (2013).
54. Zeng, P., Zhou, X. & Huang, S. Prediction of gene expression with cis-SNPs using mixed models and regularization methods. *BMC Genomics* **18**, 368, https://doi.org/10.1186/s12864-017-3759-6 (2017).
55. Wheeler, H. E. *et al.* Survey of the Heritability and Sparse Architecture of Gene Expression Traits across Human Tissues. *PLoS Genet.* **12**, e1006423, https://doi.org/10.1371/journal.pgen.1006423 (2016).
56. Price, A. L. *et al.* Single-Tissue and Cross-Tissue Heritability of Gene Expression Via Identity-by-Descent in Related or Unrelated Individuals. *PLoS Genet.* **7**, e1001317, https://doi.org/10.1371/journal.pgen.1001317 (2011).
57. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098, https://doi.org/10.1038/ng.3367 (2015).
58. Laird, N. M. & Ware, J. H. Random-effects models for longitudinal data. *Biometrics*, 963–974 (1982).
59. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A Tool for Genome-wide Complex Trait Analysis. *Am. J. Hum. Genet.* **88**, 76–82, https://doi.org/10.1016/j.ajhg.2010.11.011 (2011).
60. Gilad, Y., Rifkin, S. A. & Pritchard, J. K. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet.* **24**, 408–415 (2008).
61. Pai, A. A., Pritchard, J. K. & Gilad, Y. The Genetic and Mechanistic Basis for Variation in Gene Regulation. *PLoS Genet.* **11**, e1004857, https://doi.org/10.1371/journal.pgen.1004857 (2015).
62. Bryois, J. *et al.* Cis and Trans Effects of Human Genomic Variants on Gene Expression. *PLoS Genet.* **10**, e1004461, https://doi.org/10.1371/journal.pgen.1004461 (2014).
63. Stram, D. O. & Lee, J. W. Variance Components Testing in the Longitudinal Mixed Effects Model. *Biometrics* **50**, 1171–1177, https://doi.org/10.2307/2533455 (1994).
64. Lin, X. Variance component testing in generalised linear models with random effects. *Biometrika* **84**, 309–326, https://doi.org/10.1093/biomet/84.2.309 (1997).
65. Chen, Y. & Liang, K. Y. On the asymptotic behaviour of the pseudolikelihood ratio test statistic with boundary problems. *Biometrika* **97**, 603–620, https://doi.org/10.1093/biomet/asq031 (2010).
66. Self, S. G. & Liang, K.-Y. Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests under Nonstandard Conditions. *J. Am. Stat. Assoc.* **82**, 605–610, https://doi.org/10.1080/01621459.1987.10478472 (1987).
67. Liang, K. Y. & Self, S. G. On the Asymptotic Behaviour of the Pseudolikelihood Ratio Test Statistic. *J. R. Stat. Soc. Ser. B.* **58**, 785–796, https://doi.org/10.1093/biomet/asq031 (1996).
68. Harville, D. A. Bayesian inference for variance components using only error contrasts. *Biometrika* **61**, 383–385, https://doi.org/10.1093/biomet/61.2.383 (1974).
69. Patterson, H. D. & Thompson, R. Recovery of interblock information when block sizes are unqual. *Biometrika* **58**, 545–555, https://doi.org/10.2307/2334389 (1971).
70. Kwee, L. C., Liu, D., Lin, X., Ghosh, D. & Epstein, M. P. A Powerful and Flexible Multilocus Association Test for Quantitative Traits. *Am. J. Hum. Genet.* **82**, 386–397, https://doi.org/10.1016/j.ajhg.2007.10.010 (2008).
71. Zhou, J. J., Hu, T., Qiao, D., Cho, M. H. & Zhou, H. Boosting Gene Mapping Power and Efficiency with Efficient Exact Variance Component Tests of SNP Sets. *Genetics*, in press, https://doi.org/10.1534/genetics.116.190454 (2016).
72. Corbeil, R. R. & Searle, S. R. Restricted Maximum Likelihood (REML) Estimation of Variance Components in the Mixed Model. *Technometrics* **18**, 31–38, https://doi.org/10.1080/00401706.1976.10489397 (1976).
73. Crainiceanu, C. M. & Ruppert, D. Likelihood ratio tests for goodness-of-fit of a nonlinear regression model. *J. Multivariate Anal.* **91**, 35–52, https://doi.org/10.1016/j.jmva.2004.04.008 (2004).
74. Crainiceanu, C. M. & Ruppert, D. Likelihood ratio tests in linear mixed models with one variance component. *J. R. Stat. Soc. Ser. B.* **66**, 165–185, https://doi.org/10.1111/j.1467-9868.2004.00438.x (2004).
75. Crainiceanu, C., Ruppert, D., Claeskens, G. & Wand, M. P. Exact likelihood ratio tests for penalised splines. *Biometrika* **92**, 91–103, https://doi.org/10.1093/biomet/92.1.91 (2005).
76. Pinheiro, J. C. & Bates, D. *Mixed-Effects Models in S and S-PLUS*. 2nd edn, (Springer, 2009).
77. Lippert, C. *et al.* Greater power and computational efficiency for kernel-based association testing of sets of genetic variants. *Bioinformatics* **30**, 3206–3214, https://doi.org/10.1093/bioinformatics/btu504 (2014).
78. Greven, S., Crainiceanu, C. M., Küchenhoff, H. & Peters, A. Restricted Likelihood Ratio Testing for Zero Variance Components in Linear Mixed Models. *J. Comput. Graph. Statist.* **17**, 870–891, https://doi.org/10.1198/106186008x386599 (2008).
79. Robinson, G. K. That blup is a good thing: the estimation of random effects. *Stat. Sci.* **6**, 15–32 (1991).
80. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252, https://doi.org/10.1038/ng.3506 (2016).
81. Duong, D. *et al.* Using genomic annotations increases statistical power to detect eGenes. *Bioinformatics* **32**, i156–i163 (2016).
82. Lowe, W. L. & Reddy, T. E. Genomic approaches for understanding the genetics of complex disease. *Genome Res.* **25**, 1432–1441, https://doi.org/10.1101/gr.190603.115 (2015).
83. Lappalainen, T. Functional genomics bridges the gap between quantitative genetics and molecular biology. *Genome Res.* **25**, 1427–1431, https://doi.org/10.1101/gr.190983.115 (2015).
84. Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* **16**, 197–212, https://doi.org/10.1038/nrg3891 (2015).
85. Stranger, B. E. *et al.* Patterns of cis regulatory variation in diverse human populations. *PLoS Genet.* **8**, e1002639, https://doi.org/10.1371/journal.pgen.1002639 (2012).
86. AC't Hoen, P. *et al.* Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nat. Biotechnol.* **31**, 1015–1022, https://doi.org/10.1038/nature12531 (2013).
87. Stegle, O., Parts, L., Durbin, R. & Winn, J. A Bayesian Framework to Account for Complex Non-Genetic Factors in Gene Expression Levels Greatly Increases Power in eQTL Studies. *PLoS Comput. Biol.* **6**, e1000770, https://doi.org/10.1371/journal.pcbi.1000770 (2010).
88. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507, https://doi.org/10.1038/nprot.2011.457 (2012).
89. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774, https://doi.org/10.1101/gr.135350.111 (2012).
90. Guan, Y. & Stephens, M. Practical Issues in Imputation-Based Association Mapping. *PLoS Genet.* **4**, e1000279, https://doi.org/10.1371/journal.pgen.1000279 (2008).
91. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65, https://doi.org/10.1038/nature11632 (2012).
92. Delaneau, O., Marchini, J. & Zagury, J. F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181, https://doi.org/10.1038/nmeth.1785 (2012).

93. Delaneau, O., Zagury, J. F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6, https://doi.org/10.1038/nmeth.2307 (2013).
94. Delaneau, O., Howie, B., Cox, A. J., Zagury, J. F. & Marchini, J. Haplotype estimation using sequencing reads. *Am. J. Hum. Genet.* **93**, 687–696, https://doi.org/10.1016/j.ajhg.2013.09.002 (2013).
95. Howie, B. N., Donnelly, P. & Marchini, J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genet.* **5**, e1000529, https://doi.org/10.1371/journal.pgen.1000529 (2009).
96. Claringbould, A., de Klein, N. & Franke, L. The genetic architecture of molecular traits. *Current Opinion in Systems Biology* **1**, 25–31, https://doi.org/10.1016/j.coisb.2017.01.002 (2017).
97. Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32**, 283–285, https://doi.org/10.1093/bioinformatics/btv546 (2016).
98. Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95**, 535–552 (2014).
99. Beck, S., Geraghty, D., Inoko, H. & Rowen, L. Complete sequence and gene map of a human major histocompatibility complex. *Nature* **401**, 921–923, https://doi.org/10.1038/44853 (1999).
100. Matzaraki, V., Kumar, V., Wijmenga, C. & Zhernakova, A. The MHC locus and genetic susceptibility to autoimmune and infectious diseases. *Genome Biol.* **18**, 76, https://doi.org/10.1186/s13059-017-1207-1 (2017).
101. Zhu, M. *et al.* Fine mapping the MHC region identified four independent variants modifying susceptibility to Chronic Hepatitis B in Han Chinese. *Hum. Mol. Genet.* **25**, 1225–1232, https://doi.org/10.1093/hmg/ddw003 (2016).
102. Zhou, F. *et al.* Deep sequencing of the MHC region in the Chinese population contributes to studies of complex disease. *Nat. Genet.* **48**, https://doi.org/10.1038/ng.3576 (2016).
103. Nica, A. C. *et al.* The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet.* **7**, e1002003, https://doi.org/10.1371/journal.pgen.1002003 (2011).
104. Dimas, A. S. *et al.* Common regulatory variation impacts gene expression in a cell type–dependent manner. *Science* **325**, 1246–1250, https://doi.org/10.1126/science.1174148 (2009).
105. Gerrits, A. Expression quantitative trait loci are highly sensitive to cellular differentiation state. *PLoS Genet.* **5**, e1000692, https://doi.org/10.1371/journal.pgen.1000692 (2009).

Acknowledgements

Our study makes use of data generated by Wellcome Trust Case Control Consortium (WTCCC). A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the WTCCC project was provided by the Wellcome Trust under award 076113 and 085475. We are also grateful to the Geuvadis consortium for making the Geuvadis gene expression data available to us. We would like to thank Dr. Xiang Zhou at the University of Michigan for great supports and the useful comments from reviewers. This work was supported by the National Natural Science Foundation of China [grant number 81402765], the Statistical Science Research Project from National Bureau of Statistics of China [grant number 2014LY112], and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD) for Xuzhou Medical University.

Author Contributions

P.Z. analyzed the simulated and real data; P.Z. drafted the manuscript; T.W. participated in the data analysis; P.Z. and S.H. conceived the idea and revised the manuscript. All authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-017-15055-8.

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.