


RESEARCH

Open Access



Incorporating genetic similarity of auxiliary samples into eGene identification under the transfer learning framework

Shuo Zhang^{1†}, Zhou Jiang^{1†} and Ping Zeng^{1,2,3,4,5,6*} 

Abstract

Background The term eGene has been applied to define a gene whose expression level is affected by at least one independent expression quantitative trait locus (eQTL). It is both theoretically and empirically important to identify eQTLs and eGenes in genomic studies. However, standard eGene detection methods generally focus on individual cis-variants and cannot efficiently leverage useful knowledge acquired from auxiliary samples into target studies.

Methods We propose a multilocus-based eGene identification method called TLegene by integrating shared genetic similarity information available from auxiliary studies under the statistical framework of transfer learning. We apply TLegene to eGene identification in ten TCGA cancers which have an explicit relevant tissue in the GTEx project, and learn genetic effect of variant in TCGA from GTEx. We also adopt TLegene to the Geuvadis project to evaluate its usefulness in non-cancer studies.

Results We observed substantial genetic effect correlation of cis-variants between TCGA and GTEx for a larger number of genes. Furthermore, consistent with the results of our simulations, we found that TLegene was more powerful than existing methods and thus identified 169 distinct candidate eGenes, which was much larger than the approach that did not consider knowledge transfer across target and auxiliary studies. Previous studies and functional enrichment analyses provided empirical evidence supporting the associations of discovered eGenes, and it also showed evidence of allelic heterogeneity of gene expression. Furthermore, TLegene identified more eGenes in Geuvadis and revealed that these eGenes were mainly enriched in cells EBV transformed lymphocytes tissue.

Conclusion Overall, TLegene represents a flexible and powerful statistical method for eGene identification through transfer learning of genetic similarity shared across auxiliary and target studies.

Keywords Transfer learning framework, Joint effect test, Hierarchical modeling, Linear mixed model, Expression quantitative trait loci, Harmonic mean *P*-value

[†]Shuo Zhang and Zhou Jiang are co-first authors.

*Correspondence:

Ping Zeng

zpstat@xzhmu.edu.cn

Full list of author information is available at the end of the article



Background

In genomic studies the term “eGene” is used to define a gene whose expression level is affected by at least one independent single nucleotide polymorphism (SNP) nearby that gene (called cis-SNP); the corresponding SNP is called expression quantitative trait locus (eQTL) [1–4]. Identification of eGenes is imperative because genes are the major molecular unit in many biological processes, are interpretable and allow for subsequent network and pathway analyses [3]. The eQTL study provides not only the set of cis-variants related to gene expression, but also the set of eGenes for which eQTLs are identified [5–7], both of which exhibit an important implication regarding functional roles of significant loci discovered in genome-wide association studies (GWAS) in influencing diseases and intermediate phenotypes [8, 9].

Further, the finding that a trait-related SNP or gene detected in GWAS is also an eQTL or eGene renders substantial evidence for causality of this variant or gene. For example, in the GWAS of inflammatory bowel disease (IBD) [10], Repnik et al. [11] utilized eQTL mapping to analyze associated loci and confirmed several genes (e.g., *SLC22A5* and *ORMDL3*) involved in the pathogenesis of IBD. By integrating genetic associations from the GWAS of major depressive disorder (MDD) and brain eQTL data, Zhong et al. [12] discovered some risk variants contributed to MDD susceptibility through affecting the expression of *FLOT1*, providing new insights into the etiology of this disorder.

The standard method of identifying eGene is to examine the association of cis-SNPs of a given gene with its expression level to assess whether any of them are significant. The permutation-based multiple testing correction is required to properly account for the linkage disequilibrium (LD) among variants, which is computationally intensive although novel improvements have been recently proposed [3, 13]. Due to the issue of multiple comparisons, a relatively small P value might not be sufficiently small to reach the significance level. Thus, the standard analysis is often underpowered for eGene detection [2, 14].

Alternatively, we consider the discovery of eGene from a statistical perspective of variant-set association analysis by evaluating the joint influence of all cis-SNPs on gene expression. SNP-set analysis has been widely employed in genomic studies [15–17], where a set of variants defined a priori within a gene or other genetic unit are analyzed collectively to examine their joint effects on a disease or phenotype, so that the power of eGene detection is acceptable even individual eQTLs cannot be detected. Therefore, compared to the standard analysis above, the SNP-set based method is expected to be more powerful to identify eGene because it aggregates multiple weakly

correlated signals and reduces the burden of multiple comparisons [15].

However, the current SNP-set based approach is likely still sub-optimal if there exists additional knowledge that is informative for eGene discovery in target samples. For example, eQTLs are shared across tissues [18, 19], it is feasible to incorporate such sharing to boost power for association analysis [20–22] or improve accuracy for gene expression prediction in a specific tissue [23, 24]. In addition, it has been demonstrated that leveraging functional genomic annotations of variants (e.g., distance from transcription start site or certain histone modification) or utilizing cross-tissue genetic similarity can also increase power of eGene and eQTL detection [1, 2].

To integrate genetic information available from external studies, we here propose a multilocus-based eGene identification method by borrowing the idea of transfer learning [25–29] as well as the idea of SNP-set association analysis [15–17]. Particularly, within the transfer learning framework, we refer to individuals under analysis as target samples, and individuals of different but closely related studies as auxiliary samples. For efficient knowledge transferring, we assume the effect of cis-SNP in the target study is analogous to that in auxiliary studies, and suppose the former could be predictive by the cis-SNP effect of auxiliary samples. Consequently, our eGene identification consists of two components: the first component represents the indirect influence of auxiliary study after transfer learning, and the second component represents the direct effect of target study. We refer to the proposed eGene identification statistical framework as TLegene. Further, even no eQTLs are discovered in target samples, significant eGenes are still likely identified due to the indirect influence of auxiliary samples. If the target samples are strongly associated to the auxiliary samples, these identified eGenes are likely biologically meaningful.

Specifically, to implement our method, we first make a novel decorrelation modification to generate two independent statistics for each of the two components [20]; then we can easily construct a unified joint test based on the two uncorrected statistics through various combination strategies including optimal weighted linear combination (TLegene-oScore), adaptive weighted linear combination (TLegene-aScore), and Fisher’s combination (TLegene-fScore). To further enhance power, we employ the recently developed harmonic mean P -value method (TLegene-HMP) [30] to aggregate the strength of the three joint test methods. Finally, we apply TLegene to eGene identification for ten TCGA cancers (Table 1) which have explicit relevant tissues available from the GTEx project [18], and learn genetic effect of SNP in TCGA from GTEx. We also adopt TLegene to the

Table 1 Descriptive statistics of the ten TCGA cancers after combining GTEx

Cancer	n_0	n_1	m_0	m_1	Age	Female/male	Stage/grade (1/2/3/4/5)	Tissue in GTEx	n_2	m_2	k_0	k_1
ACC	97	75	9,473,821	4,592,516	47.6 ± 16.5	50/25	8/33/16/18/0	Adrenal gland	175	8,886,529	7371	6897
BRCA	1283	736	10,639,477	2,211,750	58.8 ± 13.0	736/0	138/408/174/11/5	Breast mammary	251	8,886,322	7714	4236
COAD	570	201	16,825,606	3,873,035	66.0 ± 13.0	97/104	34/78/62/27/0	Colon transverse	246	8,879,795	7716	6221
LIHC	469	166	12,269,510	3,359,173	62.7 ± 14.1	73/93	79/44/39/4/0	Liver	153	8,871,933	7143	5318
LUAD	577	384	15,199,217	3,607,888	65.9 ± 9.9	213/171	216/90/62/16/0	Lung	383	8,853,514	7781	666
LUSC	765	344	11,830,021	3,130,896	67.1 ± 8.8	94/250	176/117/48/3/0	Lung	383	8,835,520	7781	5464
OV	758	455	5,635,755	1,373,814	60.2 ± 11.4	455/0	9/19/353/74/0	Ovary	122	8,798,367	7499	410
PAAD	223	159	14,099,808	4,410,547	65.6 ± 10.8	69/90	20/130/4/5/0	Pancreas	220	8,764,053	7237	6709
STAD	544	249	12,286,585	3,183,096	64.8 ± 10.2	97/152	33/75/128/13/0	Stomach	237	8,700,105	7586	5482
UCEC	605	368	17,131,516	3,747,929	64.4 ± 10.8	368/0	239/33/79/17/0	Uterus	101	8,886,529	7532	6642

n_0 : the initial sample size in TCGA; n_1 : the sample size after quality control; n_2 : the sample size in GTEx; m_0 : the initial number of SNPs in TCGA; m_1 : the number of shared SNPs between TCGA and GTEx; m_2 : the initial number of SNPs in GTEx; k_0 : the number of genes after combination; k_1 : the number of genes after quality control; ACC: adrenocortical cancer; BRCA: breast cancer; COAD: colon cancer; LIHC: liver cancer; LUAD: lung adenocarcinoma; LUSC: lung squamous cell carcinoma; OV: ovarian cancer; PAAD: pancreatic cancer; STAD: stomach cancer; UCEC: endometrial cancer

Geuvadis project [31] to further evaluate its usefulness in non-cancer studies. Overall, in line with the simulations, we identified more candidate eGenes with TLegene than the methods that did not consider knowledge transfer across target and auxiliary studies, and demonstrated that TLegene was powerful in real-data applications.

Methods

SNP-set based eGene identification within linear mixed model

In TLegene we analyze one gene at each time. Suppose that there are n individuals and m cis-SNPs denoted by $\mathbf{G}=(g_1, \dots, g_m)$ for a given gene in the target study; we aim to assess whether the expression level (denoted by \mathbf{e}) of a specific gene is affected by its local variants. To examine such relation, we construct a linear mixed model [32, 33]

$$\mathbf{e} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{G}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (1)$$

where \mathbf{X} stands for the design matrix of p covariates, with $\boldsymbol{\alpha}=(\alpha_1, \dots, \alpha_p)$ the fixed effect vector; $\boldsymbol{\beta}=(\beta_1, \dots, \beta_m)$ is the random effect vector for these SNPs, $\beta_j \sim N(0, \tau)$ ($j=1, \dots, m$); and $\boldsymbol{\varepsilon}=(\varepsilon_1, \dots, \varepsilon_n)$ is the normal residual vector. Under this modeling specification, evaluating whether the focused gene is an eGene is equivalent to examining the null hypothesis $H_0: \tau=0$. The variance component-based score test (denoted by Score) is often employed as it is powerful across distinct settings [15, 20, 34].

Transfer learning via genetic similarity integration

To make efficient use of existing auxiliary data resources that are closely analogous to the target samples, we precede TLegene using transfer learning techniques [35–37]. First, let $\boldsymbol{\gamma}=(\gamma_1, \dots, \gamma_m)$ be the vector of known SNP effects obtained from summary statistics of auxiliary studies; then, we assume the genetic effect $\boldsymbol{\beta}$ in the target study can be predicted by $\boldsymbol{\gamma}$ in a given auxiliary study [20]

$$\beta_j = \gamma_j \times \theta + b_j, j = 1, \dots, m, \quad (2)$$

where θ is the indirect effect of auxiliary study, and b_j ($j=1, \dots, m$) is the direct effect not completely interpreted by auxiliary data. We still assume $b_j \sim N(0, \tau)$. Here, we are attempting to learn $\boldsymbol{\beta}$ based on auxiliary effect estimate. Finally, plugging (2) into (1), we obtain the TLegene model

$$\begin{aligned} \mathbf{e} &= \mathbf{X}\boldsymbol{\alpha} + \mathbf{G}(\boldsymbol{\gamma} \times \theta + \mathbf{b}) + \boldsymbol{\varepsilon}, b_j \sim N(0, \tau) \\ &= \mathbf{X}\boldsymbol{\alpha} + (\mathbf{G}\boldsymbol{\gamma}) \times \theta + \mathbf{G}\mathbf{b} + \boldsymbol{\varepsilon} \end{aligned} \quad (3)$$

where $\mathbf{G}\boldsymbol{\gamma}$ is a weighted genetic score which is also called burden component [38, 39], with θ quantifying its association with the expression level. Under the TLegene modeling framework, the null hypothesis turns into

$$H_0 : \theta = 0 \text{ and } \mathbf{b} = 0 \Leftrightarrow H_0 : \theta = 0 \text{ and } \tau = 0. \quad (4)$$

This is a joint test which requires simultaneously assessing the significance of both fixed effects and random effects: the first part of H_0 evaluates the indirect influence of auxiliary samples, whereas the second part assesses the direct impact of target samples. If $\theta=0$, model (3) reduces to $\mathbf{e} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{G}\mathbf{b} + \boldsymbol{\varepsilon}$, testing the effects of all cis-SNPs equal to zero ($H_0: \mathbf{b}=0$) is equivalent to examining the variance of \mathbf{b} equal to zero ($H_0: \tau=0$), which is a special case of the joint hypothesis test given in (4), and is particularly powerful when only the target effects are present [40]. The code for implementing the hypothesis test of TLegene is freely available at <https://github.com/biostatpzeng/TLegene>.

Joint test in TLegene

Combination of two independent score tests

We here employ the score test to examine the joint null hypothesis given in because this test method successfully avoids estimating the variance parameter under the alternative and is thus computationally efficient. We can easily obtain the respective score statistics for θ and τ under the null, with the score statistic for θ following a χ^2 distribution with one degree of freedom and the score statistic for τ following a mixture of χ^2 distribution [40]; however, the two statistics are statistically correlated if there are no any additional specifications [20–22]. Therefore, it is not straightforward to derive their joint null distribution.

To overcome this challenge, we decorrelate the two score statistics so that they could be asymptotically independent. Specifically, we first derive the score statistic of θ under H_0 (i.e., $\theta=0$ and $\tau=0$) as usual, but we next derive the score statistic of τ under the null of only $\tau=0$ without restricting $\theta=0$. By doing this, it guarantees that the two score statistics are uncorrelated [20–22]. The decorrelation greatly simplifies the derivation of joint test statistic, and independence itself offers various possibilities of aggregating the two test statistics so that we can maximize the ability to target different types of alternatives. Finally, we construct the joint test statistic by combining the two unrelated statistics through several combination methods such as optimally weighted linear combination (TLegene-oScore), adaptively weighted linear combination (TLegene-aScore), and Fisher's combination (TLegene-fScore) [20–22]; technical details of the three combination methods are given in Additional file 1.

Aggregation of three combination-based joint test methods

The three combination-based joint test methods (i.e., TLegene-oScore, TLegene-aScore, and TLegene-fScore) have distinct advantages and would show higher power under respective modeling assumptions. To circumvent

the difficulty of selecting an optimal one, we aggregate their strengths via the recently developed harmonic mean P -value method (TLegene-HMP) to generate an omnibus test [30]

$$T = 1 / \left(\sum_{t=1}^3 \frac{1}{P_t} \right), P = \int_{\frac{1}{T}}^{\infty} f_x \left(x | \log T + 0.874, \frac{\pi}{2} \right) dx \quad (5)$$

where P_t ($t=1, 2, 3$) denotes the P value yielded from each of these methods, and f_x denotes the Landau distribution probability density function. It has been demonstrated that HMP is robust against positive dependency among combined P values [30, 41].

Simulations for type I error control and power evaluation

We now carried out simulation studies to evaluate the performance of type I error control and power of TLegene. To this aim, we extracted common SNPs within an LD block of genotypes from the 1000 Genomes Project ($n=503$) [42] and the Geuvadis project ($n=465$) [31]. We randomly selected m SNPs (with m following a uniform distribution from 20 to 50) and 165 individuals from the 1000 Genomes Project to generate gene expression in the auxiliary study; among these selected SNPs, 30%, 50% or 70% were null, while the remaining had non-zero effects following a normal distribution with a mean zero and a particular variance so that the gene expression phenotypic variance explained (PVE) by SNPs would be 30% or 50%.

Second, in the target study we created gene expression using 300 individuals randomly selected from the Geuvadis project with the same set of selected SNPs, but calculated the effect as $\beta = \gamma \times \theta + b$, with b having a normal distribution with a mean zero and a variance τ . Two independent covariates were also generated (i.e., X_1 is binary and X_2 is continuous) in both target and auxiliary samples, each having an effect of 0.50. To evaluate type I error control, we set $\theta=0$ and $\tau=0$ with 10^5 replications. To evaluate power, we specified $\theta=0, 0.1, 0.2, 0.3$ or 0.4 , and $\tau=0, 0.02$ or 0.04 (with at least one of θ and τ being non-zero) with 10^3 replications.

To assess the performance of power of TLegene when there existed obvious differences in the sample sizes of the auxiliary and target studies, we performed our simulations with 400 randomly selected individuals from the 1000 Genomes Project as the auxiliary samples and 100 randomly selected individuals from the Geuvadis project as the target samples. We also conducted our simulations with 100 randomly selected individuals from the 1000 Genomes Project as the auxiliary samples and 400 randomly selected individuals from the Geuvadis project as the target samples. Other simulation settings were

analogous to those as done before, but only $\theta=0.3$ or/and $\tau=0.04$ were considered.

Real data applications

TCGA datasets and quality control

We applied TLegene to multiple TCGA cancers to identify eGenes in various tumor tissues. We only focused on cancers for which there existed an explicit tissue available from GTEx [18]; thus, TCGA was our target study and GTEx was our auxiliary study. To avoid the influence of ethnic heterogeneity, we only contained patients of European ancestry. We selected several clinical covariates such as age, gender as well as the tumor pathological stage (Table 1), because these variables could be obtained for the majority of TCGA patients [20, 43–45]. We would choose the clinical stage when the tumor pathological stage was unavailable for some tumors (e.g., OV). For each cancer we only retained samples of primary tumor tissues, and imputed missing values via the multivariate imputation by chained equation method. For genotypes of each tumor in TCGA, we performed quality control and imputation, with the details described elsewhere [20, 46]. The sample size ranged from 75 for ACC to 455 for OV, and the number of SNPs ranged from 1,373,814 for OV to 4,592,516 for ACC.

GTEx summary statistics and the alignment with TCGA

For each cancer, we obtained summary statistics data of the related tissue from GTEx (version 7) [18]. The sample size ranged from 101 for uterus to 383 for lung, and the number of SNPs ranged from 8,700,105 for liver to 8,886,529 for adrenal gland. Then, we carried out stringent quality control (Table 1): (i) reserved SNPs with $MAF > 0.05$; (ii) excluded non-biallelic SNPs and those with strand-ambiguous alleles; (iii) excluded SNPs without rs ID or removed duplicated variants; (iv) removed SNPs not in the TCGA; (v) removed SNPs whose alleles did not match those in TCGA; (vi) aligned the effect allele of SNP between TCGA and GTEx. Finally, the sample size ranged from 75 for ACC to 455 for OV, the number of shared SNPs ranged from 1,373,814 for OV to 4,592,516 for ACC, and the final number of genes included ranged from 4236 for BRCA to 6897 for ACC. The details of used datasets are described in Table 1.

Correlation of SNP effects of SNPs between GTEx and TCGA

To get an initial insight of the correlation of two types of SNP effects between TCGA and GTEx, we first generated the marginal effect of SNP in TCGA through a single-marker linear model by regressing the expression of every gene on each of its cis-SNP in TCGA while adjusting for cancer-specific covariates such as age and tumor stage [14]. Then, we performed a linear regression for

SNP effects between TCGA and GTEx to characterize their relation. For numerical stability, we only focused on genes with at least five SNPs.

Traditional method of identifying eGene

For every eGene discovered by TLegene across the cancers, we also performed the traditional linear regression for eGene identification by examining the association of each cis-SNP with the expression level. We adjusted for the same covariates as those in TLegene and applied Bonferroni's method to explain the multiple test issue.

Geuvadis project

We further applied TLegene to the Geuvadis project [31] to identify eGenes in non-cancer studies. The Geuvadis project contains gene expression measurements in lymphoblastoid cell lines for 465 individuals. Following previous work [32, 47, 48], we mainly analyzed protein-coding genes and lincRNAs defined according to GENCODE (version 12) [49]. We removed zero-count low-expressed genes in at least half of the individuals, obtaining 15,810 genes. Then, in terms of the previous

study [50], we performed PEER normalization to remove confounding effects and unwanted variation. All individuals in Geuvadis were sequenced for their genotypes in the 1000 Genomes Project. Here, the Geuvadis individuals were our target samples, and the individuals with the cells EBV transformed lymphocytes tissue from the GTEx project were our auxiliary samples. A total of 7269 genes and the number of 3,124,631 shared SNPs were finally included.

Results

Type I error control and power evaluation

First, we showed that all tests, including Score (i.e., the variance-component based score test), TLegene-oScore, TLegene-aScore, TLegene-fScore and TLegene-HMP, could maintain correct control of type I error (Fig. 1). We next compared the power of these tests under distinct alternative scenarios. To save space, here we only presented the estimated power under the scenario where the PVE in the auxiliary study was set to 0.3 or 0.5, θ (the effect of the auxiliary study) was set to 0 or 0.1, and τ (the variance of the direct effect of target study) was set to 0

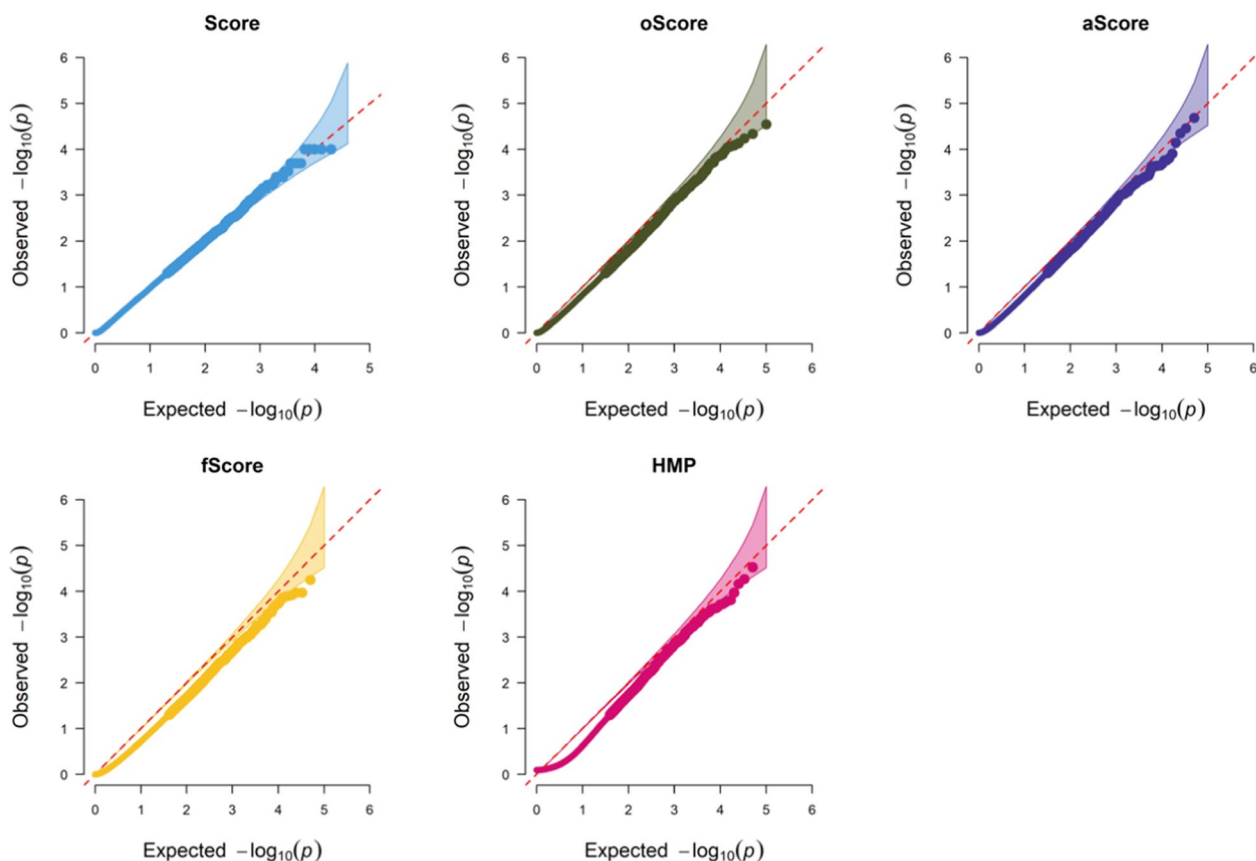


Fig. 1 QQ plots evaluating the performance of type I error control for Score, TLegene-oScore, TLegene-aScore, TLegene-fScore, and TLegene-HMP under the null in which both θ and τ were zero

or 0.02. The results for other scenarios were displayed in Additional file 1: Figs. S1–S9.

As shown in Fig. 2, when the sample size of the target study was 300 and the sample size of the auxiliary study was 165, we found that Score was powerful when only the target effect was present (e.g., $\theta=0$ and $\tau=0.02$), but suffered from power reduction when only the indirect auxiliary impact existed (e.g., $\theta=0.4$ and $\tau=0$) (Additional file 1: Fig. S6). In contrast, compared to Score, the three joint tests (i.e., TLegene-oScore, TLegene-aScore and TLegene-fScore) and the omnibus test (i.e., TLegene-HMP) were better when both the target and auxiliary effects existed (e.g., $\theta=0.3$ and $\tau=0.04$) (Additional file 1: Fig. S7). Moreover, across all scenarios of power evaluation, TLegene-HMP was more powerful or comparable compared to other methods.

When the sample sizes of the target and auxiliary studies were 100 and 400, respectively, we observed that the three combined tests and TLegene-HMP were better compared to Score when both the target and auxiliary influences existed (Additional file 1: Fig. S8). However, when the sample size of the auxiliary study was 100 and the sample size of the target study was 400, we did not observe that TLegene showed a substantially higher power compared to Score. This was likely due to the small sample size of the auxiliary study (Additional file 1: Fig. S9), which implied little learnable information from auxiliary samples to target samples and resulted in high uncertainty of auxiliary genetic effects during transfer learning.

Correlation of cis-SNP effects for each gene between TCGA and GTEx

In the real application, we first assessed the relation of cis-SNP effects for all genes between TCGA and GTEx,

with the estimated correlation summarized in Table 2. We found that the effects of the two sets of SNPs were substantially dependent for a great number of genes in each cancer. On average, most of genes (~75.4%) (ranging from 72.0% for OV to 77.5% for LUAD) had a significant regression coefficient (false discover rate [FDR] < 0.05), and a small proportion of genes (~5.2%) had a coefficient of determination (R^2) larger than 0.10, suggesting the SNP effects in GTEx did possess the ability to inform the SNP effects for some genes in TCGA.

In addition, we recognized that the regression coefficients were positive for some cancers but negative for other cancers (Fig. 3A), indicating the distinct influence of SNPs on the regulation of gene expression. Particularly, there were 96 genes whose regression coefficients were significant across all the ten cancers in TCGA (Fig. 3B). In short, the relatively high correlation of SNP

Table 2 Summary information of cis-SNPs for the ten cancers and the correlation of cis-SNP effects for each gene in TCGA and GTEx

Cancer	Median	M (%)	$R^2 > 0.10$ (%)
ACC	3096	5227 (75.5)	237 (3.4)
BRCA	2878	3191 (75.3)	265 (6.3)
COAD	3478	4724 (69.8)	283 (4.2)
LIHC	3119	4018 (73.5)	191 (3.5)
LUAD	3534	4701 (77.5)	232 (3.8)
LUSC	3313	4149 (75.9)	289 (5.3)
OV	1571	3176 (70.9)	766 (17.1)
PAAD	3560	5067 (75.5)	203 (3.0)
STAD	3397	4176 (76.2)	174 (3.2)
UCEC	3324	4663 (73.8)	137 (3.4)

Median: the median number of cis-SNPs across genes; R^2 : the determination coefficient of the cis-SNPs effects for each gene in the linear regression; M : the number of genes whose regression coefficient is significant (FDR < 0.05)

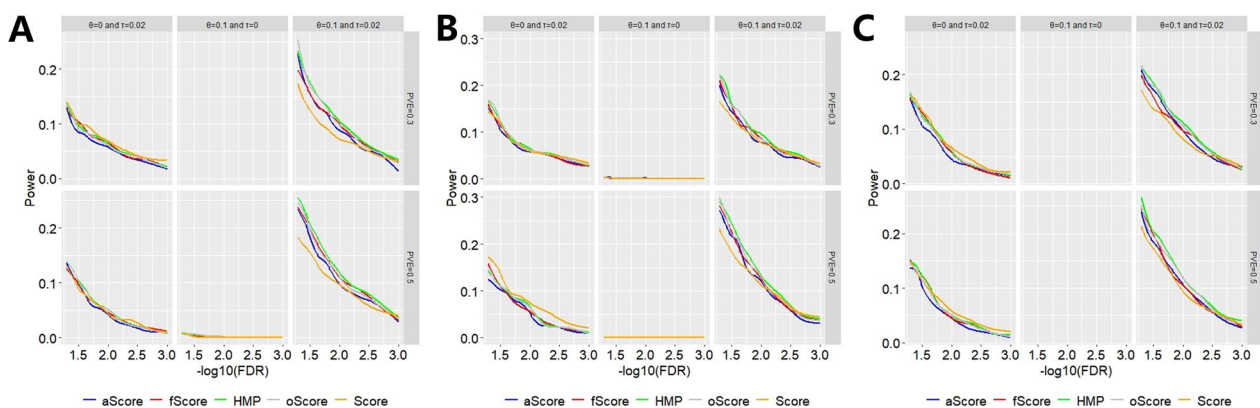


Fig. 2 Comparison of power for the five test methods under the alternative scenarios. Here, the PVE in the auxiliary study was set to 0.3 (top) or 0.5 (bottom), the sample size of the target study was 300 and the sample size of the auxiliary study was 165, $\theta=0.1$ or/and $\tau=0.02$. **A** 30% of SNPs were null; **B** 50% of SNPs were null; **C** 70% of SNPs were null

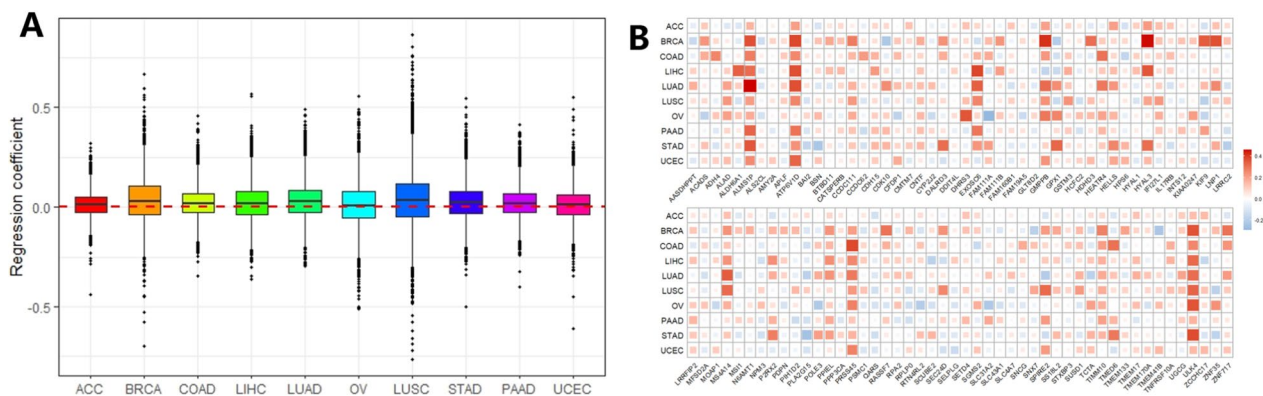


Fig. 3 **A** Distribution of regression coefficient for each gene when regressing the SNP effect in TCGA on that in GTEx. **B** Summary of 96 genes whose estimated regression coefficients simultaneously significant (FDR < 0.05) across all cancers; the magnitude of the regression coefficient is represented by the density and size of the color

effects between TCGA and GTEx demonstrated substantial similarity in expression regulation. Therefore, it was worthwhile to learn SNP effects of TCGA through GTEx to improve power for eGene identification.

Correlation evaluation in TCGA cancer

As described above, the effects of SNPs in GTEx were predictive and informative for the genetic influence of variants in TCGA; we could thus reasonably assume that smaller *P* values would be generated when implementing TLegene. Consequently, we expected higher detection rate of eGenes (*P* < 0.05) for specific genes with significant regression coefficients compared to those with insignificant ones. To validate this conjecture, for each cancer we classified all genes into four different groups by regression coefficients (whether FDR < 0.05) and the association results of the four tests constructed in TLegene (whether *P* < 0.05) (Additional file 1: Tables S1–S4).

Taking COAD as an example, there were 4724 (= 602 + 4211) genes with significant regression coefficients (FDR < 0.05), whereas 1497 (= 66 + 1431) with non-significant regression coefficients (FDR > 0.05); of these genes, the *P* values of 602 (12.7% = 602/4724) and 66 (4.4% = 66/1497) genes in TLegene-oScore were less than 0.05, indicating that TLegene-oScore had an approximate three-fold higher likelihood (2.9 = 12.7/4.4) of identifying an eGene after informative transfer learning. We further employed the χ^2 test to formally evaluate the difference in detection rate (e.g., 12.7% vs. 4.4%), and observed significant evidence in detection rate for nearly all cancers in TCGA and significant increase in detection rate for TLegene (Fig. 4).

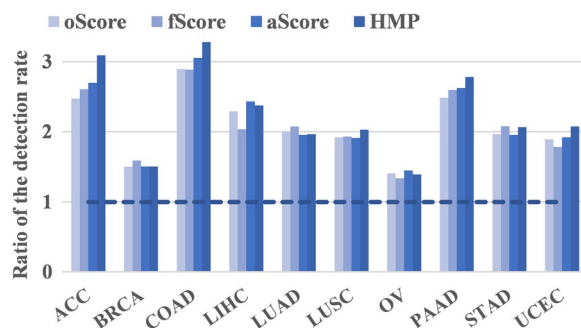


Fig. 4 Improvement in detection rate for genes with significant regression coefficients for all the ten TCGA cancers after learning SNP effect of TCGA from those of GTEx. The improvement was calculated by the ratio of the detection rate of genes with significant regression coefficients to the detection rate of genes with non-significant regression coefficients; thus, a ratio > 1 indicates an increase in detection rate

Discovered eGenes and their characteristics

Detected eGenes by TLegene

The number of eGenes discovered by TLegene is summarized in Table 3. Particularly, among these cancers only few eGenes (Bonferroni-corrected *P* < 0.05) were discovered for ACC, OV, and UCEC. Totally, 169 distinct eGenes were identified, of which 88 were identified only for one cancer, while 81 eGenes were detected for at least one type of ten TCGA cancers (Additional file 1: Table S5). Specifically, one eGene was shared by eight cancers, three eGenes were shared by seven cancers, five eGenes were shared by six cancers, six eGenes were shared by five cancers, seven eGenes were shared by four cancers, 20 eGenes were shared by three cancers, and 39 eGenes were shared by two cancers (Additional file 1: Fig. S10). Furthermore, we found that TLegene-HMP identified 325 eGenes across all the ten TCGA cancers, with

Table 3 Number of significant genes identified by TLegene for all the ten TCGA cancers

Cancer	TLegene				G_1	G_2	Linear regression (%)	PancanQTL (%)
	oScore	aScore	fScore	HMP				
ACC	2	2	2	2	2	2	2 (100)	2 (100)
BRCA	72	73	85	83	68	87	74 (85.1)	80 (92.0)
COAD	39	40	35	39	34	41	34 (82.9)	36 (87.8)
LIHC	13	13	9	12	9	13	8 (61.5)	13 (100)
LUAD	68	67	65	69	60	71	54 (76.1)	69 (97.2)
LUSC	39	40	40	40	37	42	32 (76.2)	40 (95.2)
OV	9	9	10	10	9	10	7 (70.0)	10 (100)
PAAD	42	43	40	41	36	47	38 (80.9)	41 (87.2)
STAD	21	22	24	24	21	24	20 (83.3)	22 (91.7)
UCEC	5	5	4	5	3	6	5 (83.3)	4 (66.7)
Total	310	314	314	325	279	343	274 (79.9)	317 (92.4)

G_1 : the number of eGenes simultaneously identified by all the four TLegene tests; G_2 : the number of eGenes identified by any of the four tests; the last two columns denote the number (proportion) of eGenes replicated by the traditional eGene identification method using linear regression or PancanQTL

more discoveries compared to the other TLegene tests. We also performed Score [51, 52], but failed to detect any eGenes.

For the identified eGenes of each cancer, we carried out the traditional method for eGene discovery via linear regression to identify the association of each cis-SNP in a given gene with its expression level. The results showed that on average 79.9% of the eGenes identified by TLegene could be replicated by the traditional approach (Table 3 and Additional file 1: Fig. S11).

We also tried to validate the eGenes discovered by TLegene (G_2 in Table 2) with those identified eGenes in PancanQTL [53]. PancanQTL aimed to comprehensively provide cis-eQTLs and trans-eQTLs in 33 cancer types from TCGA, which allowed us to obtain validated eGenes (http://gong_lab.hzau.edu.cn/PancanQTL/). We found that that on average 92.4% of the eGenes identified by TLegene could be repeated by PancanQTL (Table 3 and Additional file 1: Fig. S11).

We further selected only significant cis-SNPs ($FDR < 0.05$) in explicit tissues available from GTEx to estimate the correlation of SNP effects between TCGA and GTEx for these identified eGenes for each cancer. The results showed that, except a few eGenes, R^2 of eGenes estimated only with significant cis-SNPs were much higher than that of eGenes estimated with insignificant cis-SNPs. Taking ACC as an example, two eGenes (*ERAP2* and *LRRC37A2*) were identified, the R^2 was 0.3 or 0.4 estimated with only significant cis-SNPs, respectively, which was larger than the value (0.1 or 0.2) estimated with insignificant cis-SNPs (Additional file 1: Fig. S12).

Characteristics of eGenes and functional enrichment analysis

Previous studies provided evidence for some of these identified eGenes to support their connections with certain cancers, with some examples given in Additional file 1. Moreover, we performed KEGG and GO enrichment analyses for these eGenes using the clusterProfiler package [54] (Additional file 1), and identified several enriched pathways for BRCA, LUAD, LUSC, PAAD and STAD (Additional file 1: Figs. S13–S14). We also conducted functional analysis with FUMA [55]. However, except for the eGenes found in BRCA, which were enriched in whole blood tissue, as well as the eGenes detected in STAD, which were enriched in brain putamen basal ganglia tissue, we did not find that the expression levels of these identified genes were significantly enriched in the most relevant tissue for the corresponding cancer (Additional file 1: Figs. S15–S16); see Additional file 1 for more information.

Applying TLegene to Geuvadis

When applying TLegene to the Geuvadis data, we found that TLegene-HMP identified 329 eGenes, which was slightly less than the number of eGenes identified by TLegene-fScore (340), but was much more than the number of eGenes identified by TLegene-oScore (221) and TLegene-aScore (288). We also performed the score test, but only identified 27 eGenes. Furthermore, the KEGG and GO enrichment analyses and the functional analysis with FUMA [55] identified several enrichment pathways (Additional file 1: Fig. S17), and showed these eGenes were mainly enriched in cells EBV transformed lymphocytes tissue (Additional file 1: Fig. S18).

Discussions

Summary of our proposed method and real data applications

In this paper, we have proposed a powerful eGene identification method called TLegene by efficiently transferring auxiliary information into target task [35–37]. The efficient utilization of existing eQTL knowledge acquired from different but related problems can not only enhance power but also save time and cost for additional data collection. We derived two separate score test statistics for the auxiliary effect and the target effect, respectively, and carried out novel combination to construct three joint tests in TLegene; however, it is hard to know which of them are optimal in practice. We thus sought to further aggregate the advantages of these joint tests. The challenge emerged due to the non-negligible positive dependence among test statistics of the three methods because they were implemented on the same dataset with the similar logic [20, 30, 41]. The minimum P -value and permutation methods can be used but they are either computationally intensive or difficult to conduct since the correlation structure is unknown. To overcome this difficulty, we employed HMP (i.e., TLegene-HMP) to generate an omnibus test. The advantage of HMP is that it allows us to aggregate correlated P -values obtained from distinct tests into a single well-calibrated P -value without the knowledge of correlation structure while maintaining correct control of type I error [30, 41].

In our real application, by examining the genetic effects of SNP between GTEx and TCGA, we observed substantial similarity between the two types of impacts for many genes. This offered pivotal foundation for transferring knowledge acquired from GTEx to TCGA for power improvement when identifying eGenes. As expected, we revealed that leveraging this similarity could identify substantially more eGenes which were otherwise not detected if not transferring such knowledge. We also applied TLegene to the Geuvadis project, and showed that TLegene identified more eGenes and that these eGenes were mainly enriched in cells EBV transformed lymphocytes tissue, which further demonstrated the usefulness of TLegene in non-cancer studies.

However, we noted that many of eGenes identified by TLegene in the TCGA project were not enriched in the same cancer tissue. We here discussed this point from multiple aspects. First, it is important to highlight that we have shown supportive evidence for these discovered relationships between the identified eGenes and TCGA cancers (Additional file 1), and most of the eGenes could be replicated by the traditional method or PancanQTL [53]. Furthermore, we found that some eGenes not replicated by PancanQTL had been demonstrated to be likely associated with certain cancers. For example, *HLA-L*

showed strong evidence of association with lung cancer [56], *PSORS1C1* was implicated in adenocarcinoma at the gastroesophageal junction [57]. Therefore, these eGenes were still of great value for further cancer research.

Second, when applying TLegene to the Geuvadis data, we found that TLegene-HMP identified 329 eGenes and that these eGenes were mainly enriched in cells EBV transformed lymphocytes tissue, which demonstrated the usefulness of TLegene in non-cancer studies. Moreover, compared to the number of eGenes identified in Geuvadis, the number of eGenes identified in TCGA was much less, which implied that more TCGA eGenes were not discovered yet and may be a possible reason why the identified eGenes in TCGA were not enriched in cancer-specific tissues.

Third, eGenes likely play functional roles in only a small fraction of biological processes and pathways involved in cancer development and progression [58, 59], and may be also involved in other processes or pathways that are not directly related to cancers. For example, the discovered eGenes may be associated with normal tissue development, cell homeostasis, or other biological functions that indirectly promote cancer susceptibility or progress [60, 61]. Therefore, the lack of enrichment in cancer-associated tissues does not negate the potential involvement of eGenes in the cancer biology.

Comparison with previous work

TLegene distinguishes itself from previous approaches in four aspects. First, TLegene is based on a set of cis-SNPs rather than individual variants. This SNP-set based method has been shown to possess higher power in many cases compared to single-marker analysis [15].

Second, TLegene is different from Func-eGene [2] which attempted to improve the power of eGene detecting with own genomic functional annotations of variants rather than auxiliary information. In addition, Func-eGene analyzed individual cis-variants and was time-consuming; thus, it was not yet applicable for simultaneously handling a great many annotations. In contrast, integrating multiple auxiliary studies into TLegene is conceptually and practically easy.

Third, from the statistical and methodological perspective, TLegene is related to RECOV [1]. However, RECOV was proposed to utilize information shared across distinct tissues with the aim to identify eGene within any of tissues under consideration, which is different from our objective that we hope to discover eGene within one special target tissue by exploiting knowledge acquired from different but possibly related auxiliary samples. Moreover, RECOV also focused on eGene identification with individual variants.

Fourth, the most pronounced distinction is that, as far as we know, TLegene is among the first to explicitly construct the eGene identification method under the transfer learning framework [25–29, 35–37]. Actually, the pervasive genetic similarity of shared across distinct tissues [18], populations [62–69] and studies [70–78] provides solid foundation for eGene identification by exploiting existing genomic knowledge via transfer learning.

Furthermore, from a Bayesian point of view [79], the genetic overlap between the target and auxiliary studies was integrated via a prior function in TLegene. The effect similarity of many genes means informative and predictive, which would produce more accurate parameter estimation and more powerful hypothesis test method. Compared to the classical Bayesian model, the main difference of TLegene is that we adopted a probabilistic approach [80], instead of sampling techniques or variational methods [33], to estimate unknown parameters and assess the hypothesis test. In addition, in TLegene we ignored the uncertainty of the SNP effects estimated from auxiliary populations.

Future study directions

Our proposed method is not without limitations. Obviously, TLegene did not consider how to further pinpoint eQTLs within a discovered eGene. The step-down procedure can be applied to determine which cis-variants likely drive the association [81]. Indeed, using the step-down procedure after detecting eGenes, we observed that approximately 90% of discovered eGenes had only one independent eQTL and the remaining had at least two uncorrelated eQTLs, indicating allelic heterogeneity of gene expression [82]. However, the step-down procedure is not suitable for eGene if the gene only includes cis-SNPs with considerably weak effect, where no single eQTL would be picked out. We reserve the topic of further eQTL determination in TLegene as an important direction for future study. Finally, as a methodological study, our work cannot perfectly validate these discovered eGenes for various cancers although many of the eGenes detected by TLegene could be replicated by the traditional method or PanCanQTL. Therefore, our findings warranted further validations by extensive experimental studies.

Conclusions

TLegene represents a flexible and powerful statistical method for eGene identification through transfer learning of genetic similarity shared across auxiliary and target studies.

Abbreviations

GWAS Genome-wide association study

eQTL	Expression quantitative trait loci
SNP	Single nucleotide polymorphisms
GTEX	Genotype-tissue expression
FDR	False discover rate
PVE	Phenotypic variance explained
LD	Linkage disequilibrium
HMP	Harmonic mean <i>P</i> -value

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12967-024-05053-6>.

Additional file 1. Table S1. Number of genes determined by whether the *P*-values of regression coefficients and TLegene-oScore are less than 0.05. **Table S2.** Number of genes determined by whether the *P*-values of regression coefficients and TLegene-fScore are less than 0.05. **Table S3.** Number of genes determined by whether the *P*-values of regression coefficients and TLegene-aScore are less than 0.05. **Table S4.** Number of genes determined by whether the *P*-values of regression coefficients and TLegene-HMP are less than 0.05. **Table S5.** eGenes which identified by TLegene in the all ten TCGA cancers. **Figure S1–S9.** Comparison of power for the five test methods under the alternative scenarios. **Figure S10.** Upset plot represents the number of shared eGenes across the ten TCGA cancers. **Figure S11.** Bar plot represents the percentage of replicated eGenes by the traditional method using linear regression and PanCanQTL, respectively. **Figure S12.** R^2 distribution of SNP effects of eGenes identified in TCGA cancers. **Figure S13.** Result of the KEGG enrichment analysis of eGenes for COAD, LUAD and PAAD. **Figure S14.** Result of the GO enrichment analysis of eGenes for LUSC, LUAD, BRCA, COAD and PAAD. **Figure S15.** Enrichment of differentially expressed ones of all identified eGenes in terms of expression level across 54 GTEx tissues in BRCA and STAD. **Figure S16.** Enrichment of differentially expressed ones of all identified eGenes in terms of expression level across 54 GTEx tissues in COAD, LUAD, LUSC and PAAD. **Figure S17.** Result of the GO enrichment analysis of eGenes identified in the Geuvadis project. **Figure S18.** Enrichment of differentially expressed ones of all identified eGenes in terms of expression level across 54 GTEx tissues in Geuvadis.

Acknowledgements

We thank Geuvadis, TCGA and GTEx for the sharing of datasets analyzed in our work; these datasets can be available at <https://xenabrowser.net/> and <https://www.gtexportal.org/>. The data analyses in the present study were carried out with the high-performance computing cluster that was supported by the special central finance project of local universities for Xuzhou Medical University. We thank the Editor, the Associate Editor, and two anonymous Reviewers for their thorough and constructive comments which substantially improved our manuscript.

Author contributions

PZ conceived the idea for the study. PZ and SZ obtained the data. SZ and ZJ cleared up the datasets; SZ and ZJ performed the data analyses. PZ, ZJ and SZ interpreted the results of the data analyses. PZ, ZJ and SZ wrote the manuscript.

Funding

The research of Ping Zeng was supported in part by the National Natural Science Foundation of China (82173630 and 81402765), the Youth Foundation of Humanity and Social Science funded by Ministry of Education of China (18YJC910002), the Natural Science Foundation of Jiangsu Province of China (BK20181472), the China Postdoctoral Science Foundation (2018M630607 and 2019T120465), the QingLan Research Project of Jiangsu Province for Young and Middle-aged Academic Leaders, the Six-Talent Peaks Project in Jiangsu Province of China (WSN-087), and the Training Project for Youth Teams of Science and Technology Innovation at Xuzhou Medical University (TD202008). The research of Shuo Zhang was supported in part by Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX22_2960).

Availability of data and materials

All data generated or analyzed during this study are included in this published article and its Additional file.

Declarations**Ethics approval and consent to participate**

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Department of Biostatistics, School of Public Health, Xuzhou Medical University, Xuzhou 221004, Jiangsu, China. ²Center for Medical Statistics and Data Analysis, Xuzhou Medical University, Xuzhou 221004, Jiangsu, China. ³Key Laboratory of Human Genetics and Environmental Medicine, Xuzhou Medical University, Xuzhou 221004, Jiangsu, China. ⁴Key Laboratory of Environment and Health, Xuzhou Medical University, Xuzhou 221004, Jiangsu, China. ⁵Xuzhou Engineering Research Innovation Center of Biological Data Mining and Healthcare Transformation, Xuzhou Medical University, Xuzhou 221004, Jiangsu, China. ⁶Jiangsu Engineering Research Center of Biological Data Mining and Healthcare Transformation, Xuzhou Medical University, Xuzhou 221004, Jiangsu, China.

Received: 27 January 2023 Accepted: 1 March 2024

Published online: 09 March 2024

References

- Duong D, Gai L, Snir S, Kang EY, Han B, Sul JH, Eskin E. Applying meta-analysis to genotype-tissue expression data from multiple tissues to identify eQTLs and increase the number of eGenes. *Bioinformatics*. 2017;33:i67–74.
- Duong D, Zou J, Hormozdiari F, Sul JH, Ernst J, Han B, Eskin E. Using genomic annotations increases statistical power to detect eGenes. *Bioinformatics*. 2016;32:i156–63.
- Sul JH, Raj T, de Jong S, de Bakker PI, Raychaudhuri S, Ophoff RA, Stranger BE, Eskin E, Han B. Accurate and fast multiple-testing correction in eQTL studies. *Am J Hum Genet*. 2015;96:857–68.
- Derks EM, Thorp JG, Gerring ZF. Ten challenges for clinical translation in psychiatric genetics. *Nat Genet*. 2022;54:1457–65.
- Brem RB, Kruglyak L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc Natl Acad Sci USA*. 2005;102:1572–7.
- Gilad Y, Rifkin SA, Pritchard JK. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet*. 2008;24:408–15.
- The GTEx Consortium. Human genomics. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*. 2015;348:648–60.
- Manolio TA. Genomewide association studies and assessment of the risk of disease. *N Engl J Med*. 2010;363:166–76.
- Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, et al. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*. 2019;47:D1005–d1012.
- Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, Lee JC, Schumm LP, Sharma Y, Anderson CA, et al. Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*. 2012;491:119–24.
- Repnik K, Potočnik U. eQTL analysis links inflammatory bowel disease associated 1q21 locus to ECM1 gene. *J Appl Genet*. 2016;57:363–72.
- Zhong J, Li S, Zeng W, Li X, Gu C, Liu J, Luo XJ. Integration of GWAS and brain eQTL identifies FLOT1 as a risk gene for major depressive disorder. *Neuropsychopharmacology*. 2019;44:1542–51.
- Davis JR, Fresard L, Knowles DA, Pala M, Bustamante CD, Battle A, Montgomery SB. An efficient multiple-testing adjustment for eQTL studies that accounts for linkage disequilibrium between variants. *Am J Hum Genet*. 2016;98:216–24.
- Zeng P, Zhao Y, Qian C, Zhang L, Zhang R, Gou J, Liu J, Liu L, Chen F. Statistical analysis for genome-wide association study. *J Biomed Res*. 2015;29:285–97.
- Shao Z, Wang T, Qiao J, Zhang Y, Huang S, Zeng P. A comprehensive comparison of multilocus association methods with summary statistics in genome-wide association studies. *BMC Bioinform*. 2022;23:359.
- Berrandou T-E, Balding D, Speed D. LDK-GBAT: fast and powerful gene-based association testing using summary statistics. *Am J Hum Genet*. 2022;110(1):23–9.
- Li A, Liu S, Bakshi A, Jiang L, Chen W, Zheng Z, Sullivan PF, Visscher PM, Wray NR, Yang J, Zeng J. mBAT-combo: a more powerful test to detect gene-trait associations from GWAS data. *Am J Hum Genet*. 2023;110:30–43.
- The GTEx Consortium. The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science*. 2020;369:1318–30.
- Qi T, Wu Y, Zeng J, Zhang F, Xue A, Jiang L, Zhu Z, Kemper K, Yengo L, Zheng Z, et al. Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. *Nat Commun*. 2018;9:2282–2282.
- Lu H, Wei Y, Jiang Z, Zhang J, Wang T, Huang S, Zeng P. Integrative eQTL-weighted hierarchical Cox models for SNP-set based time-to-event association studies. *J Transl Med*. 2021;19:418.
- Su YR, Di C, Bien S, Huang L, Dong X, Abecasis G, Berndt S, Bezieau S, Brenner H, Caan B, et al. A mixed-effects model for powerful association tests in integrative functional genomics. *Am J Hum Genet*. 2018;102:904–19.
- Sun J, Zheng Y, Hsu L. A unified mixed-effects model for rare-variant association in sequencing studies. *Genet Epidemiol*. 2013;37:334–44.
- Zhou D, Jiang Y, Zhong X, Cox NJ, Liu C, Gamazon ER. A unified framework for joint-tissue transcriptome-wide association and Mendelian randomization analysis. *Nat Genet*. 2020;52:1239–46.
- Hu Y, Li M, Lu Q, Weng H, Wang J, Zekavat SM, Yu Z, Li B, Gu J, Muchnik S, et al. A statistical framework for cross-tissue transcriptome-wide association analysis. *Nat Genet*. 2019;51:568–76.
- Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, Xiong H, He Q. A comprehensive survey on transfer learning. *Proc IEEE*. 2021;109:43–76.
- Weiss K, Khoshgoftar TM, Wang D. A survey of transfer learning. *J Big Data*. 2016;3:9.
- Niu S, Liu Y, Wang J, Song H. A decade survey of transfer learning (2010–2020). *IEEE Trans Artif Intell*. 2020;1:151–66.
- Yang Q. Big data, lifelong machine learning and transfer learning. In: *Proceedings of the sixth ACM international conference on Web search and data mining*; Rome, Italy. Association for Computing Machinery; 2013. p. 505–6.
- Pan SJ, Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng*. 2010;22:1345–59.
- Wilson DJ. The harmonic mean p-value for combining dependent tests. *Proc Natl Acad Sci USA*. 2019;116:1195–200.
- Lappalainen T, Sammeth M, Friedländer MR, Hoen PA, Monlong J, Rivas MA, González-Porta M, Kurbatova N, Griebel T, Ferreira PG, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*. 2013;501:506–11.
- Zeng P, Zhou X, Huang S. Prediction of gene expression with cis-SNPs using mixed models and regularization methods. *BMC Genom*. 2017;18:368.
- Zeng P, Zhou X. Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. *Nat Commun*. 2017;8:456.
- Zeng P, Zhao Y, Liu J, Liu L, Zhang L, Wang T, Huang S, Chen F. Likelihood ratio tests in rare variant detection for continuous phenotypes. *Ann Hum Genet*. 2014;78:320–32.
- Bastani H. Predicting with proxies: transfer learning in high dimension. *Manag Sci*. 2021;67:2964–84.
- Li S, Cai TT, Li H. Transfer learning for high-dimensional linear regression: prediction, estimation and minimax optimality. *J R Stat Soc Ser B Stat Methodol*. 2022;84:149–73.

37. Zhao Z, Fritsche LG, Smith JA, Mukherjee B, Lee S. The construction of cross-population polygenic risk scores using transfer learning. *Am J Hum Genet.* 2022;109:1998–2008.
38. Wang T, Qiao J, Zhang S, Wei Y, Zeng P. Simultaneous test and estimation of total genetic effect in eQTL integrative analysis through mixed models. *Brief Bioinform.* 2022;23: bbac038.
39. Lee S, Abecasis Gonçalo R, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet.* 2014;95:5–23.
40. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet.* 2011;89:82–93.
41. Zeng P, Dai J, Jin S, Zhou X. Aggregating multiple expression prediction models improves the power of transcriptome-wide association studies. *Hum Mol Genet.* 2021;30:939–51.
42. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature.* 2015;526:68–74.
43. Yu X, Wang T, Huang S, Zeng P. How can gene-expression information improve prognostic prediction in TCGA cancers: an empirical comparison study on regularization and mixed Cox models. *Front Genet.* 2020;11:920.
44. Yu X, Xiao L, Zeng P, Huang S. Jackknife model averaging prediction methods for complex phenotypes with gene expression levels by integrating external pathway information. *Comput Math Methods Med.* 2019;2019:2807470.
45. Zhang J, Lu H, Zhang S, Wang T, Zhao H, Guan F, Zeng P. Leveraging methylation alterations to discover potential causal genes associated with the survival risk of cervical cancer in TCGA through a two-stage inference approach. *Front Genet.* 2021;12: 667877.
46. Gao Y, Wei Y, Zhou X, Huang S, Zhao H, Zeng P. Assessing the relationship between leukocyte telomere length and cancer risk/mortality in UK biobank and TCGA datasets with the genetic risk score and mendelian randomization approaches. *Front Genet.* 2020;11: 583106.
47. Nagpal S, Meng X, Epstein MP, Tsoi LC, Patrick M, Gibson G, De Jager PL, Bennett DA, Wingo AP, Wingo TS, Yang J. TIGAR: an improved Bayesian tool for transcriptomic data imputation enhances gene mapping of complex traits. *Am J Hum Genet.* 2019;105:258–66.
48. Zeng P, Wang T, Huang S. Cis-SNPs set testing and PrediXcan analysis for gene expression data using linear mixed models. *Sci Rep.* 2017;7:15237.
49. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res.* 2012;22:1760–74.
50. 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.* 2012;7:500–7.
51. Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet.* 2010;86:929–42.
52. Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. *Am J Hum Genet.* 2013;92:841–53.
53. Gong J, Mei S, Liu C, Xiang Y, Ye Y, Zhang Z, Feng J, Liu R, Diao L, Guo AY, et al. PancanQTL: systematic identification of cis-eQTLs and trans-eQTLs in 33 cancer types. *Nucleic Acids Res.* 2018;46:D971–d976.
54. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS.* 2012;16:284–7.
55. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun.* 2017;8:1826.
56. Zhou Y, Zhou X, Sun J, Wang L, Zhao J, Chen J, Yuan S, He Y, Timofeeva M, Spiliopoulou A, et al. Exploring the cross-cancer effect of smoking and its fingerprints in blood DNA methylation on multiple cancers: a Mendelian randomization study. *Int J Cancer.* 2023;153:1477–86.
57. Zhong C, Wu C, Lin Y, Lin D. Refined expression quantitative trait locus analysis on adenocarcinoma at the gastroesophageal junction reveals susceptibility and prognostic markers. *Front Genet.* 2023;14:1180500.
58. Jung I, Messing E. Molecular mechanisms and pathways in bladder cancer development and progression. *Cancer Control.* 2000;7:325–34.
59. Deng L, Meng T, Chen L, Wei W, Wang P. The role of ubiquitination in tumorigenesis and targeted drug discovery. *Signal Transduct Target Ther.* 2020;5:11.
60. Shi H, Xu H, Chai C, Qin Z, Zhou W. Integrated bioinformatics analysis of potential biomarkers for pancreatic cancer. *J Clin Lab Anal.* 2022;36: e24381.
61. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, et al. Integrative analysis of 111 reference human epigenomes. *Nature.* 2015;518:317–30.
62. Liu JZ, van Sommeren S, Huang HL, Ng SC, Alberts R, Takahashi A, Ripke S, Lee JC, Jostins L, Shah T, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet.* 2015;47:979–86.
63. Ikeda M, Takahashi A, Kamatani Y, Okahisa Y, Kunugi H, Mori N, Sasaki T, Ohmori T, Okamoto Y, Kawasaki H, et al. A genome-wide association study identifies two novel susceptibility loci and trans population polygenicity associated with bipolar disorder. *Mol Psychiatry.* 2018;23:639–47.
64. Bigdeli TB, Ripke S, Peterson RE, Trzaskowski M, Bacanu SA, Abdellaoui A, Andlauer TF, Beekman AT, Berger K, Blackwood DH, et al. Genetic effects influencing risk for major depressive disorder in China and Europe. *Transl Psychiatry.* 2017;7: e1074.
65. Guo J, Bakshi A, Wang Y, Jiang L, Yengo L, Goddard ME, Visscher PM, Yang J. Quantifying genetic heterogeneity between continental populations for human height and body mass index. *Sci Rep.* 2021;11:5240.
66. Brown BC, Ye CJ, Price AL, Zaitlen N, Network AGE. Transethnic genetic-correlation estimates from summary statistics. *Am J Hum Genet.* 2016;99:76–88.
67. Veturi Y, de los Campos G, Yi N, Huang W, Vazquez AI, Kühnel B. Modeling heterogeneity in the genetic architecture of ethnically diverse groups using random effect interaction models. *Genetics.* 2019;211:1395–407.
68. Lu HJ, Wang T, Zhang JH, Zhang SO, Huang SP, Zeng P. Evaluating marginal genetic correlation of associated loci for complex diseases and traits between European and East Asian populations. *Hum Genet.* 2021;140:1285–97.
69. Lam M, Chen C-Y, Li Z, Martin AR, Bryois J, Ma X, Gaspar H, Ikeda M, Benyamin B, Brown BC. Comparative genetic architectures of schizophrenia in East Asian and European populations. *Nat Genet.* 2019;51:1670–8.
70. Carlson CS, Matisse TC, North KE, Haiman CA, Fesinmeyer MD, Buyske S, Schumacher FR, Peters U, Franceschini N, Ritchie MD, et al. Generalization and dilution of association results from European GWAS in populations of non-European ancestry: the PAGE study. *PLoS Biol.* 2013;11: e1001661.
71. Waters KM, Stram DO, Hassanein MT, Le Marchand L, Wilkens LR, Maskarinec G, Monroe KR, Kolonel LN, Altschuler D, Henderson BE, Haiman CA. Consistent association of type 2 diabetes risk variants found in Europeans in diverse racial and ethnic groups. *PLoS Genet.* 2010;6: e1001078.
72. Hindorf LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA.* 2009;106:9362–7.
73. Marigorta UM, Navarro A. High trans-ethnic replicability of GWAS results implies common causal variants. *PLoS Genet.* 2013;9: e1003566.
74. Kraft P, Zeggini E, Ioannidis JP. Replication in genome-wide association studies. *Stat Sci.* 2009;24:561–73.
75. Li Y, Lan L, Wang Y, Yang C, Tang W, Cui G, Luo S, Cheng Y, Liu Y, Liu J, Jin Y. Extremely cold and hot temperatures increase the risk of diabetes mortality in metropolitan areas of two Chinese cities. *Environ Res.* 2014;134:91–7.
76. Gurdasani D, Barroso I, Zeggini E, Sandhu MS. Genomics of disease risk in globally diverse populations. *Nat Rev Genet.* 2019;20:520–35.
77. Chen J, Spracklen CN, Marenne G, Varshney A, Corbin LJ, Luan J, Willems SM, Wu Y, Zhang X, Horikoshi M, et al. The trans-ancestral genomic architecture of glycemic traits. *Nat Genet.* 2021;53:840–60.
78. de Candia TR, Lee SH, Yang J, Browning BL, Gejman PV, Levinson DF, Mowry BJ, Hewitt JK, Goddard ME, O'Donovan MC, et al. Additive genetic variation in schizophrenia risk is shared by populations of African and European descent. *Am J Hum Genet.* 2013;93:463–70.
79. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. 2nd ed. New York: Chapman and Hall; 2003.

80. Yuan Z, Zhu H, Zeng P, Yang S, Sun S, Yang C, Liu J, Zhou X. Testing and controlling for horizontal pleiotropy with probabilistic Mendelian randomization in transcriptome-wide association studies. *Nat Commun.* 2020;11:3861.
81. Sun R, Hui S, Bader GD, Lin X, Kraft P. Powerful gene set analysis in GWAS with the generalized Berk-Jones statistic. *PLoS Genet.* 2019;15: e1007530.
82. Jansen R, Hottenga J-J, Nivard MG, Abdellaoui A, Laport B, de Geus EJ, Wright FA, Penninx BWJH, Boomsma DI. Conditional eQTL analysis reveals allelic heterogeneity of gene expression. *Hum Mol Genet.* 2017;26:1444–51.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.