Published in final edited form as:

Genet Epidemiol. 2020 November ; 44(8): 798–810. doi:10.1002/gepi.22341.

An enhanced machine learning tool for *cis*-eQTL mapping with regularization and confounder adjustments

Kang K. Yan¹, Hongyu Zhao², Joseph T. Wu¹, Herbert Pang¹

¹School of Public Health, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, China

²Department of Biostatistics, Yale University, New Haven, Connecticut

Abstract

Many expression quantitative trait loci (eQTL) studies have been conducted to investigate the biological effects of variants in gene regulation. However, these eQTL studies may suffer from low or moderate statistical power and overly conservative false-discovery rate. In practice, most algorithms for eQTL identification do not model the joint effects of multiple genetic variants with weak or moderate influence. Here we present a novel machine-learning algorithm, lasso least-squares kernel machine (LSKM-LASSO) that model the association between multiple genetic variants and phenotypic traits simultaneously with the existence of nongenetic and genetic confounding. With a more general and flexible framework for the estimation of genetic confounding, LSKM-LASSO is able to provide a more accurate evaluation of the joint effects of multiple genetic variants. Our simulations demonstrate that our approach outperforms three state-of-the-art alternatives in terms of eQTL identification and phenotype prediction. We then apply our method to genotype and gene expression data of 11 tissues obtained from the Genotype-Tissue Expression project. Our algorithm was able to identify more genes with eQTL than other algorithms. By incorporating a regularization term and combining it with least-squares kernel machine, LSKM-LASSO provides a powerful tool for eQTL mapping and phenotype prediction.

Keywords

cis-eQTL mapping; gene expression; least-squares kernel machine; multiple variants; penalized; population structure

Correspondence Herbert Pang, School of Public Health, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, China. herbpang@hku.hk.

AUTHOR CONTRIBUTIONS

K. Y. and H. P. conceived and designed this study, obtained and interpreted the data, and drafted the manuscript. K. Y. performed the data analysis. H. Z. and J. T. W. interpreted the data and revised the manuscript. All authors read and approved the final manuscript.

CONFLICT OF INTERESTS

The authors declare that there are no conflict of interests.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

1 | INTRODUCTION

A large number of genome-wide association studies (GWAS; Ozaki et al., 2002) have been conducted to investigate the biological effects of variants in regulating disease or organismal phenotypes. However, the majority of statistically significantly associated variants identified in GWAS are located in noncoding protein regions of the genome (Manolio, 2010; Maurano et al., 2012). These variants are typically very far away from the closest gene with a distance varied from thousands to several hundreds of thousands of base pairs (Spielmann & Mundlos, 2016; Symmons et al., 2014), therefore impeding the explanation of their biological effects (Baresic, Nash, Dahoun, Howes, & Lenhard, 2019). The ENCODE project (ENCODE Project Consortium, 2012; Qu & Fang, 2013) revealed that many of the noncoding GWAS hits fall in the annotated functional region which are likely to be associated with gene regulation. Some cardiovascular disease studies also revealed the regulatory function of long noncoding RNA (Dykes & Emanueli, 2017; Ruhle & Stoll, 2016). Consequently, expression quantitative trait loci (eQTLs), the genetic association analysis of these variants and quantitative gene expression, has flourished as one of the research directions in the last decade (Nica & Dermitzakis, 2013; Rockman & Kruglyak, 2006). Generally, an eQTL analysis involves the discovery of genetic variants that explain the variations in the expression level of genes and offer promise to reveal the connections between gene expression and genetic polymorphisms. Therefore, eQTL studies provide us with an opportunity to learn the molecular and mechanisms of genetic variants in regulating the heritable phenotypes. The first achievement for eQTLs can be seen in budding yeast study, in which it was demonstrated differential gene expression could be segregated by parental genotypes (Brem, Storey, Whittle, & Kruglyak, 2005). Several eQTL studies that focused on human lymphoblastoid cell lines (LCL; Liang et al., 2013; Mackay, Stone, & Ayroles, 2009; Pickrell et al., 2010) during the past several years have identified over 7,000 strong local and distant eQTLs for 14,177 genes. More recently, Mangravite et al. (2013) discovered multiple disease-specific eQTLs that encode the rate-limiting enzyme in synthesis and statin-induced myopathy. However, current eQTL studies may have limitations like low or moderate statistical power, conservative false discovery rate and overestimated effect sizes (Huang, Ritchie, Brozynska, & Inouye, 2018).

The large number of genetic variants and that of genes (usually over 20,000 genes) in a genome-wide eQTL analysis in the human genome will undoubtedly result in a large scale of searching space. Therefore, eQTL studies that aim to seek a small number of causal variants in its high-dimensional search space yield a needle-in-a-haystack problem and bring us unique computational and statistical challenges. In practice, the general eQTL analysis in human genetics can be divided into three main categories: (a) univariate-association based, (b) Bayesian model, and (c) penalized multivariable regression, that we will discuss briefly below.

The most common one is the univariate-association testing-based methods that typically perform a correlation or regression analysis on the genetic variant and the expression of a gene. It generally assumes additive genetic effects with all variants are independent with each other, and the independence among all gene expressions. As a result, one can carry out a separate association analysis for these gene-variants pairs. A multiple testing correction

will then be used to adjust for the bias introduced by a large number of repeated correlation or regression analysis. Several tools based on univariate-association testing-based methods like matrix eQTL (Shabalin, 2012) and FastQTL (Ongen, Buil, Brown, Dermitzakis, & Delaneau, 2016) are publicly available and widely used in eQTL studies. There are several drawbacks for these eQTL methods. First, they tend to be too conservative because of the large number of gene-variants pairs in multiple testing adjustment. Second, they are unable to evaluate the joint effects of multiple genetic variants that individually have a weak or moderate impact. Third, these statistically significant variants are likely to be in linkage disequilibrium with true causal variants, rather than having a biological function themselves when an intricate linkage disequilibrium structure is presented. Another category based on Bayesian inference allows us to assign random parameters with appropriate prior distributions and to make the inferential framework more intuitive and straightforward. Typically, the Bayesian structure model used in eQTL mapping will include observed data (gene expression and variants) and a set of intermediate variables with no fixed value but particular prior distribution. The gene expression can be modelled by univariate regression or multivariable regression with appropriate distribution that accounts for linkage disequilibrium structure. Moreover, the Bayesian hierarchical model can be easily extended to cross-population (Pickrell, 2014; Wen, Lee, Luca, & Pique-Regi, 2016; Wen, Luca, & Pique-Regi, 2015) or multitissue (Li, Shabalin, Rusyn, Wright, & Nobel, 2018) setting owing to the flexibility of hierarchical structure. One downside of Bayesian inference model is that it could be computationally intensive because of the model complexity and the use of Markov Chain Monte Carlo (Scott-Boyer et al., 2012) for posterior distribution sampling. Because of the large number of variants involved in eQTL studies and the computational burden of Bayesian model, enumerating all possible combinations of these variants becomes impossible, most Bayesian-based methods assume only one causal variant per locus. Even some of these algorithms relax the constraint to three causal variants per locus, such assumption may not be realistic for real data (Albert, Bloom, Siegel, Day, & Kruglyak, 2018; Zeng et al., 2017). The last category is the penalized multivariable regression-based methods that have been proposed to account for the sparsity of genetic variants contributed to the gene expression. Lasso-based penalty term (Kim & Xing, 2012; Z. Wang, Xu, & Shi, 2014) as one of the most widely used in penalized multivariable regression-based eQTL mapping can lead to a sparse set of variants identified as eQTLs by shrinking the majority of regression coefficients towards zero. More recently, Rakitsch et al. proposed a method that combines the mixed linear model with lasso for the correction of population structure in eQTL analysis (Rakitsch, Lippert, Stegle, & Borgwardt, 2013). This class of approaches assumes that the majority of variants in the genome have negligible effects on the expression of genes. Its hypothesis is more biologically plausible than the other two categories described above. However, one drawback of lasso-based penalty is that it tends to select only one variant from a group when the group of variants share high linkage disequilibrium. Second, there are no optimal strategies for the tuning of the penalty parameter. Thus, the choice of the penalty parameters is tuned via cross-validation which is computationally expensive.

As far as we know, existing eQTL methods that consider fixed effects and genetic confounding simultaneously while evaluating the joint effects of multiple genetic variants

are very limited, and they typically adjust the genetic confounding via linear mixed model. None of them estimates the genetic confounding in a kernel smoothing way. Here, we present a novel machine-learning algorithm, a penalty-based multivariable regression model lasso least-squares kernel machine (LSKM-LASSO) for discovering the association between genetic variants and traits with the existence of nongenetic and genetic confounding. With the combination of multidimensional smoothing approach and the penalized term, our penalized multivariable regression-based method, LSKM-LASSO can provide a more general and flexible framework for the estimation of genetic confounding and thus give a more accurate evaluation of the joint effects of multiple genetic variants.

In this article, we first introduce the underlying assumptions of our proposed algorithm and then assess its performance both from the ability of eQTL identification and phenotype prediction by applying it to a series of simulated data under different scenarios and to real data retrieved from Genotype-Tissue Expression (GTEx) Consortium (GTEx Consortium, 2013, 2015; GTEx Consortium et al., 2017) in release V7. We further benchmark our algorithm LSKM-LASSO against three other state-of-the-art alternatives.

2 | MATERIALS AND METHODS

2.1 | Lasso least-squares Kernel machine

Denote the response phenotype observations vector with n observations as $\mathbf{Y} = (y_1, \dots, y_n)^T$, the matrix of fixed effects, including the intercept and other known nongenetic covariates, as $\mathbf{X}_{n \times d} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T$ where $\mathbf{x}_i = (x_{i1}, \dots, x_{id})$, and the genotype array that contains p single nucleotide polymorphisms (SNPs) as $\mathbf{G}_{n \times p} = (\mathbf{g}_1, \dots, \mathbf{g}_n)^T$ with $\mathbf{g}_i = (g_{i1}, \dots, g_{ip})$. Different encodings for $\mathbf{G}$ could be used based on the corresponding genetic structure, for example, a common and widely accepted practice to convert the alleles at an single-nucleotide polymorphism (SNP) locus in human genome systems is to count the number of minor alleles, “0” means both the two alleles at this locus are major alleles, and “2” means the two alleles are minor alleles. We assume that the phenotype trait model with an additive effect can be formulated as follows:

$$\mathbf{Y} = \underbrace{\mathbf{X}\boldsymbol{\alpha}}_{\text{fixed effects}} + \underbrace{\mathbf{G}\boldsymbol{\beta}}_{\text{genetic effects}} + \underbrace{\mathbf{F}(\mathbf{P})}_{\text{genetic confounding}} + \underbrace{\boldsymbol{\Psi}}_{\text{noise}}, \quad (1)$$

where $\boldsymbol{\alpha} \in \mathbb{R}^{d \times 1}$ and $\boldsymbol{\beta} \in \mathbb{R}^{p \times 1}$ are vectors of the corresponding regression coefficients of $\mathbf{X}$ and $\mathbf{G}$, $\mathbf{F}$ denote an unknown centered smooth function that captures the unobservable polygenic background, $\mathbf{P} = (p_1, \dots, p_n)^T$ is the population or pedigree structure corresponding to $\mathbf{Y}$ and $\boldsymbol{\Psi} \sim \mathcal{N}(0, \sigma_e^2 \mathbf{I})$ represents the independent random Gaussian noise. Equation (1) will reduce to the standard linear regression if $\mathbf{F}(\mathbf{P}) = \mathbf{0}$ and it will reduce to LSKM regression (Manolio, 2010; Suykens, 2002) when no fixed effects and genetic effects exist.

Suppose $\mathbf{K}$ is a Gram-matrix of a positive definite kernel function $(\bullet, \bullet)$, and the unknown smooth function $\mathbf{F}$ falls in the function space $\mathcal{H}_k$ generated by a kernel function k . Under Mercer’s theorem (Cortes & Vapnik, 1995), the smooth function can be expressed as a linear combination of the kernel function $k(\bullet, \bullet)$ under some regularity conditions (Liu, Lin, &

Ghosh, 2007), that is $F(P) = \sum_{i=1}^j \gamma_i k(P, p_i)$, where $\gamma = (\gamma_1, \dots, \gamma_j)^T$. In reality, it is more convenient to specify the unknown centered smooth function F with such representation, because the explicit functions or features of F could be too complex to specify. Here, we treat $F(P)$ as a random function following the mean zero Gaussian process with covariance K . The estimation of regression coefficients α , β and smooth function $F(P)$ can be worked out by maximizing the penalized-likelihood function of Equation (1).

$$\begin{aligned} \mathcal{L}(\alpha, \beta, F, \lambda) = & -\frac{1}{2}(Y - X\alpha - G\beta - F)^T(Y - X\alpha \\ & - G\beta - F) - \frac{1}{2}\lambda\|F\|_{\mathcal{H}_k}^2, \end{aligned} \quad (2)$$

where λ is a tuning parameter that deals with the weight between the goodness of fit and model complexity. Note that the entries of the Gram matrix of K is given by $K_{ij} = k(p_i, p_j)$, hence the objective function can be reformatted as

$$\begin{aligned} \mathcal{L}(\alpha, \beta, \gamma, \lambda) = & -\frac{1}{2}(Y - X\alpha - G\beta - K\gamma)^T(Y - X\alpha \\ & - G\beta - K\gamma) - \frac{1}{2}\lambda\gamma^T K \gamma. \end{aligned} \quad (3)$$

In practice, the number of SNPs can greatly exceed the number of samples when performing the eQTLs analysis, which makes it difficult to solve this likelihood function and hard to identify candidate causal SNPs. To overcome these problems, we use penalized LSKM to perform the variable selection by introducing a lasso penalty on the regression coefficients of genetic effects β . The final objective is to find a set of α , β , γ that maximize our revised likelihood function

$$\begin{aligned} \mathcal{L}(\alpha, \beta, \gamma, \lambda, \eta) = & -\frac{1}{2}(Y - X\alpha - G\beta - K\gamma)^T(Y - X\alpha \\ & - G\beta - K\gamma) - \frac{1}{2}\lambda\gamma^T K \gamma - \eta\|\beta\|_1. \end{aligned} \quad (4)$$

In reality, when we try to infer the unknown genetic confounding, the two most commonly used kernel functions are the linear kernel $k(x_i, x_j) = x_i^T x_j + c$, and the Gaussian kernel

$k(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2})$, where c and σ^2 are their tuning parameters, respectively. For

example, in the penalized linear mixed model (LMM-LASSO) proposed by Rakitsch et al. (2013), the linear kernel between the SNPs, also known as the realized relationship matrix, was adopted to correct for genetic confounding effects. And Heslot and Jannink (2015) demonstrated that the Gaussian kernel could potentially better reconstruct the real genetic architecture of the trait for genome-wide association studies (GWAS) using linear mixed model in some scenarios, as the Gaussian kernel can better control false positive results and increase power.

2.2 | Convergence criteria and parameters

Learning the regularization parameters $\Theta = \{\lambda, \eta\}$ and estimating the effect size α, β, γ jointly with the above penalized likelihood function (Equation (4)) can be expressed as a nonconvex optimization problem due to its nondifferentiability. Here we adopt an alternative strategy to optimize the objective function with given λ and η .

Step 1. For fixed β , the reduced objective function becomes differentiable, where $\tilde{Y} = Y - G\beta$.

$$\mathcal{L}(\alpha, \gamma, \lambda) = -\frac{1}{2}(\tilde{Y} - X\alpha - K\gamma)^T(\tilde{Y} - X\alpha - K\gamma) - \frac{1}{2}\lambda\gamma^T K\gamma. \quad (5)$$

Differentiating $\mathcal{L}(\alpha, \gamma, \lambda)$ with respect to α and γ , and set the corresponding value to zero, we can solve α and γ with the following formulas. Details could be found in the Supporting Information.

$$\gamma = \lambda^{-1}(I + \lambda^{-1}K)^{-1}(\tilde{Y} - X\alpha). \quad (6)$$

$$\alpha = (X^T(I + \lambda^{-1}K)^{-1}X)^{-1}X^T(I + \lambda^{-1}K)^{-1}\tilde{Y}. \quad (7)$$

Step 2. For fixed α and γ , the reduced objective function is equivalent to the cost function of a general lasso regression with response $\hat{Y} = Y - X\alpha - K\gamma$ and predictor G .

$$\mathcal{L}(\beta, \lambda) = -\frac{1}{2}(\hat{Y} - G\beta)^T(\hat{Y} - G\beta) - \eta\|\beta\|_1 \quad (8)$$

Step 3. Repeat Steps 1 and 2, until the convergence of the objective function $\mathcal{L}(\alpha, \beta, \gamma, \lambda, \eta)$. In our experiment, the convergence is reached when the absolute difference of our objective function between two successive iterations is less than a prespecified minimum acceptable tolerance, for example, $1e-5$.

In our experiment, the candidate pool of the penalty coefficient on genetic confounding λ is set to $\{0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 5, 10\}$. For each candidate value of λ , we first initialized the values of α, γ by setting the genetic effects β equal to zeros, then with new response $\hat{Y} = Y - X\alpha - K\gamma$ and predictor G , we adopted the method described in Friedman, Hastie, and Tibshirani (2010) to find the minimum value of η such that all β equal to zeros. This value of η will be denoted as η_0 , and a decreasing sequence of candidate values of η was generated with a geometric series formula $\eta_{i+1} = k\eta_i$ for $i = 0, 1, \dots, m$. According to Yang's suggestion (C. Yang et al., 2014), $k = 0.95$ and $m = 20$ were used under our experimental setting. The optimal regularization parameters $\Theta = P\{\lambda, \eta\}$ was selected by cross-validation to maximize the prediction performance.

2.3 | Simulation

To assess the performance of our algorithm in detecting true eQTLs for gene expression, we performed a series of simulations under different scenarios. The genotype data used in our simulation were obtained from the GTEx Consortium (2013) release V7, and these sequencing data were based on whole genome sequencing from 635 donors (G. T. Consortium, 2015). For each simulation experiment, we considered a random subset of the 635 donors and 100,000 SNPs as population or pedigree structure. To investigate the impact of sample size, we set the number of subjects in each subset to 200, 350, and 500, respectively. Then we randomly drew 1,000 SNPs with a minor allele frequency (MAF) 0.1 from this set. The set of 1,000 SNPs were used to simulate the corresponding gene expression level based on a model with additive genetic effects. We followed the procedures below:

$$\mathbf{y} = \sum_{i=1}^p \mathbf{x}_i \alpha_i + \sum_{j=1}^q \mathbf{g}_j \beta_j + \boldsymbol{\mu} + \boldsymbol{\varepsilon}, \quad (9)$$

where $\mathbf{x}_i$ is the i th nongenetic covariate, $\mathbf{g}_j$ is the j th SNP variate,

$$\alpha_i \sim \mathcal{N}(0, \sigma_\alpha^2), \beta_j \sim \mathcal{N}(0, \sigma_\beta^2), \boldsymbol{\mu} \sim \mathcal{N}(0, \sigma_g^2 \boldsymbol{\Sigma}), \\ \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma_e^2 \mathbf{I}),$$

$\boldsymbol{\Sigma}$ is the genetic similarity matrix which captures the population or pedigree structure through kernel functions, p is the number of nongenetic covariates, and q is the number of causal SNPs. The first two parts are additive effects of nongenetic and genetic confounding, and the latter two elements are population or pedigree structure and random noise.

We fixed $p = 2$, $\sigma_\alpha^2 = 0.5$, $\sigma_\beta^2 = 0.375$, and $\sigma_e^2 = 0.5$ under our experiment setting. The true causal SNPs are randomly sampled from the 1,000 SNPs, and $\boldsymbol{\Sigma}$ is constructed as a realized relationship matrix from the 100,000 SNPs by using the linear kernel. For the two nongenetic covariates, one is generated as a binary variable and the other one is a continuous variable. To evaluate the performance of our algorithm under different strengths of population structure and the complexity of simulated genetic models, we varied the number of causal SNPs to be 1, 5, 10, 20, 30, 40, and 50, respectively, and the ratio of variance contributed from population structure against random noise, that is $\frac{\sigma_g^2}{\sigma_e^2}$, from 0.25 to 2 by adjusting q and σ_g^2 accordingly.

2.4 | Data sets and comparisons with other top methods

We performed *cis*-eQTL analysis with real data retrieved from the GTEx Consortium (2013, 2015; GTEx Consortium et al., 2017) release V7. In the analysis, we considered fixed covariates, normalized gene expression and genotype data generated from whole-genome sequencing. The fixed covariates consist of the top three genotyping principal components, a set of covariates identified using the probabilistic estimation of expression residuals (PEER) method (Stegle, Parts, Piipari, Winn, & Durbin, 2012), genotype platform, and sex. We

followed the same procedures mentioned in GTEx documentation for *cis*-eQTL analysis. Only variants with MAF 0.01 were considered and the mapping windows was set to 1 megabase upstream and downstream of the transcription start site. In our experiment, we only carried out *cis*-eQTL analysis for human tissues with sample size greater than 150. Considering the computational burden and tissues characteristics, we further narrowed down the tissue types to those that belong to important human systems, namely the cardiovascular, endocrine, digestive, and respiratory systems. The 11 tissues are adrenal gland, artery aorta, artery coronary, artery tibial, heart atrial appendage, heart left ventricle, liver, lung, stomach, pancreas, and thyroid.

We compared the performance of our machine learning lasso least-squares kernel machine, denoted as LSKM-LASSO, with one univariate association testing algorithm and two machine learning algorithms both in simulation and 11 tissues real data. The *cis*-eQTL analysis algorithms compared include FastQTL (Ongen et al., 2016), a QTL mapping tool that performs univariate association testing in a fast and flexible manner, LMM-LASSO (Rakitsch et al., 2013), a penalized linear mixed model that allows both multilocus mapping and correction for confounding effects, and mixed random forest (Mixed RF; Stephan, Stegle, & Beyer, 2015) that simultaneously accounts for population structure and captures nonlinear genetic effects in a nonparametric way. Both FastQTL and Mixed RF have executable file or python package. We implemented LSKM-LASSO and LMM-LASSO in R 3.5.3. The R package of LSKM-LASSO is available at the website: <https://web.hku.hk/~herbpang/LassoLSKM.html>. In our experiment, we assessed the performance both from the ability of eQTL identification and phenotype prediction. The eQTL identification can be treated as a classification problem, so we compared the proportion of true eQTLs identified by alternative methods and generated the confusion matrix to calculate classification F1 score performance metric. F1 score considers precision and recall rate simultaneously as it can be regarded as the harmonic mean of precision and recall. We further examined the power of eQTL detection by checking the number of eGenes (genes whose expressions are associated with at least one eQTL) discovered through these algorithms. We explored the proportion of gene expression variance explained by all additive effects terms (R^2) and the proportion of genetic variance explained by sparse genetic effects terms (PGE; Zhou, Carbonetto, & Stephens, 2013) for phenotype prediction.

3 | RESULTS

3.1 | Identification of causal variants

We assessed the ability of our method and other methods in detecting true associations between genotype and phenotype with simulated data under different scenarios. As described in the simulation part, we mimic the genetic confounding by borrowing the key characteristics from real population structure generated with genotypes from the GTEx Consortium. FastQTL, as a univariate association testing algorithm, is unable to account for population structure. To make a fair comparison, we corrected the genetic confounding for FastQTL by calculating the top 10 principal components from the genotype data as a set of covariates. The *p* value threshold for significant association between traits and variants was set at 0.01 after false-discovering rate (FDR) correction (Benjamini & Hochberg, 1995).

SNPs with adjusted p values less than the threshold were chosen as causal variants for the corresponding trait. For Mixed RF, we treated all variants that had the highest selection frequency (Lee, Sun, Kustra, & Bull, 2008; Michaelson, Alberts, Schughart, & Beyer, 2010) as the causal variants.

To explore the influence of different genetic architectures, we considered different simulation settings by varying the number of causal SNPs, the variance contributed from population structure, and the sample size. We simulated 100 independent traits for each simulation setting. In Figure 1, we present the results of the identification of true causal variants for these algorithms. The left panel corresponds to the complexity of the trait, which is the different number of causal SNPs, and the middle panel corresponds to the magnitude of confounding due to population structure. Both σ_g^2 and σ_e^2 are fixed as 1 when the complexity of the trait is varied. Similarly, we fixed the number of causal SNPs to 50 when the magnitude of population structure was varied. From these results, we can see that our method LSKM-LASSO performed better than other methods across different scenarios. A clear pattern can be found in the left panel of Figure 1. As the complexity of traits increased (i.e., the number of causal variants increased), the number of true causal variants identified by these algorithms also increased. On average, the median percentage of true causal variants successfully identified by LSKM-LASSO under alternative complexity of the trait could reach 30%. While the median percentages of true causal variants identified for LMM-LASSO, Mixed RF, and FastQTL were approximately 8%, 5%, and 10%, respectively. However, the number of true causal variants identified by LSKM-LASSO and LMM-LASSO decreased when the magnitude population structure increased, which can be seen from the panel B of Figure 1. The right panel of Figure 1 shows the relationship between the number of true SNPs identified and the sample size. We observed that the number of true SNPs identified increased as the sample size increased. The performance of Mixed RF and FastQTL under such simulation setting was relatively stable compared with LMM-LASSO and LSKM-LASSO. In our experiments, we noted that FastQTL occasionally failed to identify causal SNPs after multiple testing correction. Thus, traits with no causal variants discovered by FastQTL were excluded when calculating its performance measurements. Another interesting phenomenon observed in our simulation is that the minimum values of boxplots for LMM-LASSO are zeros under our simulation settings. After checking the results for LMM-LASSO, we found that the causal variants identified by LMM-LASSO had a lower chance to be true causal variants. In other words, the precision rate of LMM-LASSO was low under our simulation settings.

We also generated the confusion matrix to calculate classification measurements for each trait as an alternative metric to evaluate the ability of causal variants identification for the different methods. The results of the F1 score for different complexity of traits, different magnitude of confounding due to population structure and different sample size were reported as bar charts from the left to right panel in Figure 2, respectively. The black error bar on each bar represents the 95% confidence interval by percentile bootstrap. For the F1 score presented in Figure 2, we observed the same pattern as Figure 1. The performance of LMM-LASSO and LSKM LASSO increased as the number of causal variants associated with traits increased, or the sample size increased or the magnitude population structure

decreased. However, the performance of Mixed RF and FastQTL was relatively less sensitive to the complexity of traits or the magnitude of confounding due to population structure when compared with LMM-LASSO and LSKM-LASSO. We also found that the complexity of traits had a positive effect on the variability of F1 score for these algorithms. The variance of F1 score for different algorithms decreased as the number of causal variants increased. This is indicated by the narrowing width of the 95% confidence interval as the complexity of traits increased. When the magnitude population structure increased, the variance of F1 score for LMM-LASSO and LSKM-LASSO became larger. However, its influence on F1 score for Mixed RF and FastQTL was not too prominent.

3.2 | Phenotype prediction and genetic variance explained by sparse genetic effects terms

Apart from evaluating our method in terms of the ability to detect causal variants, the ability to explain the phenotypic variation from genotype data should be assessed as well. So far, most studies predicted phenotype from genotypes through the use of linear regression or regularized regression. Since phenotype prediction is linked to heritability estimation, it is natural to incorporate population or kinship structure with linear regression or regularized regression when estimating phenotype. Several studies pointed out that the heritable component of phenotype variability could not be fully estimated from single-marker based association methods. Basically, single-marker based association approaches are conceptually inappropriate for phenotype prediction, and simply using all significant variants chosen from these univariate methods as predictors might not be a good strategy for phenotype prediction. One of the reasons could be single-marker based methods are unable to evaluate the joint effects of multiple genetic variants that individually have weak or moderate effects. Thus, we excluded FastQTL from our comparison and only compared our methods LSKM-LASSO with LMM-LASSO and Mixed RF in respect of phenotype prediction. The three methods that based on regularized LSKM, regularized linear mixed model or mixed forest regression trees are able to inspect the joint effects of causal variants.

The prediction accuracy of the three algorithms was assessed with one training and testing split for all simulated traits under different settings. We randomly splitted the simulated data into training and testing set at the ratio of 7:3. Then the corresponding model was trained with training data and evaluated on the testing set. We considered the squared correlation coefficient (R^2) that describe the proportion of phenotypes variance explained by prediction values in the testing set. Despite the fact that the numbers of predictors are different in the models derived from the three algorithms, we did not use the adjusted R^2 that adjusted for the number of predictors in a model. This is because adjusted R^2 could be negative when the number of explanatory variables is greater than the number of observations.

Figure 3 shows the box plots of R^2 for measuring phenotypes variance explained under different simulation settings. The left panel is the results for different complexity of traits, and the middle panel is the results for the different magnitude of confounding due to population structure and the right panel is the results for different sample size. Our method LSKM-LASSO was the best performer in terms of the proportion of variance explained for all scenarios. The performance of LMM-LASSO and Mixed RF were quite similar, and both

were not as good as that of LSKM-LASSO. We did not observe any clear patterns from these figures and the impact of the complexity of traits and the magnitude population structure on the performance of three algorithms are not significant. For each algorithm, we performed a paired t test for the R^2 of two successive setting and none of them was statistically significant. However, as the confounding variance from population structure increased, the variability of R^2 for LSKM-LASSO became larger. Generally speaking, the confounding variance from population structure might have a negative effect on the performance of our algorithm, LSKM-LASSO when considering the ability of causal variants identification and phenotype prediction.

3.3 | Application to GTEx data

We carried out *cis*-eQTL analysis with phenotype data and corresponding genotype data that retrieved from the GTEx consortium. The data were preprocessed with the same procedures of GTEx *cis*-eQTL analysis as described in Section 2.

Unlike simulations, we do not know the true causal variants for real data. Therefore, it is impossible to assess the power of causal variants identification by calculating classification measurements like precision, or specificity. Increasing the power of eQTL detection means we can identify more causal variants for one trait, which is equivalent to detecting more genes that have eQTLs under some situations. However, the variants that have the highest selection frequency in Mixed RF were defined as causal variants in our study. Therefore, all genes will be identified as eGenes, which means that the true-positives and false-positives will be at the extreme. Therefore, Mixed RF was excluded from this comparison. Figure 4 presents the results of eGenes detection for algorithms FastQTL, LMM-LASSO, and LSKM-LASSO. In general, tissues that have more observations would have a greater number of eGenes discovered by the three algorithms. This is in line with our perception that increasing sample size will increase the power of eQTL identification. Compared with single-marker-based association approach FastQTL, algorithms that are able to inspect the joint effects of causal variants like LMM-LASSO and LSKM-LASSO could identify more eGenes. This is especially true for our algorithm LSKM-LASSO. For tissues with smaller sample size, for example, adrenal gland, artery coronary and liver, LSKM-LASSO could identify at least 50% more eGenes than single-marker based association approach FastQTL.

We further evaluated the ability of phenotype prediction in real data by exploring the proportion of genetic variance explained by sparse genetic effects terms (PGE). PGE measures how well we can predict the trait by using causal variants alone. We did not consider R^2 for real data in our experiment. Because it reflects the ability of how well we can predict the trait by using all predictors that include nongenetics confounding and genetic variants. The estimated genetic effects and genetic confounding effects are required to calculate PGE. Mixed RF based on mixed forest regression trees could not provide the prediction value contributed from genetic variants and genetic confounding. FastQTL based on single-marker association approach is unable to yield an adequate estimation of the heritability components for phenotype variability. Thus, we did not consider Mixed RF and FastQTL for phenotype prediction comparison. The detailed results of the proportion of genetic variance explained by causal variants for GTEx data are shown in Figure 5. On

average, variants that were identified by LSKM-LASSO from eGenes could explain more genetic variance when compared with those of LMM-LASSO for each tissue. When we took the sample size for each tissue into account, we found that tissues with smaller sample size would have a more considerable variability for PGE, which can be seen from the width of 95% confidence interval shown in the error bar. Meanwhile, the difference of PGE between LMM-LASSO and LSKM-LASSO became more significant for tissues with fewer observations.

4 | DISCUSSION

In this article, we have developed a penalized LSKM-LASSO for discovering the association between genetic variants and traits in the presence of nongenetic and genetic confounding. This approach has two appealing characteristics: (a) LSKM, as a multidimensional smoothing approach, provides a more general and flexible approach for the estimation of genetic confounding and (b) penalized multimarker regression that can model the joint effects of multiple genetic variants. Thus, the key contributions of LSKM-LASSO are twofold, first the effect sizes of genetic variants, genetic or nongenetic confounding can be estimated simultaneously, and second the joint effects of causal genetic variants can be better detected in complex genetic architectures with the presence of weak and/or moderate genetic effects.

LSKM-LASSO has better performance than other existing algorithms in identifying eQTLs as demonstrated in simulations and real data. It is widely accepted that a large proportion of phenotypic variance can be explained by a number of genetic variants jointly, while the individual effects of these genetic variants could be weak. Thus, penalized regression approaches could be a natural way to estimate the joint effects of genetic variants in eQTL mapping. There is a vast amount of literatures on complex multivariable models with penalized term that have been utilized for capturing the joint effects of genetic variants on one or multiple genes (Kim, Sohn, & Xing, 2009; Kim & Xing, 2012). None of these approaches combines the LSKM that estimate the genetic confounding with a multidimensional smoothing function and penalized joint regression model simultaneously. In our experiments, we have demonstrated that LSKM-LASSO is able to detect more causal variants and more eGenes in simulation and real data. Moreover, it has a lower sensitivity when compared with other algorithms in the simulation. By accounting for nongenetic and genetic confounding simultaneously, LSKM-LASSO is able to increase the power of eQTL detection. The performance of LMM-LASSO is better than that of Mixed RF, but it not as good as FastQTL. One of the reasons is because LMM-LASSO was not able to adjust for nongenetic covariates. We should also point out that FastQTL occasionally failed to identify causal SNPs after p value correction even when there are true causal variants in our simulations. This is particularly true when the number of causal variants is low and their corresponding effect sizes are small. Algorithms that use single-variant association analysis, for example, FastQTL, can have such pitfall because of insufficient power to detect true causal variants or to detect the variants that individually have weak or moderate effects but have significant joint effects with other genetic variants (Pearson & Manolio, 2008; J. Wang et al., 2013; J. Yang et al., 2010). Our method LSKM-LASSO can mitigate the problem by modeling all genetic variants in an additive fashion to detect the joint effects. In our

simulation, we observed that the performance of LSKM-LASSO in eQTL identification decreased when the magnitude of confounding that contributed from population structure increased. However, the performance of FastQTL was not influenced by it. This may be caused by the way FastQTL deals with the genetic confounding effects. LSKM-LASSO employed a kernel function to capture the impacts of population structure, and FastQTL used the top 10 principal components calculated from genotype data to adjust for it.

Results from both simulation and real data showed that LSKM-LASSO performed better than LMM-LASSO and Mixed RF in terms of phenotype prediction. The average values for R^2 of LSKM-LASSO under different scenarios are significantly higher than the other two methods. There could be two reasons for this. First, the total variance explained by predictors can be affected by these nongenetic confounding, and existing approaches LMM-LASSO and Mixed RF are unable to adjust for nongenetic variables. Second, the variants identified by LSKM-LASSO are more likely to be true causal variants. Therefore, the variance explained by these variants potentially will be considerably higher. When applied to real data, we observed that the estimated values of PGE are quite different across tissues for the same genes. This could potentially be due to the existence of tissue-specific eQTLs and the variability of effects sizes for eQTLs across tissues. These results demonstrated that the combination of regularization term and linear squares kernel machine that accounts for both genetic confounding and nongenetic confounding would improve the prediction of phenotype from genotyping data.

Although our proposed model LSKM-LASSO performed well in a number of scenarios, the penalized LSKM does have some limitations. First of all, due to the nonrandom association of alleles at different loci, the level of collinearity among these genetic variants could be high. Consequently, the variants selected by the L_1 regularization penalty term in our model could be false-positives but share high linkage disequilibrium with true causal variants. Potentially, addressing the linkage disequilibrium structure of genetic variants in eQTL mapping could increase the accuracy of eQTL detection and reduce the false-positive rate. Second, unlike Mixed RF that assesses the genetic effects in a nonparametric way, our model is an additive linear regression model. The model assumptions will be violated when there are nonlinear correlations. However, we can still address it to some extent. For example, choosing an appropriate kernel function to reflect the nonlinear relationships and adding interaction terms of genetic variants to approximate the nonlinear correlations among genetic variants. However, this may increase the number of epistatic interactions terms among genetic variants and therefore result in enormous computational challenges and low statistical power (Crawford, Zeng, Mukherjee, & Zhou, 2017; Fish, Capra, & William, 2016). Third, the performance of LSKM-LASSO is influenced by the magnitude of confounding from population structure. Adding principal components as covariates might be helpful for this as many studies demonstrated that principal component analysis could be an effective procedure for population stratification adjustment. In summary, our analysis revealed that jointly considering the nongenetic and genetic confounding as well as the genetic variants in an additive manner can be a better way for unravelling the genotype–phenotype relationships and explaining the heritability of complex traits. At a subsequent time, we may extend our method to multitissue settings that can strengthen the power by

borrowing sharing patterns across tissues and can provide an insight into tissue-specific genetic mechanisms.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health (commonfund.nih.gov/GTEx). Additional funds were provided by the NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Donors were enrolled at Biospecimen Source Sites funded by NCI \Leidos Biomedical Research, Inc., subcontracts to the National Disease Research Interchange (10XS170), Roswell Park Cancer Institute (10XS171), and Science Care, Inc. (X10S172). The Laboratory, Data Analysis, and Coordinating Center (LDACC) was funded through a contract (HHSN268201000029C) to the The Broad Institute, Inc., Biorepository operations were funded through a Leidos Biomedical Research, Inc., subcontract to Van Andel Research Institute (10ST1035). Additional data repository and project management were provided by Leidos Biomedical Research, Inc. (HHSN26120-0800001E). The Brain Bank was supported supplements to University of Miami grant DA006227. Statistical Methods development grants were made to the University of Geneva (MH090941 and MH101814), the University of Chicago (MH090951, MH090937, MH101825, and MH101820), the University of North Carolina—Chapel Hill (MH090936), North Carolina State University (MH101819), Harvard University (MH090948), Stanford University (MH101782), Washington University (MH101810), and to the University of Pennsylvania (MH101822). The data sets used for the analyses described in this manuscript were obtained from dbGaP at <http://www.ncbi.nlm.nih.gov/gap> through dbGaP accession number phs000424.v7.p2. H. Zhao was partially supported by NIH grant R01 GM134005 and NSF grant DMS 1902903.

Funding information

National Cancer Institute

DATA AVAILABILITY STATEMENT

The single-tissue *cis*-eQTL results that support the findings of this study are publicly available at <https://www.gtexportal.org/home/datasets>. The raw genotype and gene expression data (protected) used in the conduct of this study can be request on dbGaP with accession number phs000424.v7.p2.

REFERENCES

- Albert FW, Bloom JS, Siegel J, Day L, & Kruglyak L (2018). Genetics of trans-regulatory variation in gene expression. *eLife*, 7, 7 10.7554/eLife.35471
- Baresic A, Nash AJ, Dahoun T, Howes O, & Lenhard B (2019). Understanding the genetics of neuropsychiatric disorders: The potential role of genomic regulatory blocks. *Molecular Psychiatry*, 25, 6–18. 10.1038/s41380-019-0518-x [PubMed: 31616042]
- Benjamini Y, & Hochberg Y (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1), 289–300. Retrieved from <http://www.jstor.org/stable/2346101>
- Brem RB, Storey JD, Whittle J, & Kruglyak L (2005). Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature*, 436(7051), 701–703. 10.1038/nature03865 [PubMed: 16079846]
- Cortes C, & Vapnik V (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. 10.1007/bf00994018
- Crawford L, Zeng P, Mukherjee S, & Zhou X (2017). Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. *PLOS Genetics*, 13(7), e1006869 10.1371/journal.pgen.1006869 [PubMed: 28746338]

- Dykes IM, & Emanueli C (2017). Transcriptional and post-transcriptional gene regulation by long non-coding RNA. *Genomics, Proteomics & Bioinformatics/Beijing Genomics Institute*, 15(3), 177–186. 10.1016/j.gpb.2016.12.005
- ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), 57–74. 10.1038/nature11247 [PubMed: 22955616]
- Fish AE, Capra JAB, & William S (2016). Are interactions between cis-regulatory variants evidence for biological epistasis or statistical artifacts? *American Journal of Human Genetics*, 99(4), 817–830. 10.1016/j.ajhg.2016.07.022 [PubMed: 27640306]
- Friedman J, Hastie T, & Tibshirani R (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 1–22. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/20808728> [PubMed: 20808728]
- GTEX Consortium (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6), 580–585. 10.1038/ng.2653 [PubMed: 23715323]
- GTEX Consortium (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6), 580–585. 10.1038/ng.2653 [PubMed: 23715323]
- GTEX Consortium (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235), 648–660. 10.1126/science.1262110 [PubMed: 25954001]
- GTEx Consortium, Laboratory Data Analyst & Coordinating Center, NIH Program Management, Biospecimen Collection, Pathology, ... eQTL Manuscript Working Group (2017). Genetic effects on gene expression across human tissues. *Nature*, 550(7675), 204–213. 10.1038/nature24277 [PubMed: 29022597]
- Heslot N, & Jannink JL (2015). An alternative covariance estimator to investigate genetic heterogeneity in populations. *Genetics, Selection, Evolution*, 47, 93. 10.1186/s12711-015-0171-z
- Huang QQ, Ritchie SC, Brozynska M, & Inouye M (2018). Power, false discovery rate and Winner's Curse in eQTL studies. *Nucleic Acids Research*, 46(22), e133. 10.1093/nar/gky780 [PubMed: 30189032]
- Kim S, Sohn KA, & Xing EP (2009). A multivariate regression approach to association analysis of a quantitative trait network. *Bioinformatics*, 25(12), i204–212. 10.1093/bioinformatics/btp218 [PubMed: 19477989]
- Kim S, & Xing EP (2012). Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eQTL mapping. *Annals of Applied Statistics*, 6(3), 1095–1117. 10.1214/12-AOAS549
- Lee SS, Sun L, Kustra R, & Bull SB (2008). EM-random forest and new measures of variable importance for multi-locus quantitative trait linkage analysis. *Bioinformatics*, 24(14), 1603–1610. 10.1093/bioinformatics/btn239 [PubMed: 18499695]
- Li G, Shabalín AA, Rusyn I, Wright FA, & Nobel AB (2018). An empirical Bayes approach for multiple tissue eQTL analysis. *Biostatistics*, 19(3), 391–406. 10.1093/biostatistics/kxx048 [PubMed: 29029013]
- Liang L, Morar N, Dixon AL, Lathrop GM, Abecasis GR, Moffatt MF, & Cookson WO (2013). A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. *Genome Research*, 23(4), 716–726. 10.1101/gr.142521.112 [PubMed: 23345460]
- Liu D, Lin X, & Ghosh D (2007). Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics*, 63(4), 1079–1088. 10.1111/j.1541-0420.2007.00799.x [PubMed: 18078480]
- Mackay TF, Stone EA, & Ayroles JF (2009). The genetics of quantitative traits: Challenges and prospects. *Nature Reviews Genetics*, 10(8), 565–577. 10.1038/nrg2612
- Mangravite LM, Engelhardt BE, Medina MW, Smith JD, Brown CD, Chasman DI, ... Krauss RM (2013). A statin-dependent QTL for GATM expression is associated with statin-induced myopathy. *Nature*, 502(7471), 377–380. 10.1038/nature12508 [PubMed: 23995691]
- Manolio TA (2010). Genomewide association studies and assessment of the risk of disease. *New England Journal of Medicine*, 363(2), 166–176. 10.1056/NEJMra0905980

- Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, ... Stamatoyannopoulos JA (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099), 1190–1195. 10.1126/science.1222794 [PubMed: 22955828]
- Michaelson JJ, Alberts R, Schughart K, & Beyer A (2010). Data-driven assessment of eQTL mapping methods. *BMC Genomics*, 11, 502 10.1186/1471-2164-11-502 [PubMed: 20849587]
- Nica AC, & Dermitzakis ET (2013). Expression quantitative trait loci: Present and future. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 368(1620), 20120362 10.1098/rstb.2012.0362 [PubMed: 23650636]
- Ongen H, Buil A, Brown AA, Dermitzakis ET, & Delaneau O (2016). Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics*, 32(10), 1479–1485. 10.1093/bioinformatics/btv722 [PubMed: 26708335]
- Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, ... Tanaka T (2002). Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nature Genetics*, 32(4), 650–654. 10.1038/ng1047 [PubMed: 12426569]
- Pearson TA, & Manolio TA (2008). How to interpret a genome-wide association study. *Journal of the American Medical Association*, 299(11), 1335–1344. 10.1001/jama.299.11.1335 [PubMed: 18349094]
- Pickrell JK (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *American Journal of Human Genetics*, 94(4), 559–573. 10.1016/j.ajhg.2014.03.004 [PubMed: 24702953]
- Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, ... Pritchard JK (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(7289), 768–772. 10.1038/nature08872 [PubMed: 20220758]
- Qu H, & Fang X (2013). A brief review on the Human Encyclopedia of DNA Elements (ENCODE) project. *Genomics, proteomics & bioinformatics/Beijing Genomics Institute*, 11(3), 135–141. 10.1016/j.gpb.2013.05.001
- Rakitsch B, Lippert C, Stegle O, & Borgwardt K (2013). A Lasso multi-marker mixed model for association mapping with population structure correction. *Bioinformatics*, 29(2), 206–214. 10.1093/bioinformatics/bts669 [PubMed: 23175758]
- Rockman MV, & Kruglyak L (2006). Genetics of global gene expression. *Nature Reviews Genetics*, 7(11), 862–872. 10.1038/nrg1964
- Ruhle F, & Stoll M (2016). Long non-coding RNA databases in cardiovascular research. *Genomics, proteomics & bioinformatics/Beijing Genomics Institute*, 14(4), 191–199. 10.1016/j.gpb.2016.03.001
- Scott-Boyer MP, Imholte GC, Tayeb A, Labbe A, Deschepper CF, & Gottardo R (2012). An integrated hierarchical Bayesian model for multivariate eQTL mapping. *Statistical Applications in Genetics and Molecular Biology*, 11(4), 10.1515/1544-6115.1760
- Shabalín AA (2012). Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, 28(10), 1353–1358. 10.1093/bioinformatics/bts163 [PubMed: 22492648]
- Spielmann M, & Mundlos S (2016). Looking beyond the genes: The role of non-coding variants in human disease. *Human Molecular Genetics*, 25(R2), R157–R165. 10.1093/hmg/ddw205 [PubMed: 27354350]
- Stegle O, Parts L, Piipari M, Winn J, & Durbin R (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols*, 7(3), 500–507. 10.1038/nprot.2011.457 [PubMed: 22343431]
- Stephan J, Stegle O, & Beyer A (2015). A random forest approach to capture genetic effects in the presence of population structure. *Nature Communications*, 6, 7432 10.1038/ncomms8432
- Suykens JAK (2002). *Least squares support vector machines*, River Edge, NJ: World Scientific.
- Symmons O, Uslu VV, Tsujimura T, Ruf S, Nassari S, Schwarzer W, ... Spitz F (2014). Functional and topological characteristics of mammalian regulatory domains. *Genome Research*, 24(3), 390–400. 10.1101/gr.163519.113 [PubMed: 24398455]
- Wang J, Bansal AT, Martin M, Germer S, Benayed R, Essioux L, ... Platt A (2013). Genome-wide association analysis implicates the involvement of eight loci with response to tocilizumab for the treatment of rheumatoid arthritis. *Pharmacogenomics Journal*, 13(3), 235–241. 10.1038/tpj.2012.8

- Wang Z, Xu J, & Shi X (2014). Finding alternative expression quantitative trait loci by exploring sparse model space. *Journal of Computational Biology*, 21(5), 385–393. 10.1089/cmb.2014.0026 [PubMed: 24689773]
- Wen X, Lee Y, Luca F, & Pique-Regi R (2016). Efficient integrative multi-SNP association analysis via deterministic approximation of posteriors. *American Journal of Human Genetics*, 98(6), 1114–1129. 10.1016/j.ajhg.2016.03.029 [PubMed: 27236919]
- Wen X, Luca F, & Pique-Regi R (2015). Cross-population joint analysis of eQTLs: Fine mapping and functional annotation. *PLOS Genetics*, 11(4), e1005176 10.1371/journal.pgen.1005176 [PubMed: 25906321]
- Yang C, Li C, Chen M, Chen X, Hou L, & Zhao H (2014). A penalized linear mixed model for genomic prediction using pedigree structures. *BMC Proceedings*, 8(Suppl 1), S67 10.1186/1753-6561-8-S1-S67. Genetic Analysis Workshop 18 Vanessa Olmo. [PubMed: 25519399]
- Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, ... Visscher PM (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7), 565–569. 10.1038/ng.608 [PubMed: 20562875]
- Zeng B, Lloyd-Jones LR, Holloway A, Marigorta UM, Metspalu A, Montgomery GW, ... Gibson G (2017). Constraints on eQTL fine mapping in the presence of multisite local regulation of gene expression. *G3*, 7(8), 2533–2544. 10.1534/g3.117.043752 [PubMed: 28600440]
- Zhou X, Carbonetto P, & Stephens M (2013). Polygenic modeling with bayesian sparse linear mixed models. *PLOS Genetics*, 9(2), e1003264 10.1371/journal.pgen.1003264 [PubMed: 23408905]

FIGURE 1.

The number of true SNPs identified of alternative methods on a different number of causal SNPs and magnitude population structure for simulated data sets. The proposed LSKM-LASSO is compared with the LMM-LASSO, Mixed RF, and FastQTL under a different number of causal SNPs, magnitude population structure and sample size for simulated data sets. LMM-LASSO, linear mixed model lasso; LSKM-LASSO, lasso based least-squares kernel machine; Mixed RF, mixed random forest; SNP, single-nucleotide polymorphism

FIGURE 2.

F1 score of alternative methods on a different number of causal SNPs and magnitude population structure for simulated data sets. The proposed LSKM-LASSO is compared with the LMM-LASSO, Mixed RF and FastQTL under a different number of causal SNPs, magnitude population structure and sample size for simulated data sets. LMM-LASSO, linear mixed model lasso; LSKM-LASSO, lasso based least-squares kernel machine; Mixed RF, mixed random forest; SNP, single-nucleotide polymorphism

FIGURE 3.

R^2 of alternative methods on a different number of causal SNPs and magnitude population structure for simulated data sets. The proposed LSKM-LASSO is compared with the LMM-LASSO and Mixed RF under a different number of causal SNPs, magnitude population structure and sample size for simulated data sets. LMM-LASSO, linear mixed model lasso; LSKM-LASSO, lasso based least-squares kernel machine; Mixed RF, mixed random forest; SNP, single-nucleotide polymorphism

FIGURE 4.

The number of eGenes identified across tissues of GTEx data on alternative methods. The proposed LSKM-LASSO is compared with the LMM-LASSO and FastQTL in terms of the power to detect eGenes for tissues in GTEx data. GTEx, genotype-tissue expression; LMM-LASSO, linear mixed model lasso; LSKM-LASSO, lasso based least-squares kernel machine

FIGURE 5.

The proportion of genetic variance explained by activated SNPs across tissues of GTEx data on alternative methods. The proposed LSKM-LASSO is compared with the LMM-LASSO. The proportion of genetic variance explained by activated SNPs defines the ability of how well one could predict the genetic variance with SNPs identified. GTEx, genotype-tissue expression; LMM-LASSO, linear mixed model lasso; LSKM-LASSO, lasso based least-squares kernel machine; SNP, single-nucleotide polymorphism