frontiers in Genetics

METHODS
published: 17 November 2021
doi: 10.3389/fgene.2021.690926

LORSEN: Fast and Efficient eQTL Mapping With Low Rank Penalized Regression

Cheng Gao[1], *Hairong Wei*[2] *and Kui Zhang*[1]*

[1]Department of Mathematical Sciences, Michigan Technological University, Houghton, MI, United States, [2]College of Forest Resources and Environmental Science, Michigan Technological University, Houghton, MI, United States

Characterization of genetic variations that are associated with gene expression levels is essential to understand cellular mechanisms that underline human complex traits. Expression quantitative trait loci (eQTL) mapping attempts to identify genetic variants, such as single nucleotide polymorphisms (SNPs), that affect the expression of one or more genes. With the availability of a large volume of gene expression data, it is necessary and important to develop fast and efficient statistical and computational methods to perform eQTL mapping for such large scale data. In this paper, we proposed a new method, the low rank penalized regression method (LORSEN), for eQTL mapping. We evaluated and compared the performance of LORSEN with two existing methods for eQTL mapping using extensive simulations as well as real data from the HapMap3 project. Simulation studies showed that our method outperformed two commonly used methods for eQTL mapping, LORS and FastLORS, in many scenarios in terms of area under the curve (AUC). We illustrated the usefulness of our method by applying it to SNP variants data and gene expression levels on four chromosomes from the HapMap3 Project.

Keywords: eQTL mapping, proximal gradient method, cis-eQTL, *trans*-eQTL, penalized regression

OPEN ACCESS

Edited by:
Qi Yan,
Columbia University, United States

Reviewed by:
Rong Zhang,
Amgen , United States
Chi-Yang Chiu,
University of Tennessee Health Science Center (UTHSC), United States

***Correspondence:**
Kui Zhang
kuiz@mtu.edu

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *04 April 2021*
Accepted: *08 October 2021*
Published: *17 November 2021*

Citation:
Gao C, Wei H and Zhang K (2021) LORSEN: Fast and Efficient eQTL Mapping With Low Rank Penalized Regression. Front. Genet. 12:690926. doi: 10.3389/fgene.2021.690926

1 INTRODUCTION

With rapid advancements in sequencing technologies and high-throughput technologies, a large number of single nucleotide polymorphism (SNP) data and gene expression data have become available. This allows us to investigate the associations between SNP genotypes and gene expression levels. Expression quantitative trait loci (eQTLs) are those genetic variants that can explain variation in gene expression levels and can help to elucidate the underlying genetic mechanisms of human complex traits (Albert and Kruglyak, 2015). eQTL mapping aims to identify eQTLs associated with genes of interest (Hu et al., 2015; Banerjee et al., 2021). In general, eQTLs are classified into two types: *cis*-eQTLs (or local eQTLs) and *trans*-eQTLs (or distant eQTLs) (Cookson et al., 2009). *cis*-eQTLs refer to the genetic variants that functionally act on local genes and are physically located close to the target genes. *trans*-eQTLs are those genetic variants that functionally act on distant genes residing on the same or different chromosome and are physically located far from the target genes. It is worth mentioning that *trans*-eQTLs account for a large proportion of heritability of gene expression levels, though *trans* effects are usually weaker than *cis* effects in humans (Cookson et al., 2009).

In fact, gene expression levels observed are not only regulated by genetic variants but also influenced by non-genetic factors which are known or hidden, for example, batch effects. Therefore, in eQTL mapping, how to account for confounding factors is an important issue and can influence the detection power of eQTL mapping. Up to now, a number of methods have been proposed to

account for confounding factors in eQTL mapping, for example, PANAMA (Fusi et al., 2012), PEER (Stegle et al., 2010), LORS (Yang et al., 2013), HEFT (Gao et al., 2014), LMM-EH-PS (Listgarten et al., 2010) and ECCO (Yue et al., 2020). Another challenge in eQTL mapping is that the number of SNPs involved is usually very large (Yang et al., 2013). This not only results in heavy computational burden for estimating model parameters but also generally results in reduced detection power if all SNPs are included in eQTL mapping. This is because the signal-to-noise ratio (SNR) is very low, meaning only a very small portion of SNPs that are actually associated with gene expression levels. To overcome this problem, a number of SNP screening procedures (Wang et al., 2011; Yang et al., 2013; Jeng et al., 2020) and variable selection techniques (Fan and Lv, 2008) that aim to reduce the number of SNPs and only keep informative SNPs in eQTL mapping have been developed. More importantly, a number of methods based on the penalized regression have been developed to model such sparsity of eQTLs (Lee and Xing, 2012; Yang et al., 2013; Cheng et al., 2014; Jeng et al., 2020).

LORS, a method based on the low rank sparse regression, was proposed for eQTL mapping in (Yang et al., 2013). LORS is based on a linear model with gene expression levels as response variables and SNP genotypes as predictors. To model the sparsity of regression coefficients, LORS poses the L_1 penalty on the regression coefficient matrix. In addition, LORS includes one unknown matrix with the nuclear norm penalty to account for variations caused by non-genetic factors. Yang et al. (2013) applied the coordinate descent algorithm to optimize the objective function and estimate the model parameters. A SNP screening method, called LORS-Screening, was also developed to reduce the number of SNPs involved in the subsequent joint modeling, thus reduce the computational burden greatly. Similar to LORS, FastLORS (Jeng et al., 2020) employs the same low rank sparse regression model that is used in LORS. Different from LORS, FastLORS uses generic proximal gradient algorithm to optimize the objective function and estimate the model parameters. Moreover, Jeng et al. (2020) proposed a SNP screening method based on the Higher Criticism (HC) statistic, called HC-Screening.

To improve the detection power of eQTL mapping, a number of methods have been developed to incorporate the structure information from SNP variants data and gene expression levels, for example, clustering based on gene expression levels (Kendziorski et al., 2006; Chun and Keles, 2009) and gene regulatory networks (Rakitsch and Stegle, 2016), into eQTL mapping. A number of studies have shown that such structure information from SNP variants data and gene expression levels can be effectively used in penalized regression to boost the detection power of eQTL mapping (Chen et al., 2012; Kim and Xing, 2012, 2009). For example, the graph-regularized dual lasso (GDL) proposed by (Cheng et al., 2014) can simultaneously integrate the correlation structures among SNPs and gene expression levels. Through extensive experimental evaluations, Cheng et al. (2014) showed that GDL significantly outperformed the existing method for eQTL mapping. Similar to GDL, the graph-guided fused lasso (GFlasso) proposed by (Lee and Xing, 2012) can also consider the structure of the genetic variants and the structure of the gene expression levels. As a penalized regression method, GFlasso also inherits the benefits from the group lasso. Lee and Xing (2012) showed that GFlasso was able to detect weak association signals between the genetic variants and the gene expression levels.

However, there are some drawbacks for most of the aforementioned methods. First, if two SNPs are highly correlated with each other, and one SNP is associated with some genes, but the other SNP is not associated with them, we should not expect that these two SNPs have similar coefficients for those genes. Similarly, if some SNPs are classified into one group, we should not expect that the SNPs within the same group have similar coefficients for common genes. Second, the group structures of SNP data and gene expression data are usually identified by performing clustering on the data, however, clustering is an unsupervised leaning approach, the number of clusters is usually artificially determined. When we use the resulting clusters of SNPs and gene expressions to design the penalty term, it may lead to loss of detection power and even spurious associations. Third, complicated design of penalty term in penalized regression modeling can result in untractable computational bottleneck, especially when dealing with a large volume of data.

To overcome such limitations of existing methods for eQTL mapping, we proposed a novel method, LOw Rank Sparse regression with Elastic Net penalty, abbreviated as LORSEN. Different from LORS (Yang et al., 2013) and FastLORS (Jeng et al., 2020), we applied the Elastic Net penalty to the association coefficients instead of the L_1 penalty in LORSEN. In addition, we used the low rank approximation to account for non-genetic factors in LORSEN (Yang et al., 2013). There are several advantages to use the Elastic Net penalty instead of the L_1 penalty (Tibshirani, 1996). First, when the number of SNPs p is much larger than the sample size n, theoretically, the methods based on the L_1 penalty can only yield at most n non-zero coefficients. This can lead to the substantial loss of detection power in eQTL mapping since the number of samples is generally much smaller than the number of eQTLs in gene expression studies. Second, when several eQTLs are in linkage disequilibrium (LD), the methods based on the L_1 penalty can only select one of them. In theory, the Elastic Net penalty can overcome these two drawbacks. For the estimation of the model parameters in LORSEN, we developed an efficient optimization algorithm based on the proximal gradient method (Parikh and Boyd, 2014). Our algorithm allows us to perform the eQTL mapping for a large number of SNPs and genes. We evaluated and compared the performance of LORSEN with LORS and FastLORS using extensive simulation studies as well as the HapMap3 data.

2 MATERIAL AND METHODS

2.1 Model

We assume that the genotypes for p SNPs and the gene expression levels for q genes over n samples are collected. Let X denote the $n \times p$ matrix of SNP genotypes coded in an additive manner, and

Y denote the $n \times q$ matrix of gene expression levels. To model the association between SNPs and gene expressions, we can use the following multivariate linear model as proposed in (Yang et al., 2013):

$$Y = XB + L + \mathbf{1}\mu^T + e, \tag{1}$$

where B is a $p \times q$ matrix for the regression coefficients, $\mathbf{1}$ is a n-dimensional all-ones vector, μ is a q-dimensional vector for the intercepts in the regression model, e is a $n \times q$ matrix for the error terms and each element in e has a normal distribution with zero mean and variance σ^2, all e_{ij} are independent, L is a $n \times q$ matrix which is introduced to account for variations caused by non-genetic factors.

For the convenience of description, we first introduce the following notations used in this paper. For a n-dimensional vector v with the elements $v_i (i = 1, \ldots, n)$: the L_1 norm of v is defined as $\|v\|_1 = \Sigma_{i=1}^n |v_i|$ (the sum of absolute values of the elements) and the L_2 norm (also called the Euclidean norm) of v is defined as $\|v\|_2 = \sqrt{\Sigma_{i=1}^n v_i^2}$ (the squared root of the sum of squares of the elements), respectively. For a $m \times n$ matrix M with the elements $M_{ij} (i = 1, \ldots, m; j = 1, \ldots, n)$, the Frobenius norm of M is defined as $\|M\|_F = \sqrt{\Sigma_{i=1}^m \Sigma_{j=1}^n M_{ij}^2}$ (the squared root of the sum of squares of the elements); the nuclear norm $\|M\|_* = \Sigma_{i=1}^r \sigma_i$, where $\sigma_1, \cdots, \sigma_r$ are the singular values of M and r is the rank of M; and the L_1 norm of M is defined as $\|M\|_1 = \Sigma_{i=1}^m \Sigma_{j=1}^n |M_{ij}|$ (the sum of absolute values of the elements).

In this paper, we follow the same sparsity assumptions used in (Yang et al., 2013). First, we assume that there are only a small number of non-genetic factors that influence the gene expression levels globally, not locally. Second, we assume that there are only a small fraction of SNPs that influence the gene expression levels. This assumption implies that the regression coefficient matrix B is sparse. Yang et al. (2013) proposed the following LORS procedure to estimate B, L, μ by solving the optimization problem

$$\min_{B,L,\mu} \quad \frac{1}{2}\|Y - XB - L - \mathbf{1}\mu^T\|_F^2 + \rho\|L\|_* + \lambda\|B\|_1, \tag{2}$$

where ρ and λ are regularization (tuning) parameters that control the rank of L and the sparsity of B, respectively. When L and μ are fixed, the optimization problem becomes a least absolute shrinkage and selection operator (Lasso) (Tibshirani, 1996) problem with respect to B. As pointed out in (Zou and Hastie, 2005), the Lasso has some limitations that affect its usefulness. First, when $n < p$ (the number of samples is smaller than the number of SNPs), the Lasso selects at most n SNPs. In the context of eQTL mapping, there are usually a small number of samples available. Even though the proportion of SNPs that are associated with the gene expression levels is small, it is highly likely that the number of SNPs associated with the gene expressions can still be larger than the number of samples. In this case, the L_1 penalty on B will fail to identify some SNPs that are associated with the gene expressions. Second, the Lasso tends to select only one variable among a group of highly correlated variables. This can be problematic in eQTL mapping. For example, if two SNPs are in high linkage disequilibrium and both of them are associated with gene expressions, only one SNP will be selected by the Lasso. Furthermore, if two SNPs are in high linkage disequilibrium and only one of them is associated with gene expressions, the selected SNP by the Lasso may not even be associated with gene expressions.

The use of the Elastic Net penalty (Zou and Hastie, 2005) instead of the L_1 penalty on B can overcome the limitations of the Lasso. Therefore, we propose the following optimization problem to estimate B, L, μ:

$$\min_{B,L,\mu} \quad \frac{1}{2}\|Y - XB - L - \mathbf{1}\mu^T\|_F^2 + \rho\|L\|_* + \lambda_1\|B\|_1 + \frac{\lambda_2}{2}\|B\|_F^2, \tag{3}$$

where ρ, λ_1 and λ_2 are non-negative tuning parameters. For real data sets, it is quite possible that some entries in Y are unobserved (missing). In such scenarios, the missing data will not be used in (**Eq. 3**). As used in (Yang et al., 2013), we use Ω to index the observed entries in Y. Specifically, Ω is a $n \times q$ matrix with the entry

$$\Omega_{ij} = \begin{cases} 0, & Y_{ij} \quad missing \\ 1, & otherwise. \end{cases} \tag{4}$$

Then we define the projection of a matrix A onto Ω as $\tilde{A} = P_\Omega(A) = \Omega \odot A$, where A has the same dimension as Ω and $\odot$ represents Hadamard product, that is, $\tilde{A}_{ij} = A_{ij} \times \Omega_{ij}$. Based on the observed data, the optimization problem becomes

$$\min_{B,L,\mu} \quad \frac{1}{2}\|P_\Omega(Y - XB - L - \mathbf{1}\mu^T)\|_F^2 + \rho\|L\|_* + \lambda_1\|B\|_1 + \frac{\lambda_2}{2}\|B\|_F^2. \tag{5}$$

2.2 Theory and Algorithm

To solve the optimization problem in (**Eq. 5**) efficiently, we developed a fast and efficient algorithm based on proximal gradient method (Parikh and Boyd, 2014).

We first describe the proximal gradient method for a general optimization problem

$$\min_x \quad f(x) = g(x) + h(x), \tag{6}$$

where $g(x)$ is a convex and differentiable function, $h(x)$ is a closed proper convex which means $h(x)$ is a convex function, the epigraph of $h(x)$ is closed and $h(x) < +\infty$ for at least one x and $h(x) > -\infty$ for every x. Furthermore, we assume that $\nabla g(x)$, the gradient of $g(x)$, is Lipschitz continuous with constant ℓ, which implies that $\nabla^2 g(x) \preceq \ell\mathbf{I}$. Two symmetric matrices of the same dimensions A and B have the relationship $A \preceq B$, if $B - A$ is positive semidefinite. Then we have

$$f(x) = g(x) + h(x) \leqslant g(x_0) + \langle \nabla g(x_0), x - x_0 \rangle + \frac{1}{2t}\|x - x_0\|^2 + h(x), \quad t \in (0, \frac{1}{\ell}], \tag{7}$$

where x_0 is an arbitrary point in the domain of $f(x)$ and $\langle \cdot, \cdot \rangle$ represents the inner product of two vectors. Instead of using the optimization problem (**Eq. 6**), we focus on minimizing an upper bound of the objective function, that is,

Algorithm 1 | FISTA with constant step size.

Input: $t_L, t_B, t_\mu, \widetilde{L}_1 = L_0 \in \mathbb{R}^{n\times q}, \widetilde{B}_1 = B_0 \in \mathbb{R}^{p\times q}, \widetilde{\mu}_1 = \mu_0 \in \mathbb{R}^{q\times 1}, t_1 = 1$, the maximum number of iterations N, Ω

Output: optimal feasible solutions L^*, B^*, μ^*

for $k = 1$ to N **do**

$L_k \leftarrow \mathrm{S}_{t_L\rho}(\widetilde{L}_k - t_L\Omega \odot (X\widetilde{B}_k + \mathbf{1}\widetilde{\mu}_k^T + \widetilde{L}_k - Y))$

$B_k^1 \leftarrow \widetilde{B}_k - t_B X^T(\Omega \odot (X\widetilde{B}_k + \mathbf{1}\widetilde{\mu}_k^T + L_k - Y))$

$B_k^2 \leftarrow \mathrm{sign}(\mathrm{B}_k^1) \odot (|B_k^1| - \lambda_1 J)_+$

for $j = 1$ to q **do**

$B_k[,j] \leftarrow \{1 - \frac{\lambda_2}{max\{\|B_k^2[,j]\|_2, \lambda_2\}}\} B_k^2[,j]$

end for

$\mu_k \leftarrow \widetilde{\mu}_k - t_\mu(\Omega \odot (XB_k + \mathbf{1}\widetilde{\mu}_k^T + L_k - Y))^T\mathbf{1}$

$t_{k+1} \leftarrow (1+\sqrt{1+4t_k^2})/2$

$\widetilde{L}_{k+1} \leftarrow \mathrm{L}_k + \frac{t_k-1}{t_{k+1}}(L_k - L_{k-1})$

$\widetilde{B}_{k+1} \leftarrow \mathrm{B}_k + \frac{t_k-1}{t_{k+1}}(B_k - B_{k-1})$

$\widetilde{\mu}_{k+1} \leftarrow \mu_k + \frac{t_k-1}{t_{k+1}}(\mu_k - \mu_{k-1})$

if stopping criteria is satisfied **then**

break;

end if

end for=0

$$\min_x \quad g(x_0) + \langle \nabla g(x_0), x - x_0\rangle + \frac{1}{2t}\|x - x_0\|^2 + h(x), \quad t \in (0, \frac{1}{\ell}], \tag{8}$$

which can be interpreted as an application of majorization-minimization algorithm (Parikh and Boyd, 2014). The optimization problem in (**Eq. 8**) is equivalent to the following optimization problem:

$$\min_x \quad \frac{1}{2t}\|x - (x_0 - t\nabla g(x_0))\|^2 + h(x). \tag{9}$$

Problem (**Eq. 9**) can be solved with an iterative procedure: given the value of x at the k-th iteration, i.e., x_k, the value of x at the $k+1$-th iteration, x_{k+1} can be updated by the following formula

$$\begin{aligned} x_{k+1} &= \underset{x}{argmin}\ \frac{1}{2t}\|x - (x_k - t\nabla g(x_k))\|^2 + h(x) \\ &= Prox_{t,h}(x_k - t\nabla g(x_k)), \end{aligned}$$

where $Prox(\cdot)$ is called proximal operator. The iterative process is repeated until the stopping criterion is satisfied or the maximum number of iterations is reached.

To solve the optimization problem (**Eq. 5**), we adopted an alternating optimization approach that is similar to the method in (Yang et al., 2013). Note that in the following part, t_L, t_B, and t_μ are like t used in problem (**Eq. 9**) and correspond to the variables L, B, and μ, respectively.

First, for fixed B and μ, (**Eq. 5**) becomes

$$\min_L \quad \frac{1}{2}\|Y - XB - \mathbf{1}\mu^T - L\|_F^2 + \rho\|L\|_*. \tag{10}$$

In the setting of optimization problem (**Eq. 10**), $\frac{1}{2}\|Y - XB - \mathbf{1}\mu^T - L\|_F^2$ plays the role of $g(x)$ and $\rho\|L\|_*$ plays the role of $h(x)$ in (**Eq. 6**). By Lemma 1 (**Appendix A**), at the $k+1$-th iteration, we have

$$\begin{aligned} L_{k+1} &= Prox_{t_L,\rho\|\cdot\|_*}(L_k - t_L(XB_k + \mathbf{1}\mu_k^T + L_k - Y)) \\ &= S_{t_L\rho}(L_k - t_L(XB_k + \mathbf{1}\mu_k^T + L_k - Y)), \end{aligned}$$

where $S_{t_L\rho}(\cdot)$ is the singular value shrinkage operator (please refer to the **Appendix A**), t_L is the step size which can be constant or be determined by backtracking line search.

Second, for fixed L and μ, then (**Eq. 5**) becomes

$$\min_B \quad \frac{1}{2}\|Y - XB - L - \mathbf{1}\mu^T\|_F^2 + \lambda_1\|B\|_1 + \frac{\lambda_2}{2}\|B\|_F^2, \tag{11}$$

where t_B is the step size which can be constant or be determined by backtracking line search. By Lemmas 2 and 3 and Theorem 1 (**Appendix A**), we can update B_{k+1} accordingly:

$$\begin{aligned} B_{k+1}^a &= B_k - t_B X^T(XB_k + \mathbf{1}\mu_k^T + L_{k+1} - Y) \\ B_{k+1}^b &= Prox_{t_B,\lambda_1\|\cdot\|_1}(B_{k+1}^a) \\ &= sign(B_{k+1}^a) \circledcirc (|B_{k+1}^a| - \lambda_1 J)_+ \\ B_{k+1}[,j] &= Prox_{t_B,\lambda_2\|\cdot\|_2}(B_{k+1}^b[,j]) \\ &= \{1 - \frac{\lambda_2}{max\{\|B_{k+1}^b[,j]\|_2, \lambda_2\}}\} B_{k+1}^b[,j], \quad j = 1, 2, \ldots, q, \end{aligned}$$

where J is a all-ones $p \times q$ matrix, $B[, j]$ is the j-th column of matrix B and is a p-dimensional vector, $\gamma_+ = \max\{\gamma, 0\}$, the maximum of γ and 0, $|B^a_{k+1}|$, $\text{sign}(B^a_{k+1})$, and $(|B^a_{k+1}| - \lambda_1 J)_+$ are all elementwise operations.

Third, for fixed L and B, the proximal gradient method reduces to the gradient descent method with respect to μ because there is no penalty on μ. At the $k + 1$-th iteration, we have

$$\mu_{k+1} = \mu_k - t_\mu (XB_{k+1} + \mathbf{1}\mu_k^T + L_{k+1} - Y)^T \mathbf{1}.$$

To accelerate the computational speed, we used the accelerated proximal gradient method. Specifically, we applied the fast iterative shrinkage-thresholding algorithm (FISTA) (Beck and Teboulle, 2009) which keeps the simplicity of the iterative shrinkage-thresholding algorithms (ISTA) but has an improved rate $O(1/k^2)$, where k indexes the iteration. In FISTA, the step size can be constant or be determined by backtracking line search. The algorithm to solve LORSEN with FISTA is described in Algorithm 1. For simplicity, here, only the detailed algorithm with the constant step size is described, but the algorithm using the step size determined by backtracking line search is also provided in our R program (https://github.com/gaochengPRC/LORSEN).

2.3 Parameter Tuning

For parameter tuning, we mainly followed the idea described in (Yang et al., 2013). Specifically, we divided the entries of Ω into training entries and testing entries such that training entries and testing entries include roughly the same number of 1's. We define two matrices Ω_1 and Ω_2 such that they have the same dimensions as Ω, Ω_1 contains all training entries and Ω_2 contains all testing entries. Furthermore, we have $\Omega = \Omega_1 + \Omega_2$ and $\Omega_1 \circledcirc \Omega_2 = 0$. For the consistency, we re-parameterized λ_1 and λ_2 as $\lambda \cdot \alpha$ and $\lambda \cdot (1 - \alpha)$, respectively. So the optimization problem (**Eq. 5**) becomes

$$\min_{B,L,\mu} \quad \frac{1}{2}\|Y - XB - L - \mathbf{1}\mu^T\|_F^2 + \rho\|L\|_* + \lambda\left(\alpha\|B\|_1 + \frac{1-\alpha}{2}\|B\|_F^2\right). \tag{12}$$

This form is the same as that in glmnet (Friedman et al., 2010).

Given the values of parameters (ρ, α, λ), we solve the following optimization problem

$$\min_{B,L,\mu} \quad \frac{1}{2}\|P_{\Omega_1}(Y - XB - L - \mathbf{1}\mu^T)\|_F^2 + \rho\|L\|_* + \lambda\left(\alpha\|B\|_1 + \frac{1-\alpha}{2}\|B\|_F^2\right). \tag{13}$$

The solutions are $B(\rho, \alpha, \lambda)$, $L(\rho, \alpha, \lambda)$ and $\mu(\rho, \alpha, \lambda)$, then we evaluate the parameters by calculating the prediction error

$$Err(\rho, \alpha, \lambda) = \frac{1}{2}\|P_{\Omega_2}(Y - XB(\rho, \alpha, \lambda) - L(\rho, \alpha, \lambda) - \mathbf{1}\mu(\rho, \alpha, \lambda)^T)\|_F^2. \tag{14}$$

The grid search over three parameters may be too computationally intensive. Therefore, we first found an optimal value for ρ, $\hat{\rho}$, which minimizes the prediction error as shown in (Yang et al., 2013) by means of Lemmas 1 and 4 (**Appendix A**). Please refer to (Yang et al., 2013) to find the details about how to find the optimal value of ρ, $\hat{\rho}$. Once the optimal value of ρ, $\hat{\rho}$ is obtained, we selected a value of α from a sequence sequentially, thereafter, we performed one-dimensional grid search for λ for each α. Specifically, we generated a sequence of λ values with length n_λ decreasing from $\lambda_{max}(\hat{\rho}, \alpha)$ to $\epsilon\lambda_{max}(\hat{\rho}, \alpha)$ on the log scale with equal space, where $\lambda_{max}(\hat{\rho}, \alpha)$ is defined as the smallest λ such that $B(\hat{\rho}, \alpha, \lambda(\hat{\rho}, \alpha))$ is a zero matrix. $\lambda_{max}(\hat{\rho}, \alpha)$ is derived as $\frac{1}{\alpha} \max_{i=1,2,\ldots,p} \max_{j=1,2,\ldots,q} |\langle X_i, Y_j \rangle|$ from coordinate-descent algorithm (Friedman et al., 2007), where X_i is the i-th column of X, and Y_j the j-th column of Y. In our R program, we set $\epsilon = 0.02$, $n_\lambda = 50$ and $S_\alpha := (0.2, 0.4, 0.6, 0.8, 0.9)$. The optimal parameters were $(\hat{\rho}, \alpha, \hat{\lambda}(\hat{\rho}, \alpha))$ that minimize the prediction error. The optimal feasible solutions of B, L, and μ were then obtained based on the set of optimal tuning parameters.

TABLE 1 | Simulation scenarios.

Chromosome	#Causal SNPs	Scenario	Method	Screening
Chr 1	60	weak-dense	FastLORS	LORS
	200	strong-sparse	LORSEN	HC
	400		LORS	
Chr 1 + Chr 21	45 + 15	weak-dense	FastLORS	LORS
	150 + 50	strong-sparse	LORSEN	HC
	300 + 100		LORS	

2.4 Single Nucleotide Polymorphism Ranking and Joint Modeling

The procedure to select the set of optimal tuning parameters is computationally intensive. Therefore, as it is discussed in (Yang et al., 2013), it may not be computationally tractable to directly apply such method to the large-scale data sets that contain a large number of gene expression levels and SNPs. A commonly used strategy to reduce such computational burden is to choose a subset of SNPs and then only use them in the subsequent eQTL analysis. In this paper, we used and evaluated two existing methods for the pre-selection of informative SNPs: LORS-Screening (Yang et al., 2013) and Higher Criticism Screening (HC-Screening) (Jeng et al., 2020). For LORS-Screening, we first obtained the initial estimate of β_i's by solving

$$\min_{\beta_i, L, \mu} \quad \frac{1}{2}\|Y - X_i\beta_i^T - L - \mathbf{1}\mu^T\|_F^2 + \rho\|L\|_*, \tag{15}$$

where X_i is the i-th column of X, β_i is a q-dimensional vector for the coefficient of the i-th SNP on q genes, $i = 1, 2, \ldots, p$. For each gene, we selected the top n SNPs in terms of the absolute values of association coefficients, then we obtained the union of selected SNPs for each gene as the final set of SNPs to be involved in the joint modeling. For HC-Screening, we first obtained association coefficients as above, then calculated the standardized estimates of coefficients. For each SNP, the Higher Criticism (HC) statistic (Donoho and Jin, 2004) is calculated based on the standardized

TABLE 2 | Results of the HC-Screening and the LORS-Screening with ten replicates for each simulation scenario.

Chromosome	#Causal SNPs	Scenario	Screening	Average #Slected SNPs	Average #Selected Causal SNPs
Chr 1	60	weak-dense	LORS	1,017	43
			HC	165	7
		strong-sparse	LORS	1,023	60
			HC	165	9
	200	weak-dense	LORS	1,036	130
			HC	165	20
		strong-sparse	LORS	1,095	199
			HC	165	28
	400	weak-dense	LORS	1,045	237
			HC	165	39
		strong-sparse	LORS	1,142	346
			HC	165	44
Chr 1 + Chr 21	45 + 15	weak-dense	LORS	1,044	46
			HC	165	7
		strong-sparse	LORS	1,065	60
			HC	165	10
	150 + 50	weak-dense	LORS	1,064	136
			HC	165	20
		strong-sparse	LORS	1,123	199
			HC	165	28
	300 + 100	weak-dense	LORS	1,064	244
			HC	165	37
		strong-sparse	LORS	1,188	361
			HC	165	44

estimates of coefficients. Then we selected the top n SNPs in terms of the p-values of HC statistics.

2.5 Simulation Design

Our simulation is similar to that described in (Jeng et al., 2020). We first downloaded the genotype data of Chromosome 1 and Chromosome 21 for CEU samples from HapMap3, the third phase of the International HapMap Project (https://www.genome.gov/10001688/international-hapmap-project). CEU samples refer to Utah residents with Northern and Western European ancestry from the CEPH collection. After the quality-control (please refer to Real Data Analysis section), the genotype data of 13,815 SNPs of Chromosome 1 and 2,607 SNPs of Chromosome 21 for $n = 165$ samples were retained in analysis. To simulate gene expression levels for $q = 200$ genes over $n = 165$ samples, we first simulated non-genetic effects of $k = 15$ hidden factors. We randomly generated nk random numbers from $N(0, 1)$ to form a $n \times k$ matrix H, then let $\Sigma = HH^T$. U_j's were simulated from $N(0, 0.1^*\Sigma)$, $j = 1, 2, \ldots, q$ and stacked by column to form a $n \times q$ matrix U. e_j's were simulated from $N(0, I)$ as random noise for j-th gene expression and combined by column to form a $n \times q$ random noise matrix e. Then the expression data of q genes over n samples were simulated by $Y = XB + U + e$, where X is the $n \times p$ genotype data matrix. We set the total number of SNPs $p = 2000$, the number of causal SNPs as 60, 200, or 400. Each causal SNP randomly influences $m = 10$ (or 50) genes. We simulated nonzero genetic effects from a uniform distribution. For the "weak-dense" scenario, each causal SNP affects $m = 50$ randomly selected genes and the corresponding values in B were simulated from a uniform distribution between 0.25 and 0.75. For the "strong-sparse" scenario, each causal SNP affects $m = 10$ randomly selected genes and the corresponding values in B were simulated from a uniform distribution between 1.5 and 2. The different simulation scenarios are summarized in **Table 1**.

3 RESULTS

3.1 Simulation Results

The number of selected SNPs and the number of selected causal SNPs from two screening methods under different simulation scenarios are summarized in **Table 2**. Several conclusions emerge from **Table 2**. First, when the number of samples is much smaller than the number of SNPs and the number of causal SNPs is larger than the number of samples, HC-Screening is seemingly not an appropriate screening tool. This is because the number of causal SNPs retained after the HC-Screening is much smaller than the actual number of causal SNPs, resulting in possible power loss in subsequent analysis. Second, even when the number of causal SNPs is smaller than the number of samples, from **Table 2**, we still observed that the LORS-Screening retains more causal SNPs than the HC-Screening. Of course, the HC-Screening reduces much computational burden especially when the number of samples is much smaller than the number of SNPs.

The area under the curve (AUC) was used to compare the performance between LORSEN and two existing methods, LORS (Yang et al., 2013) and FastLORS (Jeng et al., 2020). For each

TABLE 3 | The average AUC and 95% confidence interval without the SNP screening with ten replicates for each simulation scenario. SNPs are only from chromosome 1. For each simulation scenario, the highest AUC is in bold.

Scenario	Method	#Causal SNPs		
		60	200	400
weak-dense	FastLORS	0.514 (0.511, 0.517)	0.582 (0.580, 0.584)	0.581 (0.580, 0.582)
	LORSEN	**0.651** (0.648, 0.654)	**0.649** (0.647, 0.651)	**0.630** (0.629, 0.631)
	LORS	0.502 (0.499, 0.505)	0.514 (0.512, 0.516)	0.515 (0.514, 0.516)
strong-sparse	FastLORS	0.762 (0.755, 0.769)	**0.840** (0.837, 0.843)	**0.810** (0.807, 0.813)
	LORSEN	0.823 (0.817, 0.829)	0.834 (0.831, 0.837)	0.774 (0.771, 0.777)
	LORS	**0.824** (0.818, 0.830)	0.819 (0.815, 0.823)	0.754 (0.751, 0.757)

TABLE 4 | The average AUC and 95% confidence interval without the SNP screening with ten replicates for each simulation scenario. SNPs are from chromosome 1 and chromosome 21. For each simulation scenario, the highest AUC is in bold.

Scenario	Method	#Causal SNPs		
		60	200	400
weak-dense	FastLORS	0.530 (0.527, 0.533)	0.567 (0.565, 0.569)	0.575 (0.574, 0.576)
	LORSEN	**0.658** (0.655, 0.661)	**0.679** (0.677, 0.681)	**0.625** (0.624, 0.626)
	LORS	0.503 (0.500, 0.506)	0.510 (0.508, 0.512)	0.514 (0.513, 0.515)
strong-sparse	FastLORS	0.774 (0.767, 0.781)	**0.826** (0.822, 0.830)	**0.813** (0.810, 0.816)
	LORSEN	**0.814** (0.807, 0.821)	0.810 (0.806, 0.814)	0.788 (0.785, 0.791)
	LORS	0.813 (0.806, 0.820)	0.801 (0.797, 0.805)	0.756 (0.753, 0.759)

TABLE 5 | The average AUC and 95% confidence interval with the SNP screening with ten replicates for each simulation scenario. SNPs are only from chromosome 1. For each simulation scenario, the highest AUC is in bold.

Scenario	#Causal SNPs	Method	Screening	
			HC	LORS
weak-dense	60	FastLORS	0.514 (0.511, 0.517)	0.596 (0.593, 0.599)
		LORSEN	**0.515** (0.512, 0.518)	**0.618** (0.615, 0.621)
		LORS	0.503 (0.500, 0.506)	0.541 (0.538, 0.544)
	200	FastLORS	**0.512** (0.510, 0.514)	0.583 (0.581, 0.585)
		LORSEN	0.511 (0.509, 0.513)	**0.592** (0.590, 0.594)
		LORS	0.502 (0.500, 0.504)	0.519 (0.517, 0.521)
	400	FastLORS	**0.510** (0.509, 0.511)	**0.557** (0.556, 0.558)
		LORSEN	0.509 (0.508, 0.510)	0.547 (0.546, 0.548)
		LORS	0.502 (0.501, 0.503)	0.511 (0.510, 0.512)
strong-sparse	60	FastLORS	**0.565** (0.557, 0.573)	0.900 (0.895, 0.905)
		LORSEN	0.558 (0.550, 0.566)	**0.903** (0.898, 0.908)
		LORS	0.560 (0.552, 0.568)	0.897 (0.892, 0.902)
	200	FastLORS	**0.552** (0.548, 0.556)	**0.894** (0.891, 0.897)
		LORSEN	0.544 (0.540, 0.548)	**0.894** (0.891, 0.897)
		LORS	0.543 (0.539, 0.547)	0.874 (0.871, 0.877)
	400	FastLORS	**0.536** (0.533, 0.539)	**0.797** (0.794, 0.800)
		LORSEN	0.523 (0.520, 0.526)	0.782 (0.779, 0.785)
		LORS	0.528 (0.525, 0.531)	0.738 (0.735, 0.741)

scenario, we repeated the simulation ten times. We considered the joint modeling of multiple SNPs and multiple gene expression levels with the SNP screening and without the SNP screening. The results without the SNP screening before the eQTL mapping under different simulation scenarios are presented in **Tables 3**, **4**.

From **Tables 3**, **4**, we can see that the average AUC of LORSEN is uniformly larger than those of LORS and FastLORS in the weak-dense scenarios across different number of causal SNPs no matter the SNPs are from single chromosome (Chr 1) or two chromosomes (Chr 1 + Chr 21). For the strong-sparse

TABLE 6 | The average AUC and 95% confidence interval with the SNP screening with ten replicates for each simulation scenario. SNPs are from chromosome 1 and chromosome 21. For each simulation scenario, the highest AUC is in bold.

Scenario	#Causal SNPs	Method	Screening	
			HC	LORS
weak-dense	60	FastLORS	0.518 (0.515, 0.521)	0.606 (0.603, 0.609)
		LORSEN	**0.518** (0.515, 0.521)	**0.629** (0.626, 0.632)
		LORS	0.505 (0.502, 0.508)	0.544 (0.541, 0.547)
	200	FastLORS	**0.512** (0.510, 0.514)	0.591 (0.589, 0.593)
		LORSEN	**0.512** (0.510, 0.514)	**0.615** (0.613, 0.617)
		LORS	0.503 (0.501, 0.505)	0.524 (0.522, 0.526)
	400	FastLORS	**0.510** (0.509, 0.511)	**0.563** (0.562, 0.564)
		LORSEN	0.507 (0.506, 0.508)	0.556 (0.555, 0.557)
		LORS	0.501 (0.500, 0.502)	0.511 (0.510, 0.512)
strong-sparse	60	FastLORS	**0.570** (0.562, 0.578)	0.891 (0.886, 0.896)
		LORSEN	0.563 (0.555, 0.571)	**0.906** (0.901, 0.911)
		LORS	0.564 (0.556, 0.572)	0.891 (0.886, 0.896)
	200	FastLORS	**0.553** (0.549, 0.557)	**0.904** (0.901, 0.907)
		LORSEN	0.547 (0.543, 0.551)	**0.904** (0.901, 0.907)
		LORS	0.544 (0.540, 0.548)	0.883 (0.880, 0.886)
	400	FastLORS	**0.534** (0.531, 0.537)	**0.821** (0.818, 0.824)
		LORSEN	0.524 (0.521, 0.527)	0.813 (0.810, 0.816)
		LORS	0.525 (0.522, 0.528)	0.765 (0.762, 0.768)

scenarios, FastLORS achieves the relatively larger AUC than LORS and LORSEN. For a fixed number of causal SNPs, each method achieves the larger AUC value in the stong-sparse scenario than in the weak-dense scenario. For each method under each simulation scenario, the AUCs in **Tables 3**, **4** are similar, implying that each of three methods has the similar power to detect *cis*-eQTLs and *trans*-eQTLs.

The results with the SNP screening before eQTL mapping under different simulation scenarios are presented in **Tables 5**, **6**. As we have mentioned, the LORS-Screening keeps more SNPs in the analysis, thus retains more causal SNPs than the HC-Screening does. Each method with the LORS-Screening has the larger AUC values than it with the HC-Screening. From **Tables 5**, **6**, we can see that the AUC values of methods with the HC-Screening are quite close to 0.5, which indicates that the HC-Screening can essentially lead to the loss of power of methods. With the LORS-Screening, similar to the non-screening cases, LORSEN has better performance than LORS and FastLORS in the weak-dense scenarios and LORSEN and FastLORS perform similarly and slightly better than LORS in the strong-sparse scenarios. Finally, we find that for the weak-dense scenarios, each method without the SNP screening before joint modeling achieves the larger AUC values than it with the SNP screening. However, for the strong-sparse scenarios, each method with the LORS-Screening before joint modeling achieves the larger AUC values than it without the SNP screening. This may be due to that there are a large number of SNP-gene pairs with the weak association effects in the weak-dense scenarios and many causal SNPs may not be selected by the pre-screening methods. So, in the weak-dense scenarios with the use of pre-screening methods, the computational cost and the detection power can be reduced at the same time. In the strong-sparse scenarios, there are a smaller number of SNP-gene pairs with the stronger association effects than in the weak-dense scenarios, and it is expected that most of the causal SNPs will be selected by the pre-screening methods. Therefore, for the strong-sparse scenarios, the use of pre-screening methods reduce the computational cost while still retain the high detection power.

Our simulation results showed that LORSEN is more powerful to identify weak signals, while it does not have obvious advantage in identifying strong signals compared to LORS. Therefore, we performed additional simulation studies in which the causal variants have mixed weak and strong effects. Specifically, the half of the causal variants had the weak effects and their effects were generated from a uniform distribution between 0.25 and 0.75, while the other half of the causal variants had the strong effects and their effects were generated from a uniform distribution between 1.5 and 2. The number of causal SNPs was set as 60, 200, or 400. The number of genes affected by one causal SNP was set as 50. The AUCs and corresponding 95% confidence intervals are presented in **Supplementary Table S1**. From the results in **Supplementary Table S1**, we can see that LORSEN has the overall highest detection power when the number of causal SNPs is large. It is well known that the rare variants play an important role in the etiology of human complex diseases. Therefore, it is necessary to assess the performance of eQTL mapping methods when most of causal variants are rare. We conducted simulations in which the proportion of rare causal variants was set to be 50 and 75%. Here, the variants with minor allel frequency (MAF) less than 0.03 were considered as the rare variants. The number of causal variants was set as 200. The results from different simulation scenarios (weak-dense and strong-sparse) are presented in **Supplementary Table S2**. From **Supplementary Table S2**, we can see that when the proportion

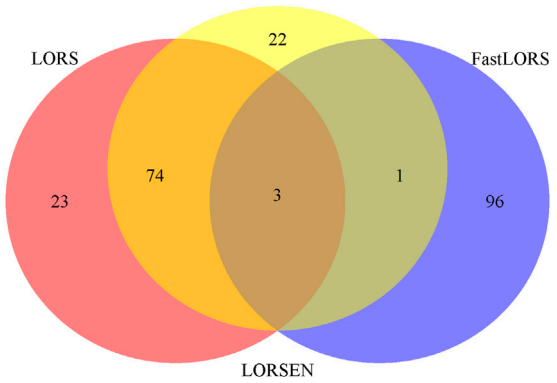

FIGURE 1 | The top 100 SNP-probe pairs identified by FastLORS, LORSEN, and LORS on Chromosome 3.

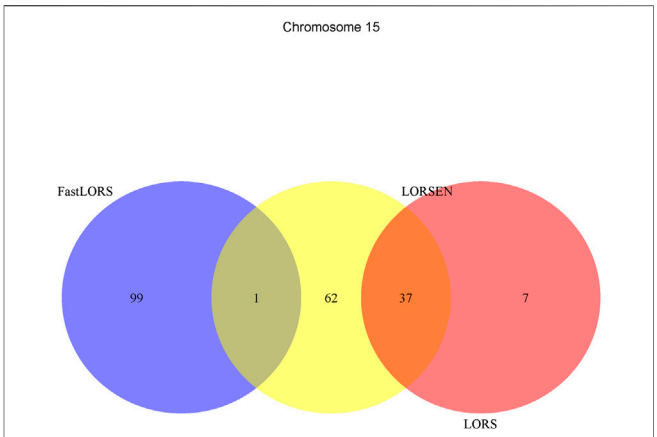

FIGURE 2 | The top 100 SNP-probe pairs identified by FastLORS, LORSEN, and LORS on Chromosome 15.

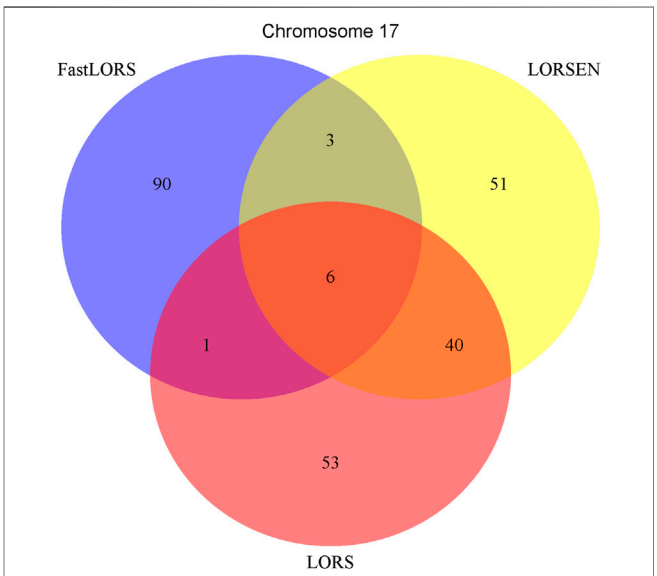

FIGURE 3 | The top 100 SNP-probe pairs identified by FastLORS, LORSEN, and LORS on Chromosome 17.

of causal rare variants is 50%, the AUCs of FastLORS are slightly higher than the AUCs of LORSEN. However, when the proportion of causal rare variants is 75%, the AUCs of LORSEN are at least 10% higher AUCs than the AUCs of FastLORS and about 20% higher than the AUCs of LORS. Our results show that LORSEN has the higher power in detecting rare causal variants. To see how the detection power of LORSEN is affected by the positive and negative effects, we conducted simulations in which the half of the causal variants had the positive effects on genes and the other half of the causal variants had the negative effects on genes. The results from different simulation scenarios (weak-dense and strong-sparse with 60, 200, and 400 causal variants) are presented in **Supplementary Table S3**. From **Supplementary Table S3**, we can see that LORSEN achieves the highest AUCs in almost all simulation scenarios, which implies that the detection power of LORSEN is not affected by the effect directions of causal variants.

In addition to AUC, a commonly used measure to assess the performance of methods for eQTL mapping, we also reported the false positive rates (FPRs) based on four thresholds for the regression coefficients: 0, 10^{-12}, 10^{-6}, 10^{-4}. From the **Supplementary Figures S1–S3**, we can see that FastLORS has the highest FPRs in almost all scenarios, and the FPRs of FastLORS are quite sensitive to the thresholds: the FPRs of FastLORS decrease dramatically for large thresholds. LORS has the smallest FPRs in all simulation scenarios. For LORSEN, it has the small and comparable FPRs with LORS when the effects of the causal variants are all weak or are a mixture of weak and strong effects. LORSEN has the large FPRs when the effects of the causal variants are all strong.

A number of conclusions emerge from the results based on our extensive simulation studies. First, the HC-Screening method retains much smaller number of SNPs than the LORS-Screening method. Second, when all the SNPs are not filtered with the SNP screening method and are used in the analysis, LORSEN is the most powerful method to identify weak signals, while it does not have obvious advantage in identifying strong signals compared to LORS and FastLORS. LORSEN still performs the best with the mixture of the strong and weak effects when the number of causal variants is large. Third, when the SNPs are first filtered with the HC-Screening method, FastLORS performs the best in all simulation scenarios. With the LORS-Screening method, LORSEN has the highest detection power in most of simulation scenarios. Fourth, LORSEN outperforms FastLORS and LORS when a large portion of the causal SNPs are rare and when the causal variants have a mixture of positive and negative effects.

3.2 Real Data Analysis Results

To illustrate our method in real data analysis, we also applied LORS-LORSEN (LORSEN with the LORS-Screening), LORS-LORS (LORS with the LORS-Screening) and HC-FastLORS

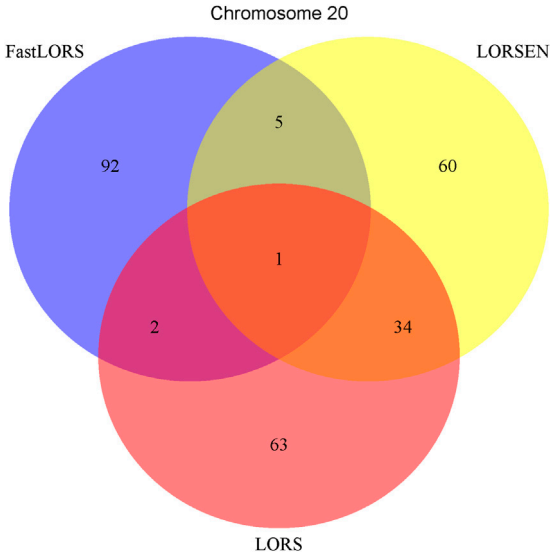

FIGURE 4 | The top 100 SNP-probe pairs identified by FastLORS, LORSEN, and LORS on Chromosome 20.

(FastLORS with the HC-Screening) to the HapMap3 data. Here, we focused on Asian samples (CHB and JPT) in the HapMap3 data and selected four chromosomes for the analysis. SNP genotype data and gene expression data are publicly available, and can be downloaded from ftp://ftp.ncbi.nlm.nih.gov/hapmap/genotypes/hapmap3_r3/plink_format/and http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-264/, respectively. Because the set of samples with the SNP genotype data and the set of samples with the gene expression data are slightly different, we only kept the samples that have both the SNP genotype data and the gene expression data in the analysis. We removed SNPs with missing values, and performed the LD pruning using PLINK with its default parameters (window size: 50; moving window increment: five SNPs; cutoff value of R^2: 0.5). After the data pre-processing, a total of 160 samples (CHB: 79; JPT: 81) were included in analysis. The number of SNPs and the number of genes with the expression used in the analysis on chromosome 3 are 4,086 and 1,075, on chromosome 15 are 2,235 and 612, on chromosome 17 are 2,226 and 1,098, on chromosome 20 are 1,863 and 606, respectively. Since the significance tests generally cannot be performed for the penalization based regression models, we focused on the top 100 SNP-probe pairs with the largest absolute regression coefficients. From the Venn diagrams (**Figures 1**–**4**), we

TABLE 7 | Top ten detected SNP-probe pairs for chromosome 3. The SNP-probe pairs that are confirmed in seeQTL database are in bold.

Method	SNP	Probe (gene)	Association coefficient	Distance	Class
HC-FastLORS	rs13084976	ILMN_1657373 (LEPREL1)	0.0430	188.72 mb	distant
	rs17029694	ILMN_1657373 (LEPREL1)	0.0424	188.49 mb	distant
	rs12494696	ILMN_1812093 (UTS2D)	0.0322	189.72 mb	distant
	rs2322212	ILMN_1756501 (ST6GAL1)	0.0310	184.74 mb	distant
	rs17029694	ILMN_1708743 (NT5DC2)	0.0303	49.86 mb	distant
	rs2322212	ILMN_1686920 (CCDC58)	0.0300	120.03 mb	distant
	rs7647780	ILMN_1762084 (DNASE1L3)	0.0292	57.51 mb	distant
	rs1516347	ILMN_1726020 (LOC652670)	0.0278	75.49 mb	distant
	rs13061928	ILMN_1692261 (EPHB1)	0.0273	133.55 mb	distant
	rs1377213	ILMN_1698934 (CMTM7)	0.0270	26.76 mb	distant
LORS-LORSEN	rs1505587	ILMN_1657373 (LEPREL1)	0.3336	127.69 mb	distant
	rs6807033	ILMN_1787750 (CD200)	0.2796	4.163 kb	local
	rs11914577	ILMN_1700967 (C3orf59)	0.2245	113.51 kb	local
	rs1403719	ILMN_1771599 (PLOD2)	0.1963	25.06 mb	distant
	rs628267	ILMN_1760509 (EOMES)	0.1941	302.30 kb	
	rs4016435	**ILMN_1757350 (CTNNB1)**	**0.1908**	**27.772 kb**	**local**
	rs16839507	**ILMN_1761058 (ACAD11)**	**0.1856**	**117.942 kb**	**local**
	rs693430	**ILMN_1657708 (MGLL)**	**0.1796**	**86.074 kb**	**local**
	rs693430	**ILMN_1707310 (MGLL)**	**0.1710**	**47.617 kb**	**local**
	rs1498090	ILMN_1793724 (C3orf31)	0.1662	58.605 kb	local
LORS-LORS	rs1505587	ILMN_1657373 (LEPREL1)	1.2549	127.69 mb	distant
	rs6807033	ILMN_1787750 (CD200)	0.5621	4.163 kb	local
	rs4857653	ILMN_1700967 (C3orf59)	0.3640	16.16 mb	distant
	rs11914577	ILMN_1700967 (C3orf59)	0.2984	113.514 kb	local
	rs1403719	ILMN_1771599 (PLOD2)	0.2824	25.06 mb	distant
	rs628267	ILMN_1760509 (EOMES)	0.2439	302.302 kb	
	rs4016435	**ILMN_1757350 (CTNNB1)**	**0.2404**	**27.772 kb**	**local**
	rs16839507	**ILMN_1761058 (ACAD11)**	**0.2338**	**117.942 kb**	**local**
	rs3773014	ILMN_1762084 (DNASE1L3)	0.2268	29.187 kb	local
	rs1799977	ILMN_1688392 (ZBED2)	0.2234	75.77 mb	distant

TABLE 8 | The SNP-probe pairs found in seeQTL database out of the top ten SNP-probe pairs for chromosomes 3, 15, 17, and 20, respectively.

Chromosome	SNP	Probe (gene)	Method	Information from GTEx
3	rs4016435	ILMN_1757350 (CTNNB1)	LORS-LORSEN, LORS-LORS	Not found in GTEx
3	rs16839507	ILMN_1761058 (ACAD11)	LORS-LORSEN, LORS-LORS	Multiple hits for eQTLs and sQTLs
3	rs693430	ILMN_1657708 (MGLL)	LORS-LORSEN	Not found in GTEx
3	rs693430	ILMN_1657708 (MGLL)	LORS-LORSEN	Not found in GTEx
15	rs7162538	ILMN_1784364 (STARD5)	LORS-LORSEN, LORS-LORS	Multiple hits for eQTLs and sQTLs
15	rs1347069	ILMN_1795822 (DIS3L)	LORS-LORSEN, LORS-LORS	Multiple hits for eQTLs and sQTLs
15	rs2292114	ILMN_1795524 (C15orf44)	LORS-LORS	Not found in GTEx
17	rs4968140	ILMN_1706959 (TIMM22)	HC-FastLORS	Not found in GTEx
17	rs4251704	ILMN_1773352 (CCL5)	LORS-LORSEN	A single hit for sQTLs
17	rs17657522	ILMN_1697227 (USP36)	LORS-LORSEN, LORS-LORS	Multiple hits for eQTLs and sQTLs
17	rs4968140	ILMN_1706959 (TIMM22)	LORS-LORSEN, LORS-LORS	Not found in GTEx
17	rs9905601	ILMN_1750511 (NT5C3L)	LORS-LORS	Not found in GTEx
20	rs16989514	ILMN_1721128 (TOMM34)	LORS-LORSEN, LORS-LORS	Multiple hits for eQTLs and sQTLs
20	rs6041750	ILMN_1702237 (FKBP1A)	HC-FastLORS, LORS-LORSEN, LORS-LORS	Multiple hits for eQTLs and sQTLs

notice that there is a large overlap between the eQTLs identified by LORS-LORS and LORS-LORSEN. However, there is a small overlap between the eQTLs identified by HC-FastLORS and LORS-LORS (or LORS-LORSEN). For example, among the top 100 SNP-probe pairs identified on Chromosome 3 (**Figure 1**), LORS-LORS and LORS-LORSEN share 77 SNP-probe pairs in common, while LORS-LORSEN and HC-FastLORS only share four SNP-probe pairs in common and LORS-LORS and HC-FastLORS share three SNP-probe pairs in common. This observation is consistent with the observation from (Jeng et al., 2020) which also noticed that there is a small overlap between the SNP-probe pairs identified by LORS-LORS and HC-FastLORS. Additionally, as adopted in (Jeng et al., 2020), we classified the detected eQTL as local if the physical distance between the SNP and the probe midpoint is less than 250 kb or as distant if the distance is greater than 5 mb following the criterion described in (Westra et al., 2013). For each chromosome, we report our findings on the top ten identified SNP-probe pairs in **Table 7** and **Supplementary Tables S4–S6** (see **Supplementary Material**). From **Table 7**, we can see that the SNPs in the top ten SNP-probe pairs identified by HC-FastLORS are all *trans*-eQTLs. As a comparison, seven SNPs in the top ten SNP-probe pairs identified by LORS-LORSEN are *cis*-eQTLs and two SNPs are *trans*-eQTLs. Five SNPs in the top ten SNP-probe pairs identified by LORS-LORS are *cis*-eQTLs and four SNPs are *trans*-eQTLs. LORS-LORSEN and LORS-LORS share seven SNP-probe pairs while LORS-LORSEN and LORS-LORS do not share any SNP-probe pair with HC-FastLORS. In addition, the coefficients obtained from HC-FastLORS are ten-fold smaller than those obtained from LORS-LORSEN and LORS-LORS. This indicates that the findings of LORS-LORSEN and LORS-LORS may be more convincing.

To further validate our findings, we searched an existing database called seeQTL (Xia et al., 2011). seeQTL (https://seeqtl.org/) records the eQTLs identified from a meta-analysis (consensus eQTLs) from the HapMap human lymphoblastoid cell lines. A total of fourteen SNP-probe pairs were found in seeQTL and were listed in **Table 8**. Among them, two SNP-probe pairs were identified by HC-FastLORS only, three were identified by LORS-LORSEN only, two were identified by LORS-LORS only, seven were identified by both LORS-LORSEN and LORS-LORS, and one was identified by all three methods. To further validate these fourteen SNP-probe pairs, we searched the eQTL web-browser (http://www.gtexportal.org/home/) built by the Genotype-Tissue Expression Project (GTEx) (https://www.genome.gov/Funded-Programs-Projects/Genotype-Tissue-Expression-Project) to see if those SNP-probe (gene) pairs are listed as the eQTLs and/or sQTLs (splicing quantitative trait locus). A total of seven SNP-probe pairs were also found in GTEx and were presented in **Table 8**. Among seven SNP-probe pairs found both in seeQTL and GTEx, one SNP-probe pair was identified by all three methods, five SNP-probe pairs were identified by both LORS-LORSEN and LORS-LORS, and one SNP-probe pair was identified by LORS-LORSEN only.

A number of conclusions emerge from the results based on HapMap3 data. First, there is a large overlap between the SNP-probe pairs identified by LORS-LORS and LORS-LORSEN but there is a small overlap between the SNP-probe pairs identified by HC-FastLORS and LORS-LORS (or LORS-LORSEN). Second, LORS-LORS and LORS-LORSEN perform similarly and have higher detection power than HC-FastLORS since LORS-LORS and LORS-LORSEN have identified more SNP-probe pairs that are also found in seeQTL and GTEx. Third, five out of seven SNP-probe pairs identified by both LORS-LORS and LORS-LORSEN and found in seeQTL are also found in GTEx, thus it may be beneficial to combine the results from multiple methods to generate a list of SNP-probe pairs for further investigation.

4 DISCUSSION

As more human gene expression data become available, fast and efficient statistical and computational methods are needed to fully take advantage of such data to investigate the relationship between genetic variants and gene expression levels to further reveal the genetic mechanisms that underlie human complex diseases. However, most existing methods are built on small-scale samples and not applicable to human-size datasets. In this paper, we proposed a new low rank penalized regression method (LORSEN) to detect eQTLs. We developed a fast and efficient algorithm to solve optimization problems arising from our methods based on proximal gradient methods. Comprehensive simulation studies showed that LORSEN outperformed two

commonly used methods, LORS and FastLORS, in many simulation scenarios. From our simulation results, we can briefly conclude that, first, LORSEN is more powerful in detecting eQTLs which are rare and/or have weak effects. This is especially an appealing advantage since it is expected that a portion of causal variants are rare and/or have the weak effects in the real world. Second, LORSEN is more powerful when some causal variants have the positive effects and the other causal variants have the negative effects.

Since there are a large number of SNPs and genes to be included in the eQTL mapping and it is expected that only a small portion of SNPs will affect the gene expression levels, a number of pre-screening methods have been developed. In this paper, we used the LORS-Screening (Yang et al., 2013) and the HC-Screening (Jeng et al., 2020). We found that the HC-Screening retained much smaller number of SNPs than the LORS-Screening. Both the LORS-Screening and the HC-Screening can reduce the computational cost, but they may also reduce the detection power in the eQTL mapping, depending on the association patterns between SNPs and gene expression levels. Since we do not know such association patterns in real studies, we should be cautious to apply such pre-screening methods.

There are several limitations for LORSEN. First, as a method based on the penalized regression model, we can rank the SNP-gene pairs in terms of the regression coefficients obtained from LORSEN, but cannot perform the significance test. Second, the computational time of LORSEN depends on many factors such as the number of candidate values of hyperparameters, the initial values of hyperparameters, and the number of samples. The computation was performed parallelly using software R (verson 4.1.1) and 16 cores on a server with 64 Intel(R) Xeon(R) Gold 6130 CPUs @ 2.10 GHz. From **Supplementary Table S7**, we can see that, as expected, LORSEN costs much more time in parameter tuning than other two methods due to the exhaustive grid search. The grid search is easy to be implemented but is computationally intensive. It may not be feasible for large scale data. A more efficient strategy is desirable.

It has shown that the incorporation of the SNP correlation and the gene interaction network can potentially increase the power of detecting eQTLs (Kim and Xing, 2009; Chen et al., 2012; Kim and Xing, 2012; Cheng et al., 2014). We expect that our method can be improved if we use the prior knowledge of correlation structures of SNPs and genes to refine the penalty terms in optimization problems.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

CG and KZ conceived the study, CG developed the method and algorithms, CG developed the R program LORSEN, CG performed simulation studies and real data analysis, CG drafted the article, CG, HW, and KZ revised the article. All the authors read and approved the final article.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.690926/full#supplementary-material

REFERENCES

Albert, F. W., and Kruglyak, L. (2015). The Role of Regulatory Variation in Complex Traits and Disease. *Nat. Rev. Genet.* 16, 197–212. doi:10.1038/nrg3891

Banerjee, S., Simonetti, F. L., Detrois, K. E., Kaphle, A., Mitra, R., Nagial, R., et al. (2021). Tejaas: Reverse Regression Increases Power for Detecting Trans-eqtls. *Genome Biol.* 22, 142. doi:10.1186/s13059-021-02361-8

Beck, A., and Teboulle, M. (2009). A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM J. Imaging Sci.* 2, 183–202. doi:10.1137/080716542

Cai, J.-F., Candès, E. J., and Shen, Z. (2010). A Singular Value Thresholding Algorithm for Matrix Completion. *SIAM J. Optim.* 20, 1956–1982. doi:10.1137/080738970

Chen, X., Shi, X., Xu, X., Wang, Z., Mills, R., Lee, C., et al. (2012). "A Two-Graph Guided Multi-Task Lasso Approach for Eqtl Mapping," in Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics. Editors N. D. Lawrence and M. Girolami (La Palma, Canary Islands: PMLR), 208–217. vol. 22 of Proceedings of Machine Learning Research, April 21-23, 2012, La Palma, Canary Islands.

Cheng, W., Zhang, X., Guo, Z., Shi, Y., and Wang, W. (2014). Graph-regularized Dual Lasso for Robust Eqtl Mapping. *Bioinformatics* 30, i139–i148. doi:10.1093/bioinformatics/btu293

Chun, H., and Keleş, S. (2009). Expression Quantitative Trait Loci Mapping with Multivariate Sparse Partial Least Squares Regression. *Genetics* 182, 79–90. doi:10.1534/genetics.109.100362

Cookson, W., Liang, L., Abecasis, G., Moffatt, M., and Lathrop, M. (2009). Mapping Complex Disease Traits with Global Gene Expression. *Nat. Rev. Genet.* 10, 184–194. doi:10.1038/nrg2537

Donoho, D., and Jin, J. (2004). Higher Criticism for Detecting Sparse Heterogeneous Mixtures. *Ann. Stat.* 32, 962–994. doi:10.1214/009053604000000265

Fan, J., and Lv, J. (2008). Sure independence Screening for Ultrahigh Dimensional Feature Space. *J. R. Stat. Soc. Ser. B (Statistical Methodology)* 70, 849–911. doi:10.1111/j.1467-9868.2008.00674.x

Fan, Y., Zhu, H., Song, Y., Peng, Q., and Zhou, X. (2020). Efficient and Effective Control of Confounding in Eqtl Mapping Studies through Joint Differential Expression and Mendelian Randomization Analyses. *Bioinformatics* 37, 296–302. doi:10.1093/bioinformatics/btaa715

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* 33, 1–22. doi:10.18637/jss.v033.i01

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise Coordinate Optimization. *Ann. Appl. Stat.* 1, 302–332. doi:10.1214/07-aoas131

Fusi, N., Stegle, O., and Lawrence, N. D. (2012). Joint Modelling of Confounding Factors and Prominent Genetic Regulators Provides Increased Accuracy in Genetical Genomics Studies. *Plos Comput. Biol.* 8, e1002330. doi:10.1371/journal.pcbi.1002330

Gao, C., Tignor, N. L., Salit, J., Strulovici-Barel, Y., Hackett, N. R., Crystal, R. G., et al. (2014). Heft: Eqtl Analysis of many Thousands of Expressed Genes while Simultaneously Controlling for Hidden Factors. *Bioinformatics* 30, 369–376. doi:10.1093/bioinformatics/btt690

Hanley, J. A., and McNeil, B. J. (1982). The Meaning and Use of the Area under a Receiver Operating Characteristic (Roc) Curve. *Radiology* 143, 29–36. doi:10.1148/radiology.143.1.7063747

Hu, Y.-J., Sun, W., Tzeng, J.-Y., and Perou, C. M. (2015). Proper Use of Allele-specific Expression Improves Statistical Power Forcis-eQTL Mapping with RNA-Seq Data. *J. Am. Stat. Assoc.* 110, 962–974. doi:10.1080/01621459.2015.1038449

Jeng, X. J., Rhyne, J., Zhang, T., and Tzeng, J.-Y. (2020). Effective SNP Ranking Improves the Performance of eQtl Mapping. *Genet. Epidemiol.* 44, 611–619. doi:10.1002/gepi.22293

Kendziorski, C. M., Chen, M., Yuan, M., Lan, H., and Attie, A. D. (2006). Statistical Methods for Expression Quantitative Trait Loci (Eqtl) Mapping. *Biometrics* 62, 19–27. doi:10.1111/j.1541-0420.2005.00437.x

Kim, S., and Xing, E. P. (2009). Statistical Estimation of Correlated Genome Associations to a Quantitative Trait Network. *Plos Genet.* 5, e1000587. doi:10.1371/journal.pgen.1000587

Kim, S., and Xing, E. P. (2012). Tree-guided Group Lasso for Multi-Response Regression with Structured Sparsity, with an Application to Eqtl Mapping. *Ann. Appl. Stat.* 6, 1095–1117. doi:10.1214/12-AOAS549

Lee, S., and Xing, E. P. (2012). Leveraging Input and Output Structures for Joint Mapping of Epistatic and Marginal Eqtls. *Bioinformatics* 28, i137–i146. doi:10.1093/bioinformatics/bts227

Listgarten, J., Kadie, C., Schadt, E. E., and Heckerman, D. (2010). Correction for Hidden Confounders in the Genetic Analysis of Gene Expression. *Proc. Natl. Acad. Sci.* 107, 16465–16470. doi:10.1073/pnas.1002425107

Mazumder, R., Hastie, T., and Tibshirani, R. (2010). Spectral Regularization Algorithms for Learning Large Incomplete Matrices. *J. Mach. Learn. Res.* 11, 2287–2322.

Parikh, N., and Boyd, S. (2014). Proximal Algorithms. *FNT in Optimization* 1, 127–239. doi:10.1561/2400000003

Rakitsch, B., and Stegle, O. (2016). Modelling Local Gene Networks Increases Power to Detect Trans-acting Genetic Effects on Gene Expression. *Genome Biol.* 17, 33. doi:10.1186/s13059-016-0895-2

Stegle, O., Parts, L., Durbin, R., and Winn, J. (2010). A Bayesian Framework to Account for Complex Non-genetic Factors in Gene Expression Levels Greatly Increases Power in Eqtl Studies. *Plos Comput. Biol.* 6, e1000770. doi:10.1371/journal.pcbi.1000770

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B (Methodological)* 58, 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x

Wang, P., Dawson, J. A., Keller, M. P., Yandell, B. S., Thornberry, N. A., Zhang, B. B., et al. (2011). A Model Selection Approach for Expression Quantitative Trait Loci (Eqtl) Mapping. *Genetics* 187, 611–621. doi:10.1534/genetics.110.122796

Westra, H.-J., Peters, M. J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., et al. (2013). Systematic Identification of Trans Eqtls as Putative Drivers of Known Disease Associations. *Nat. Genet.* 45, 1238–1243. doi:10.1038/ng.2756

Xia, K., Shabalin, A. A., Huang, S., Madar, V., Zhou, Y.-H., Wang, W., et al. (2011). Seeqtl: a Searchable Database for Human Eqtls. *Bioinformatics* 28, 451–452. doi:10.1093/bioinformatics/btr678

Yang, C., Wang, L., Zhang, S., and Zhao, H. (2013). Accounting for Non-genetic Factors by Low-Rank Representation and Sparse Regression for Eqtl Mapping. *Bioinformatics* 29, 1026–1034. doi:10.1093/bioinformatics/btt075

Yu, Y.-L. (2013). On Decomposing the Proximal Map. *Adv. Neural Inf. Process. Syst.* 26, 91–99.

Zou, H., and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *J. R. Stat. Soc. B* 67, 301–320. doi:10.1111/j.1467-9868.2005.00503.x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

APPENDIX A

Lemma 1: For each $\tau \geqslant 0$ and $Y \in \mathbb{R}^{n_1 \times n_2}$, the solution of

$$\min_X \quad \frac{1}{2}\|X - Y\|_F^2 + \tau\|X\|_* \tag{16}$$

is $S_\tau(Y) := US_\tau(\Sigma)V^T (= Prox_{\tau\|\cdot\|_*}(Y))$, where $S_\tau(\Sigma) = diag(\{(\sigma_i - \tau)_+\})$, $Y = U\Sigma V^T$, the singular value decomposition of matrix Y, $\Sigma = diag(\{\sigma_i\}_{1 \leqslant i \leqslant r})$, r is the rank of Y. $S_\tau(\cdot)$ is called singular value shrinkage operator.

Proof: see (Cai et al., 2010) or (Mazumder et al., 2010).

Lemma 2: For each fixed non-negative λ and $v \in \mathbb{R}^n$, the solution of

$$\min_x \quad \frac{1}{2}\|x - v\|_2^2 + \frac{\lambda}{2}\|x\|_2^2 \tag{17}$$

is $(Prox_{\frac{\lambda}{2}\|\cdot\|_2^2}(v))_i = sign(v_i)(|v_i| - \lambda)_+$, $i = 1, 2, \ldots, n$, known as the (elementwise) soft thresholding operator.

Proof: see (Parikh and Boyd, 2014).

Lemma 3: For each fixed non-negative ρ and $v \in \mathbb{R}^n$, the solution of

$$\min_x \quad \frac{1}{2}\|x - v\|_2^2 + \rho\|x\|_1 \tag{18}$$

is $Prox_{\rho\|\cdot\|_1}(v) = (1 - \frac{\rho}{max\{\|v\|_2, \rho\}})v$.

Proof: see (Parikh and Boyd, 2014).

Lemma 4: (soft-impute algorithm)

For the optimization problem

$$\min_X \quad \frac{1}{2}\|P_\Omega(Y - X)\|_F^2 + \tau\|X\|_*$$
$$= \min_X \quad \frac{1}{2}\|[P_\Omega(Y) + P_{\Omega^\perp}(X)] - X\|_F^2 + \tau\|X\|_*,$$

the optimization solution can be obtained via updating X using $X \leftarrow S_\tau(P_\Omega(Y) + P_{\Omega^\perp}(X))$ with an arbitrary initialization.

Proof: see (Mazumder et al., 2010).

Theorem 1: A sufficient condition for $Prox_{f+g} = Prox_f \circ Prox_g$ is $\forall\, x \in \mathcal{H}, \partial g(Prox_f(x)) \supseteq \partial g(x)$, where $\mathcal{H}$ represents Hilbert space and $\circ$ represents composition of two operators.

Proof: see (Yu, 2013).

Details of Confidence Interval of AUC

We followed the method used in (Hanley and McNeil, 1982) to calculate the 95% confidence interval (CI) of AUC. Let $\widehat{AUC}$ and $Var(\widehat{AUC})$ denote the sample mean and the estimated variance of AUCs from ten replicates, respectively, the 95% CI of average AUC was calculated using the following formula:

$$\overline{\widehat{AUC}} \pm 1.96\sqrt{Var(\widehat{AUC})/10}. \tag{19}$$

We used the following formula (Hanley and McNeil, 1982) to calculate $Var(\widehat{AUC})$:

$$Var(\widehat{AUC}) = \frac{q_0 + (n_1 - 1)q_1 + (n_2 - 1)q_2}{n_1 n_2}, \tag{20}$$

where $q_0 = \widehat{AUC}(1 - \widehat{AUC})$, $q_1 = \frac{\widehat{AUC}}{2 - \widehat{AUC}} - \widehat{AUC}^2$, $q_2 = \frac{2\widehat{AUC}^2}{1 + \widehat{AUC}} - \widehat{AUC}^2$, n_1 is the number of true positives, and n_2 is the number of true negatives.