BMC Bioinformatics

RESEARCH

Open Access

Tisslet tissues-based learning estimation for transcriptomics

Ahmed Miloudi[1†], Aisha Al-Qahtani[2*†], Thamanna Hashir[3], Mohamed Chikri[1*] and Halima Bensmail[2*]

†Ahmed Miloudi and Aisha Al-Qahtani These authors equally contributed to this work.

*Correspondence: aialqahtani@hbku.edu.qa; mohamed.chikri@usmba.ac.ma; hbensmail@hbku.edu.qa

[1] Faculty of Medicine and Pharmacy-FUSMBA, Fes, Morocco
[2] Qatar Computing Research Institute, HBKU, Doha, Qatar
[3] Carnegie Mellon University-Qatar, Doha, Qatar

Abstract

In the context of multi-omics data analytics for various diseases, transcriptome-wide association studies leveraging genetically predicted gene expression hold promise for identifying novel regions linked to complex traits. However, existing methods for multi-tissue gene expression prediction often fail to account for tissue-tissue expression interactions, limiting their accuracy and effectiveness. This research addresses the challenge of predicting gene expression across multiple tissues by incorporating tissue-tissue expression correlations based on a nonlinear multivariate model. Our findings demonstrate that this model excels in estimating tissue-tissue interactions and accurately predicting missing data. These results have significant implications for multi-omics data analytics and transcriptome-wide association studies, suggesting a novel approach for identifying regions associated with complex traits.

Keywords: Transcriptomics, Multiple-tissues, Machine learning, Sparse covariance matrix, EQTL, Likelihood estimator

Introduction

eQTL (expression Quantitative Trait Loci) studies explore the relationship between genetic variants, such as SNPs (Single Nucleotide Polymorphisms), and gene expression levels. By identifying how specific SNPs influence gene expression, researchers can gain insights into the genetic basis of complex traits and diseases. In the context of transcriptomics, which involves the comprehensive analysis of RNA transcripts, eQTL mapping is particularly valuable (Fig. 1). When considering multiple tissues, incorporating tissue-tissue correlations becomes essential, as gene expression can be influenced by interactions across different tissues. This approach allows for a more holistic understanding of gene regulation and the identification of novel regions associated with complex traits, enhancing the precision and effectiveness of transcriptome-wide association studies. Transcriptome-wide association studies (TWAS) have therefore gained prominence in the field of genomics and genetics for their potential to uncover genetic variants associated with complex traits. These studies leverage gene expression data to bridge the gap between genetic variations and phenotypic traits. Although TWAS holds great promise, the accuracy of gene expression

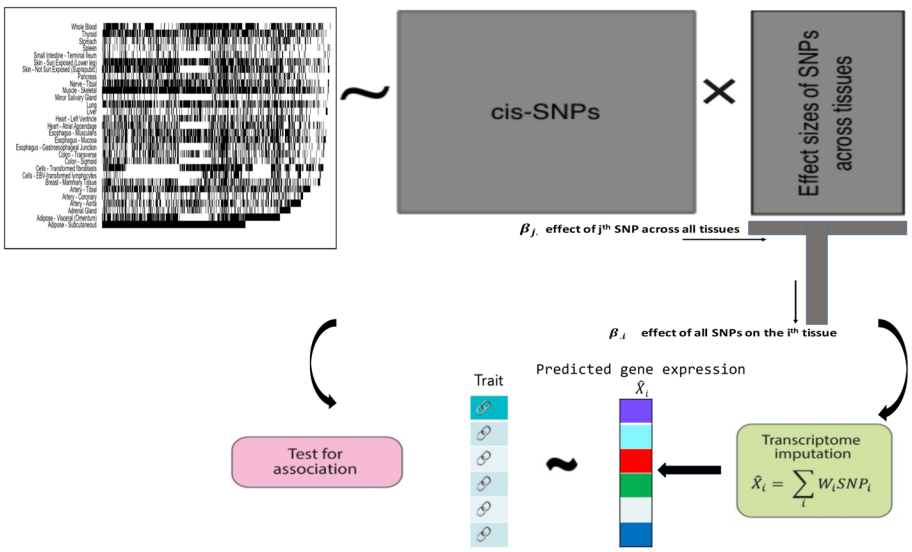

Fig. 1 Overview of TISSLET method. The inputs required for TISSLET is a matrix of gene expression for several tissues. We also use the subjects genotypes with sample matched measured expression. TISSLET's weights output are calculated based on the CEM algorithm using measured gene expression and provide the weights and covariance structure of tissues. For full details, see the Materials and Methods section

prediction across multiple tissues remains a challenge. Previous research addressed the challenge of multi-tissue gene expression (Grinberg and Wallace, 2021) [1]

Machine learning algorithms have been used to predict eQTL regulation of a gene by analyzing adjacent genetic variants [2, 3]. Researchers have used genetically predicted gene expression models to conduct transcriptome-wide association studies (TWAS) and identify novel areas linked to complex traits. However, many of these regions lack a GWAS relationship within 1Mb.

There are various benefits to such analyses: Leveraging gene expression enriches possible trait-associated SNPs, whereas joint eQTL modeling improves overall association strength and reduces the number of tests from millions to roughly 20,000 genes.

Leveraging shared eQTLs across tissues enhances eQTL discovery and gene expression imputation accuracy, leading to more powerful transcriptome-wide association analyses [4–6]. Hu et al. [7] and Molstad et al. [8] developed a penalized regression strategy for collaborative modeling of eQTLs. The penalty encourages shared eQTLs between tissues. Using genotype and expression data from the Genotype-Tissue Expression (GTEx) project, multi-tissue eQTL models significantly increase gene association identification and imputation accuracy compared to single-tissue techniques.

While Hu et al.'s technique [7] does not account for tissue-specific gene expression correlation in combined eQTL modeling, Molstad et al. [8] uses cross-tissue imputation with EM algorithm. Both approach uses linear model to derive their algorithm. Recent research suggests that accounting for metric based on skewness improves variable selection and prediction accuracy and can identify regulators or genes in large patient cohorts. A recent study showed that significant correlation was detected between expression skewness and the top 500 genes corresponding to the most significant differential DNA methylation occurring in the promotor regions in TCGA [9]

Recent research suggests that accounting for tissue-tissue correlation in high-dimensional penalized multivariate response linear regression improves variable selection and prediction accuracy. This phenomena may be explained by a seemingly unrelated regression interpretation of high-dimensional sparse multivariate response linear regression. Moreover, certain tissue types are more difficult to obtain due to biological and cost constraints. Using tissue-tissue correlation with skewness assumption enhances gene expression prediction accuracy, particularly for small data numbers.

In this paper, we propose an approach (TISSLET) which not only impute missing gene expression using cross information from several tissue, but also estimate tissue-tissue correlation. We calculate a joint eQTL weights while imputing missing gene expression using a skewed normal modeling. The technique by which our approach works is straightforward. Measuring expression in one tissue can provide a reliable estimate of expression in the other, especially if their expression is substantially correlated. Ignoring tissue-tissue correlation can significantly reduce gene expression prediction accuracy. Our methodology offers several advantages over previous approaches for multi-tissue joint eQTL mapping, including:

1. Incorporation of a full tissue-tissue correlation matrix in the model, rather than assuming a diagonal matrix, which reveals cross-tissue expression dependencies that eQTLs alone cannot explain.
2. Efficient estimation of eQTL weights by modeling cross-tissue associations.
3. Relaxation of the normality assumption by allowing for a heavy-tailed error distribution, assuming that errors follow a multivariate skewed distribution.

Figure 1 gives an overview of TISSLET method. This figure demonstrates the training of the imputation model. The inputs required for TISSLET is a matrix of gene expression for several tissues. We also use the subjects genotypes with sample matched measured expression. TISSLET's weights output are calculated based on the CEM algorithm using measured gene expression and provide the weights and covariance structure of tissues. For full details, see the Materials and Methods section.

Materials and methods.

Regression models are commonly used to map the relationship between SNPs and gene expression levels in eQTL studies. Traditional models often assume normality of errors, which may not capture the true distribution of gene expression data. Introducing a skewed model with a cross-tissue expression based on SNP genotypes can provide a more accurate representation by allowing for heavy-tailed error distributions, improving the precision of eQTL mapping and the identification of genetic influences on gene expression.

Let $\mathbf{x}_i \in \mathbb{R}^p$ represent the genotypes of p SNPs (both centered and normalized) and let $\mathbf{y}_i \in \mathbb{R}^q$ represent the vector of centered and normalized measured expression in q tissues for the i^{th} subject for a specific gene within a certain distance (e.g. 500 kb) or less away from the gene of interest. We assume that gene expression represents a realization of the random vector for the i^{th} subject:

$$\mathbf{y}_i \sim \boldsymbol{\beta}^t \mathbf{x}_i + \boldsymbol{\epsilon}_i, \qquad \boldsymbol{\epsilon}_i \sim \mathbb{SN}(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}). \tag{2.1}$$

where $\mathbb{SN}_q$ denotes the q-dimensional multivariate skewed normal distribution, where $\boldsymbol{\beta} \in \mathbb{R}^{p\times q}$ is the unknown regression coefficient matrix (i.e., eQTL weights), and $\boldsymbol{\Sigma}^{-1} \in \mathbb{R}^{q\times q}$ is the cross-tissue error precision (inverse covariance) matrix. We further assume that $\boldsymbol{\epsilon}_i$ is independent of $\boldsymbol{\epsilon}_j$ for all $i \neq j$.

The skewed distribution $\mathbb{SN}$ can reveal the asymmetric information when the observations such as gene expression are skewed [10]. Asymmetrical distributions have an additional shape parameter $\boldsymbol{\Lambda} \in \mathbb{R}^{q\times q}$ to represent the direction of the asymmetry of the density. If the skewness in observations is ignored, inferences with symmetric distributions may result in biased or even misleading conclusions.

Penalized skew-normal log-likelihood

In appendix (7.2) we outline in details the derivation of the model parameters (weights and cross-correlation matrices) under the normality assumption. This approach uses the ECM algorithm for parameter estimation. Algorithm 1 gives the pseudocode of the approach using normality assumption.

Algorithm 1 Regularized ECM Algorithm with normality

Initialize $\beta^{(1)} \in \mathbb{R}^{p\times q}$ and $\Omega^{(1)} \in \mathbb{R}^{q\times q}$.
Set $k = 1$
while not converged **do**
  $\Omega^{(k+1)} \leftarrow \underset{\Omega \in \mathbb{R}^{q\times q}}{\arg\min} \left[\frac{1}{n} \sum_{i=1}^{n} \mathtt{tr} \left\{ \mathbf{S}_i(\beta^{(\mathbf{k})}, \boldsymbol{\Sigma}^{(\mathbf{k})}) \, \boldsymbol{\Omega} \right\} - \log|\Omega| + \mathcal{R}(\Omega) \right]$
  $A \leftarrow \left(\widetilde{y}_i^{(k)} - \beta^T x_i \right)^T \left(\widetilde{y}_i^{(k)} - \beta^T x_i \right)$
  $\beta^{(k+1)} \leftarrow \underset{\beta \in \mathbb{R}^{p\times q}}{\arg\min} \left[\frac{1}{n} \sum_{i=1}^{n} A^T \Omega^{(k+1)} A + \mathcal{R}(\beta) \right]$
  Update $\mu_i^{(k+1)}, V_i^{(k+1)}$ for $i \in [n]$ as in (7.1) to (7.5) in the appendix
  Construct $S(\beta^{(k+1)}, \Sigma^{(k+1)})$
  Set $k \leftarrow k + 1$
end while

Similarly, using equation (2.1) and assuming that the errors ϵ's $\overset{\text{i.i.d}}{\sim} \mathbb{SN}(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda})$, for $1 \leq i \leq n$, and assuming that gene expression represents a realization of the random vector for the i^{th} subject:

$$\begin{aligned} \left(\mathbf{y}_i | \boldsymbol{\beta}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}\right) &\sim \mathbb{N}_p\left(\boldsymbol{\beta}^t \mathbf{x}_i + \boldsymbol{\Lambda}\boldsymbol{\tau}, \boldsymbol{\Sigma}\right) \\ \boldsymbol{\tau} &\sim \mathbb{TN}_p\left(\mathbf{0}, \mathbb{I}_p\right) \end{aligned}$$

where $\mathbf{x}_i = (\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,p})$ is the vector of covariates (here $\mathbf{x}_i$ is the i^{th} individus genotype), $\boldsymbol{\beta}^t = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)^t$, is an unknown matrix of mean regression coefficient and $\mathbb{TN}$ is a truncated normal distribution.

On the other hand, since gene expression matrix has random missing values, we consider that $\mathbf{y}_i's$ are partially observed with an arbitrary missing pattern. In order to set up estimating equations for multivariate data with possible missing values, we separate $\mathbf{y}_i(q \times 1)$ into two components $(\mathbf{y}_i^o, \mathbf{y}_i^m)$ accordingly, where $\mathbf{y}_i^o(q_i^o \times 1)$ is the observed component and $\mathbf{y}_i^m((q - q_i^o) \times 1)$ is the missing component. Further, we introduce two missingness indicator matrices, denoted by $\mathbf{O}_i$ and $\mathbf{M}_i$ henceforth, corresponding to $\mathbf{y}_i$ such that $\mathbf{y}_i^o = \mathbf{O}_i\mathbf{y}_i$ and $\mathbf{y}_i^m = \mathbf{M}_i\mathbf{y}_i$, respectively. More specifically, $\mathbf{O}_i(q_i^o \times q)$ and $\mathbf{M}_i((q - q_i^o) \times q)$ are submatrices extracted from the rows of an identity matrix of order q, $\mathbf{I}_q$, corresponding to row positions of $\mathbf{y}_i^o$ and $\mathbf{y}_i^m$ in $\mathbf{y}_i$, respectively. When $\mathbf{y}_i$ is fully observed, $\mathbf{O}_i = \mathbf{I}_q$ and $\mathbf{M}_i$ is null. Meanwhile, it is easy to verify that: $\mathbf{y}_i = \mathbf{O}_i^T\mathbf{y}_i^o + \mathbf{M}_i^T\mathbf{y}_i^m$ and $\mathbf{O}_i^T\mathbf{O}_i + \mathbf{M}_i^T\mathbf{M}_i = \mathbf{I}_q$. Using appendix (7.3) and Bayes Theorem the negative log-likelihood for the observations $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$

$$\begin{aligned}&\mathcal{L}\big(\Theta|\mathbf{y}^o, \mathbf{y}^m, \tau\big) = \log p(\mathbf{y}^m|\mathbf{y}^o, \boldsymbol{\tau})p(\boldsymbol{\tau}|\mathbf{y}^o) = \\ &\frac{1}{2}\sum_{i=1}^{n}\big\{\log|\boldsymbol{\Omega}|) \ + \big(\mathbf{y}_i - \boldsymbol{\beta}^t\mathbf{x}_i - \boldsymbol{\Lambda}\boldsymbol{\tau}\big)\boldsymbol{\Omega}\big(\mathbf{y}_i - \boldsymbol{\beta}^t\mathbf{x}_i - \boldsymbol{\Lambda}\boldsymbol{\tau}\big) + \boldsymbol{\tau}_i\boldsymbol{\tau}_i^t\big\}\end{aligned} \tag{2.2}$$

Then the expected conditional log likelihood $\mathbf{Q}(\boldsymbol{\beta}, \boldsymbol{\Omega}, \boldsymbol{\Lambda})$ is expressed as

$$\mathbf{E}[\mathcal{L}\big(\boldsymbol{\beta}, \boldsymbol{\Omega}, \boldsymbol{\Lambda}|\mathbf{y}^o, \mathbf{y}^m, \boldsymbol{\tau}\big)] \propto \frac{1}{n}\sum_{i=1}^{n}\mathrm{tr}\{\mathbf{R}_i(\boldsymbol{\beta}, \boldsymbol{\Sigma})\boldsymbol{\Omega}\} - \log|\boldsymbol{\Omega}|$$

where

$$\mathbf{R}_i(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}) = \mathbb{E}\Big[\big(\mathbf{y}_i - \boldsymbol{\xi} - \boldsymbol{\Lambda}\boldsymbol{\tau}_i)(\mathbf{y}_i - \boldsymbol{\xi} - \boldsymbol{\Lambda}\boldsymbol{\tau}_i\big)^t|\mathbf{y}_i^o, \boldsymbol{\Theta}\Big] \tag{2.3}$$

Regularization for precision matrix $\boldsymbol{\Omega}$

We regularize the entries of $\boldsymbol{\Omega}$ using ℓ_1 penalties, then the regularized conditional log likelihood $\mathbf{Q}$ is expressed as:

$$\mathbf{Q}(\boldsymbol{\beta}, \boldsymbol{\Omega}, \boldsymbol{\Lambda}) + \mathcal{R}(\boldsymbol{\Omega}) = \mathbf{Q}(\boldsymbol{\beta}, \boldsymbol{\Omega}, \boldsymbol{\Lambda}) + \lambda_1\sum_{j=1}^{q}\sum_{k=1}^{p}|\boldsymbol{\Omega}_{j,k}|$$

For sufficiently large tuning parameter values λ_1, the penalty results in precision matrix estimations $\boldsymbol{\Omega}$ with all off-diagonal entries equal to zero. The penalty assumes that some entries are equal to zero. In multi-tissue joint eQTL mapping, a zero in the (j,k)th entry of $\boldsymbol{\Omega}$ indicates independent expression in the jth and kth tissues, given expression in all other tissues and p SNP genotypes (1). We do this for two reasons: first, when $p > n$ (which is the case for practically every gene we test, more SNPs than observations), without penalizing the diagonals, a perfect fit can occur in the M-step and not in E step (see Algorithm 2): Recent literature, has shown that precision matrix estimators employed in predictive models exhibit this behavior [11]. Next we summarize the main ECM Algorithm 2. Algorithm 3 gives a detailed version of Algorithm in appendix 7.3):

Algorithm 2 Regularized ECM Algorithm with non-normality

1: Initialize $\beta^{(1)} \in \mathbb{R}^{p\times q}$ and $\Omega^{(1)} \in \mathbb{R}^{q\times q}$.
2: Set $k = 1$.
3: **while** not converged **do**
4: Compute $\mathbf{Q}(\beta, \mathbf{\Omega}, \mathbf{\Lambda}|\beta^{(k)}, \mathbf{\Omega}^{(k)}, \mathbf{\Lambda}^{(k)})$
5: $\Omega^{(k+1)} = \underset{\Omega \in \mathbb{R}^{q\times q}}{\arg\min}\left[\mathtt{tr}(R^{(k)}\Omega) - \log|\Omega| + \mathcal{R}(\Omega)\right]$
6: $\xi^{(k+1)} = \underset{\xi \in \mathbb{R}^{p\times q}}{\arg\min}\left[\mathtt{tr}(R^{(k)}\Omega)\right]$
7: $\hat{\beta}^{(k+1)} = \hat{\xi}^{(k+1)}\mathbf{x}_i/(\mathbf{x}_i\mathbf{x}_i^t)$
8: If the previous objective value Q has not converged, update $k = k + 1$ and return to Step 3.
9: **end while**

Remark 1: Step 4 of Algorithm 2 can be expressed

$$\mathbf{\Omega}^{(k+1)} = \underset{\mathbf{\Omega} \in \mathbb{R}^{q\times q}}{\arg\min}\left[\mathtt{tr}(\mathbf{R}^{(k)}\mathbf{\Omega}) - \log|\mathbf{\Omega}| + \lambda_1 \sum_{j=1}^{q}\sum_{k=1}^{p}|\mathbf{\Omega}_{j,k}|\right] \tag{2.4}$$

Conveniently, (2.4) is exactly the optimization problem for computing the ℓ_1-penalized normal log-likelihood precision matrix estimator with input sample covariance matrix $\mathbf{R}(.)$. Many efficient algorithms and software packages exist for computing (2.4). We can use our R software BiGQUIC to solve (2.4) which is available at BiGQUIC [12]

Remark 2. Step 5 of Algorithm 2 can be expressed as:

$$\boldsymbol{\xi}^{(k+1)} = \underset{\boldsymbol{\xi} \in \mathbb{R}^{p\times q}}{\arg\min}\frac{1}{n}\sum_{i=1}^{n}\left\{(\mathbf{y}_i^{(k)} - \boldsymbol{\xi} - \mathbf{\Lambda}\boldsymbol{\tau}_i)\mathbf{\Omega}^{(k+1)}(\mathbf{y}_i^{(k)} - \boldsymbol{\xi} - \mathbf{\Lambda}\boldsymbol{\tau}_i)^t\right\}$$

Using appendix (7.3), we have

$$\hat{\boldsymbol{\xi}}^{(k+1)} = \frac{1}{n}\left(\sum_{i=1}^{n}\hat{\mathbf{y}}_i^{(k)} - \hat{\mathbf{\Lambda}}^{(k)}\sum_{i=1}^{n}\hat{\boldsymbol{\eta}}_i^{(k)}\right)$$

where

$$\hat{\mathbf{y}}_i = \mathbf{\Omega}^{-1}\mathbf{S}_i^{oo}\mathbf{y}_i + \left(\mathbf{I} - \mathbf{\Omega}^{-1}\mathbf{S}_i^{oo}\right)\left(\hat{\boldsymbol{\xi}} + \hat{\mathbf{\Lambda}}\hat{\boldsymbol{\eta}}_i\right)$$

Illustrations

Simulation study

In this section, we utilize 80 replications of data generated from multivariate regression, where the sample size is $n = 50, 150$, $p = 22$, and $q = 24$. The choice of q is made to align with the dimension of the regression models applied to the GTEx data. In each

replication, a sparse matrix $\mathbf{B}$ is generated using the element-wise product of three matrices: $\mathbf{B} = \mathbf{W} \odot \mathbf{K} \odot \mathbf{Q}$ where $\odot$ is the elementwise product, $(\mathbf{W})_{ij} \sim N(0,1)$ and $(\mathbf{K})_{ij} \sim Bernouilli(s_1)$ and each row of $\mathbf{Q}$ consists of either all 1's or all 0's with a success probability of 1's equal to s_2.

By generating $\mathbf{B}$ in this manner, we anticipate that $(1 - s_2)p$ predictors will be irrelevant for all q responses, and each predictor will be relevant for $s_1 q$ of all the response variables. An $n \times p$ predictor matrix $\mathbf{X}$ is also generated with rows drawn independently from $N(\mathbf{0}, \mathbf{\Sigma_X})$, where $(\mathbf{\Sigma_X})_{ij} = 0.7^{|i-j|}$, following the approach of Yuan and Lin (2007) [13] and Peng et al. (2010) [14]. We consider an AR(1) covariance structure for the scale matrix of the errors, which is $\mathbf{\Sigma} = \boldsymbol{\rho}^{|i-j|}$.

Lastly, every row of the error matrix $\mathbf{E}$ is independently sampled from a multivariate skew-Normal distribution $\mathbb{SN}(\mathbf{0}, \mathbf{\Sigma}, \mathbf{\Lambda})$, and the response matrix $\mathbf{Y}$ is formed as $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$. To reduce computation time, we independently generate validation data (sample size $n = 50$) within each replication to estimate the prediction error for the algorithms, akin to performing a K-fold cross-validation for the algorithm, as described in Rothman et al. (2010) [15].

We consider 36 different combinations of $\mathbf{\Lambda}$; $\boldsymbol{\rho}$; s_1 and s_2 from the following ranges:(1) $\boldsymbol{\rho} = \{0, 0.4, 0.8\}$, (2) $\mathbf{\Lambda} = diag(-1, 1, -1, \ldots, 1)$ or $\mathbb{1}_q$, where $\mathbb{1}_q$ is a column vector of ones, (3) $s_1 = 0.1, 0.5$, and (4) $s_2 = 1$. Tuning parameters is selected from the set $\{2^a, a \in 0; \pm 1, \ldots, \pm 5\}$.

Results on simulation study

In the simulation study, we measure the overall performance of various methods in terms of the mean squared prediction error (PE). We have computed the prediction errors (PE) values of the entropy loss functions of the estimators of $\mathbf{\Omega}$ for the 80 simulated datasets

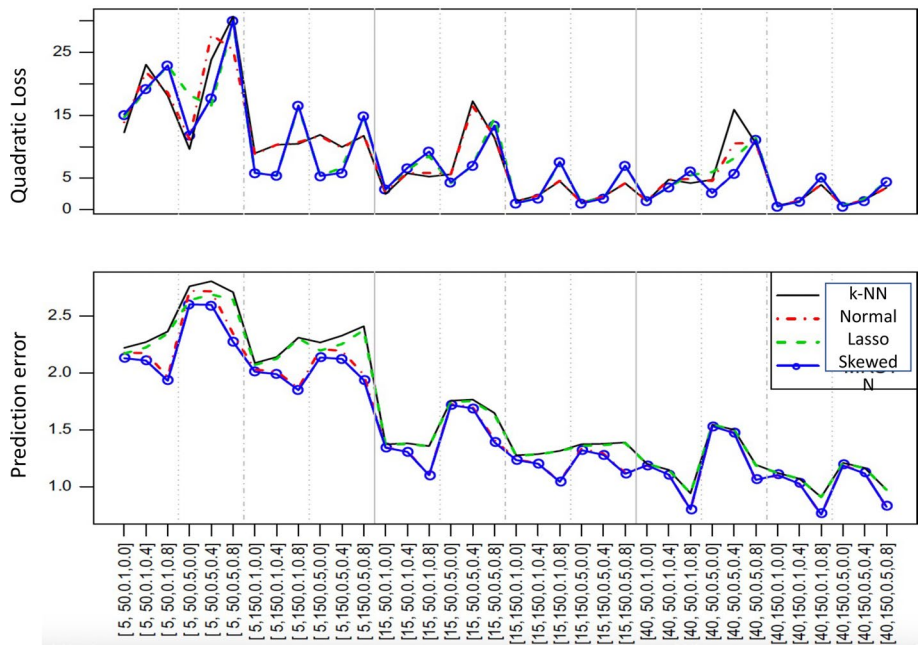

Fig. 2 Prediction errors (PE) values of the entropy loss functions of the estimators of $\mathbf{\Omega}$ for the 80 simulated datasets using the k-NN, Normal, the G-Lasso and the TISSLET (skewed) algorithm

Table 1 TPR/TNR for the matrix $\mathbf{\Omega}$ averaged over 80 replications with $s_1 = 0.1$, $s_2 = 1$ and $\lambda = (1; 1; 1; \ldots; 1)^T$

Sample size	Correlation	Normal	Skewed Normal
n=50	$\rho = 0.0$	78.24/78.64	80.43/80.20
	$\rho = 0.4$	79.26/79.39	82.10/78.60
	$\rho = 0.8$	86.88/76.79	88.27/76.04
n=150	$\rho = 0.0$	86.73/77.68	88.55/76.72
	$\rho = 0.4$	88.41/74.60	89.66/74.93
	$\rho = 0.8$	92.16/74.27	93.35/73.43

Table 2 TPR/TNR for the matrix $\mathbf{\Omega}$ averaged over 80 replications with $s_1 = 0.5$, $s_2 = 1$ and $\lambda = (1; 1; 1; \ldots; 1)^T$

Sample size	Correlation	Normal	Skewed Normal
n=50	$\rho = 0.0$	88.16/37.44	88.89/35.43
	$\rho = 0.4$	87.99/38.92	88.52/37.42
	$\rho = 0.8$	90.16/39.12	91.36/34.34
n=150	$\rho = 0.0$	92.88/43.17	92.97/43.58
	$\rho = 0.4$	93.21/42.20	93.27/41.60
	$\rho = 0.8$	95.76/36.93	95.97/36.81

using the k-NN, Normal, the G-lasso and the TISSLET algorithm (Fig. 2). To assess the significance of the difference in prediction error between the Normal and Tisslet algorithms, we conducted a Wilcoxon signed-rank test ($R_i = Sign(D_i) - rank(|D_i|)$ where D_i is the difference of both predictions. The results indicated a statistically significant difference, with a p-value of 0.0124 at the 5% significance level ($\alpha = 0.05$).

While the ℓ_1 loss is used to evaluate the performance of $\mathbf{\Omega}$, the sparsity recognition performance of $\mathbf{\Omega}$ is measured by the true positive rate (TPR) as well as the true negative rate (TNR) defined as

$$\mathrm{TPR}(\hat{\mathbf{\Omega}}, \mathbf{\Omega}) = \frac{\#\{(i,j);\, \hat{\mathbf{\Omega}}_{ij} \neq 0 \text{ and } \mathbf{\Omega}_{ij} \neq 0\}}{\#\{(i,j);\, \mathbf{\Omega}_{ij} \neq 0\}}$$

$$\mathrm{TNR}(\hat{\mathbf{\Omega}}, \mathbf{\Omega}) = \frac{\#\{(i,j);\, \hat{\mathbf{\Omega}}_{ij} = 0 \text{ and } \mathbf{\Omega}_{ij} = 0\}}{\#\{(i,j);\, \mathbf{\Omega}_{ij} = 0\}}$$

The corresponding true positive rates (TPR) and true negative rates (TNR) for $\mathbf{\Omega}$ are reported in Tables 1 and 2. From these tables it is evident that a slightly higher TPR is accompanied by a lower TNR for our algorithm Tisslet algorithm versus Molstad et al. algorithm (normal assumption). We have also compared the numerical performance of the algorithms and the G-Lasso method obtained from Friedman et al. (2022) [16] by setting Λ to be $\mathbf{0}$ for the 36 combinations. In general, it turns out that for higher error correlations ($\boldsymbol{\rho} = 0.8$) the mean square error of these methods is somewhat lower compared to the K-NN method, but the estimators of $\mathbf{\Omega}$ are improved considerably. Finally, the computational times of the two algorithms are computed.

The computational times for the four algorithms were evaluated and compared. On average, the CPU time ratios of Normal, K-NN, and Lasso to our algorithm,

Tisslet, were 3.79, 4.72 and 2.17 respectively. These computations were performed on the Panther cluster, equipped with 2.5 GHz processors. The differences in computational time are attributed to their computational complexities, which are generally $O(np)$, for all except for K-NN, which has a complexity of $O(kn^2p)$. Lasso exhibits a lower ratio compared to the other algorithms because it penalizes small values in the covariance matrix, effectively reducing the time spent on its computation.

Data generating study for SPATC1L gene

We conducted extensive numerical experiments to examine how the number of shared eQTLs, population R^2 (also known as narrow-sense heritability), and how tissue-tissue correlation structure affect the performance of various methods for estimating eQTL weights across multiple tissues. To closely replicate the conditions of joint eQTL mapping in GTEx data, we obtained whole-genome sequencing SNP genotype data for all SNPs within 500kb of the SPATC1L gene for 620 subjects from the GTEx dataset [17]. After removing highly correlated SNPs (see Data Preparation section), we were left with $p = 1178$ SNP genotypes. For each replication, we then generated $n = 620$ subjects' expressions in $q = 29$ tissues. Denoting $\mathbf{x}_i \in \mathbf{R}^p$ as the SNP genotypes for the i^{th} subject, we generated $\mathbf{y}_i \in \mathbb{R}^q$ as a realization of the random vector $\boldsymbol{\beta}_*^t \mathbf{x}_i + \boldsymbol{\epsilon}$ for $i = [n]$, where $\boldsymbol{\beta} \in \mathbb{R}^{p\times q}$ are the eQTL weights and errors are independent and identically distributed as $\mathbb{SN}_q(0, \boldsymbol{\Lambda}, \boldsymbol{\Omega}_*^{-1})$ (Same protocol as [8] and [7]). For five hundred independent replications in each scenario, we randomly divided the $n = 620$ subjects into a training set of size $n_{\text{train}} = 400$, a validation set of size 110, and a testing set of size 110.

To create missingness, in each independent replication, we generated $\mathbf{y}$ as follows: first, we created a matrix $\mathbf{B} \in \mathbb{R}^{p\times q}$ with entries that were independent $\mathcal{N}(0, 1)$. Then, we generated a matrix $\mathbf{S} \in \mathbb{R}^{p\times q}$ to be a matrix whose rows are either all zero or all one: we randomly selected s rows to be nonzero, where $s \in [20]$. Given $\mathbf{S}$, we then generated a matrix $\mathbf{U} \in \mathbb{R}^{p\times q}$ so that each of the q columns has $20 - s$ randomly selected entries equal to one only from entries which are zero in S and all others equal to zero. We calculated $\mathbf{y} = \mathbf{B} \odot \mathbf{S} + \mathbf{B} \odot \mathbf{U}$, where $\odot$ denotes the element-wise product. This construction ensured that each tissue has twenty total eQTLs, s of which are shared across all tissue types. We considered $s = \{5, 10, 15, 18, 20\}$ in the simulations presented in this section. It is important to note that due to high linkage disequilibrium, many SNPs are highly correlated, resulting in a larger number of SNPs associated with gene expression. Figure 3 gives a summary of the protocol for generating data.

In addition, we randomly introduced missing values to both the training and validation set responses with a missing probability of 0.55, which corresponds to the missing rate in the GTEx gene expression data. For each method, we trained the model on the training data, chose tuning parameters using the validation data, and evaluated prediction and variable selection accuracy on the testing data.

To construct the covariance matrix $\boldsymbol{\Omega}$, we build it to have a block-diagonal structure and to control the R^2. Specifically, we consider a covariance structure for the scale matrix of the errors, which is $\boldsymbol{\Sigma}_{ij} = \rho^{|i-j|}$, for $i \in [20, 20]$ and $\boldsymbol{\Sigma}_{ij} = (\rho + 0.2)^{|i-j|}$, for $i \in [10, 10]$.

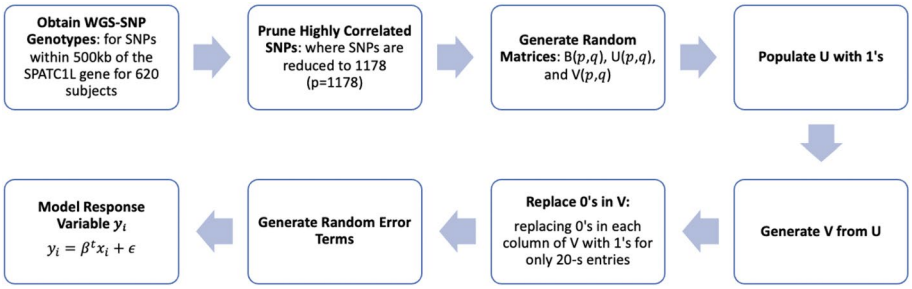

Fig. 3 Overview of the Data Pre-processing Steps. Input WGS SNP data then prune highly correlated SNPs. Create U and V and generate B. Generate error annd fit the model

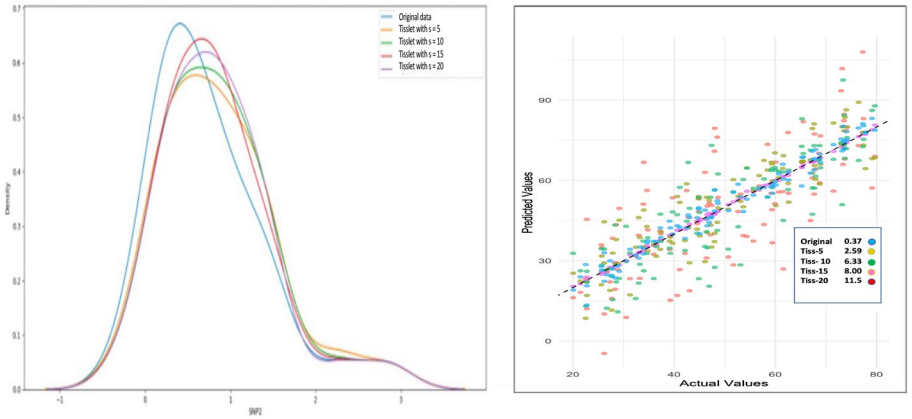

Fig. 4 Left: comparing distribution of original versus Tisslet imputed SNP_2 using several choice of s (5, 10, 15, 20) distributions. Right: providing the Mean absolute error (MAE)

Results on SPATC1L gene study

Comparative Analysis of Covariance Approaches: We evaluated the s-TISSLET imputation method's performance at different values of s for imputing missing data generated in section (3.3). Figure 4 shows that despite the varying s values, the imputed distributions closely align with the original distribution. This indicates that the imputation method maintains the overall structure of the gene expression data, which is critical for the reliability of subsequent analyses.

Dataset size and prediction accuracy. In the assessment of the TISSLET framework's performance, a nuanced relationship between dataset size (N) and R^2 scores became apparent (Fig. 5a). An increase in N was correlated with a decline in R^2 scores, underscoring the challenges in sustaining prediction accuracy with the expansion of data volume inherent in multi-omics studies. Despite this, the TISSLET framework displayed a remarkable resilience in R^2 scores across varying levels of skewness in gene expression data (Fig. 5b), attesting to its robustness and adaptability to the diverse data distributions encountered in multi-omics research. Further illustrating its efficacy, the TISSLET method demonstrated superior predictive accuracy when compared with other imputation methods such as MEANimputer, k-NN and iterative

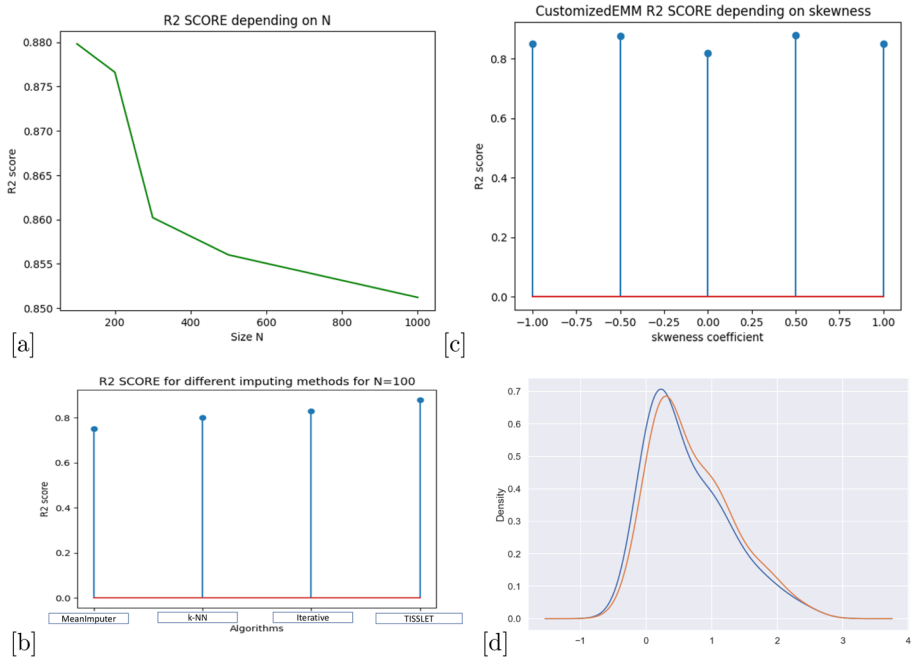

Fig. 5 **a** Relationship between dataset size and R^2 scores. TISSLET (CEM) R^2 scores across skewness coefficients. **b** *A bar graph displaying the TISSLET framework's R^2 scores, indicating consistent predictive accuracy despite varying levels of skewness* $\{-1.0; -0.5; 0; 0.5; 1.0\}$ *in gene expression data.* **c** R^2 score comparison among imputing methods for $N = 100$. **d** *Bar chart illustrating R^2 scores for different imputing methods, highlighting the superior performance of the TISSLET framework over MEANimputer, k-Nearest Neighbor (k=2), and iterative methods*

imputation for multi-tissue gene expression prediction (Fig. 5c), reinforcing its suitability for complex genomic analyses.

Impact of the number of causal variants in simulation study: We extended our simulation study to assess the impact of a larger number of causal variants, ranging from 5 to 125, across a wide range of heritability levels. We found that increasing the number of causal variants had very little effect on the predictive performance of TISSLET. We believe this is because the expected imputation accuracy primarily depends on the total heritability explained by the causal SNPs (Fig. 6).

Performance of fitting the skewed normal versus normal on predicted gene: Finally the density plot provides a visual comparison between the distributions of gene expression predictions using two [8] and our method of imputation. The curves are closely aligned and have a similar shape, peaking around a value of 1, which suggests that both methods yield relatively similar predictions in terms of their distribution. However, one curve is slightly shifted towards the right indicating that our approach may predict slightly higher expression values compared to the one that uses Normality assumption (Fig. 5d).

Results of for SPATC1L gene study: In this section, we present results with tissue-tissue correlation for several errors, where $\boldsymbol{\rho} \in \{0; 0.1; 0.3; 0.5; 0.7\}$ varying. In this setting, we observe that our method, TISSLET performs better than all realistic competitors: only G-Lasso, outperforms slightly our method when ρ is less or equal than 0.015. As one would expect, when expression is nearly uncorrelated ($\rho = 0$), our method

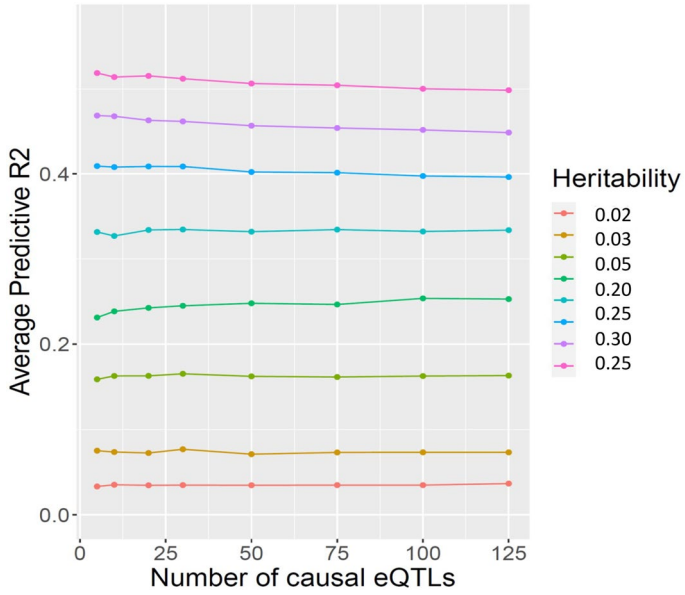

Fig. 6 Impact of number of causal variants in simulation study

TISSLET performs better than all the others 3 methods that does implicitly assumes no tissue-tissue correlation. Remarkably, when ρ is greater than or equal to 0.3, TISSLET outperforms even the normal method which assume normality and missingness. In fact, the prediction accuracy of TISSLET increases as ρ increases, whereas all other methods, which do not explicitly model tissue-tissue correlation, have prediction accuracy remaining constant or slightly decreasing as ρ increases. This demonstrates the benefit of not only accounting for tissue-tissue correlation in multi-tissue joint eQTL mapping but also assume the correct distribution of the gene expression when expression across tissue types can be reasonably assumed to be conditionally dependent (Fig. 7).

Furthermore, the heatmap of gene expressions presented in Fig. 8 illustrates a detailed example of a specific gene. In this case, out of the eight eQTLs (expression quantitative trait loci) that have been identified for this gene, seven of these eQTLs are consistently shared across all 29 examined tissues, demonstrating a uniform regulatory effect. Additionally, one of the eQTLs is shared across 28 of the 29 tissues, indicating a high level of commonality with a single tissue exception. This comprehensive visualization underscores the widespread influence of these eQTLs on gene expression across a broad range of tissue types, highlighting their significant role in the regulatory network.

Conclusion and discussion

Our results suggest that the proposed TISSLET framework can robustly handle missing data in multi-tissue gene expression studies. The framework's ability to retain accuracy across different imputation parameters, such as 'k', and its superior performance compared to other imputation methods, underscores its potential in improving

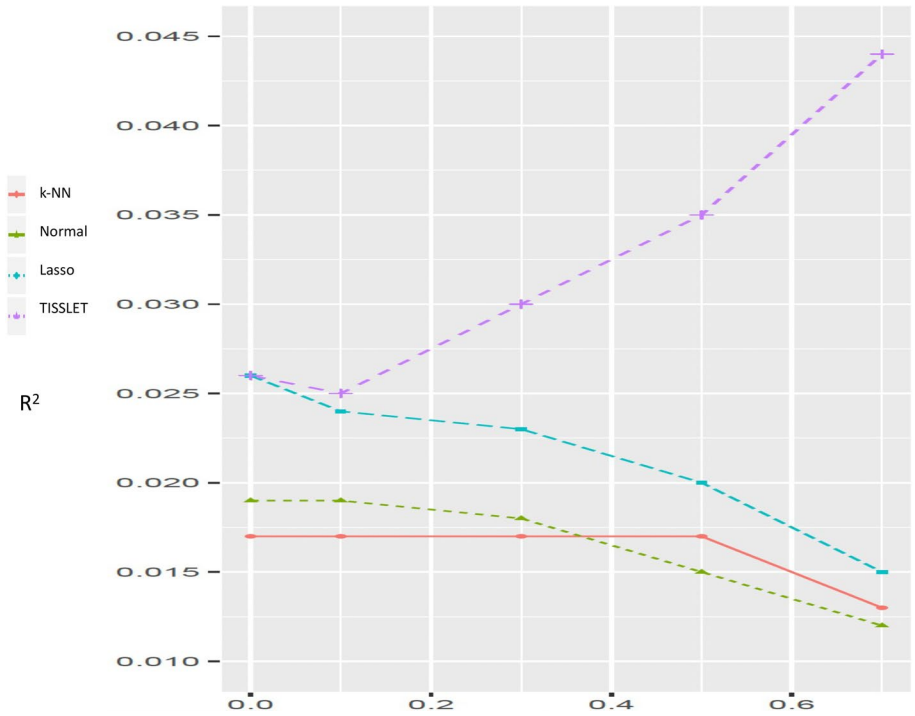

Fig. 7 Average test R^2 for 4 competing methods where ρ the correlation of the errors varies from 0.0 to 0.8

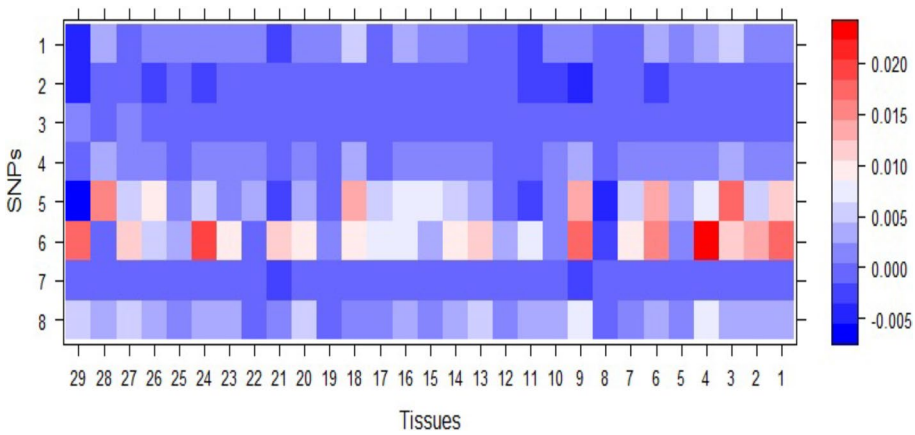

Fig. 8 A heatmap depicting the expression of SNPs per tissue, with darker blue shades indicating weaker relationships

complex trait predictions in multi-omics studies. The consistency of R^2 scores regardless of skewness and dataset size provides evidence of the model's stability. However, the observed decline in prediction accuracy with increasing dataset size suggests that further optimization of the framework may be necessary to manage the complexities of large-scale multi-omics data.

In conclusion, our findings advocate for the TISSLET framework's application in multi-omics studies to improve the prediction of complex traits. Its ability to accurately account for tissue-tissue correlations and withstand data distribution asymmetries positions it as a powerful tool in the advancement of personalized medicine. As the complexity of multi-omics studies grows, the development of robust statistical machine learning like TISSLET is crucial for unveiling the genetic foundations of complex traits and advancing our understanding of gene expression dynamics.

While the TISSLET framework marks a significant advancement in gene expression prediction across multiple tissues, this study demonstrates its precision and stability, showcasing its high accuracy in predictions. However, the framework is still in its foundational stages and requires further refinement. The next steps involve optimizing the model to enhance its performance, particularly for large-scale datasets, thereby fully realizing its potential in complex multi-omics research and personalized medicine.

Appendix

Definition and notation

- $\boldsymbol{\beta}_{.,m_i} = \mathbf{M}_i \boldsymbol{\beta}^t_{.,i}$ $\quad$ $\boldsymbol{\beta}_{.,o_i} = \mathbf{O}_i \boldsymbol{\beta}^t_{.i}$
- $\boldsymbol{\Sigma}_{o_i} = \mathbf{O}_i \boldsymbol{\Sigma} O^t_i$ $\quad$ $\boldsymbol{\Sigma}_{o_i} \in \mathbf{R}^{(q_0 \times q_0)}$
- $\boldsymbol{\Sigma}_{m_i} = \mathbf{M}_i \boldsymbol{\Sigma} \mathbf{M}^t_i$ $\quad$ $\boldsymbol{\Sigma}_{m_i} \in \mathbf{R}^{(q-q_0)\times(q-q_0)}$
- $\boldsymbol{\Sigma}_{o_i,m_i} = \mathbf{O}_i \boldsymbol{\Sigma} \mathbf{M}^t_i$ $\quad$ $\boldsymbol{\Sigma}_{o_i,m_i} \in \mathbf{R}^{(q_0 \times (q-q_0))}$
- $\boldsymbol{\Sigma}_{m_i,o_i} = \mathbf{M}_i \boldsymbol{\Sigma} \mathbf{O}^t_i$ $\quad$ $\boldsymbol{\Sigma}_{m_i,o_i} \in \mathbf{R}^{(q-q_0)\times q_0}$
- $\mathbf{y}_i^{m_i} \sim \mathbf{N}_{q-q_0}(\boldsymbol{\beta}^T_{.,m_i}\mathbf{x}_i, \boldsymbol{\Sigma}_{m_i})$
- $\mathbf{y}_i^{o_i} \sim \mathbf{N}_{q_0}(\boldsymbol{\beta}^T_{.,o_i}\mathbf{x}_i, \boldsymbol{\Sigma}_{o_i})$

Penalized normal log-likelihood

Using Aaron et al. (2020), then can summarize the derivation of EM for estimating missing gene expression $\hat{y}_i$, eQTL weights β and cross-tissue correlation Ω.

$$\mathbf{y}_i^{m_i} | y_i^{o_i}, \boldsymbol{\Theta} \sim \mathbf{N}(\mu_{\mathbf{i}}, \mathbf{V}_i)$$

where

$$\boldsymbol{\mu}_i = \boldsymbol{\beta}^T_{.,m_i}\mathbf{x}_i + \boldsymbol{\Sigma}_{m_i,o_i}\boldsymbol{\Sigma}^{-1}_{o_i}\left(\mathbf{y}^{o_i} - \boldsymbol{\beta}^T_{.,o_i}\mathbf{x}_{\mathbf{i}}\right) \tag{5.1}$$

$$\mathbf{V}_i = \boldsymbol{\Sigma}_{m_i} - \boldsymbol{\Sigma}_{m_i,o_i}\boldsymbol{\Sigma}^{-1}_{o_i}\boldsymbol{\Sigma}_{o_i,m_i} \tag{5.2}$$

Then the negative log-likelihood of the **conditional** multivariate normal distribution for $\mathbf{y}_{i,m_i} \mid \mathbf{y}_{i,o_i}$ is proportional to

$$\frac{1}{n}\sum_i \left(\mathbf{y_i}^{o_i} - \boldsymbol{\beta}^T_{.,o_i}\mathbf{x}_i\right)^T \left(\mathbf{y_i}^{m_i} - \boldsymbol{\beta}^T_{.,m_i}\mathbf{x}_i\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{y_i}^{o_i} - \boldsymbol{\beta}^T_{.,o_i}\mathbf{x}_i\right)^T \left(\mathbf{y_i}^{m_i} - \boldsymbol{\beta}^T_{.,m_i}\mathbf{x}_i\right)^T + \log|\boldsymbol{\Sigma}|$$

The negative log likelihood is useful in the EM steps because we needed to calculate the expectation of the log-likelihood $Q(\Theta)$:

$$\mathbf{Q}(\boldsymbol{\Theta}) = \mathtt{tr}\left\{\mathbf{S}_{(\boldsymbol{\beta},\boldsymbol{\Sigma})}\boldsymbol{\Omega}\right\} - \log\det\boldsymbol{\Omega}$$

We regularize the entries of $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$ using l_1 penalties, and estimate them by minimizing the penalties likelihood

$$\mathbf{Q}(\boldsymbol{\Theta}) = \mathtt{tr}\left\{\mathbf{S}_{(\boldsymbol{\beta},\boldsymbol{\Sigma})}\boldsymbol{\Omega}\right\} - \log\det\boldsymbol{\Omega} + \lambda_1\|\Omega\|_1 + \lambda_2\sum_{jk}|\beta_{jk}|$$

where the empirical covariance matrix

$$\begin{aligned}
&\mathbf{S}_{(\boldsymbol{\beta},\boldsymbol{\Sigma})} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Gamma}_i \\
&\boldsymbol{\Gamma}_i = \begin{pmatrix} \boldsymbol{\Gamma}_{o_i,o_i} & \boldsymbol{\Gamma}_{o_i,m_i} \\ \boldsymbol{\Gamma}_{m_i,o_i} & \boldsymbol{\Gamma}_{m_i,m_i} \end{pmatrix} \\
&\boldsymbol{\Gamma}_{o_i,o_i} = (\mathbf{y}_{i,o_i}^T - \mathbf{x}_i^T\boldsymbol{\beta}_{\cdot,o_i})^T(\mathbf{y}_{i,o_i}^T - \mathbf{x}_i^T\boldsymbol{\beta}_{\cdot,o_i})
\end{aligned} \tag{5.3}$$

$$\boldsymbol{\Gamma}_{m_i,m_i} = (\boldsymbol{\mu}_{i,m_i}^T - \mathbf{x}_i^T\boldsymbol{\beta}_{\cdot,m_i})^T(\boldsymbol{\mu}_{i,m_i}^T - \mathbf{x}_i^T\boldsymbol{\beta}_{\cdot,m_i}) + \mathbf{V}_{m_i} \tag{5.4}$$

$$\boldsymbol{\Gamma}_{m_i,o_i} = (\mathbf{y}_{i,o_i}^T - \mathbf{x}_i^T\boldsymbol{\beta}_{\cdot,o_i})^T(\boldsymbol{\mu}_i - \mathbf{x}_i^T\boldsymbol{\beta}_{\cdot,m_i}) \tag{5.5}$$

Then the algorithm is summarized as the following:

Penalized skew-normal log-likelihood

$$\mathcal{L}\left(\Theta|\mathbf{y}^o,\mathbf{y}^m,\boldsymbol{\tau}\right) = \log p(\mathbf{y}^m|\mathbf{y}^o,\boldsymbol{\tau})p(\boldsymbol{\tau}|\mathbf{y}^o) \tag{5.6}$$

$$\mathcal{L}(\boldsymbol{\Theta}|\mathbf{y}) = \frac{1}{2}\sum_{i=1}^{n}\{\log(\det\boldsymbol{\Omega}^{-1}) + (\mathbf{y}_i - \boldsymbol{\beta}^t\mathbf{x}_i - \boldsymbol{\Lambda}\boldsymbol{\tau})\boldsymbol{\Omega}(\mathbf{y}_i - \boldsymbol{\beta}^t\mathbf{x}_i - \boldsymbol{\Lambda}\boldsymbol{\tau}) + \boldsymbol{\tau}_i\boldsymbol{\tau}_i^t\} \tag{5.7}$$

where $\boldsymbol{\Theta} = (\boldsymbol{\beta}, \boldsymbol{\Omega}, \boldsymbol{\Lambda})$ represents all unknown parameters. In the E-step of ECM, we need to calculate the Q-function, which is the conditional expectation of the complete data log-likelihood function (7) given the observed data $\mathbf{y}^o$ and the current estimate $\hat{\Theta}^k = (\hat{\beta}, \hat{\Sigma}, \hat{\Lambda})$. Herein, the term $-\frac{1}{2}E(\tau_j^t\tau_j|\mathbf{y}_j^0, \Theta^k)$ can be omitted because it does not include any parameters. Therefore, we have

$$\mathbf{Q}\left(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(k)}\right) = \frac{1}{2}\sum_{i=1}^{n}\left[\log|\boldsymbol{\Omega}| - \mathbb{E}\left(\mathbf{Z}_i\mathbf{Z}_i^t\right)\right] \tag{5.8}$$

Using the formula $\mathbb{E}\left(\mathbf{Z}\mathbf{Z}^t\right) = \mathtt{Var}(\mathbf{Z}) + \mathbb{E}(\mathbf{Z})\mathbb{E}(\mathbf{Z})^t$, we have:

$$\mathbb{E}\left(\mathbf{Z}\mathbf{Z}^t\right) = \mathtt{Var}\left[\left(\mathbf{y} - \xi - \boldsymbol{\Lambda}\boldsymbol{\tau}\right)\boldsymbol{\Omega}^{1/2}\right] + \left(\mathbb{E}(\mathbf{y}|\mathbf{y}^o) - \boldsymbol{\xi} - \boldsymbol{\Lambda}\hat{\boldsymbol{\eta}}\right)\left(\mathbb{E}(\mathbf{y}|\mathbf{y}^o) - \xi - \boldsymbol{\Lambda}\hat{\boldsymbol{\eta}}\right)\boldsymbol{\Omega} \tag{5.9}$$

$$= \mathtt{Var}\left[\left(\hat{\mathbf{y}} - \xi - \boldsymbol{\Lambda}\boldsymbol{\tau}\right)\boldsymbol{\Omega}^{1/2}\right] + \mathtt{tr}\left(\hat{\mathbf{y}} - \boldsymbol{\xi} - \boldsymbol{\Lambda}\hat{\boldsymbol{\eta}}\right)\left(\hat{\mathbf{y}} - \boldsymbol{\xi} - \boldsymbol{\Lambda}\hat{\boldsymbol{\eta}}\right)^t\boldsymbol{\Omega}$$

Using iterative expectation and $\mathbf{O}_i^t\mathbf{O}_i(\mathbf{I}_p - \boldsymbol{\Sigma}\mathbf{S}_i^{oo}) = 0$, we have

$$\mathbb{E}(\mathbf{y}_i^m|\mathbf{y}_i^o) = \mathbf{M}_i\left(\boldsymbol{\xi} + \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\Sigma}\mathbf{S}_i^{oo}(\mathbf{y}_i - \boldsymbol{\xi} - \boldsymbol{\Lambda}\boldsymbol{\eta}_i)\right)$$

$$cov(\mathbf{y}_i, \mathbf{y}_i) = \mathbf{M}_i(\mathbf{I}_p - \boldsymbol{\Sigma}\mathbf{S}_i^{oo})(\boldsymbol{\Sigma} + \boldsymbol{\Lambda}(\boldsymbol{\Psi}_i - \boldsymbol{\eta}_i\boldsymbol{\eta}_i^t)\boldsymbol{\Lambda}(\mathbf{I}_p - \mathbf{S}_i^{oo}\boldsymbol{\Sigma}))\mathbf{M}_i^t$$

$$\mathbb{E}(\mathbf{y}_i\boldsymbol{\tau}^t|\mathbf{y}_i^o) = \mathbf{M}_i(\mathbf{I}_p - \boldsymbol{\Sigma}\mathbf{S}_i^{oo})(\boldsymbol{\xi}\boldsymbol{\eta}_i^t + \boldsymbol{\Lambda}\boldsymbol{\Psi}_i) + \boldsymbol{\Sigma}\mathbf{S}_i^{oo}\mathbf{y}_i\boldsymbol{\eta}_i^t$$

After simplification we have

$$\mathbf{Q}(\boldsymbol{\Theta}|\hat{\boldsymbol{\Theta}}) = \frac{1}{2}\sum_i\left\{\mathtt{tr}\left(\mathbf{R}_{(\boldsymbol{\xi},\boldsymbol{\Lambda})}\boldsymbol{\Omega}^{-1}\right) - \log\det(\boldsymbol{\Omega})\right\} + \boldsymbol{\lambda}_1\|\boldsymbol{\Omega}\|_1 + \boldsymbol{\lambda}_2\sum_{jk}|\boldsymbol{\beta}_{jk}| \tag{5.10}$$

$$\hat{\mathbf{R}}_i\left(\boldsymbol{\xi}, \boldsymbol{\Lambda}\right) = \left((\mathbf{I}_p - \hat{\boldsymbol{\Omega}}^{-1}\hat{\boldsymbol{S}}_i^{oo}\hat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}\right)\left(\hat{\boldsymbol{\Psi}}_i - \hat{\boldsymbol{\eta}}_i\hat{\boldsymbol{\eta}}_i^t\right) \times \left((\boldsymbol{I} - \hat{\boldsymbol{\Omega}}^{-1}\hat{\mathbf{S}}_i^{oo}\hat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}\right)$$

$$+\left(\mathbf{I} - \hat{\boldsymbol{\Omega}}^{-1}\hat{\mathbf{S}}_i^{oo}\right)\hat{\boldsymbol{\Omega}}^{-1} \; + \; \left(\hat{\mathbf{y}}_i - \boldsymbol{\xi} - \boldsymbol{\Lambda}\hat{\boldsymbol{\eta}}_i\right)\left(\hat{\mathbf{y}}_i - \boldsymbol{\xi} - \boldsymbol{\Lambda}\hat{\boldsymbol{\eta}}_i\right)$$

$$\hat{\mathbf{S}}_i^{oo} = \mathbf{O}_i^t\left(\mathbf{O}_i\hat{\boldsymbol{\Omega}}^{-1}\mathbf{O}_i\right)^{-1}\mathbf{O}_i$$
$$\hat{\boldsymbol{\eta}}_i = \mathbb{E}\left(\boldsymbol{\tau}_i|\mathbf{y}_o, \boldsymbol{\Theta}\right)$$
$$\hat{\boldsymbol{\Psi}}_i = \mathbb{E}\left(\boldsymbol{\tau}_i\boldsymbol{\tau}_i^t|\mathbf{y}_o, \boldsymbol{\Theta}\right)$$

The predicted gene expression $\hat{\mathbf{y}}_i$ is given by

$$\hat{\mathbf{y}}_i = \mathbb{E}(\mathbf{y}_i|\mathbf{y}^o, \boldsymbol{\Theta}) = \boldsymbol{\Omega}^{-1}\mathbf{S}_i^{oo}\mathbf{y}_i + \left(\mathbf{I} - \boldsymbol{\Omega}^{-1}\mathbf{S}_i^{oo}\right)\left(\hat{\boldsymbol{\xi}} + \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\eta}}_i\right)$$

Using those derivation, Algorithm becomes:

Algorithm 3 Regularized EM Algorithm with skewed-normality

Initialize: $\mathbf{\Omega}^0$, ξ^0, $\mathbf{\Lambda}^0$ and η^0.
Set $k = 1$.
E-Step:
while not converged **do**
  $\mathbf{\Upsilon}_i^k \leftarrow I - \hat{\mathbf{\Omega}}^{-1(k)}\hat{\mathbf{S}}_i^{oo(k)}$
  $\hat{\mathbf{y}}_i^{(k+1)} \leftarrow \hat{\mathbf{\Omega}}^{(k)}\hat{\mathbf{S}}^{oo(k)}\mathbf{y}_i^{(k)} + \mathbf{\Upsilon}^k(\hat{\xi}^{(k)} + \hat{\mathbf{\Lambda}}^{(k)}\hat{\eta}_i^{(k)})$
  $\mathbf{C}_i = \mathbf{O}_i^t\mathbf{\Omega}^{oo}\mathbf{O}_i$
  $\tau^{(k+1)}|\mathbf{y}_i^{(k)} \leftarrow TN_q\Big(\hat{\mathbf{\Lambda}}^{(k)}\mathbf{C}_i^{oo(k)}(\hat{\mathbf{y}}_i^{(k)} - \hat{\xi}^{(k)}), (I - \hat{\mathbf{\Lambda}}^{(k)}\mathbf{C}^{oo(k)}\mathbf{\Lambda}^{(k)}), \mathbb{R}_+^q\Big)$
  $\hat{\eta}^{(k+1)} \leftarrow \mathbb{E}(\tau|\mathbf{y}_i)$
  $\hat{\psi}^{(k+1)} \leftarrow \texttt{var}(\tau|\mathbf{y}_i) + \hat{\eta}^{(k)}\hat{\eta}^{t(k)}$
end while
CM Step 1:
$\hat{\xi}^{(k+1)} \leftarrow \frac{1}{n}\Big(\sum_{i=1}^n \hat{\mathbf{y}}_i^{(k)} - \hat{\mathbf{\Lambda}}^{(k)}\sum_{i=1}^n \hat{\eta}_i^{(k)}\Big)$
CM-Step 2:
$\hat{\mathbf{\Omega}}^{(k+1)} \leftarrow \frac{1}{n}\sum_{i=1}^n \mathbf{R}_i^{(k)}$
For $i := 1$ to n
$\mathbf{R}_i^{(k+1)} \leftarrow \Big(\mathbf{\Upsilon}_i^k\hat{\mathbf{\Lambda}} - \mathbf{\Lambda}\Big)\Big(\hat{\mathbf{\Psi}}_i - \hat{\eta}_i\hat{\eta}_i{}^t\Big)\Big(\mathbf{\Upsilon}_i^k\hat{\mathbf{\Lambda}} - \mathbf{\Lambda}\Big)^t + \Upsilon_i^k\hat{\mathbf{\Omega}}^{-1} + (\hat{\mathbf{y}}_i - \xi - \mathbf{\Lambda}\hat{\eta}_i)(\hat{\mathbf{y}}_i - \xi - \mathbf{\Lambda}\hat{\eta}_i)$
End
CM-Step 3:
$\hat{\mathbf{\Lambda}}^{(k+1)} \leftarrow diag\Big\{(\hat{\mathbf{\Omega}}^{(k)} \odot \sum_{i=1}^n \hat{\mathbf{\Psi}}^{(k)})^{-1}(\hat{\mathbf{\Omega}}^{(k+1)} \odot \sum_{i=1}^n \hat{\mathbf{\Phi}}_i^{(k)})\mathbb{1}_p\Big\}$

For $i := 1$ to n $\hat{\mathbf{\Phi}}_i^{(k+1)} \leftarrow (\hat{\mathbf{\Psi}}^{(k)} - \hat{\eta}^{(k)}\hat{\eta}^{(k)t})\hat{\mathbf{\Lambda}}^{(k)}(\mathbf{I} - \hat{\Omega}^{(k)-1}\hat{S}^{oo(k)}) + \hat{\eta}^{(k)}(\hat{\mathbf{y}}_i - \hat{\xi}^{(k)})$
End

Precision matrix estimator

To estimate the inverse of the covariance matrix, several thresholding approaches have been proposed. The two most popular ones that we propose to use in this paper are GLASSO, available in R, that was proposed par Friedman et al. (2007) [16] and CLIME that was proposed by Cai et al. (2011) [18] If we use CLIME from Cai and Zhou approach, their equation emphasizes the fact that the solution may not be unique. Such nonuniqueness always occurs since $\Sigma = \mathbf{X}\mathbf{X}^T$ where $\mathbf{X} \in \mathbb{R}^{n\times p}$ is the sample data matrix and rank$(\mathbf{X}) \leq n \ll p$ which is our case. However, there is no guarantee that the thresholding estimator is always positive definite [19]. Although the positive definite property is guaranteed in the asymptotic setting with high probability, the actual estimator can be an indefinite matrix, especially in real data analysis.

In the following we will describe quickly the argument and parameters of implementing BigQuiC in R.

Previous ad-hoc process for association SNPs with tissues

1. Create a copy of the genotype data matrix for each tissue. Only keep the samples that are matched with the corresponding gene expression and covariates data.
2. Remove SNPs with low minor allele frequency (MAF).

3. Remove genes with consistently low expression level.
4. For each tissue, load the tailored genotype data, gene expression data, covariates data, and gene/SNP location files into R.
5. To do a single tissue analysis, we will feed the data into MatrixEQTL to conduct single-tissue eQTL analysis and obtain summary statistics (i.e., t-statistics).
6. Using the cis window size to be 1 kb (i.e., 1×105) or 1 Mb (i.e., 1×106) and the p value threshold to be 1 in order to output summary statistics for all cis gene-SNP pairs then we convert each summary statistic t to a correlation r in each tissue $r = \frac{1}{\sqrt{df+t^2}}$ where df is the degree of freedom (the number of samples minus the number of covariates minus two) in the corresponding tissue.
7. Further we convert the correlations to z-statistics using Fisher transformation $z = \frac{1}{2}\sqrt{df-1}\log(\frac{1+r}{1-r})$.
8. Then we get the list of common gene-SNP pairs across all tissues and curate the obtained z-statistics into a matrix where each row corresponds to a gene-SNP pair in the common list and each column corresponds to a tissue. The z-statistics matrix is all we need for subsequent analyses.

Acknowledgements
We extend our gratitude to Dr. Molstad from the University of Minnesota for his invaluable assistance in providing access to his code and offering detailed explanations.

Author Contributions
The mathematical model was developed by H.B., writing the theoretical part and overview revision of the paper. M.C. Revised the paper. The author A.M. worked on running the mathematical model and producing the results. The author T.H. worked on producing the results in the earlier stage. The author A.A. worked on writing the results and discussion part with H.B. and A.M.; The corresponding author for submitting the article is A.A..

Funding
Not Applicable.

Availability of data and materials
Not Applicable.

Declarations

Ethics approval and consent to participate
Not Applicable.

Competing interests
The authors have no Conflict of interest to declare that are relevant to the content of this article.

Received: 19 September 2024 Accepted: 27 December 2024
Published online: 25 February 2025

References

1. Grinberg NF, Wallace C. Multi-tissue transcriptome-wide association studies. Genetic Epidemiology. 2021;45(3):324–37.
2. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, Eyler AE, Denny JC, Consortium G, Nicolae DL, et al. A gene-based association method for mapping traits using reference transcriptome data. Nature genetics. 2015;47(9):1091–8.
3. Manor O, Segal E. Genoexp: a web tool for predicting gene expression levels from single nucleotide polymorphisms. Bioinformatics. 2015;31(11):1848–50.
4. Lage K, Hansen NT, Karlberg EO, Eklund AC, Roque FS, Donahoe PK, Szallasi Z, Jensen TS, Brunak S. A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. Proceedings of the National Academy of Sciences. 2008;105(52):20870–5.
5. Bai Y, Wang X, Hou J, Geng L, Liang X, Ruan Z, Guo H, Nan K, Jiang L. Identification of a five-gene signature for predicting survival in malignant pleural mesothelioma patients. Frontiers in Genetics. 2020;11:899.

6. Zhang N, Li P-c, Liu H, Huang T-c, Liu H, Kong Y, Dong Z-c, Yuan Y-h, Zhao L-l, Li J-h. Water and nitrogen in-situ imaging detection in live corn leaves using near-infrared camera and interference filter. Plant Methods. 2021;17:1–11.
7. Hu Y, Li M, Lu Q, Weng H, Wang J, Zekavat SM, Yu Z, Li B, Gu J, Muchnik S, et al. A statistical framework for cross-tissue transcriptome-wide association analysis. Nature genetics. 2019;51(3):568–76.
8. Molstad AJ, Sun W, Hsu L. A covariance-enhanced approach to multi-tissue joint eqtl mapping with application to transcriptome-wide association studies. The annals of applied statistics. 2021;15(2):998.
9. Church BV, Williams HT, Mar JC. Investigating skewness to understand gene expression heterogeneity in large patient cohorts. BMC bioinformatics. 2019;20:1–14.
10. Hosking JRM, Wallis JR, Wood EF. Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. Technometrics. 1985;27(3):251–61.
11. Neal RM, Hinton GE. A view of the em algorithm that justifies incremental, sparse, and other variants. In: Learning in Graphical Models, 1998;pp. 355–368. Springer, ???
12. Hsieh C-J, Sustik MA, Dhillon IS, Ravikumar PK, Poldrack R. Big & quic: Sparse inverse covariance estimation for a million variables. Advances in neural information processing systems 2013;**26**
13. Yuan M, Lin Y. Model selection and estimation in the gaussian graphical model. Biometrika. 2007;94(1):19–35.
14. Peng C. Estimating and testing quantile-based process capability indices for processes with skewed distributions. Journal of Data Science. 2010;8(2):253–68.
15. Rothman AJ, Levina E, Zhu J. Sparse multivariate regression with covariance estimation. Journal of Computational and Graphical Statistics. 2010;19(4):947–62.
16. Friedman J TR. Hastie T: Sparse inverse covariance estimation with the graphical lasso. Biostatistics 2007;**9**(3)
17. Morgan A, Vuckovic D, Krishnamoorthy N, Rubinato E, Ambrosetti U, Castorina P, Franzè A, Vozzi D, La Bianca M, Cappellani S, et al. Next-generation sequencing identified spatc1l as a possible candidate gene for both early-onset and age-related hearing loss. European Journal of Human Genetics. 2019;27(1):70–9.
18. Cai W.L. Tony, Luo X. A constrained 1 minimization approach to sparse precision matrix estimation. Journal of the American Statistical Association. 2011;106(494):594–607.
19. Elanbari M, Rawi R, Ceccarelli M, Bouhali O, Bensmail H. Advanced computation of a sparse precision matrix hadap: A hadamard-dantzig estimation of a sparse precision matrix. In: Proceedings of the Sixth International Conference on Computational Logics, Algebras, Programming, Tools, and Benchmarking 2015. Citeseer

Publisher's Note